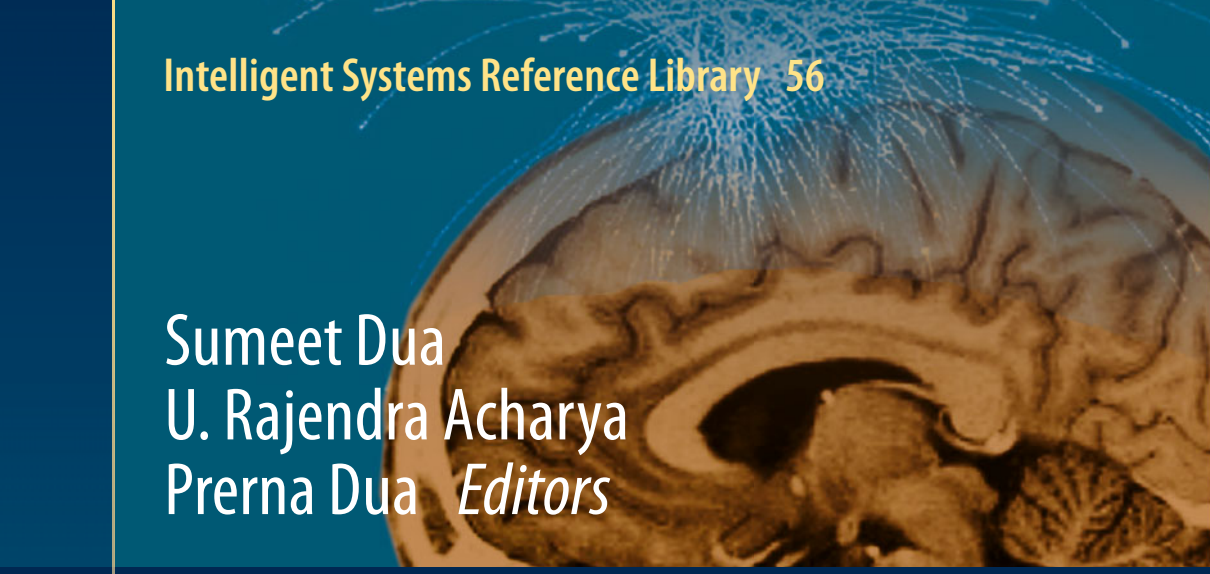
Intelligent Systems Reference Library 56



Sumeet Dua
U. Rajendra Acharya
Prerna Dua *Editors*

Machine Learning in Healthcare Informatics

Springer

Intelligent Systems Reference Library

Volume 56

Series editors

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

Lakhmi C. Jain, University of Canberra, Canberra, Australia
e-mail: Lakhmi.jain@unisa.edu.au

For further volumes:
http://www.springer.com/series/8578

About this Series

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included.

Sumeet Dua · U. Rajendra Acharya
Prerna Dua
Editors

Machine Learning in Healthcare Informatics

Springer

Editors
Sumeet Dua
Department of Computer Science
Louisiana Tech University
Ruston
USA

Prerna Dua
Department of Health Informatics and
Information Management
Louisiana Tech University
Ruston
USA

U. Rajendra Acharya
Ngee Ann Polytechnic
Singapore

ISSN 1868-4394 ISSN 1868-4408 (electronic)
ISBN 978-3-642-40016-2 ISBN 978-3-642-40017-9 (eBook)
DOI 10.1007/978-3-642-40017-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013954841

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Rapid recent advances in automated data collection routines in clinical sciences have led to a tsunami of patient-oriented data stored in distributed, heterogeneous, and large databases and datamarts. The lack of existing computing tools to enable connectivity and interoperability between these fragmented sources and scattered locations (issues concerning accessibility), and to perform machine learning on heterogeneous and highly dimensional data sources (issues concerning complexity), is an overbearing impediment, not only to healthcare sciences, but also to computational research. Moreover, the rapid deployment of high speed networks coupled with developments in knowledge discovery, bolstered by mobile technologies has amplified the emphatic demand for a unifying, coherent computing resources designed to accommodate, enhance, and empower multidisciplinary, and multi-institutional healthcare informatics research.

Healthcare data is complex, highly context-dependent, inherently heterogeneous, and high dimensional—generating an amalgamation of computing research challenges that renders the extraction of insightful knowledge through interpretation of raw data a challenging computational task. These data resources encompass a spectrum of data types ranging from free-text notes to complex image types such as position emission tomography scans. As clinical data collection technologies continue to grow and storage costs continue to fall, more complex data types such as hyperspectral images are becoming available in abundance. These diverse and prolific data sources provide an outstanding research test bed for development of the novel machine learning algorithms that are at the heart of the current data-rich but information-poor paradigm, saddling many disciplines outside of just health care. It is evident that an integrated, panoramic view of data will provide an opportunity for previously impossible clinical insights and discoveries.

The book provides a unique compendium of current and emerging machine learning paradigms for healthcare informatics. Chapters provided by established scientists in the area with the wealth of experience in the area, and have been carefully selected to reflect the diversity, complexity, and the depth and breath of this multidisciplinary area. Machine learning paradigms in healthcare informatics such as the ones presented in the chapters offer the promise of precise, objective, and accurate in-silico analysis of this emerging area using information learning routines that reveal embedded patterns, trends, and anomalies in order to create models for faster and more accurate physiological and healthcare discovery.

Chapter 1 provides an introduction to machine learning in healthcare informatics. The chapter provides an overview of the data and knowledge discovery challenges associated in the field of healthcare informatics. It introduces the challenges of machine learning in the area and the relevant areas of investigation in the area. The chapter explains the taxonomy of the healthcare informatics area and the current and provides an overview of the current efforts and emerging challenges of the Electronic Health Records (EHR) systems.

Chapter 2 discusses a machine learning approach to screen arrhythmia from normal sinus rhythm from the ECG. The methodology consists of R-point detection using the Pan-Tompkins algorithm, discrete wavelet transform (DWT) decomposition, subband principal component analysis (PCA), statistical validation of features, and subsequent pattern classification. Different classifiers used were Gaussian mixture model (GMM), error back propagation neural network (EBPNN), and support vector machine (SVM). Results indicate that the Symlet-2 wavelet basis function provided the highest accuracy in classification. Among the classifiers, SVM yields the highest classification accuracy, whereas EBPNN yields a higher accuracy than GMM.

Uncontrolled diabetes may lead to many serious complications. The result may be ketosis, which is normally due to an increase of acetone (a toxic acid product) and may lead to a situation such as diabetic coma. A fuzzy logic control system for the regulation of glucose level for diabetic patients was proposed in Chap. 3. A mathematical model describing the relationship between the human glucose level, insulin, and food was first presented. Then, a generalized fuzzy logic controller, including a set of fuzzy logic rules, is introduced to regulate glucose levels for diabetic patients. Following the fuzzy logic controller, simulation is presented. The results show that the fuzzy logic control is effective for handling the glucose level based on feedback scheme.

An integrated methodology for electrocardiogram (ECG)-based differentiation of arrhythmia and normal sinus rhythm using genetic algorithm optimized *k*-means clustering was discussed in Chap. 4. Open source databases consisting of the MIT BIH arrhythmia and MIT BIH normal sinus rhythm data were used. The methodology consists of QRS-complex detection using the Pan-Tompkins algorithm, principal component analysis (PCA), and subsequent pattern classification using the *k*-means classifier, error back propagation neural network (EBPNN) classifier, and genetic algorithm optimized *k*-means clustering. The *k*-means classifier provided an average accuracy of 91.21 % over all folds, whereas EBPNN provided a greater average accuracy of 95.79 %. In the proposed method, the *k*-means classifier is optimized using the genetic algorithm (GA), and the accuracy of this classifier is 95.79 %, which is equal to that of EBPNN.

Pixel/voxel-based machine learning (PML) is a powerful tool in computer-aided diagnosis (CAD) schemes for detection of lesions in medical images. Massive-training ANNs (MTANNs) were used for improving the performance (i.e., both sensitivity and specificity) of CAD schemes for detection of lung nodules in computer tomography (CT) and the detection of polyps in CT colonography in Chap. 5. The MTANN supervised filter is effective for enhancement

of lesions including lung nodules and colorectal polyps and suppression of non-lesions in medical images, which contributed to the improvement of the sensitivity as well as specificity in the initial lesion detection stage in CAD schemes, whereas the classification MTANNs contributed to the improvement of specificity in the false positive (FP) reduction stage in CAD schemes.

Understanding the biomechanics of the human foot during each stage of walking is important for the objective evaluation of movement dysfunction, accuracy of diagnosis, and prediction of foot impairment. In Chap. 6 Bayesian Network (BN) was used to extract the probabilistic causal information of foot function data, such as muscle activities, plantar pressures, and toe trajectories, from different types of data on human walking phases. The graphical networks extracted from the three stages of the stance phase of gait measurement data were useful for understanding the foot function of the normal walking and simulated hemiplegic walking. Thus, understanding the foot function during walking is important for further analysis of diagnostic, therapy, and training programs for foot impairment.

Successful application of machine learning in health care requires accuracy, transparency, acceptability, ability to deal with complex data, ability to deal with background knowledge, efficiency, and exportability. Rule learning is known to satisfy the above criteria. Chapter 7 introduces rule learning in health care, presents very expressive attributional rules, briefly describes the AQ21 rule learning system, and discusses three application areas in healthcare and health services research.

In the past two decades, machine learning techniques have been extensively applied for the detection of neurologic or neuropsychiatric disorders, especially Alzheimer's disease (AD) and its prodrome, mild cognitive impairment (MCI). Chapter 8 presents some of the latest developments in the application of machine learning techniques to AD and MCI diagnosis and prognosis. Discussion on how various biomarkers as well as connectivity networks can be extracted from the various modalities, such as structural T1-weighted imaging, diffusion-tensor imaging (DTI), and resting-state functional magnetic resonance imaging (fMRI), for effective diagnosis and prognosis was provided in detail.

Chapter 9 discusses several examples of how machine learning algorithms can be used to guide clinical decision making, and to generate scientific insights about these decisions. The focus of the chapter has been on rehabilitation in home care. In clinical applications, it was shown that machine learning algorithms can produce better decisions than standard clinical protocols. A "simple" algorithm such as KNN may work just as well as a more complex one such as the SVM. More importantly, it was shown that machine learning algorithms can do much more than make "black-box" predictions; they can generate important new clinical and scientific insights. This can be used to make better decisions about treatment plans for patients and about resource allocation for healthcare services, resulting in better outcomes for patients, and in a more efficient and effective healthcare system.

The widespread adoption of electronic health records in large health systems, combined with recent advances in data mining and machine methods, creates opportunities for the rapid acquisition and translation of knowledge for use in clinical practice. One area of great potential is in risk prediction of chronic progressive diseases from longitudinal medical records. Chapter 10 illustrates this potential of using a case study involving prediction of heart failure. Throughout, we discuss challenges and areas in need of further development.

Chapter 11 provides a framework to improve the physicians' diagnostic accuracy with the aid of machine learning algorithm. The resulting system is effective in predicting patient survival, and rehab/home outcome. A method has been introduced that creates a variety of reliable rules that make sense to physicians by combining CART and C4.5 and using only significant variables extracted via logistic regression. A novel method for assessment of Traumatic Brain Injury (TBI) has also been presented. The ability of such a system to assess levels of Intracranial Pressure (ICP) as well as predict survival outcomes and days in ICU, together encompasses a wholesome diagnostic tool, which can help improve patient care as well as save time and reduce cost.

One of the most crucial problems facing the U.S. government is fraud in healthcare system. Due to a large amount of data, it is impossible to manually audit for fraud. Hence, many statistical approaches have been proposed to overcome this problem. As fraud can be committed in complex and numerous ways, fraud detection is challenging, and there is a greater need for working models for fraud detection, including types of fraud that are not yet in use, as these models will not be outdated quickly. To establish a well-functioning healthcare system, it is important to have a good fraud detection system that can fight fraud that already exists and fraud that may emerge in future. In Chap. 12 an attempt has been made to classify fraud in the healthcare system, identify data sources, characterize data, and explain the supervised machine learning fraud detection models.

A migraine is a neurological disorder that can be caused by many factors, including genetic mutations, lifestyle, cardiac defects, endocrine pathologies, and neurovascular impairments. In addition to these health problems, an association between some types of migraines and increased cardiovascular risk has emerged in the past 10 years. Moreover, researchers have demonstrated an association between migraines and impaired cerebrovascular reactivity. It is possible to observe carbon dioxide dysregulation in some migraineurs, while others show a markedly decreased vasomotor reactivity to external stimuli. Therefore, the assessment of the cerebrovascular pattern of migraineurs is important both for the onset of a personalized therapy and for follow-up care. Chapter 13 discusses the analysis of hemodynamic changes during external stimulation using near-infrared spectroscopy (NIRS) signals.

The segmentation of the carotid artery wall is an important aid to sonographers when measuring intima-media thickness (IMT). Automated and completely user-independent segmentation techniques are gaining increasing importance, because they avoid the bias coming from human interactions. Chapter 14 discusses the calculation of the large and overabundant number of parameters extracted from

ultrasound carotid images and then selects a smaller subset to classify the pixels into three classes (lumen, intima-media complex, and adventitia). The selection was obtained through a feature selection method based on rough set theory. In particular, the use of QuickReduct Algorithm (QRA), the Entropy-Based Algorithm (EBR), and the Improved QuickReduct Algorithm (IQRA) was discussed.

Many authors have contributed to this book with their tremendous hard work and valuable time. We deeply thank them for their great contributions. In no particular order, they are: Roshan Joy Martis, Chandan Chakraborty, Ajoy Kumar Ray, K. Y. Zhu, W. D. Liu, Y. Xiao, Teik-Cheng Lim, Hari Prasad, Kenji Suzuki, Myagmarbayar Nergui, Jun Inoue, Murai Chieko, Wenwei Yu, Janusz Wojtusiak, Chong-Yaw Wee, Daoqiang Zhang, Luping Zhou, Pew-Thian Yap, Dinggang Shen, Mu Zhu, Lu Cheng, Joshua J. Armstrong, Jeff W. Poss, John P. Hirdes, Paul Stolee, Walter F. Stewart, Jason Roy, Jimeng Sun, Shahram Ebadollahi, Ashwin Belle, Soo-Yeon Ji, Wenan Chen, Toan Huynh, and Kayvan Najarian, Sonali Bais, Samanta Rosati, Gabriella Balestra, Filippo Molinari, Samanta Rosati, Gabriella Balestra, and Jasjit S. Suri.

Sumeet Dua
U. Rajendra Acharya
Prerna Dua

Contents

Chapter 1
Introduction to Machine Learning in Healthcare Informatics

Pradeep Chowriappa, Sumeet Dua and Yavor Todorov

Abstract Healthcare informatics, a multi-disciplinary field has become synonymous with the technological advancements and big data challenges. With the need to reduce healthcare costs and the movement towards personalized healthcare, the healthcare industry faces changes in three core areas namely, electronic record management, data integration, and computer aided diagnoses. Machine learning a complex field in itself offers a wide range of tools, techniques, and frameworks that can be exploited to address these challenges. This chapter elaborates on the intricacies of data handling the data rich filed of healthcare informatics, and the potential role of machine learning to mitigate the challenges faced.

1.1 Introduction

Healthcare informatics deals with the acquisition, transmission, processing, storage, and retrieval of information pertinent to healthcare for the early detection, early diagnosis, and early treatment of diseases [1]. The scope of healthcare informatics is confined to data associated with diseases, healthcare records, and the computational techniques associated with handling of such data. With the intent of providing affordable, quality, and seamless healthcare—traditional medical practices across the United States over the past few decades have invested on better technology and computational support to researchers, medical practitioners, and patients. These efforts have brought to the foray the benefits and importance of using computational tools for referral and prescription aids, the creation and management of electronic health records (EHR), and technological advances in digital medical imaging.

P. Chowriappa (✉) · S. Dua · Y. Todorov
Data Mining Research Laboratory (DMRL), Department of Computer Science, Louisiana Tech University, LA, USA
e-mail: pradeep@latech.edu

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_1,

For instance, studies have shown the success of Computerized Physician Order Entry (CPOE) have reduced medication errors and adverse drug events and inadvertently improved the quality of care [2]. CPOE makes patient information readily available to physicians at the time they enter the prescription for a patient [3]. It provides necessary alerts to the physician about adverse reactions that could arise specific to a patient's history. Moreover, CPOE allows the physician to track the order. This provides an additional mechanism for physicians to identify issues in a prescription and re-design it to eliminating errors.

Machine learning is a natural extension of artificial intelligence. Researchers and medical practitioners often resort to using machine learning to address complex statistical analysis. The niche of combining both healthcare data and machine learning with the goal of identifying patterns of interest is commonly referred to as healthcare informatics. The goal of healthcare informatics is therefore used to identify patterns in data, and then learn from the identified [4].

EHR systems have enabled easier access to and sharing of patient's health records between hospitals reducing the costs of healthcare manifolds. This reduction in costs has been attributed to the elimination of redundant health tests and reduction in operational costs [5]. However, with the current state of management of EHR systems makes it difficult to collate and mine clinical information for patterns of trends across various populations. With efforts such as the American Recovery and Reinvestment Act (ARRA) of 2009,[1] strides are being taken to digitize medical records to a universal format that enables the collation of medical data to large repositories. Data from these large repositories can then be used for machine learning to predict and understand trends across geographical locations [6]. Research in this area is focused on computational bottlenecks of expansion, sharing, and standardization of EHRs. The objective is to create open-access databases that are secure and can handle various forms of cyber-threat, as these databases contain confidential information of patients. Some of the prominent medical databases in the area are listed in Table 1.1. There are several challenges in creating these large data repositories of the health records (discussed in later sections) that require significant investment computation research. For instance the handling of evolving data structures in handling changing modalities of technological advances in medical devices and data generated from them.

Technological strides in medical imaging have brought about innovative means to capturing diseases such as cancers, for quicker disease prognosis [7, 8]. These advances have enabled effective detection and diagnosis of cancers. Prominent imaging modalities such as computed tomography (CT), ultrasound, and magnetic resonance imaging (MRI) have brought about minimally invasive surgery, image guided therapy, and effective monitoring of treatment response [9]. These technologies have made it possible to provide in situ anatomical data on the size, shape, and location of tumors and growths. Newer technologies such as, 3D-ultrasound, electrical impedance tomography, tomosynthesis, diffuse optical

[1] http://www.recovery.gov/About/Pages/The_Act.aspx

Table 1.1 Prominent medical databases [5]

Database	Access
NN/LM medical database	Public access—resource provides access to three databases related to biomedical and health science, these include—PubMed, Medlineplus, and Locatorplus
Allen brain atlas	Public access—a data portal that has a collection of multi-modal, multi-resolution gene expression profiles of the human and mouse brain
Alzheimer's disease neuroimaging initiative (ADNI)	Public access—web-based resource containing clinical, genetics, MRI, and PET data of individuals with Alzheimer's Disease
Australian EEG database	User access required—web-based de-identified searchable database of 18,500 EEG records
Unified medical language system® (UMLS)®	Restricted access—contains a list of known biomedical vocabularies and standards to facilitate interoperability between computer systems
Epilepsiae European database on epilepsy	Research community—database contains well documented meta data related to epilepsy
Kaiser permanente national research database	Kiaser permanente researchers and collaborating non-KP researchers—contains clinical information of approximately 3,000,000 patients of the Kaiser Foundation Health Plan
National patient care database (NPCD)	Research community—is the VHA's centralized database for integrated patient care data

tomography, diffusion-weighted magnetic resonance imaging (MRI), positron emission tomography (PET) and single-photon-emission CT (SPECT) further more provide functional activity of detected tumors, thereby revealing both location and metabolic activity of disease. These imaging platforms take advantage of molecular-targeted contrast agents to monitor the complex biochemical processes in neoplastic transformations of tumors and cancers [10].

Undeniably, the role of machine learning is paramount to the testing and development of these modalities and their practical application in a clinical setting. Machine learning in the medical imaging field manifests itself as image segmentation, image registration, image fusion, image-guided therapy, image annotation, and image database retrieval. With advances in medical imaging, there is a need for newer machine-learning algorithms/applications. Newer imaging technologies have brought about large variations and complexity to image data. It is impractical and difficult to use existing machine learning techniques to extract patterns or derive analytic solutions from newer imaging techniques. Researchers in machine learning are in pursuit of creating algorithms that scale to the changes in data. Because of its essential needs, machine learning in medical imaging is a bourgeoning field [11].

Another issue that plagues the integration of machine learning and healthcare is the use of software engineering to keep pace with the technological advances in medical data capture (multi-modal images) and advances made with machine

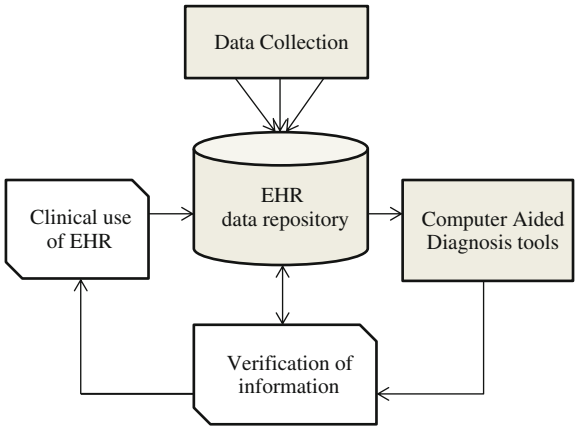


Fig. 1.1 A schematic representation of a pervasive computing engine in healthcare. The key components that constitute such a pervasive system pose independent challenges

learning algorithms. The production of quality machine learning software is currently in a state of infancy and therefore has much to learn from advances in software engineering, for instance model-driven engineering and cloud computing. A large proportion of published work in machine learning comprises investigations of algorithms and projects that are not deployed at scale in actual practice. Therefore the realization of truly scalable, robust, and reliable machine learning software's are important. Future sections of this chapter highlight the challenges in handling healthcare data, the role of machine learning and existing online health informatics tools.

1.2 Challenges

The future of healthcare lies of effective storage and distribution of patient health records in EHR data repositories. The benefits of creating an EHR data repository lies in facilitating the productivity of healthcare personnel in the delivery of quality healthcare and the optimal use of pervasive computing engine. The realization of a pervasive computing engine is conceptualized as the logical sharing of data across its components. Figure 1.1 represents the flow of information in across the components of a pervasive computing engine. These include (a) data collection, (b) an EHR data repository, and (c) computer aided diagnosis tools. In this section we emphasize challenges in the creation of a pervasive computing engine in healthcare.

1.2.1 Data Collection

Data collection and reporting is predominantly manual and largely paper based [12], and is carried out by healthcare entities. These entities endeavor to collect data in a cohesive and standardized manner. The procedure of patient data collection entails (a) the consent of patient, (b) the de-identification of data, and (c) to ensure that the data conforms to pre-set standards before the data is uploaded to a database. Though this task entails trained personal to validate the data collected it is still challenging to ensure standards are being maintained across different entities of a large healthcare system. The following are the key challenges of data collection.

1.2.1.1 Patient Consent

One of the biggest challenges in the realization of effective data collection lies on the patients/subjects consent. Several patients are concerned with the privacy of their personal information being stored over large repositories. There are several standards in place to ensure effective de-identification of patient/subject information before the data is stored on large data repositories. Typically healthcare institutions opt for the open-consent process to share de-identified information onto repositories.

1.2.1.2 Controlled Vocabulary

The biggest challenge in data collection is diversity and change/evolution of controlled medical terminologies. In the past decade, significant efforts were made to overcome the lack of standards in controlled medical terminologies (CMTs) [13]. CMTs facilitate data entry, data retrieval, data analysis, and data sharing. The overarching goal of introducing CMTs is to create efficient diagnostic decision-support systems that would send out timely alerts and reminders to medical practitioners. It also facilitates the creation of administrative systems for billing and effective administration of large healthcare facilities.

A practical example of controlled is in the case of data collection for clinical research the gathering of data is driven by variables that are relevant to "deterministic outcome" referred to as a 'research hypothesis.' The variables associated to the research hypothesis like 'patient parameters', 'data items', 'data elements', or 'questions' are gathered and represented in a cohesive manner into a data-collection form called the 'Case Report Forms' or CRFs.

1.2.1.3 Standardization

It is the responsibility of the healthcare entity to ensure that the variables on a data collection form adhere to acceptable standards. The most prominently adopted standard is that of The International Organization for Standardization/International Electro-technical Commission (ISO/IEC 11179 technical standard).[2] The ISO/IEC 11179 identifies a data element as 'that unit of data' that has a definition, identification, representation, and values that are represented as a set of attributes. These attributes include: the internal name, data type, caption presented to users, a detailed description, and a validation scheme associated such as a range check or set membership [12].

1.2.2 EHR Data Repositories

Large databases/repositories in healthcare are often acquired from a variety of sources, with a corresponding variety of design and structure. This uncertainty in the data can make linking of diverse data bases a challenge. The intended use data also poses its own set of challenges. While technical, medical, and managerial differ in intend use of data, the multifaceted nature of healthcare data calls for a multifaceted perspective of handling data [14]. Furthermore the purpose of storing EHR should systematically enable the use that machine learning strategies to mine trends in data. This is relevant considering that these databases/repositories exhibit and exponential growth in size. The following discussion highlights challenges in the creation and maintenance of EHR repositories.

1.2.2.1 Feasibility of Information Technology (IT) Infrastructure

Keeping up with technological advancements is vital to the effective utilization and maintenance of a pervasive computing resource. The IT infrastructure is an integral component of large EHR data repositories. Typically, the infrastructure should be able to scale to the growth of data that is continuously added on a regular basis.

While evaluating IT in an EHR repository, one must take into account that IT is only one part of a pervasive computing system of an organization. The objective of carrying out the feasibility of IT infrastructure is to ensure that consistent improvements are made in a timely manner to ensure longevity of the resource. Many different questions can be asked in the assessment of IT [15]. These include:

- What type of IT should be selected and employed?
- What are the key healthcare processes that the EHR repository should facilitate?

[2] http://metadata-standards.org/11179/

- What are the technical and system aspects (e.g. performance, software quality) of the IT that will bear upon its use?
- How does the IT infrastructure impact structural or process quality (time saving, data quality, clinical workflow, patient administration) with regard to different users (physicians, nurses, administrative staff)? Does it work effectively?
- How IT infrastructure impacts the quality of care?
- What are the capital and operational costs of information technology? Is it cost-effective?

1.2.2.2 Privacy Preservation and Data Integration

To understand and treat an outbreak of large proportion, the analysis of the prevalence, incidence, and risk factors of disease is crucial. To carry out such an analysis, would have a substantial ramification on policy decisions. Data from diverse repositories/databases have aggregated and integrated [16, 17]. It is vital at this juncture that private and sensitive information be handled with care. There is therefore a need for privacy preservation framework and data integration strategy in place.

In creating an EHR database/repository is a requirement to gaining the approval of the Institutional Review Board (IRB).[3] The interest of the IRB is to ensure that proper de-identification of all records is carried out before the release of data. It also ensures that the HIPPA regulations and the Helsinki declaration are adopted [18]. Moreover, it is important to ensure that despite the de-identification, the integration of data from multiple sources is not hindered and clear to the end user.

1.2.2.3 The Human Element in Creating EHR Repositories

Though current EHRs have been successfully adopted and accepted in the healthcare industry, significant challenges remain in exploring how the human element influence EMR acceptance, implementation, and use. It is believed that social intricacies and communication patterns influence the use of EHR and can be utilized to enhance the delivery of healthcare.

Researchers have indicated that communication-patterns can be characterized based on categorize of users and how individual categories users communicate with the EHR. There are roughly three categories of users namely: high, medium, and low. The users that belong to the high category are those that display high integration of EHR use with work practices. Users of this category rely highly on features of reports, flow sheets and/or tracking and treading features of the EHR. The users of the medium category display moderate integration of EHR use with

[3] http://www.thehastingscenter.org/Publications/IRB/Default.aspx/

work practices. Similarly, users of the low category rarely rely on the features offered by the EHR [19]. It is believed that understanding the communication patterns among users of an EHR can provide an understanding and achievement of a flexible EHR.

1.2.3 Computer Aided Diagnostic (CAD) Tools

The creation of sophisticated CAD tools to analyze complex biological data has been spruced by advances made in machine learning. Although results from current CAD tools are promising, there are several hurdles to overcome before these tools can be deployed in a clinical setting. Research efforts are on in the creation of computer aided prognosis and diagnosis tools that use multimodal data fusion [20]. For instance, fusing computerized image analysis with digitized patient data such genomic information for predicting outcomes and survival. The current fusion of biomedical informatics and bioinformatics techniques would propel existing CAD tools to a more patient specific diagnosis [21]. While CAD tools have proved to be instrumental in healthcare, it suffers from the following challenges:

1.2.3.1 Data Preprocessing

CAD tools extensive use patient data from diverse sources. These could be image sources such as position tomography (PET), computed tomography (CT), low-dose computed tomography (LDCT), functional magnetic resonance imaging (fMRI), and contrast-enhanced computed tomography (CE-CT) [22]. Other sources of medical data could be obtained from signaling sources such as electrocardiogram (ECG), and electroencephalogram (EEG). Typically data from medical sources suffer from noise in the form of inconsistences in measurements. This noise could significantly affect the quality performance of CAD tools. Researchers rely on data preprocessing strategies to capture features of discrimination (or interest). Significant ongoing efforts in this area rely on the creation of novel machine learning techniques for effective data preprocessing.

1.2.3.2 Effective Software Design

The creation of quality machine learning software is currently in the state of infancy. There are several challenges that need to be taken into consideration while porting machine learning algorithm to functional CAD tools. It is the responsibility of developers (in the USA) to ensure that their CAD tools conform to the standards set by the FDA. The FDA certifies both CAD tools and CAD systems for use in medical practice [4]. This requires fully traceable, auditable procedures for

software development of the kind developed by software engineers over the past decades.

To conform to software engineering it is firstly desired that the algorithms are scalable. It is desired that the machine learning algorithms are tested to handle data sets on a large scale as in actual practice. Furthermore, the algorithms should be capable of delivering reliable and accurate results. This is a challenge as these algorithms require intense testing procedures. Another desirable feature is cross platform re-usability. This requires a more formal approach to modular code abstraction, design, and specifications.

1.2.3.3 Validation and Verification

From existing research, there is a lack of consensus in the theoretical understanding especially from a non-physician's perspective of CAD tools. This renders certain CAD tools ineffective. It is believed that without a solid validation and verification scheme physicians are unfairly susceptible to accepting recommendations of CAD tools, questioning the quality of decisions made. This renders the verification and validation of paramount importance.

Several publications use statistical methods to interpret and explain the various criteria. However, with numerous clinical implementations of decision support systems for a variety of medical applications, there is a need for robust and systematic methods to verify, validate the performance of a CAD tool.

1.3 Healthcare Informatics and Personalized Medicine

Listed as one of the 14 grand engineering challenges by the US National Academy of engineering for the twenty-first century, healthcare informatics is a multidisciplinary field that derives advances from fields of biomedical engineering, data analytics, and bioinformatics to solve day to day challenges of healthcare (refer Fig. 1.2). The field of healthcare informatics is constantly evolving with advances made in three core dimensions namely: Data acquisition, health record management, and the role of machine learning (data analytics) for pattern analysis [23]. These advances are funneled to meet the healthcare goals of disease prevention and personalized disease diagnosis and treatment. In this section we focus our discussion on developments made in personalization of healthcare.

1.3.1 Future of Data Acquisition in Healthcare

New acquisition systems are being created. Traditional diagnostic tools for most diseases rely on the manifestation of visible symptoms to identify the disease

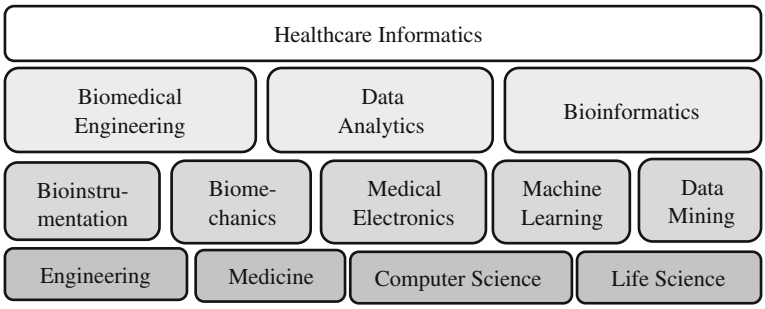


Fig. 1.2 The multidisciplinary field of healthcare informatics

affecting the patient. This approach to identification of a disease is at time too late. Research is underway in the creation of Nano-devices that allow for the detection of pathogens and diseased cells at the early stages of disease progression. Moreover, new systems are being tested and designed in the form of wearable, implantable, and ingestible devices to collect patient data with or without medical supervision [24]. Another actively pursued area of research in biomedical monitoring is body sensor networks (BSN) [25]. With the objective of improving the quality of healthcare, BNS's is a cyber-physical system (CPS) that consists of a diverse set of wearable sensors over the body of a patient. It would provide up to date information of the patient.

This opens up several challenges of data integration from devices in a dynamic environmental situation, requiring machine learning. There is a dearth of innovative machine learning algorithms that could scale to the multi-dimensional data that these systems and devices generate to help in decision support in health.

1.3.2 Patient Centric EHR

The prominence of use of EHR in healthcare has increased over the last decade. With the overall goal of using EHR is to enable exchange of information independent of the patient's lactation, the creating of such a model comes with its challenges as we take into consideration the different actors and organizations involved.

These challenges are exacerbated when we consider issues of security and efficient organization of EHR when we consider an EHR on a global scale. Furthermore the health data for each individual spans over multiple dimensions/scales from a genetic to cellular to tissue to system levels. Research efforts are on in the formation of global databases for the retrieval of relevant information to the early signaling of disease outbreaks [26].

With the prevalence of the internet, there has been an increasing trend in patients seeking to use social media sites and web-based resources to seek

healthcare information. Despite the benefits, flexibility, and ease of access to information, the prevalence of EHR on the web is stemmed by issues of security. However, there is a growing realization in the medical community of exploiting the benefits of giving the patients more liberty to access and control their own information. This is commonly referred to the transition towards a patient centric EHR.

Providing patient's access to their individual EHR is not a novel idea [26, 27]. With the objective of improving efficiency, reducing costs, and enhancing quality of healthcare and patient satisfaction the model of online communication and sharing of EHR with patients and healthcare providers has been proven beneficial.

There have been EHR systems that propose integration of data between institutions, and sharing data with patients since the beginning of the twenty-first century [28]. However, there are studies that focus on aspects of security and granting of access to EHR. It is believed that a patient-centric approach should facilitate a novice user to interpret medical information and enable the user to act appropriately. The success of any EHR system therefore lies on a design that is able to balance both ease of use and security of information being disseminated.

There are several online systems aimed at providing users the freedom to control, supervise, and recover and share their health information over the internet. Two of the prominent tools are described in the following sections.

1.3.2.1 Cambio Healthcare Systems®

Since 1993, Cambio Healthcare Systems®[4] a Swedish company has been a pioneer in distributed healthcare administration systems. Consisting of approximately a hundred office around the world providing service to about 50,000 users, the objective of creating a healthcare solution to support healthcare at every stage of a patient's life, Cambio created a product called COSMIC®. At the heart of COSMIC is the COSMIC Spider, an engine that connects a gamut of individual modules. Each module is dedicated to specific task such as care documentation, order management for both labs and referrals, e-Prescribing, patient management, resource planning, and care administration consisting of billing, digital dictation, and data warehousing.

1.3.2.2 Microsoft's® Health Vault®

Microsoft's® Health Vault®[5] an online platform for health management provides patients to "collect, store, and share" health information. It is a cloud service that has built in functionality of privacy, security, and data provenance. Health Vault®,

[4] http://www.cambio.se/

[5] https://www.healthvault.com/us/en

currently available only in the United States, supports nearly 300 applications, and connectivity to 80 health and fitness devices, such as those that measure heart rate, blood pressure, blood glucose, or peak airway flow.

1.3.3 Information Retrieval and Semantic Relationships

Rapid access to reliable information has been a perpetual need in healthcare informatics. With the proliferation of medical resources across the internet, the need for up to date medical care related information (e.g. published articles, clinical trials, news, etc.) is important to both healthcare providers and patients who prefer to be informed about their health.

The use of natural language processing (NLP) and machine learning (ML) techniques to optimize searchers and classify relevant medical information in documents [29] is not new in the field of health informatics. However, these techniques have been known to be susceptible to vocabulary mismatch. Vocabulary mismatches manifest in form of instances where relevant documents to a user's query may actually contain little or no shared terms. This hampers the performance of keyword-based retrieval. Furthermore, certain queries are inference driven requiring inferences to determine related documents. There is therefore a need for an information retrieval system capable of overcoming the mismatch between the terms found in documents and those in queries.

In the medical domain, the identification of sentences published in medical abstracts as containing or not containing information about the queried disease or treatment, and then establishing semantic relations to the prevention, cures, and side effects associated with illness and treatments, in the context as expressed in using these texts. This is brought about through domain ontologies [30].

1.3.3.1 Domain Ontology

The purpose of domain ontology in an EHR system is to represent medical terms as they apply to a medical domain. Terms pertaining to meaning and use help provide information and knowledge for better health informatics service. Of the widely used medical ontologies used in EHR, namely: the Unified Medical Language System (UMLS), Guide Line Interchange Format (GLIF), Generalized Architecture for Languages (GALEN), International Classification of Diseases (ICD), the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT®)[6] is the most predominant.

Touted for its power and efficiency in handling EHR, SNOMED CT® has been scientifically validated, and viewed as a comprehensive resource of clinical

[6] http://www.ihtsdo.org/snomed-ct/snomed-ct0/

healthcare terminology that is accepted world-wide. SNOMED CT® provides an abstraction through a hierarchical representation of encapsulated classes namely, disorders, drug, and organisms. Moreover, it covers a huge number of concepts and relationships.

As with any EHR management system, both information retrieval and information extraction are significant issues. Moreover, it is important to ensure accuracy and reporting (in a timely fashion). To overcome these issues and facilitate application development, SNOMED CT® adopts a concept-oriented and machine-readable design. Formal vocabulary in SNOMED CT® is maintained in a knowledgebase. The knowledgebase grows in an incremental fashion with the inclusion of newer domain specific concepts provided by experts in a domain.

1.4 Data Interoperability in EHR

The exchange of EHR data between institutions and care providers poses a great challenge. To facilitate effective data communication, EHR adopts an archetype standards developed by *open*EHR[7] and CEN/ISO [31]. These standards enable sharing of patient health information between healthcare providers in a multi-disciplinary environment. The objective of adopting these standards is to provide interoperability at different levels of functioning namely, within an enterprise, regionally, nationally, and globally. Moreover, it facilitates interoperability between software's and vendors. Currently the use of archetypes in the deployment of EHR's are limited [32]. However, the benefits of providing interoperability outweigh implementation challenges.

The evolution of healthcare has taken place in three avenues, namely: (a) the evolution in the knowledge base. Here rules that were relevant yesterday can become irrelevant with the addition of newer medical facts. (b) The refinement of information. With newer more focused studies and the use of better technologies, information is moving towards a finer grain with time. And (c) complexity; the relationships between facts and existing information makes the execution of a system complex. With intent to handle the evolution of healthcare the *open*EHR archetypes was proposed. *open*EHR is constructed using a two-level approach [33] that separates information structure from clinical knowledge base.

*open*EHR provides the necessary abstraction to domain experts to create content models for clinical concepts without worrying about the equipment used. These content models can then collectively make up the information system. This also facilitates EHR systems to accommodate changes in medical and health service delivery practices over time.

[7] http://www.openehr.org/

1.4.1 Archetype Modeling and openEHR

Archetypes as the name suggest, are data models designed to store clinical data and content. Archetypes are different from traditional data models in such that they focus on providing three key data functionalities that are susceptible to time and that are congruent to the clinical practice. These functionalities are semantic interoperability, semantic interpretability, and syntactic interoperability. It should be known that these three functionalities are aimed at providing free data exchange between two or more entities.

1. *Syntactic (data) interoperability*: Syntactic interoperability refers to all aspects of data representation. Here we strive to ensure that the structure and origin of data is understood by the information system. Key emphasis is place on the syntax of by which data is stored.
2. *Semantic interpretability*: Semantic interpretability ensures that concepts of clinical significance are represented efficiently. Good semantic interpretability is ensured when information can be easily interpreted by a domain expert.
3. *Semantic interoperability*: Semantic interoperability ensures data exchange on the basis of shared, pre-defined and mutually accepted meanings of terms and expressions. Semantic interoperability requires that a system understand the semantics of data requested.

The *open*EHR archetype is touted as the comprehensive open specifications for EHR systems. The *open*EHR architecture adopts a two level modeling approach to building comprehensive EHR management systems. The first level also referred to as the reference information model is pruned to carry the minimum information to carryout effective record management. This level also ensures effective data transmission between clinicians and providers thereby bring about the desired data interoperability.

At the second level, *open*EHR brings about the semantic interoperability. This is brought about by *open*EHR providing the required semantic to store/record relevant information that needs to be processed. In other words, the archetype represents domain specific concepts by providing the necessary rules or constraints applicable in the *open*EHR information models. These constraints therefore represent the valid data structures, data types, and values that are define in advance.

This two level approach offered by the *open*EHR architecture enables a clear separation between record keeping and clinical data collection, thereby isolating the challenges of record keeping that can hinder the clinical data collection and vice versa.

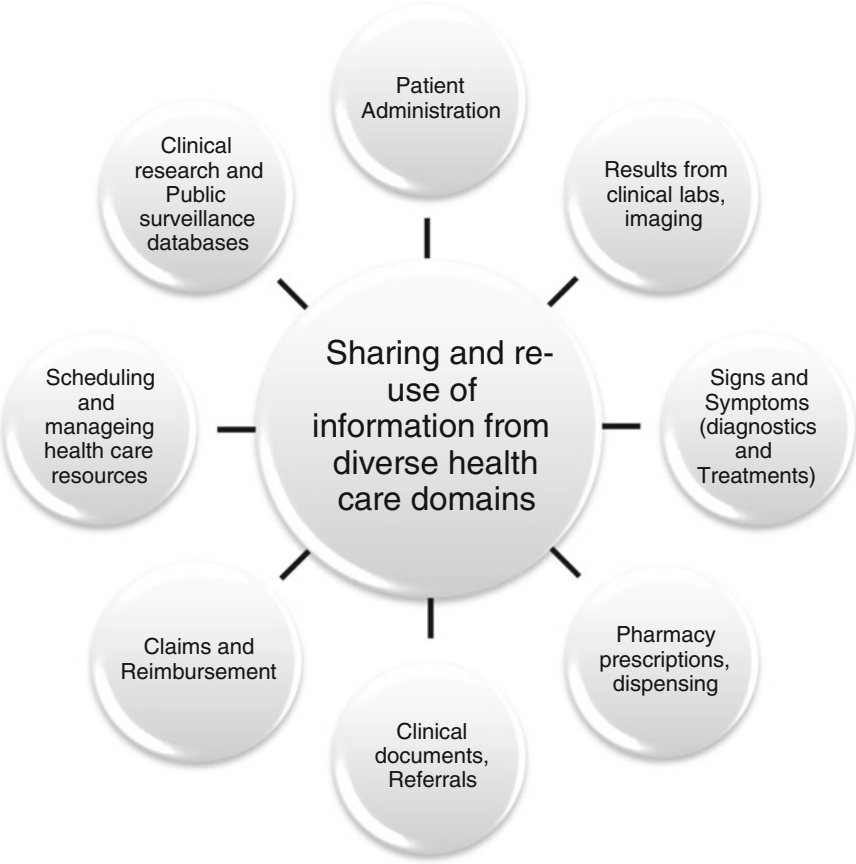


Fig. 1.3 HL7 standards and associated domains

1.4.2 Health Level 7 (HL7)

With respect to standards in healthcare data exchange and interoperability, health level seven (HL7) is one of the most prominent interchange standard for clinical data exchange, both in the US and around the world. Clinical data exchange based on HL7 spreads the gamut ranging from numerical data, coded and text observations, orders, scheduled clinical work, and exchanges of master file records (refer Fig. 1.3).

With the mission is to provide standards to improve care delivery, optimize workflow, reduce ambiguity, and enhance knowledge transfer HL7 offers standards to a wide range of domains. These include clinical, clinical genomics, administrative, clinical research, electronic claims attachments, public health, personal health, etc., to name a few.

The HL7 was initially designed and systematized using the unified service action model (USAM). This design was used to create the ordering, scheduling, and care planning capabilities, however, evolved to handle task with workflow management.

In order to support effective decision support (DS) for medical care, computerized guidelines are able to improve the quality and reduce the cost of healthcare. However, objective and decidable guidelines are expensive to define, which suggests a sharing of guidelines to reduce these expenses. The Arden Syntax, a standard to define medical logical modules, was created to facilitate guideline sharing and dissemination. However, despite the Arden's considerable acceptance in the industry, it did not lead to a broad-based guideline deployment.

The InterMed Collaboratory, an online medical collaboration facility created by Columbia, Harvard, McGill, and Stanford proposed the guideline interchange format (GLIF) to meet the same basic goals as Arden. GLIF was designed based on experience with several research guideline systems (e.g. EON), is based on an information model, it is a declarative rather than procedural language. It is designed keeping in mind complex clinical protocols.

It is believed that sharing and deployment of guidelines has been limited because of very practical reasons. Guidelines must be tied to the EHR, not bothering the user with data entry. Yet, coupling generic shared guidelines to an EHR is difficult to achieve as the structure and condition of clinical data varies across medical institutions. In order to achieve coupling, the identification of clinical variables is important and a well-known challenge. Furthermore the data in databases are not of good condition and therefore challenging to create automated decision systems that are reliable. This challenge is further exacerbated when guidelines require unique or derived keys from the data which traditional EHR does not support.

1.5 Machine Learning in Computer Aided Diagnostics (CAD)

With the intent to decrease observation oversights by clinical domain experts, computer aided diagnostics (CAD) have revolutionized medicine. CAD bridges the gap between technological advances and clinical practice by introducing newer modalities to understand diseases. These modalities include acquisition techniques such as MRI's, CT scans to name a few, and better storage technologies [34].

With the realization of various technologies that could potentially benefit clinical practice. Research in computer aided diagnostics is moving towards exploiting machine learning techniques for the following reasons.

1. *Newer forms of data*: With technological advances made in genetics, imaging, signal monitoring, and radio-frequency identifications (RFID) to name a few, medicine is moving towards personalized mentoring and treatment. This has

also created a gamut of unconventional forms of data. There is therefore a need for machine learning to scale up to the variety of data forms.
2. *The scope of statistical analysis on data*: Statistical learning is a user driven process. It largely relies on confirming a set hypothesis and is driven by a set of predefined assumptions. Moreover, statistical analysis used to carry out predictions of the general population. On the contrary, machine learning is used to generate hypotheses. It is exploratory and driven by fewer assumptions. But the one characteristic difference between statistical analysis and machine learning is that machine learning is data driven.
3. *The scalability of techniques*: With the growth of data reaching exponential rate, there is a need for algorithms and techniques that can exploit the generated data and provide predictions that can scale to the changes in data, as well as discover hidden and non-trivial observations that cannot be carried out manually.

Machine learning is built around the popular KDD process [35], consists of key steps namely the data exploratory phase, the training phase, and the validation phase. The data exploratory phase consists of feature extraction and feature selection strategies. The objective of the data exploratory phase is to discover patterns. Each pattern will therefore result in an independent hypothesis for testing. Moreover, the data exploratory phase is useful in identifying those factors that are influential and contribute towards the hypothesis.

As part of the training phase, a learning model is fit on the data using the influential factors discovered in the data exploratory phase. In the training phase, data from known classes is used to create a model, that when tested in the verification stage to would yield results in favor of the hypothesis. Unlike statistical testing, machine learning relies on domain experts and analytical skills to verify the outputs obtained. Though promising, the application of machine learning for medical diagnosis is challenging as it must meet the following criteria for success: good performance, the transparency of diagnostic knowledge, the ability to explain decisions, the ability of the algorithm to reduce the number of tests necessary to obtain reliable diagnosis, and the ability to appropriately deal with missing data.

Machine learning community has a long tradition in classical knowledge discovery applications and can be traced at least as far as the mid-1960s. Several approaches have been proposed and find their application in the clinical data analysis. The most prominently referred techniques include, neural networks (NN), support vector machines (SVM), decision trees (DT), etc. [36]. Though popular, the use of machine learning in healthcare or clinical data analysis, the following needs have to be addressed while using machine learning, namely

1. *Datasets*: Typically datasets used for analysis are plagued with missing and inappropriate values. These errors typically referred to as noise are brought about by devise errors. Care should be taken to handle noise in datasets. The application of appropriate data cleaning techniques has to be applied to data before subjecting it to learning strategies.

2. *Model selection*: For a given model, various model selection techniques can be applied to determine quality of extracted knowledge [37]. However, with the gamut of modeling techniques available—bootstrapping, hold-out techniques, and k-fold cross validation to name a few—the choice of an appropriate model becomes a challenge. An inappropriate model selection technique will result in biased and over-fit estimates of results. Therefore care should be taken while choosing an appropriate model selection technique.
3. *Feature extraction and feature selection*: It should be noted that most of the data generated from newer technologies are high dimensional datasets. For example, in the cases of positron emission tomography (PET) and single photon emission computed tomography (SPECT) [38], and functional magnetic resonance imaging (fMRI) rely on voxel identification and tracking. Voxel identification and tracking generates multi-dimensional data that requires effective feature extraction from raw data [39].

Machine learning offers a wide array of learning approaches that can be chosen to capture hidden patterns from the data [40]. We categorize these techniques into three and provide a brief overview of these categories as follows:

1.5.1 Unsupervised Approaches

The unsupervised approaches of machine learning are those that find hidden patterns or trends in data. These approaches seek to find key features that drive differentiation among data samples [41]. Also referred to as clustering, these approaches are prominently used in signal analysis frameworks. Some of the commonly referred unsupervised approaches include, spectral clustering, Gaussian mixture models [42], K-means [43], fuzzy clustering [44]. These approaches however, suffer from a bottleneck as they are subject to predetermined thresholds.

1.5.2 Supervised Approaches

Unlike unsupervised approaches, supervised approaches are model building approaches. These approaches use a preprocessed training set of sample to build a model, where each sample in the training set has a determined class label. The objective of supervised learning approaches is to determine characteristic sets of rules that can be used to discriminate between samples of different classes. For instance, decision trees (DT) is a supervised learning approach that is easy to comprehend and relatively easy to implement. However, it is challenging to apply to complex non-linear problems. Prominently used supervised approaches in healthcare include support vector machines (SVM) [45], k-nearest neighbor (k-NN) [46], Bayesian models [47] to name a few.

1.5.3 Evolutionary Approaches

Evolutionary approaches to knowledge extraction, better referred to as genetic algorithms (GA) [48] are prominently used in medical data analysis. These approaches are based on the evolutionary ideas of natural selection and genetic processes of biological organisms. As the natural populations evolve according to the principles of natural selection and "survival of the fittest", GA are able to evolve solutions to real-world problems. The objective of using GA is to find optimal solutions even in the most complex of search spaces [49].

1.5.4 Hybrid or Ensemble Approaches

The hybrid or ensemble approaches rest on the assumption that a combination multiple single models can generate effective discriminatory rules. Moreover, each of the single models has its advantages, and inherent disadvantages, that can be overcome by other models in the ensemble [50]. However, while combining different models to overcome the disadvantages of a single model can lead to issues of over-fitting [51].

1.6 Application of Machine Learning in Healthcare

It is believed that with the integration of machine learning in healthcare can bring us close to the elusive quest of improving both the efficiency and quality of medical care. However, there are challenges and opportunities in doing so. As discussed in the previous sections, machine learning provides a gamut of approaches and techniques that have cascaded to a diverse set of tools to aid in diagnostic and prognostic challenges faced in medical domains.

In this section we focus the effects of machine learning for the identification and analysis of clinical parameters in understanding disease diagnosis and disease progression. There is significant interest in the use of machine learning for the extraction of features that could potentially lead to patient specific therapy planning and support, that could eventually lead to reduction in medical costs [52]. Machine learning is also being used to suggest real time clinical monitoring of patients. This entails real time analysis of data to appropriately deal with monitoring data from different sensors or devices, and the interpretation of continuous data to be used in Intensive Care Units (ICU) [53]. On the same lines, over the past two decades, there has been considerable research effort directed towards the monitoring and classification of physical activity patterns from body-fixed sensor data [25]. This effort has been motivated by a number of important health-related applications. For example, with the trend toward more sedentary lifestyles, there is growing interest

in the link between levels of physical activity and common health problems, such as diabetes, cardiovascular diseases, and osteoporosis. As self-reported measures have been shown to be unreliable measures for activity profiling, sensor data measures are beginning to play an important role in large-scale epidemiological studies in this area. Computer aided diagnosis (CAD) and their associated tools have been instrumental in the realization of the potential of machine learning. There are a wide range of CAD tools in the area of cancer research [21]. This is attributed to the abundant data resource that can be used to develop such tools. However, there is a need in effective integration of data and knowledge from diverse data sources. These tools also lack effective validation schemas. The fastest area to adopt CAD tools is in the area of radiology. As several of these tools are in their inspection stages of development they lack comprehensive datasets to include information of a diverse set of illnesses, complications, and injuries.

Another area that can benefit from machine learning is emergency medicine [21]. Though there are few CAD tools adopted in clinical practice, existing tools have shown the potential of improving the quality of healthcare. Ongoing research in the area is focused on making these tools to address a wider variety of illnesses and trauma scenarios. The application of machine learning in cardiovascular CAD tools has not received significant success as there is a lack of comprehensive validation processes [54]. While most cardiovascular based CAD tools suffer from high false positive rates, they often help in detecting the disease at an early stage. Therefore there is a need for tools that incorporate a wider range of information to reduce the false positive rates. Just as in the areas discussed above, digital radiology finds its application in orthodentistry [55]. They enable early diagnosis of dental complication at a stage. However, the CAD tools in this area are relatively expensive and a bottle neck for wide adaptation.

1.7 Conclusion

The importance of healthcare to individuals and governments and its growing cost to the economy have contributed to the emergence of healthcare as an important area of research focal points for scholars in business and other researchers. Both the quality of healthcare and the managing of medical care costs can be benefitted from the use of pervasive computing. In addition pervasive computing is responsible for effective data collection, standardization, storage, processing, and timely communication of information to decision makers for better coordination of healthcare. Pervasive computing relies on three interrelated components namely, patient data collection and handling, effective ERM, and CAD tools. With medicinal and clinical practice moving towards the personalized, more emphasis is placed on the patient to control his medical information to reduce medical costs. There is a growth in cheaper technologies to detect, tract, and understand diseases. This chapter focuses on creating an awareness of these trends and brining to the foray the role on machine learning in the future of healthcare.

References

1. Zhang Y, Poon C (2010) Editorial note on bio, medical and health informatics. IEEE Trans Inf Technol Biomed 14(3):543–545
2. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R (2005) Can electronic medical record systems transform healthcare? potential health benefits, savings, and costs. Health Aff 24(5):1103–1117
3. Chaudry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton S, Shekelle P (2006) Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Annal Internal Med 144, E-12-E-22
4. Clifton DA, Gibbons J, Davies J, Tarassenko L (2012) Machine learning and software engineering in health informatics. In: First international workshop on realizing artificial intelligence synergies in software engineering (RAISE), Zurich, Switzerland, 5 June 2012
5. Kerr W, Lau E, Owens G, Trefler A (2012) The future of medical diagnostics: large digitized databases. Yale J Biol Med 85(3):363–377
6. van Ginneken B, Schaefer-Prokop C, Prokop M (2011) Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261(3):719–732
7. Bedard N, Pierce M, El-Naggar A, Anandasabapathy S, Gillenwater A, Richards-Kortum R (2010) Emerging roles for multimodal optical imaging in early cancer detection: a global challenge. Technol Cancer Res Treat 9(2):211–217
8. Weissleder R, Pittet M (2008) Imaging in the era of molecular oncology. Nature 452:580–589
9. Pierce M, Javier D, Richards-Kortum R (2008) Optical contrast agents and imaging systems for detection and diagnosis of cancer. Int J Cancer 123:1979–1990
10. Massoud T, Gambhir S (2007) Integrating noninvasive molecular imaging into molecular medicine: an evolving paradigm. Trends Mol Med 13(5):183–191
11. Suzuki K, Yan P, Wang F, Shen D (2012) Machine learning in medical imaging. Int J Biomed Imaging: Article ID 123727
12. Richesson R, Nadkarni P (2011) Data standards for clinical research data collection forms: current status and challenges. J Am Med Inform Assoc 18(3):341–346
13. Oliver DE, Shahar Y, Shortliffe E, Musen M (1999) Representation of change in controlled medical terminologies. Artif Intell Med 15(1):53–76
14. Chismar W (2007) Introduction to the information technology in healthcare track. In: System sciences, 2007. HICSS 2007. 40th annual Hawaii international conference on, Waikoloa
15. Ammenwerth E, Gräber S, Herrmann G, Bürkle T, König J (2003) Evaluation of health information systems-problems and challenges. Int J Med Inform 71(2–3):125–135
16. Salih R, Othmane L, Lilien L (2011) Privacy protection in pervasive healthcare monitoring systems with active bundles. In: Parallel and distributed processing with applications workshops (ISPAW), 2011 ninth IEEE international symposium on, Busan, South Korea, 2011
17. Yakut I, Polat H (2012) Privacy-preserving hybrid collaborative filtering on cross distributed data. Knowl Inf Syst 30(2):405–433
18. World Medical Association (2008) WMA declaration of Helsinki - ethical principles for medical research involving human subjects. http://www.wma.net/en/30publications/10policies/b3/. Accessed 10 Feb 2013
19. Lanham H, Leykum L, McDaniel Jr. R (2012) Same organization, same electronic health records (EHRs) system, different use: exploring the linkage between practice member communication patterns and EHR use patterns in an ambulatory care setting. J Am Med Inform Assoc 19:382–391
20. Madabhushi A, Agner S, Basavanhally A, Doyle S, Lee G (2011) Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of mult-scale, multi-modal data. Comput Med Imaging Graph 35:506–514
21. Belle A, Kon M, Najarian K (2013) Biomedical informatics for computer-aided decision support sytems: a survey. Sci World J: Article ID 769639

22. El-Baz A, Beache G, Gimel'frab G, Suzuki K, Okada K, Elnakib A, Soliman A, Abdollahi B (2012) Computer-Aided diagnosis systems for lung cancer: challenges and methodologies. Int J Biomed Imaging: Article ID 942353
23. Poon CCY, Wang MD, Bonato P, Fenstermacher DA (2013) Editorial: special issue on health informatics and personalized medicine. Biomed Eng IEEE Trans 60(1):143–146
24. Fakruddin M, Hossain Z, Afroz H (2012) Prospects and applications fo nanobiotechnology: a medical perspective. J Nanobiotechnol 10(31):1–8
25. Calhoun B, Lach J, Stankovic J, Wentzloff D, Whitehouse K, Barth A, Brown J, Li Q, Oh S, Roberts N, Zhang Y (2012) Body sensor networks: a holistic approach from silicon to users. Proc IEEE 100(1):91–106
26. Halamka J, Mandl K, Tang P (2008) Early experiences with personal health records. J Am Med Infom Assoc 15:1–7
27. Ross S, Lin C (2003) The effects of promoting patient access to medical records: a review. J Am Med Inform Assoc 10:129–138
28. Hripcsak G, Cimino J, Sengupta S (1999) WebCIS: large scale deployment of a web-based clinical information system. In: Proceedings of AMIA symposium, Washington, DC
29. Liu Z, Chu W (2007) Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Inf Retieval 10(2):173–202
30. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, Kershenbaum A, Li M, Schonberg E, Srinivas K (2007) Matching patient records to clinical trials using ontologies. Semant Web 4825:816–829
31. Schloeffel P, Beale T, Hayworth G, Heard S, Leslie H (2006) The relationship between CEN 13606, HL7, and openEHR. In: HIC 2006 bridging the digital divide: clinician, consumer and computer, Australia, health informatics society of Australia Ltd (HISA)
32. Chen R, Klein G, Sundvall E, Karlsson D, Ahlfeldt H (2009) Archetyoe-based conversion of EHR content models: pilot experience with a regional EHR system. BMC Med Inform Decis Mak 9(33):1–13
33. Garde S, Knaup P, Hovenga E, Heard S (2007) Towards semantic interoperability for electronic health records: domain knowledge governance for openEHR archetypes. Methods Inf Med 46(3):332–343
34. Doi K (2007) Computer-Aided diagnosis in medical imaging: historical review, current status, and future potential. Comput Med Imaging Graph 31(4–5):198–211
35. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39(1):27–34
36. van Ginneken B, ter Haar Romeny B, Viergever M (2001) Computer-aided diagnosis in chest radiography: a survey. Med Imaging IEEE Trans 20(12):1228–1241
37. Schorfheide F, Wolpin K (2012) On the use of holdout samples for model selection. Am Econ Rev 102(3):477–481
38. Padilla P, Lopez M, Gorriz J, Ramirez J, Salas-Gonzalez D, Alvarez I (2012) Alzheimer's Disease Neuroimaging Initiative NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's Disease. Med Imaging IEEE Trans 31(2):207–216
39. Dua S, Srinivasan P (2008) A non-voxel bsed feature extraction to detect cognitive states in fMRI. In: 30th annual international IEEE EMBS conference, Vancouver
40. Kamruzzaman J, Begg R, Sarker R (2006) Overview of artificial neural networks and their applications in healthcare. Neural Networks in Healthcare: Potential and Challenges, Idea Group Inc (IGI), pp 1
41. Michel V, Gramfort A, Varoquaux G, Eger E, Keribin C, Thirion B (2012) A supervised clustering approach for fMRI-based inference of brain states. Pattern Recogn 45(6):2041–2049
42. Lawhern V, Hairston W, McDowell K, Westerfield M, Robbins K (2012) Detection and classification of subject artifacts in EEG signals using autoregressive models. J Neurosci Methods 208(2):181–189
43. Schalk G, Brunner P, Gerhardt L, Bischof H, Wolpaw JR (2008) Brain–computer interfaces (BCIs): detection instead of classification. J Neurosci Methods 167(1):51–62

44. Majumdar K (2011) Human scalp EEG processing: various soft computing approaches. Appl Soft Comput 11(8):4433–4447
45. Ma Z, Tavares J, Jorge R, Mascarenhas T (2010) A review of algorithms for medical image segmentation and their applications to the female pelvic cavity. Comput Methods Biomech Biomed Eng 13(2):235–246
46. Peters J, Ecabert O, Meyer C, Kneser R, Weese J (2010) Optimizing boundary detection via simulated search with applications to multi-modal heart segmentation. Med Image Anal 14(1):70
47. Suk HI, Lee SW (2013) A novel bayesian framework for discriminative feature extraction in brain-computer interfaces. Pattern Anal Mach Learn IEEE Trans 35(2):286–299
48. Maulik U (2009) Medical image segmentation using genetic algorithms. Inf Technol Biomed IEEE Trans 13(2):166–173
49. McIntosh C, Hamarneh G (2011) Evolutionary deformable models for medical image segmentation: a genetic algorithm approach to optimizing learned, intuitive, and localized medial-based shape deformation. In: Stephen L, Smith, Cagnoni S (eds) Genetic and evolutionary computation: Medical Applications, Wiley, pp 46–67
50. Varol E, Gaonkar B, Erus G, Schultz R, Davatzikos C (2012) Feature ranking based nested support vector machine ensemble for medical image classification. In: 9th IEEE international symposium on biomedical imaging (ISBI)
51. Fraz M, Remagnino P, Hoppe A, Uyyanonvara B, Rudnicka A, Owen C, Barman S (2012) An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans Biomed Eng 59(9):2538–2548
52. Mansi T, Mihalef V, Sharma P, Georgescu B, Zheng X, Rapaka S, Kamen A, Mereles D, Steen H, Meder B, Katus H, Comaniciu D (2012) Data-driven computational models of heart anatomy, mechanics and hemodynamics: an integrated framework. In: 9th IEEE international symposium on biomedical imaging (ISBI)
53. Mao Y, Chen W, Chen Y, Lu C, Kollef M, Bailey T (2012) An integrated data mining approach to real-time clinical monitoring and deterioration warning. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12), New York, NY, USA
54. Caceres C, Rikli A (2012) The digital computer as an aid in the diagnosis of cardiovascular disease. Transactions of the New York Academy of Science. 23(3 series II):240–245
55. Tracy K, Dykstra B, Gakenheimer D, Scheetz J, Lacina S, Scarfe W, Farman A (2011) Utility and effectiveness of computer-aided diagnosis of dental caries. Gen Dent 59(2):136–144

Chapter 2
Wavelet-based Machine Learning Techniques for ECG Signal Analysis

Roshan Joy Martis, Chandan Chakraborty and Ajoy Kumar Ray

Abstract Machine learning of ECG is a core component in any of the ECG-based healthcare informatics system. Since the ECG is a nonlinear signal, the subtle changes in its amplitude and duration are not well manifested in time and frequency domains. Therefore, in this chapter, we introduce a machine-learning approach to screen arrhythmia from normal sinus rhythm from the ECG. The methodology consists of R-point detection using the Pan-Tompkins algorithm, discrete wavelet transform (DWT) decomposition, sub-band principal component analysis (PCA), statistical validation of features, and subsequent pattern classification. The *k*-fold cross validation is used in order to reduce the bias in choosing training and testing sets for classification. The average accuracy of classification is used as a benchmark for comparison. Different classifiers used are Gaussian mixture model (GMM), error back propagation neural network (EBPNN), and support vector machine (SVM). The DWT basis functions used are Daubechies-4, Daubechies-6, Daubechies-8, Symlet-2, Symlet-4, Symlet-6, Symlet-8, Coiflet-2, and Coiflet-5. An attempt is made to exploit the energy compaction in the wavelet sub-bands to yield higher classification accuracy. Results indicate that the Symlet-2 wavelet basis function provides the highest accuracy in classification. Among the classifiers, SVM yields the highest classification accuracy, whereas EBPNN yields a higher accuracy than GMM. The use of other time frequency representations using different time frequency kernels as a future direction is also observed. The developed machine-learning approach can be used in a web-based telemedicine system, which can be used in remote monitoring of patients in many healthcare informatics systems.

R. J. Martis (✉) · C. Chakraborty
School of Medical Science and Technology, IIT, Kharagpur, India
e-mail: roshaniitsmst@gmail.com

A. K. Ray
Department of Electronics and Electrical Communication Engineering,
IIT, Kharagpur, India

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_2,

Keywords Arrhythmia · Normal sinus rhythm · DWT · SVM · Neural network · GMM

2.1 Introduction

In the modern world cardiovascular disease (CVD) is one of the most common causes of death, and is responsible for approximately 30 % of deaths worldwide, and nearly 40 % of deaths in high-income, developed countries [1, 2]. Even though the CVD rates are declining in high-income countries, the rates are increasing in every other part of the world [1].

Generally, the sino-atrial (SA) node acts as the pacemaker of the heart, and the primary source of electrical impulse. Cardiac arrhythmia (also known as dysrhythmia) represents a heterogeneous group of conditions in which there is abnormal electrical activity in the heart. During arrhythmia, other impulse sources may dominate the sinus node and act as independent sources of impulses. Arrhythmia is one kind of CVD, which if left untreated may lead to life-threatening medical emergencies that can result in cardiac arrest, hemodynamic collapse, and sudden death. Abnormalities of both impulse formation and impulse conduction can result in cardiac arrhythmias [3]. The heartbeat interval may be regular or irregular, and may be too fast or too slow. Early intervention with appropriate therapy is recommended in many arrhythmias; if left untreated, such arrhythmias may lead more serious complications. Arrhythmias like ventricular fibrillations and ventricular flutter are imminently life-threatening.

Increasing incidence of cardiovascular disease and death has drawn attention worldwide to the research and development of methods for mass screening to provide prognostic healthcare. One of the greatest challenges for both developed and under-developed countries is the delivery of high-quality cardiac care to the entire population. The lack of sufficiently qualified cardiac experts may, however, limit individual attention for patients and force healthcare professionals to cater to critical conditions and patients requiring immediate attention. The development of automated tools to detect cardiac arrhythmias with considerable accuracy is challenging. Widespread applications of such tools by qualified nurses or paramedics trained to handle the equipment can greatly strengthen the screening programs and aid in providing mass cardiac care with scarce resources.

Electrocardiography (ECG) is a noninvasive test for recording the electric activity of the heart over time and can be captured by surface electrodes. ECG is the simplest and most specific diagnostic test for many heart abnormalities, including arrhythmia, and is especially essential in screening for heart problems. The ECG pattern obtained from a normal subject is known as a normal sinus rhythm. The assessment of alternations in the heart rhythm using an ECG is commonly used to diagnose and assess the risk of any given arrhythmia. Different computational tools and algorithms are being developed for the analysis of the

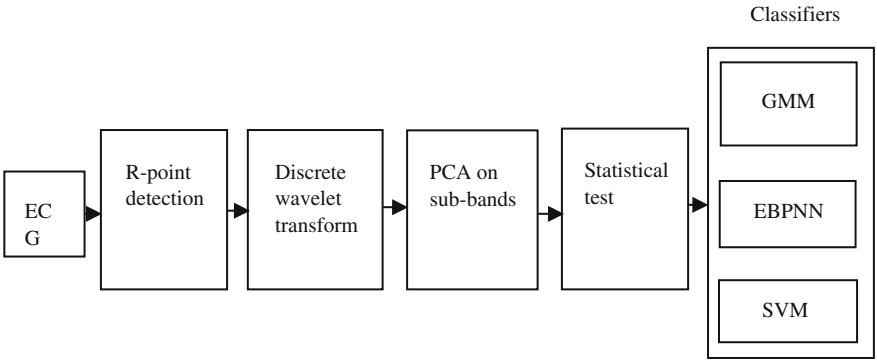


Fig. 2.1 Machine-learning approach of ECG classification into normal sinus rhythm and arrhythmia

ECG signal, and its automated diagnosis. In this chapter, the authors have made an attempt to use machine-based classification of ECG signals to sort normal sinus rhythm and arrhythmia signals into their respective classes.

Many methods for the detection of QRS complex (or the R-point) in the ECG have been proposed [4–6]. The Pan-Tompkins algorithm is commonly used because of its computational simplicity. The wavelet-based method proposed by [5], later extended by [6], can also be used for R-point detection in the ECG. The Pan-Tompkins algorithm has been used in the analysis in this chapter because of its simplicity and higher detection rate.

Few approaches for the classification of arrhythmia beats have been described in the literature [7, 8]. Most of these approaches use principal component analysis (PCA) in the time domain signal [9]. Recently, [10] gave an account of the use of PCA in DWT sub bands. Here, DWT sub-band features are compressed using PCA. Since DWT provides compact supported basis space for the signal, the PCA should provide higher compression than time domain counterparts.

2.2 Materials

In the proposed work, the open source data available at www.physionet.org from MIT BIH arrhythmia and the MIT BIH normal sinus rhythm database is used. The database is explained as follows.

2.2.1 MIT- BIH Normal Sinus Rhythm Database

The MIT-BIH normal sinus rhythm database consists of 18 long term ECG recordings of subjects referred to the Arrhythmia Monitoring Laboratory at Boston's Beth Israel Deaconess Medical Center. Subjects included in this database were found to have had no significant arrhythmias; they included five men, aged 26−45 and thirteen women, aged 20−50. The ECG data was digitized at 128 Hz.

2.2.2 MIT BIH Arrhythmia Database

The MIT BIH arrhythmia database consists of 48 half-hour excerpts of two channel ambulatory ECG data obtained from 47 subjects studied by the BIH arrhythmia laboratory between 1975 and 1979. Twenty-three recordings were randomly taken from a set of 4,000 24 h ambulatory ECG data collected from a mixed population including both inpatients (approximately 60 %) and outpatients (approximately 40 %) at the medical center. The remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias. The ECG recordings were sampled at 360 Hz per channel with an 11-bit resolution over the 10 mV range.

2.3 Methodology

Figure 2.1 depicts the machine learning approach of the proposed ECG classification system. The proposed methodology consists of an automated detection of the R-point using the Pan-Tompkins algorithm, wavelet sub-band decomposition using multiple DWT basis functions, principal component analysis (PCA) on DWT sub-bands, statistical significance tests using independent sample *t*-tests, and automated classification using three classifiers, Gaussian mixture model (GMM), error back propagation neural network (EBPNN), and support vector machine (SVM) classifiers.

Prior to R-point detection, some pre-processing is necessary to remove noise and artifacts that the signal may contain. Also, the two classes of signals (arrhythmia and normal sinus rhythm) are sampled at different rates. Therefore, re-sampling is also required.

2.3.1 Preprocessing

Since the signals considered for analysis are sampled at different rates, it is necessary to choose a common sampling rate and re-sample the signals. We have chosen 250 Hz as the common sampling rate, and both signals are re-sampled using standard re-sampling techniques [11]. Also, the signals chosen are from an open source database, and might contain noise, artifacts, and power line interference. It is, therefore, necessary to preprocess the signal. Some basic filters [12] have been used here for noise and artifact filtering.

2.3.2 R-point Detection

The R-wave in the QRS complex of ECG has a high amplitude and an easily detectable peak. The R-point is, therefore, chosen as a characteristic point for registration. A number of algorithms are being reported in the literature for the detection of R-point. The Pan-Tompkins algorithm (1985) is a popular approach for QRS detection, which is computationally simple and, hence, takes less time to run on a computer. In addition to this method, there is a method using the quadratic spline-based discrete wavelet transform [6] that detects the beats accurately, but this method is computationally exhaustive. We have chosen the Pan-Tompkins method due to its computational simplicity and ease in implementation. An extended version of the Pan-Tompkins algorithm consists of the following steps.

1. Compute the first derivative of ECG, and find its absolute value.
2. Smooth this signal by passing through a moving average filter as follows.

$$y(n) = \frac{1}{4}\{x(n) + 2x(n-1) + x(n-2)\}, \tag{2.1}$$

where$x(n)$and $y(n)$ represent the input and output of the smoothing filter.
3. Compute the derivative of the smoothened signal and its absolute value.
4. Smooth the signal obtained from step 3 using the filter in Eq. (2.1).
5. Sum the signal obtained from steps 2 and 4.
6. Threshold the signal obtained from step 5, and obtain square pulses.
7. Compensate for the group delay due to the involved filters by advancing in time.

The derivative gives the slope information, whereas smoothing removes high-frequency noise. The above operations are multistage filtering methods with a non-linear operation in between, which yields the R-point.

2.3.3 DWT Computation

Though Fourier analysis [12] is a traditional tool for the analysis of global frequencies present in the signal, it lacks in temporal resolution due to the increased frequency resolution. Some frequency resolution can be exchanged to get better time resolution. This exchange is performed by defining short duration waves called mother wavelet functions so that the given signal for analysis is projected on this basis function. In traditional Fourier transform, the data is projected on sinusoidal basis functions which extend the span of time domain, i.e., $-\infty$ to $+\infty$. The wavelet basis function [13] is parameterized by the translation 'b' and dilation 'a,' such basis function is given by,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right). \tag{2.2}$$

Equation (2.2) provides a basis for wavelet transformation. The ECG signals are decomposed for translation and dilation in order to get a multi-resolution representation. This is the case of continuous wavelet transform. This transform is made discrete using a dyadic grid scale in order to get a discrete wavelet transform (DWT) [14]. Such DWT at scale 2^{-m} and time location n is given by

$$\psi_{m,n}(t) = 2^{\frac{m}{2}} \cdot \psi(2^{-\frac{m}{2}} \cdot t - n). \tag{2.3}$$

The dyadic grid sampled DWT are generally orthonormal. Using the basis function in Eq. (2.3), the DWT can be expressed as the inner product between the ECG signal $x(t)$and the basis function as

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt \tag{2.4}$$

$T_{m,n}$ is the wavelet coefficient at scale (or dilation) m and location (or translation) n, and it provides the detail (fine information) present in the signal.

The dyadic grid sampled orthonormal discrete wavelets are associated with scaling functions and their dilation equations. The scaling function is associated with signal smoothing and has the same form as the wavelet. It is given by,

$$\emptyset_{m,n}(t) = 2^{-m/2}\emptyset(2^{-m} \cdot t - n), \tag{2.5}$$

where $\emptyset_{m,n}(t)$ has the property $\int_{-\infty}^{\infty} \emptyset_{0,0}(t)dt = 1$.

Often $\emptyset_{0,0}(t)$ is referred to as the father scaling function or father wavelet. The scaling function is orthogonal to the translations of itself, but not to dilations of itself. The smoothing of the signal (or the coarse details or the envelope of the signal) is obtained by convolving the scaling function with the signal, and the obtained samples are called approximation coefficients and are defined as

$$S_{m,n} = \int_{-\infty}^{\infty} x(t)\emptyset_{m,n}(t)dt. \tag{2.6}$$

A continuous approximation of the signal can be obtained at scale m using following equation,

$$x_m(t) = \sum_{n=-\infty}^{\infty} S_{m,n}\emptyset_{m,n}(t), \tag{2.7}$$

where $x_m(t)$ is a smooth, scaling function-dependent version of the signal at scale m. Using both approximation and wavelet (detail) coefficients, the signal can be expressed as follows

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m_0,n}\emptyset_{m_0,n}(t) + \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t). \tag{2.8}$$

From Eq. (2.8), we can see that the original continuous signal is expressed as a combination of an approximation of itself at arbitrary index, m_0 added to a succession of signal details from scales m_0 to negative infinity. The signal detail at scale m is given by,

$$d_m(t) = \sum_{n=-\infty}^{\infty} T_{m,n}\psi_{m,n}(t). \tag{2.9}$$

From Eqs. (2.7) and (2.9), we can write

$$x(t) = x_{m_0}(t) + \sum_{m=-\infty}^{\infty} d_m(t). \tag{2.10}$$

From Eq. (2.10), it easily follows that

$$x_{m-1}(t) = x_m(t) + d_m(t). \tag{2.11}$$

From Eq. (2.11), we can see that if we add the signal detail at an arbitrary scale to the signal approximation at the same scale, we get the signal approximation at an increased resolution. Hence, wavelet transformation provides multi-resolution analysis (MRA) capability.

In this work, different basis functions are used. They are Daubechies-4, Daubechies-6, Daubechies-8, Symlet-2, Symlet-4, Symlet-6, Symlet-8, Coiflet-2 and Coiflet-5. All the considered wavelet families are orthogonal.

The frequency components in each of the sub-bands are shown in Fig. 2.2. Since the sampling frequency of the signal under study is 250 Hz, the maximum frequency contained by the signal will be 125 Hz. Therefore, in the first level, approximation will consist of 0–62.5 Hz frequencies, whereas first level detail consists of 62.5–125 Hz frequencies.

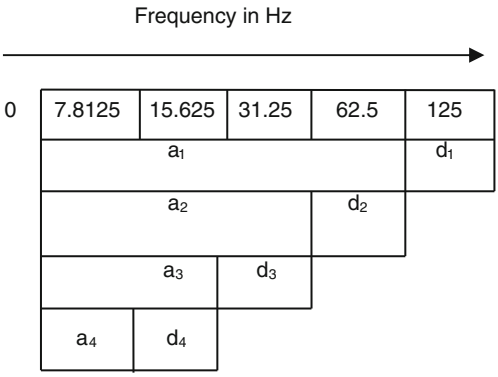


Fig. 2.2 Wavelet decomposition: Distribution of frequencies in various sub-bands

2.3.4 Sub-band Principal Component Analysis

There will be a large number of DWT coefficients in every sub-band of the ECG. If all these coefficients are considered, they will create a large computational burden on the classifier. Therefore, it is wise to represent these coefficients by fewer components. In this study, we have used PCA [15] to reduce the number of features in each of the sub-bands of interest. We identified four sub-bands based on the frequency present in the signal. The four sub-bands are 2nd-level detail, 3rd-level detail, 4th-level detail, and 4th-level approximation. Each of these sub-band wavelet coefficients is subjected to PCA, and the components are chosen such that they will contain 98 % or more of the total energy present in that sub-band.

Mathematically, PCA projects the data from the original coordinate system to a new coordinate system in which the first coordinate corresponds to the direction of maximum variance, and successive coordinates correspond to the directions in decreasing order of variance. Some directions contribute less variability, and those directions need not be preserved in our representation. In the new coordinate system, the axes are called principal components (PCs). A bound of 98 % containment of total variability of segmented ECG is used as a threshold on the total variance in all the considered PCs. PCA consists of following steps.

Compute data covariance matrix as

$$C = \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T, \tag{2.12}$$

where x_i represents the ith pattern $\overline{x}$ represents the pattern mean vector, and N is the number of patterns.

Compute the matrix V of Eigen vectors and diagonal matrix of Eigen values D as

$$V^{-1}CV = D. \tag{2.13}$$

The Eigen vectors in V are sorted in descending order of Eigen values in D, and the data is projected on these Eigen vector directions by taking the inner product between the data matrix and the sorted Eigen vector matrix.

2.3.5 Statistical Test

The DWT features in compact supported basis space provide sparser representation for ECG in sub-bands. When PCA is applied on sub-bands, it should provide higher compression, and the method is more meaningful. Therefore, it is expected for the principal components of DWT features to provide better statistical significance than time domain principal components. Both time domain features and DWT features are compared against the two classes of signals for equality of class group means using independent sample t test [16].

2.3.6 Classification

The significant DWT features obtained from statistical tests are used for subsequent pattern classification. We have used three classifiers, Gaussian mixture model (GMM), error back propagation neural network (EBPNN), and support vector machine (SVM).

2.3.6.1 Gaussian Mixture Model

We have a two-class pattern classification of ECG into normal sinus rhythm and arrhythmia classes. The GMM assumes that the features are normally distributed, and each class is characterized by its mean (μ_k) and covariance matrix (Σ_k). Since we have applied an orthogonal transformation in compact supported basis space, the features are likely to be uncorrelated. The off-diagonal elements in the covariance matrix are approximately zero. The probability density function of GMM for every sample belonging to a given class k, is given by

$$P(x_n|\omega_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x_n - \bar{x}_k)^T |\Sigma_k|^{-1}(x_n - \bar{x}_k)\right\}, \tag{2.14}$$

where

$$\bar{x}_k = \frac{1}{|X_k|} \sum_{x_n \in \omega_k} x_n \tag{2.15}$$

and

$$\Sigma_k = \frac{1}{|X_k|}\sum (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T = diag\left(\sigma_{ii}^2\right), 1 \le i \le d \tag{2.16}$$

The corresponding posterior probabilities are given by Bayes's rule as

$$P(\omega_k|x_i) = \frac{P(x_i|\omega_k)}{\sum_{k=1}^{2} P(\omega_k)P(x_i|\omega_k)} \tag{2.17}$$

Initially, the mean and covariance matrices are assigned with some random values. The values are updated using an expectation maximization (EM) algorithm and a maximum likelihood estimation method. The re-estimation formulae are as follows.

$$\widehat{\mu}_j = \frac{\sum_{i=1}^{N} x_i \cdot P\left(\omega_j|x_i\right)}{\sum_{i=1}^{N} P\left(\omega_j|x_i\right)} \tag{2.18}$$

$$\widehat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} (x_i - \widehat{\mu}_j)^2 P\left(\omega_j|x_i\right)}{\sum_{i=1}^{N} P\left(\omega_j|x_i\right)} \tag{2.19}$$

$$p\left(\widehat{\omega}_j\right) = \frac{1}{N}\sum_{i=1}^{N} P\left(\omega_j|x_i\right) \tag{2.20}$$

An initial model having parameters (μ_k, Σ_k) is assumed from the data. The EM algorithm has two steps: an E step and an M step. During the E step, the class conditional density is computed according to Eq. 2.14 and the posterior density is also computed according to Eq. 2.17. During the M step, the model parameters are re-estimated according to Eqs. 2.18−2.20. The process is continued until the new model remains almost identical to the previous model. At this point, the algorithm is said to be converged. The GMM optimizes the following objective function,

$$J = \prod_n \sum_k p(\omega_k)p(x_n|\omega_k). \tag{2.21}$$

GMM minimizes the product over all the patterns, the total class conditional density weighted with the respective prior probability.

2.3.6.2 Error Back Propagation Neural Network

An error back propagation neural network [17] is used in our study. The neural network is trained on the training set of the data such that the weights get updated recursively with respect to the patterns. This is also an optimization problem where following objective function is minimized.

$$J(\omega) = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left\{y_k(x_n,\omega) - t_k^n\right\}^2, \tag{2.22}$$

where $y_k(x_n, \omega)$ is the network response for the kth class neuron in the output layer and t_k^n is the target for kth class of nth observation feature vector.

The gradient descent method is used in the analysis to adapt the network weights. We have used adaptive serial learning from the data using minimum mean square error criterion. Once the network is trained, the test signal is fed to the neural network and the data is classified to one of the two predefined classes.

2.3.6.3 Support Vector Machine

SVM [18] is a single layer, highly nonlinear network which optimizes the class separation boundary such that the distance from the features falling in a given class to the hyperplane gets simultaneously maximized. SVM is a supervised classifier that has generalization ability [19] in the sense that it can classify an unseen pattern correctly. If $(x_i, y_i), i = 1 : N$ is the data set, x_i is the ith pattern point, and y_i is the corresponding class label, then let $c+$ and $c-$ be the centroids for two classes in binary classification problem. The classifier output will be

$$y_i = sgn((x - c) \cdot w) = sgn((x.c+) - (x.c-) + b) \tag{2.23}$$

where

$$b = \frac{1}{2}(||c\ -||^2 - ||c\ +||^2). \tag{2.24}$$

The optimal hyperplane separating the two classes and satisfying condition given in Eq. 2.23 is

$$minimize_{w,b}\frac{1}{2}\|w\|^2 \tag{2.25}$$

such that

$$y_i((w.x_i) + b) \geq 1, i = 1, \ldots N. \tag{2.26}$$

The Lagrangian dual of Eq. 2.25 is a quadratic programming problem used to find the optimal hyperplane separating the two classes.

2.3.7 k-fold Cross Validation

k-fold cross validation [20] is used for $k = 3$. Here, the total number of samples are sub-sampled into three (k) sets; one set is used for testing, whereas the other two sets are used to train the classifier. The process is repeated two more times such

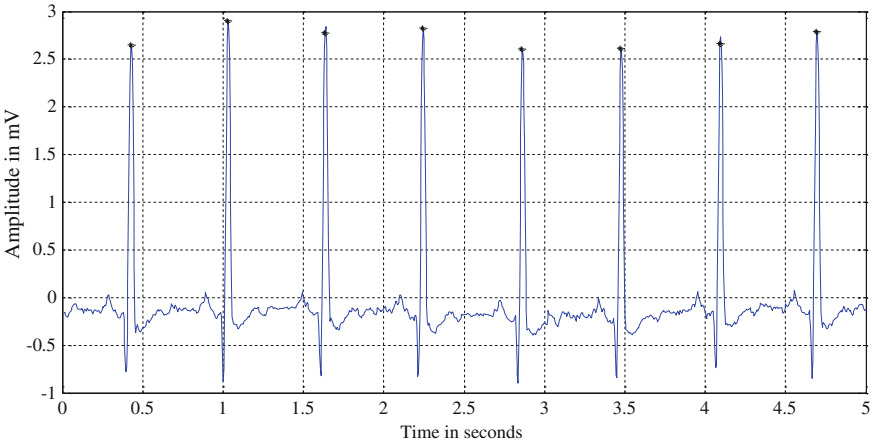


Fig. 2.3 R-point detection in normal sinus rhythm signal

that every sub-partition is used as a testing set and the rest are used for classifier training. The three accuracies are averaged to estimate the average classifier performance. Using *k*-fold cross validation, the bias in choosing the samples from the population can be overcome.

2.4 Results and Discussion

In order to apply our proposed methodology, a two-class ECG classification problem has been formulated based on the MIT BIH arrhythmia and MIT BIH normal sinus rhythm datasets (described in Sect. 2.2). The Pan-Tompkins algorithm is used to detect the R-point because of its simplicity and accuracy. The detection of the R-point is shown in Fig. 2.3, where the detected R-point is marked with a black asterisk. It can be seen from Fig. 2.3 that the Pan-Tompkins method detects the R-point with good precision. In fact, the Pan-Tompkins algorithm is a multistage filtering (differentiation, smoothing, etc.) and a nonlinear element (rectification) between the linear operations in the algorithmic steps.

Once the R-point is detected, a window (or one segment) of 200 samples is extracted by choosing 99 points on the left of the R-point, and 100 points on the right of the R-point and used for further classification. The power spectral density (psd) is computed using an autoregressive method and is plotted for a normal sinus rhythm and arrhythmia signal in Fig. 2.4. The objective of computing psd. is to identify the frequencies of interest so that they can descriminate the two kinds of beats (normal sinus rhythm and arrhythmia) distinctly. We can observe from Fig. 2.4 that frequencies in the range of 0−50 Hz can be used for that purpose. Hence, by referring to Fig. 2.2 and the graph in Fig. 2.4, it is observed that the subbands of interest are detail 2, detail 3, detail 4, and approximation 4.

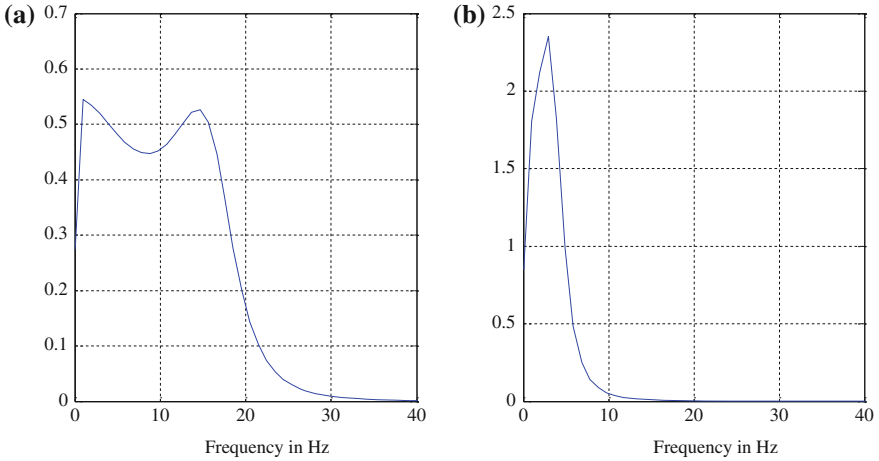


Fig. 2.4 The power spectrum of **a** Normal sinus rhythm, **b** Arrhythmia signal

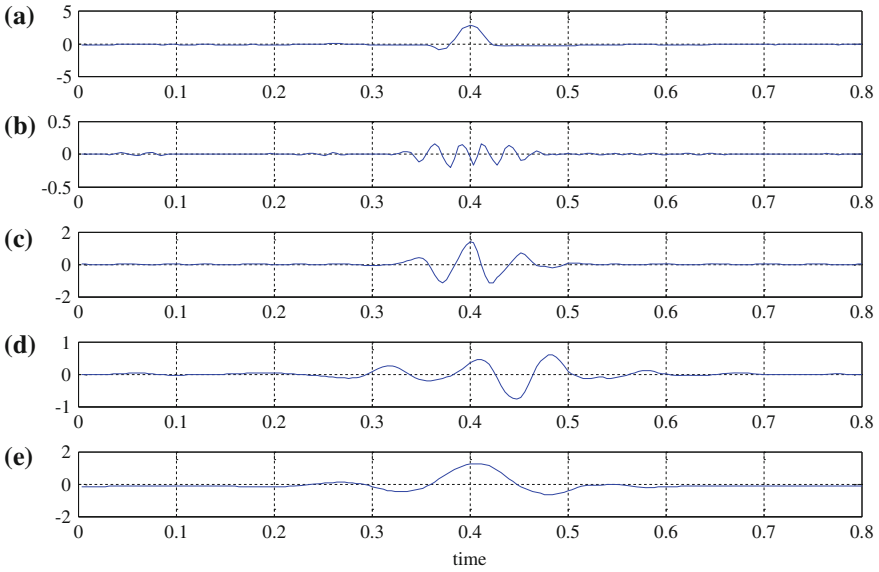


Fig. 2.5 DWT decomposition of normal sinus rhythm signal **a** Original signal, **b** Detail-2, **c** Detail-3, **d** Detail-4, **e** Approximation-4 signals

The DWT using Daubechies-4 wavelet is shown for normal sinus rhythm signal in Fig. 2.5. We can see that all the sub-bands of interest contain some signal component that can be used for performing classification. Figure 2.6 shows the DWT computed using the Daubechies-4 wavelet for an arrhythmia signal. We can see that the DWT decompositions of the two signals look different. If these

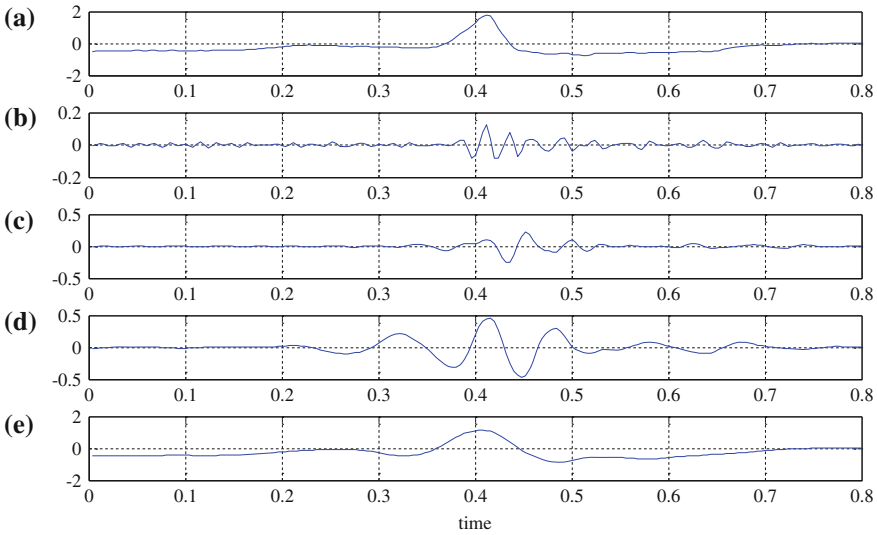


Fig. 2.6 DWT decomposition of arrhythmia signal **a** Original signal, **b** Detail-2, **c** Detail-3, **d** Detail-4, **e** Approximation-4 signals

coefficients are compressed and represented by fewer components, they can be used as features for subsequent classification. The reason for compression is that using fewer components reduces the computational burden on the classifier.

PCA is applied on each sub-band of interest and different wavelet basis functions are used. We use the Daubechies-4, Daubechies-6, Daubechies-8, Symlet-2, Symlet-4, Symlet-6, Symlet-8, Coiflet-2, and Coiflet-5 wavelet basis functions. PCA is an orthogonal transformation which maps the data into the directions of maximum variability. Since DWT is a compact supported basis function, having sparse representation, PCA on it should provide higher compression. The number of principal components is chosen so that the components contain 98 % variability of the respective sub-band. For different basis functions, the number of principal components chosen from each of the sub-bands and the total variability of the data contained is shown in Tables 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9.

The Eigen value profile for the Daubechies-4 wavelet is shown in Fig. 2.7.

After PCA, the compressed components are subjected to a statistical significance test. Based on the p-value provided by the statistical test, the significance of components is decided, and the significant features are used for further classification. The statistical significance test is performed for every basis function and the results are tabulated in Table 2.10. It is observed from Table 2.10 that the DWT domain principal components are more significant than time domain components on the basis of the statistical test.

The GMM classification for the Daubechies-6 wavelet basis function features is shown in Fig. 2.8. We can see that the log likelihood in the graph increases and becomes steady after convergence of the algorithm.

Table 2.1 DWT decomposition using Daubechies-4

Sub-band	Number of PCs	Energy in %
Detail 2	5	98.1770
Detail 3	3	99.2262
Detail 4	2	98.5930
Approximation 4	3	98.8920

Table 2.2 DWT decomposition using Daubechies-6

Sub-band	Number of PCs	Energy in %
Detail 2	3	98.1957
Detail 3	2	99.1075
Detail 4	2	99.4899
Approximation 4	2	98.3697

Table 2.3 DWT decomposition using Daubechies-8

Sub-band	Number of PCs	Energy in %
Detail 2	7	98.1937
Detail 3	2	99.4728
Detail 4	2	99.0444
Approximation 4	2	98.5753

Table 2.4 DWT decomposition using Symlet-2

Sub-band	Number of PCs	Energy in %
Detail 2	2	98.2329
Detail 3	3	99.4611
Detail 4	2	99.3243
Approximation 4	2	98.4532

Table 2.5 DWT decomposition using Symlet-4

Sub-band	Number of PCs	Energy in %
Detail 2	10	98.0661
Detail 3	2	99.0872
Detail 4	2	99.6493
Approximation 4	2	98.6254

The classification by EBPNN is shown in Fig. 2.9, which shows that the total mean-squared error reduces with epochs. We can observe that the EBPNN algorithm converges in 19 epochs for Daubechies-6 wavelet family features. Figure 2.10 shows SVM classification with linear kernel. Since the data is linearly separable, we have used only linear kernel SVM.

Table 2.6 DWT decomposition using Sym-6

Sub-band	Number of PCs	Energy in %
Detail 2	3	98.1292
Detail 3	2	98.8807
Detail 4	2	99.5307
Approximation 4	2	98.1622

Table 2.7 DWT decomposition using Sym-8

Sub-band	Number of PCs	Energy in %
Detail 2	10	96.6749
Detail 3	2	98.8332
Detail 4	2	99.6216
Approximation 4	3	98.6306

Table 2.8 DWT decomposition using Coiflet-2

Sub-band	Number of PCs	Energy in %
Detail 2	10	96.9230
Detail 3	2	99.1172
Detail 4	2	96.6110
Approximation 4	3	98.6443

Table 2.9 DWT decomposition using Coiflet-5

Sub-band	Number of PCs	Energy in %
Detail 2	10	96.0415
Detail 3	2	98.2607
Detail 4	2	99.6023
Approximation 4	2	98.5899

Table 2.11 shows classification accuracies of various schemes using different wavelet basis functions. It can be observed from Table 2.11 that EBPNN provides higher accuracy than GMM and SVM leads to the highest accuracy. Amongst various wavelet families, it can be noted that Symlet-2 consistently performs better for all the classifiers and has the highest possible accuracy.

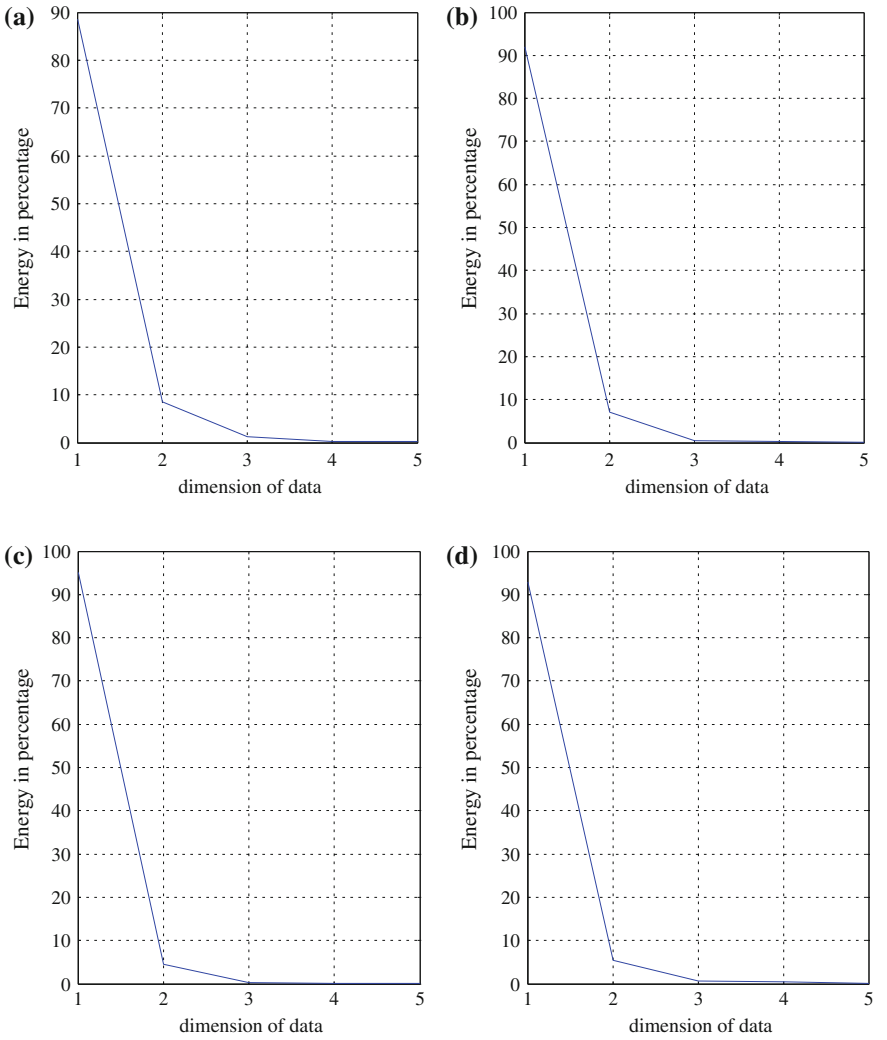


Fig. 2.7 PCA on DWT sub-bands **a** Detail 2, **b** Detail 3, **c** Detail 4, and **d** Approximation 4, decomposition using the db6 wavelet

Table 2.10 Statistical significance test for time domain and Daubechies-6 based DWT sub-band features

Time domain PCs	Statistical significance		DWT domain PCs	Statistical significance	
	t	p		t	p
1	−61.998	0.000	1	−70.9029	0.000
2	−156.68	0.000	2	−52.5739	0.000
3	0.1702	0.8650	3	−3.4508	0.000
4	−1.7762	0.0763	4	−28.7637	0.000
5	0.4157	0.6778	5	37.3751	0.000
6	0.4035	0.6867	6	−27.1491	0.000
7	−0.1840	0.8541	7	22.3390	0.000
8	−0.3165	0.7517	8	10.5710	0.000
9	0.8103	0.4181	9	−93.7320	0.000

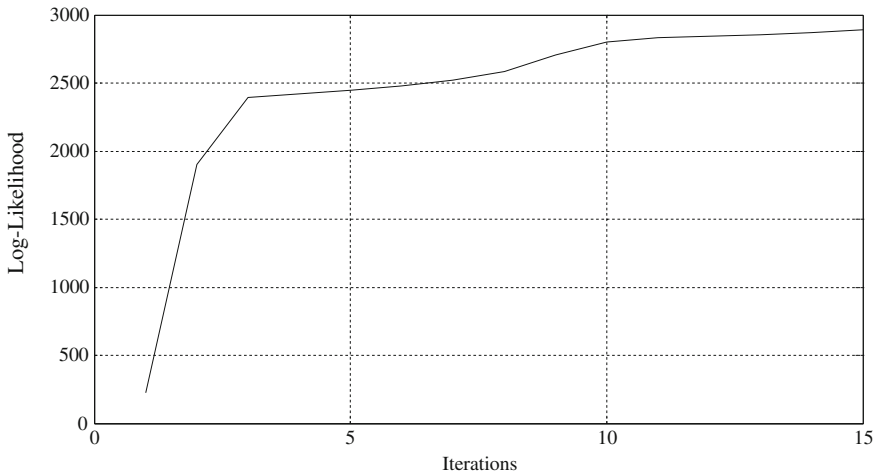


Fig. 2.8 GMM classification, log-likelihood increasing with iterations

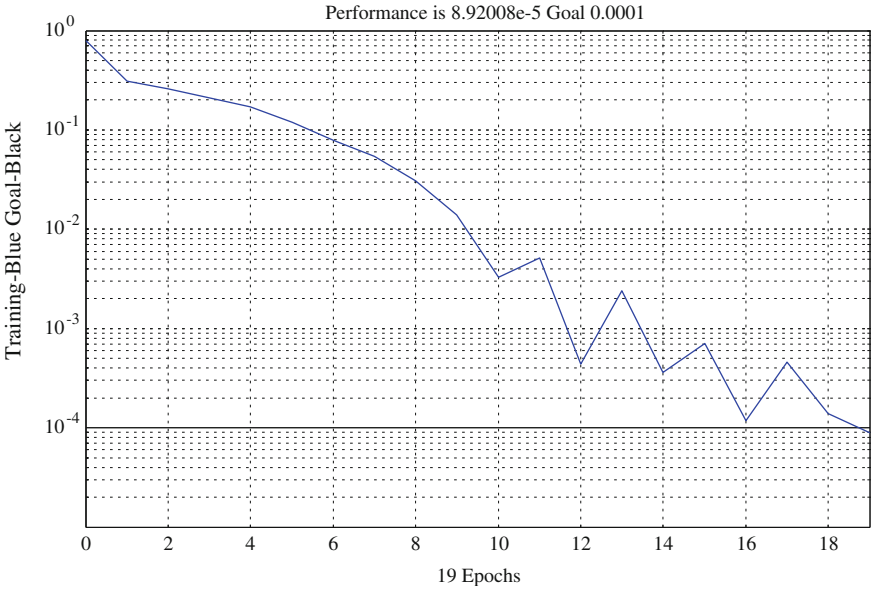


Fig. 2.9 EBPNN classification, network converging in 19 epochs for Daubechies-6 features

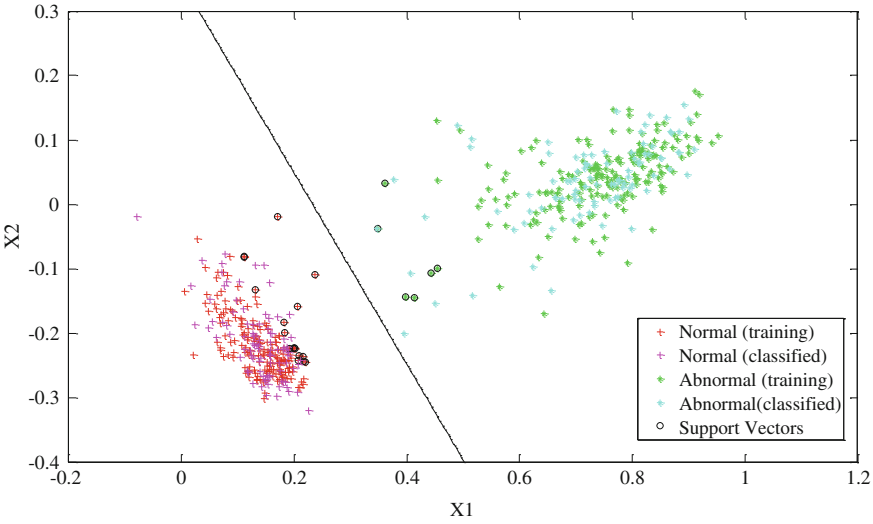


Fig. 2.10 SVM classification using linear kernel for Daubechies-6 wavelet features

Table 2.11 The average classification accuracies of various classifiers using different wavelet basis functions

Wavelet basis function	GMM	EBPNN	SVM
Db4	87.36	93.41	95.60
Db6	88.52	94.23	96.72
Db8	85.27	93.78	95.39
Sym2	89.78	95.17	97.23
Sym4	85.18	93.79	96.81
Sym 6	88.21	94.18	97.14
Sym8	84.39	93.89	96.68
Coif 2	85.91	92.96	95.95
Coif 5	85.16	92.83	95.42

2.5 Conclusion

In this chapter, a systematic approach is developed for screening arrhythmia and normal sinus rhythm from their ECG profiles. We have extracted time frequency features using various basis functions, including Daubechies, Symlet, and Coiflet wavelet families. PCA is applied on time frequency sub-band features and, in this compact supported basis space, higher compression is expected. Based on our experiments, we have determined that different basis functions distribute energy in different sub-bands in a unique way for a given wavelet. Our methodology exploits this energy distribution so that the features are well represented, thus resulting in higher accuracy. These time-frequency features are markers of disease, since these features are able to discriminate the data into two classes. As a future direction, other time-frequency representations can be used to see how the energy compaction is achieved. In addition, various other dimensionality reduction techniques can be used for performance. The machine-learning methodology given in this chapter can be used efficiently in telemedicine systems to identify abnormal events in the ECG signals so that emergency cases can be identified and such patients can be attended for critical care.

References

1. Fauci AS, Braunwald E, Kesper DL, Hauser SL, Longo DL, Jamesonn JL, Loscalzo J (2008) Harrison's principles of internal medicine, 17th edn. Mc-Graw Hill, New York
2. Park K (2005) Park's textbook of preventive and social medicine, 18th edn. Banarsidas Bhanot publishers, India
3. Guyton AC, Hall JE (2006) Textbook of medical physiology, 11th edn. W. B Saunders Co, Philadelphia
4. Pan J, Tompkins WJ (1985) A real time QRS detection algorithm. IEEE Trans Biomed Eng 32(3):230–236
5. Li C, Zheng C, Tai C (1995) Detection of ECG characteristic points using wavelet transforms. IEEE Trans Biomed Eng 42(1):21–29

6. Martinez JP, Almeida R, Olmos S, Rocha AP, Laguna P (2004) A wavelet based ECG delineator: evaluation on standard databases. IEEE Trans Biomed Eng 51(4):570–581
7. Throne RD, Jenkins JM, Winston SA, DiCarlo LA (1991) A comparison of four new time domain techniques for discriminating monomorphic ventricular tachycardia from sinus rhythm using ventricular waveform morphology. IEEE Trans Biomed Eng 38(6):561–570
8. Krasteva V, Jekova I (2007) QRS template matching for recognition of ventricular ectopic beats. Ann Biomed Eng 35(12):2065–2076
9. Martis RJ, Chakraborty C, Ray AK (2009) A two stage mechanism for registration and classification of ECG using gaussian mixture model. Pattern Recogn 42(11):2979–2988
10. Martis RJ, Krishnan MM, Chakraborty C, Pal S, Sarkar D, Mandana KM, et al (2012) Automated screening of arrhythmia using wavelet based machine learning techniques. J Med Syst 36(2):677–688
11. Vaidyanathan PP (2004) Multirate systems and filter banks. Pearson education (Asia) Pte. Ltd, Delhi
12. Oppenheim AO, Schaffer RA (2003) Discrete time signal processing. Mc-Graw Hill edition, New York
13. Addison PS (2005) Wavelet transforms and the ECG: a review. Physiol Meas 26(5): R155–199
14. Strang G, Nguyen T (1996) Wavelets and filter banks. Willesley Cambridge Press, MA
15. Duda R, Hart P, Stork D (2001) Pattern classification, 2nd edn. Wiley, New York
16. Gun AM, Gupta MK., Dasgupta B (2008) Fundamentals of statistics (Vol. I and II), 4th edn. World Press Private Ltd, Kolkata
17. Bishop C (1995) Neural networks for pattern recognition. Oxford University press, New York
18. Christianini N, Taylor JS (2000) An introduction to support vector machines and other kernel based learning methods, Cambridge university press, Cambridge
19. Gunn S (1998) Support vector machines for classification and regression, Technical report, University of Southampton
20. Schneider J (1997) Cross validation. http://www.cs.cmu.edu/∼schneide/tu5/node42.html. Accessed 15 Aug 2010

Chapter 3
Application of Fuzzy Logic Control for Regulation of Glucose Level of Diabetic Patient

K. Y. Zhu, W. D. Liu and Y. Xiao

Abstract Diabetes can lead to many complications. If a patient cannot control his or her glucose level properly, he or she may suffer serious consequences. The result may be ketosis, which is normally due to an increase of acetone (a toxic acid product) and may lead to a situation such as diabetic coma. A fuzzy logic control system for the regulation of glucose level for diabetic patients is proposed in this chapter. A mathematical model describing the relationship between the human glucose level, insulin, and food is first presented. Then, a generalized fuzzy logic controller, including a set of fuzzy logic rules, is introduced to regulate glucose levels for diabetic patients. Following the fuzzy logic controller, simulation is presented. The results show that the fuzzy logic control is effective for handling the glucose level based on feedback scheme.

3.1 Introduction

It is known that the number of diabetic patients is increasing in the world, especially in the developed countries and this increase is a challenging problem for healthcare providers in the countries concerned. Diabetes, also called diabetes mellitus or diabetes insipidus, is caused by insufficient production of insulin (either absolutely or relative to the body's needs), production of defective insulin (which is uncommon), or the inability of cells to use insulin properly. Diabetes can occur as one of two distinct types: Type I (also called as insulin dependent) or Type II (called as

K. Y. Zhu (✉) · W. D. Liu
Department of Electronic and Computer Engineering, Ngee Ann Polytechnic, Singapore 599489, Republic of Singapore
e-mail: zku2@np.edu.sg

Y. Xiao
Information and Communication Department, Shenyang University of Chemical Technology, Shenyang, China

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_3,

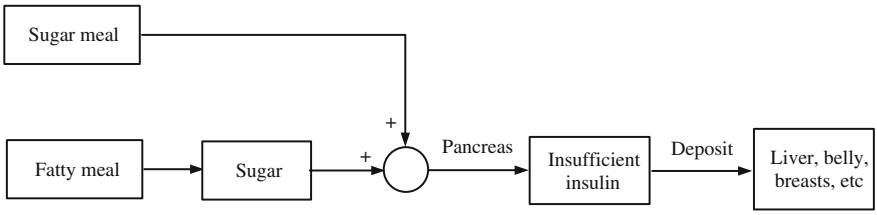


Fig. 3.1 Relationship among sugar, insulin and fat

non-insulin-dependent) diabetes. The main objective for treating diabetes is regulation of elevated blood sugars (glucose) without causing the blood sugar level to drop too low. Both types of diabetes are treated with adequate exercise and special diets. Type I diabetes is also treated with insulin, whereas Type II diabetes is treated with weight reduction in addition to adequate exercise and a special diet. When these methods are not successful, oral medications are commonly adopted. If oral medications are ineffective, insulin medications may be prescribed.

Insulin is produced by the pancreas, the key which opens the little holes on the cell membranes for the glucose to enter the cells in a normal way. Glucose is essential energy, like fuel, required for every cell of our body. However, a diabetic patient does not have enough insulin to open the little holes of the cell membranes. This lack of insulin can prevent the glucose circulating in the blood stream from entering cells. As a result, the blood glucose level increases and spills into the urine. A diabetic patient without enough insulin is like a thirsty sailor in the ocean who is surrounded by water but cannot drink it. For the diabetic his/her body cells are surrounded by sugar, but cannot consume it as they need the insulin to make the sugar enter each cell and be consumed.

Figure 3.1 shows the relationship among sugar, insulin, and fat in a diabetes patient. This figure shows why a strict diet and insulin are so important to the diabetes patients.

There are three types of insulin available to healthcare providers and diabetic patients: regular insulin, Lente insulin, and Humulin. Regular insulin is extracted from the pancreas of pork and beef. The effect from this insulin is rapid and lasts from 4 to 6 h. Lente insulin comes from pork and beef (with a type of oily substance for slower reabsorption) and its effects last longer than the effects from regular insulin. Humulin insulin is a mixture of regular and Lente insulin. A standard Humulin insulin syringe contains a mixture of 70 % Lente and 30 % regular insulin. At the present time, Human insulin is the most used because some patients will develop a resistance to insulin extracted from animals due to the minor amino acid difference between animal insulin and human insulin. We can manufacture Humulin insulin by synthesis.

Uncontrolled glucose level for diabetic patient may cause artherosclerosis, or hardening the arteries and form blockage in the circulation, affecting the cardiac arteries, brain, kidneys, liver, and feet. From the artherosclerosis, it is known that heart attacks, strokes, and liver and kidney failure frequently occur. High glucose

levels can also cause the formation of microscopic aneurysms on the retina, originating hemorrhage, and decrease of vision and can consequently cause blindness. Circulation in the feet may also decrease, leading to artery hardening, ulcers, infection, and even gangrene.

Although diabetes could cause severe damage to the human body, if the patients control the disease by diet or insulin properly, all the complications could be avoided or prevented. Research shows that we could reduce the mortality of diabetic and non-diabetic ICU (Intensive Care Unit) patients by up to 50 % [1] through tight control of blood glucose level. In order tightly control the blood glucose level, diabetic patients need to monitor their daily intake and activity strictly; this step could help maintain their blood sugar at adequate levels. Unfortunately, this strict lifestyle may cause an 'institutional' psychology, and it may be difficult to consistently maintain a strict daily regimen over several years.

We can use the devices to measure the glucose level and administer insulin, but the measure and inject are two separate procedures which have no automated interface. It is difficult for patients to perform these two procedures manually every day, and the procedures may also introduce errors due to human miscalculation and limitations. Furthermore, tightly controlling the blood glucose levels to the basal level of 4.5 $mmol^{-1}$ could significantly reduce the damage caused by long-term exposure to elevated glucose levels. The Diabetes Control and Complications Trial (DCCT), a study which followed almost 1,500 people with Type I diabetes for 10 years, proved that tight control over blood glucose levels could reduce eye disease in 62 % of patients, kidney disease in 56 % of patients, and nerve damage in 60 % of patients with Type I diabetes. On the other hand, the UK Prospective Diabetes Study (UKPDS), which followed over 5,000 patients with Type II diabetes in 23 clinics in Europe for 20 years, proved for the first time that better blood glucose control reduces the risk of eye disease by a quarter, and early kidney damage by a third in individuals with Type II diabetes. In these studies, restricting control meant keeping the blood glucose levels as close to normal as possible, which could extend the life expectation and provide protection against long-term health risks for patients [1–3].

A typical day for an insulin therapy diabetic patient might involve injecting long-acting insulin approximately three times and injecting rapid-acting insulin before meals to reduce the post-meal blood glucose spike. Moreover, most commonly available glucose sensing devices are invasive and measure the blood glucose content by a small finger-prick blood sample. The pain involved in the finger pricking may cause diabetic patients to measure blood sugar level more infrequently. Although new technologies have brought us products such as the Continuous Glucose Monitoring System (CGMS), which could provide the updated glucose level every 5 min, for up to 72 h and an insulin pump which could inject rapid-acting insulin continuously over 24 h. These two technologies will be discussed in later sections.

In 2006, the FDA approved the MiniMed Paradigm® REAL-Time Insulin Pump with Continuous Glucose Monitoring System. This treatment system is the first to provide real-time, continuous glucose monitoring. The MiniMed Paradigm REAL-

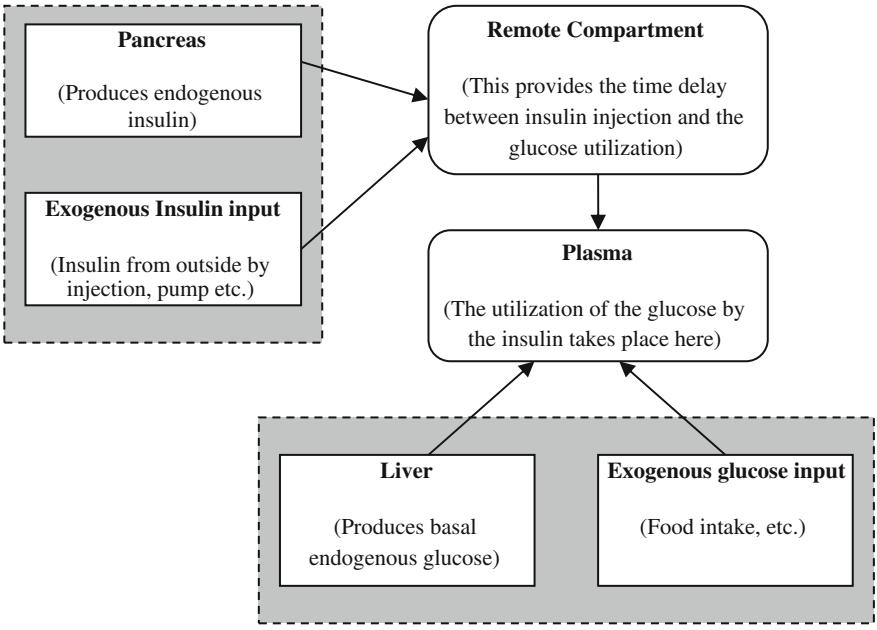


Fig. 3.2 Physiological block diagram of the modeled system

Time system is composed of two components, a REAL-Time Continuous Glucose Monitoring (CGM) System, and a MiniMed Paradigm insulin pump, but the amount of insulin injected is determined by the patients. Patients will take immediate action to improve their control of glucose control after the glucose information is displayed on the insulin pump. Integrating an insulin pump with real-time CGM is an attempt to develop a closed-loop insulin injection system that may mimic some functions of the human pancreas.

In our research, we attempted to introduce a closed-loop system based on a fuzzy logic control scheme, which could effectively control a diabetic patient's blood sugar level. This system may help patients more fully engage in the 'normal' routines of life with reduced risk of long-term adverse end-results. Figure 3.2 shows the block diagram of the modeled dynamics for the human glucose regulatory system. Our control system will be based on this glucose regulatory system. The mathematical detailed of this model will be discussed in Sect. 3.3.

3.2 Mathematical Model of Glucose Regulatory System

A simple model which could capture all the essential dynamic behaviors is presented below. This model does not require unavailable data and is suitable to a wide variety of subjects. There are also comprehensive models available, but these

models require several time points of input to generate the insulin infusion profile and are unsuitable for real-time control.

A well-known and physiologically verified model originated from the work of Bergman et al. The concept of this model is to use a remote compartment for the storage of insulin to account for the time delay between the injection of insulin and its utilization to reduce blood glucose levels. These mathematical models are:

$$\dot{G} = -p_1 G - X(G + G_B) + P(t), \tag{3.1}$$

$$\dot{X} = -p_2 X + p_3 I, \quad \text{and} \tag{3.2}$$

$$\dot{I} = -n(I + I_B) + u(t)/V_I, \tag{3.3}$$

where G is the concentration of the plasma glucose above the basal level (mmol/L), G_B is the basal level for plasma glucose concentration (mmol/L), i.e. $G + G_B$ is the total glucose in the blood plasma, where $G_B = 4.5$ typically, X is the utilization effect of insulin in a remote compartment $\left(\text{min}^{-1}\right)$, I is the concentration of the plasma insulin above basal level (mU/L), I_B is the basal level for plasma insulin concentration (mU/L), $P(t)$ is the exogenous glucose infusion rate (mmol/L/min), $u(t)$ is the exogenous insulin infusion rate (mU/L/min), p_3 is the subject dependent model parameter (mU/L/min2), V_1 is the insulin distribution volume (L), and n, p_1, p_2 are the subject dependent model parameters $\left(\text{min}^{-1}\right)$.

The parameters, p_1, p_2 and p_3 may be changed to represent different conditions of the glucose regulatory system [4]. The parameter p_1 is the fractional clearance of plasma glucose at basal insulin. For diabetic subjects,

$$p_1 = 0,\ p_2 = 0.025,\ p_3 = 0.000013 \tag{3.4}$$

The model is therefore patient specific and is adapted to each person before a controller is developed.

In some conditions, e.g. ICU patients who have direct arterial/venous lines that bypass the subcutaneous compartment, this model can be simplified as:

$$\dot{G} = -p_1 G - S_I I(G + G_B) + P(t) \tag{3.5}$$

$$\dot{I} = -n(I + I_B) + u(t)/V_I, \tag{3.6}$$

where S_I(L/mU/min) refers to patient specific parameters. In our experiments, these parameters refer only to insulin sensitivity [5–12].

The model shown in Eqs. (3.1) through (3.3) was developed to model insulin sensitivity, a measure of how efficiently the body responds to insulin input after taking a glucose tolerance test (OGTT). The model is simple but accurately represents the essential dynamics of the human glucose regulatory system for a variety of patients. These three equations represent insulin production and infusion, insulin storage in a remote compartment, and glucose input and insulin utilization in a second compartment. Equation (3.1) represents the glucose levels in the blood stream and the dynamics of its reaction with insulin. Equation (3.2)

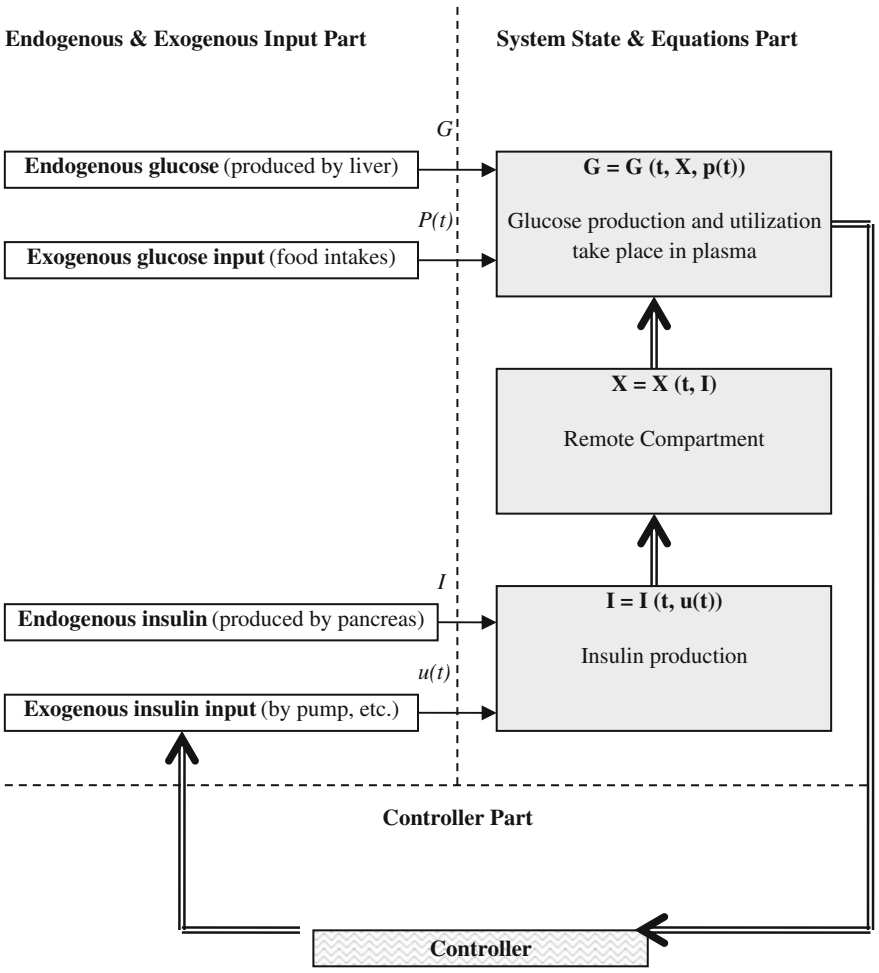


Fig. 3.3 Model of the glucose regulatory system

defines the dynamics and delay in the transport of insulin from the subcutaneous layer to the blood plasma and subsequent utilization. Insulin inputs are either endogenous from the pancreas or exogenous from a pump or injection and represented in Eq. (3.3). Figure 3.3 is graphically outlines the glucose regulatory system for the model with Eqs. (3.1)–(3.3).

The values of n, V_I, G_B, and I_B employed for all simulations are defined, for an average-weighted man as follows:

$$V_I = 12 \text{ L}, \; n = \frac{5}{54} \text{ min}^{-1}, \; G_B = \frac{4.5 \text{ mmol}}{\text{L}}, \; I_B = 15 \text{ mU/L} \tag{3.7}$$

This controller forms a simple feedback loop which employs the blood glucose level above basal, G, and its derivate, $\dot{G}$, as sensor input, and the exogenous insulin infusion rate, $u(t)$, as the control output. There are only two forms of data available to control the system: G and G. Therefore, the controller measures the output from Eq. (3.1) while directly influencing the dynamics in Eq. (3.3) via the control action. In between is the remote compartment represented in Eq. (3.2) which describes the time delay and additional dynamics.

3.3 Fuzzy Logic Control System

Fuzzy logic control systems are essentially rule-based expert systems, which consist of a set of linguistic rules in the form of "IF–THEN." The fuzzy IF–THEN rules are of the following form:

R_i: IF x_1 is F_1^i and…and x_r is F_r^i, THEN y_1 is G_1^i and…and y_m is G_m^i where F_j^i $(j = 1, 2, \ldots r)$ and G_k^i $(k = 1, 2, \ldots m)$ are labels of fuzzy sets characterized by appropriate membership function. $X = (x_1, x_2, \ldots x_r)$ and $Y = (y_1, y_2, \ldots y_m) \in V$ are input and output linguistic variables, respectively, and $i = 1, 2, \ldots u$ means the ith rule. Each of the fuzzy IF–THEN rules defines fuzzy set

$$F_1^i \times F_2^i \times \cdots F_r^i \rightarrow G_1^i + G_2^i + \cdots G_m^i, \tag{3.8}$$

where "+" represents the union of independent variables. Since the outputs of a multi-input and multi-output (MIMO) rule are independent, the general rule structure of a MIMO fuzzy system can be represented as a collection of multi-input and multi-output (MISO) fuzzy systems by decomposing the above rule into m sub-rules with F_i as the single consequence of the ith sub-rule.

Adaptive fuzzy logic control systems consist of a collection of linguistic rules, fuzzy implications, fuzzy model identifications, and an adaptive algorithm. This adaptive fuzzy logic control system can be a two-level system. The first level, or lower level, of the system is a simple fuzzy logic controller. The second level, or higher level, is the fine-tuning system that is used for processes with changing conditions. In a simple fuzzy logic control system, the measured nonfuzzy state variable is compared with a given nonfuzzy set point. Then, the crisp nonfuzzy value is converted into two fuzzy controller inputs, which are an error and a change of error. Through the inference engine and knowledge base of given rules, the expert system can obtain a linguistic value for the controller output. Because in practice it is necessary to calculate the deterministic value of the controller output, a defuzzier, which converts the output fuzzy set to a deterministic or crisp value and send this value to the final control element, is needed.

The block diagram of the fuzzy control system for regulation of glucose level is shown in Fig. 3.4. The learning rule is determined based on the errors, i.e. the error and rate of change of the error for glucose level defined by

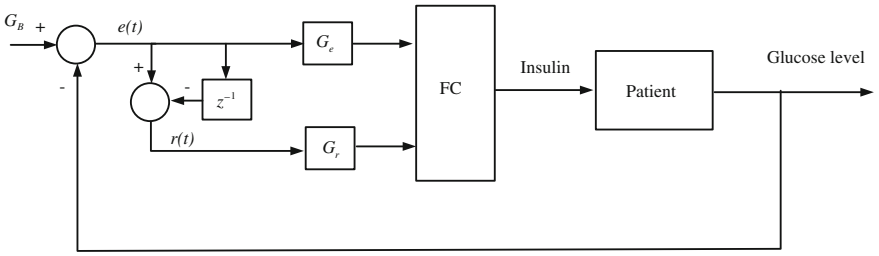


Fig. 3.4 Block diagram of fuzzy control system

$$e(t) = G(t) \tag{3.9}$$

$$r(t) = G(t) - G(t-1), \tag{3.10}$$

where $e(t)$ is the glucose level deviated from its basal level of G_B, $r(t)$ is the rate of change of error at time t.

The fuzzy logic controller is composed of the following six components. Note that $e(t)$ and $r(t)$ are scaled before fuzzification. $E(nT)$ is scaled by a scalar $G_e$while $r(t)$ is scaled by a scalar G_r. Then they are fuzzified by fuzzy sets shown in Figs. 3.5 and 3.6, where L is the interval, $|\mu_e| \leq 1$ and $|\mu_r| \leq 1$, which were memberships of $G_e e(t)$ and $G_r r(t)$, respectively.

The four fuzzy control learning rules are described linguistically and are related to fuzzy sets for an increment of insulin $\Delta u(t)$, which are listed below

Rule 1 IF $G_e e(t)$ is "positive" and $G_r r(t)$ is "positive," then $\Delta u(t)$ is "positive."

Rule 2 IF $G_e e(nT)$ is "positive" and $G_r r(nT)$ is "negative," then $\Delta u(t)$ is "zero."

Rule 3 IF $G_e e(nT)$ is "negative" and $G_r r(nT)$ is "positive," then $\Delta u(t)$ is "zero."

Rule 4 IF $G_e e(nT)$ is "negative" and $G_r r(nT)$ is "negative," then $\Delta u(t)$ is "negative."

Note that the output fuzzy sets describing $\Delta u(t)$ are three singleton fuzzy sets, "positive," "zero," and "negative," as shown in Fig. 3.7.

Learning Rule 1 shows that if the glucose level is above the basal level and is increasing, then $\Delta u(t)$ should be positive and the insulin infusion rate should be increased. Learning Rule 2 says if the glucose level is above the basal one, but is decreasing, $\Delta u(t)$ should be zero, which means the insulin infusion rate should not be changed. Learning Rule 3 states that if the glucose level is lower than the basal level and is increasing, $\Delta u(t)$ should be zero, which means the insulin infusion rate should not be changed. Learning Rule 4 indicates if the glucose level is lower than the basal level and is also decreasing, then $\Delta u(t)$ should be negative, i.e. the insulin infusion rate should be decreased. These learning rules, though very simple, represent rationally practical control strategy of human insulin infusion.

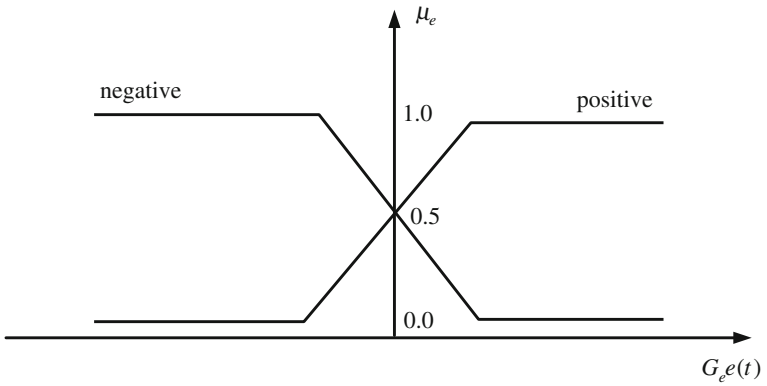


Fig. 3.5 Linear input fuzzy sets for $G_e e(t)$

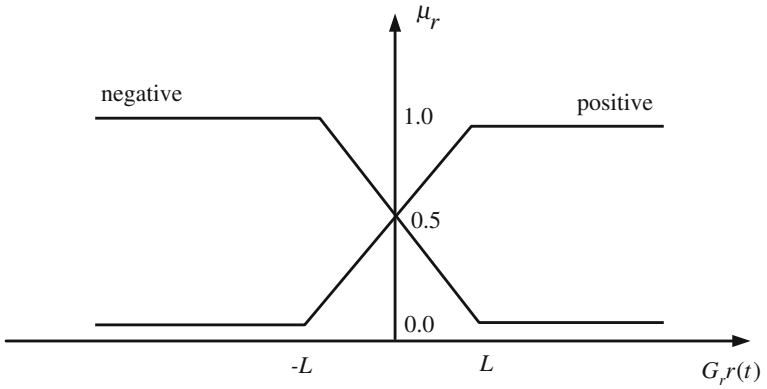


Fig. 3.6 Linear input fuzzy sets for $G_r r(t)$

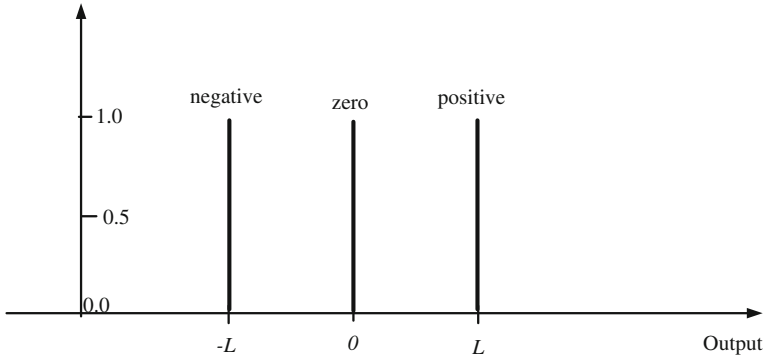


Fig. 3.7 Singleton fuzzy sets for $\Delta u(t)$, "positive," "zero," and "negative"

Consider that for the basal insulin infusion, we obtain the fuzzy logic controller

$$u(t) = \begin{cases} u_0 + \Delta u(t), & if\ u(t) \geq 0 \\ 0, & if\ u(t) < 0^{'} \end{cases} \tag{3.11}$$

where u_0 represents the basal insulin required to maintain the basal glucose level.

3.4 Simulation Study

In this section, simulations using the glucose regulatory system and controllers introduced in previous sections will be explained. First, we compare the steady state without exogenous glucose between diabetes patients and normal individuals. In our simulations, insulin infusion replaces the normal pancreatic function to help the diabetes patients' glucose concentration level at a basal infusion rate.

3.4.1 Steady-State Without Exogenous Glucose

At the steady state without exogenous glucose, patients need the basal infusion rate $u(t) = u_0$ to maintain the glucose at the desired level $G + G_B = G_B$. For the normal individuals, the relationship between glucose and the insulin infusion rate can be represented by Fig. 3.8a. However, Fig. 3.8b represents the relationship of the diabetes patients. The exogenous insulin infusion of diabetes patients mimics the pancreas secretion without food intake. The infusion rate is the basal rate at this situation which is u_0.

3.4.2 Oral Glucose Tolerance Test

The oral glucose tolerance test (OGTT) determines the state of carbohydrate metabolism and is used to recognize an early stage of diabetes mellitus. In the test, patients need to consume 400–800 kcal of glucose after more than 12 h fasting, and their responding will be observed. Upon glucose load, the concentration of glucose rises; OGTT determines the time needed for the concentration of glucose to return to normal. This test simulates the physiologic intake of food under standard conditions. This beta-cell function test represents a significant challenge to the pancreas [13–19].

The OGTT can be mathematically represented by

$$P(t) = P_m \exp\left(-a(\ln(bt) - c)^2\right), \tag{3.12}$$

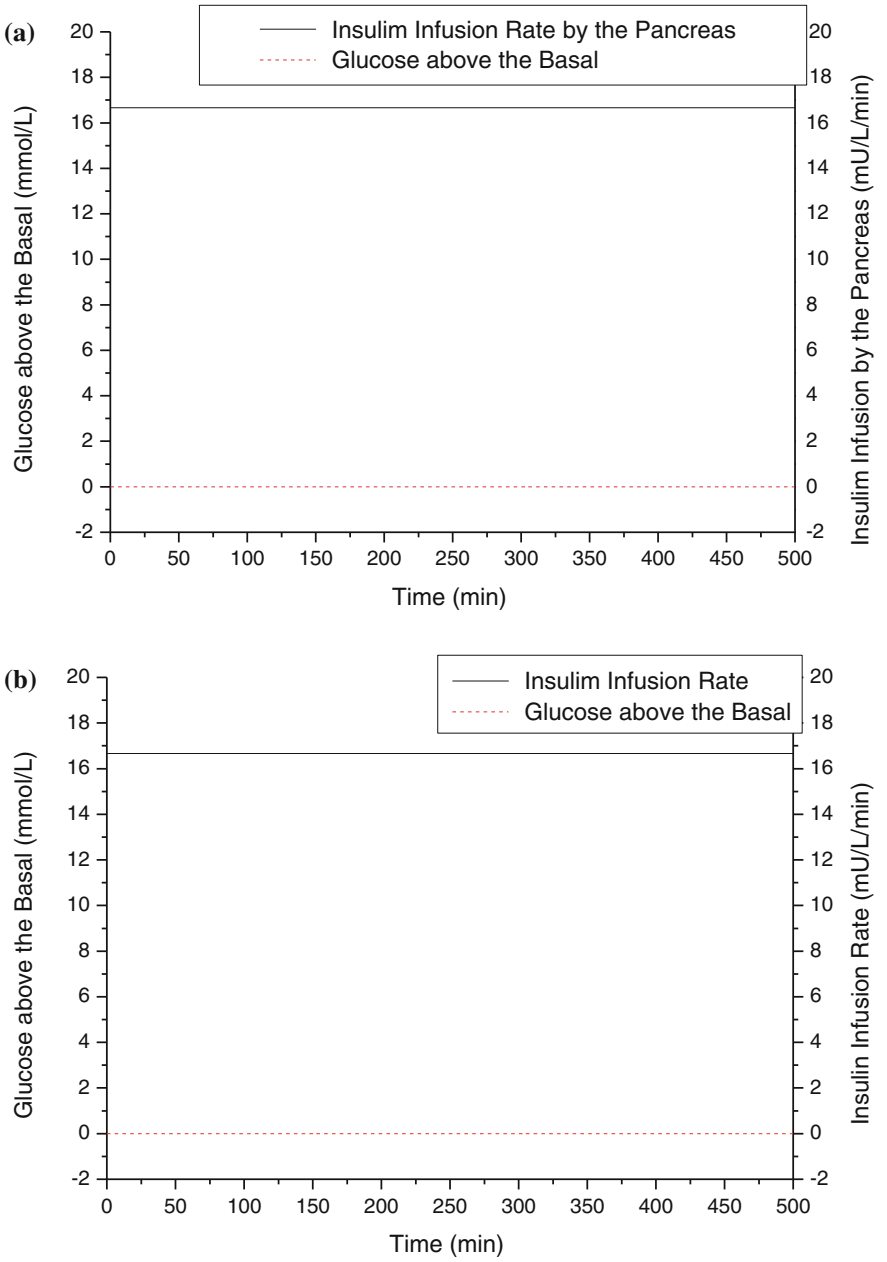


Fig. 3.8 **a** Insulin secretion and glucose level without food intake for normal individuals, **b** Exogenous insulin infusion and glucose level without food intake for diabetes patients

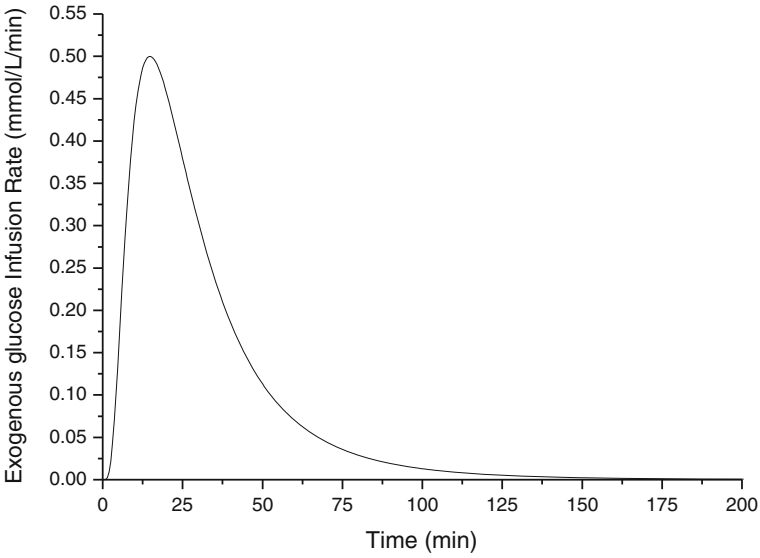


Fig. 3.9 OGTT curve under specific parameters

where the P_m is the peak value and a, b, and c are constants, which determine the slopes and curvature. It is smooth, continuously differentiable, and has zero initial conditions and easily implemented and physiologically representative. OGTT is easily modified to represent faster or slower absorption rates of exogenous glucose. Figure 3.9 represents the OGTT curve with the parameters $P_m = 0.5$, $a = 1$, $b = 0.5$ and $c = 2$.

Because a non-diabetic individual's pancreas can produce enough insulin to consume the plasma glucose and control the glucose concentration well, the glucose concentration curve is represented in Fig. 3.10a. However, for a diabetic individual, the glucose level increases and remains, as shown in Fig. 3.10b.

3.5 Glucose Level Control with Fuzzy Control System

The simulation study using the fuzzy logic control system described in the previous section is carried out for diabetic patients. In the fuzzy logic controller, the scalars G_e and G_r will be selected with different values. Figure 3.11 shows the simulation results with $\alpha = \frac{G_{eG_B}}{2}$ and $G_r = 0$, where α is with different values and in fact, α represents a gain. As we select $G_r = 0$, the rate of change of error $r(t)$ in the glucose level is ignored. If the glucose level is much lower than the basal value, it will cause unconsciousness and possible brain damage, resulting in $\alpha = 1$ provides an adequate regulation of the glucose level.

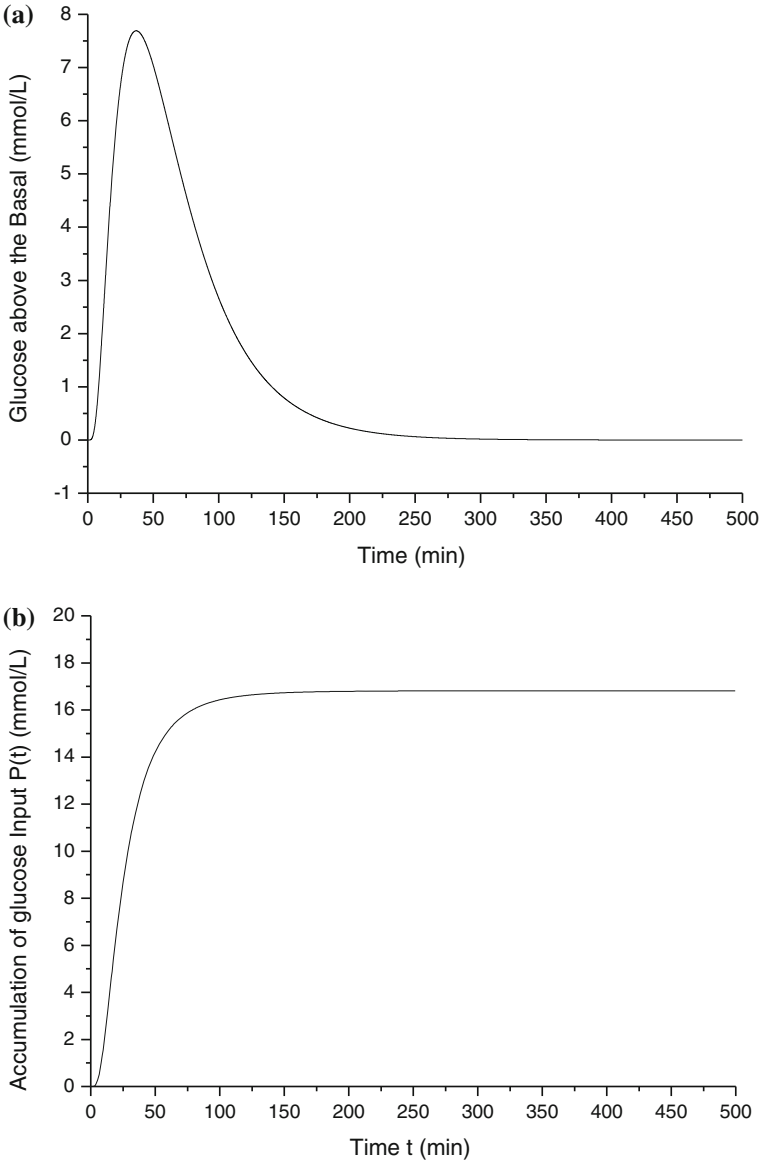


Fig. 3.10 **a** Glucose level versus time curve for a health individual, **b** Glucose level for a diabetic patient

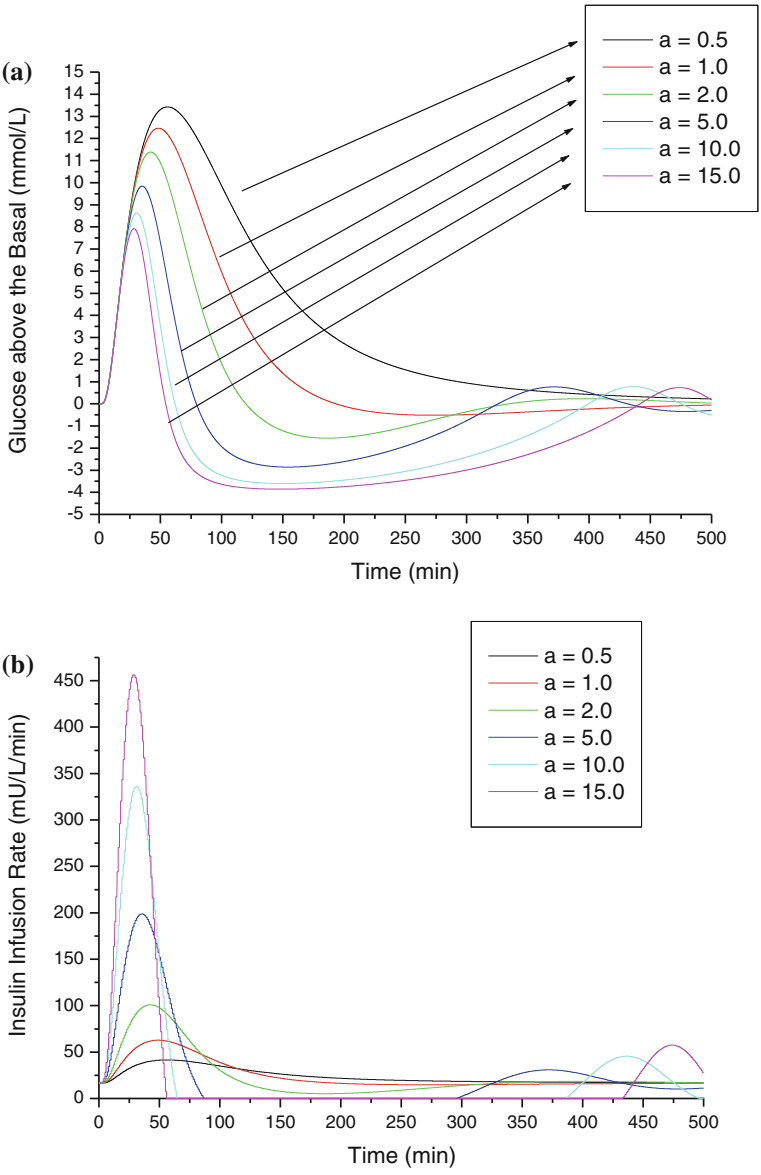


Fig. 3.11 **a** Glucose curve with different α. **b** Insulin rates with different α

Figures 3.12 and 3.13 shows the simulation results with $k_p = \frac{G_e}{2}$ and $k_d = \frac{G_r}{2}$. Figure 3.12 shows what occurs if k_p differs while k_d is fixed. Figure 3.13 shows what occurs if k_d differs and k_p differs. From the results, we can see that for $k_p = 0.8$, $k_d = 30$, and the glucose level is curbed. Thus, we achieve better control of performance.

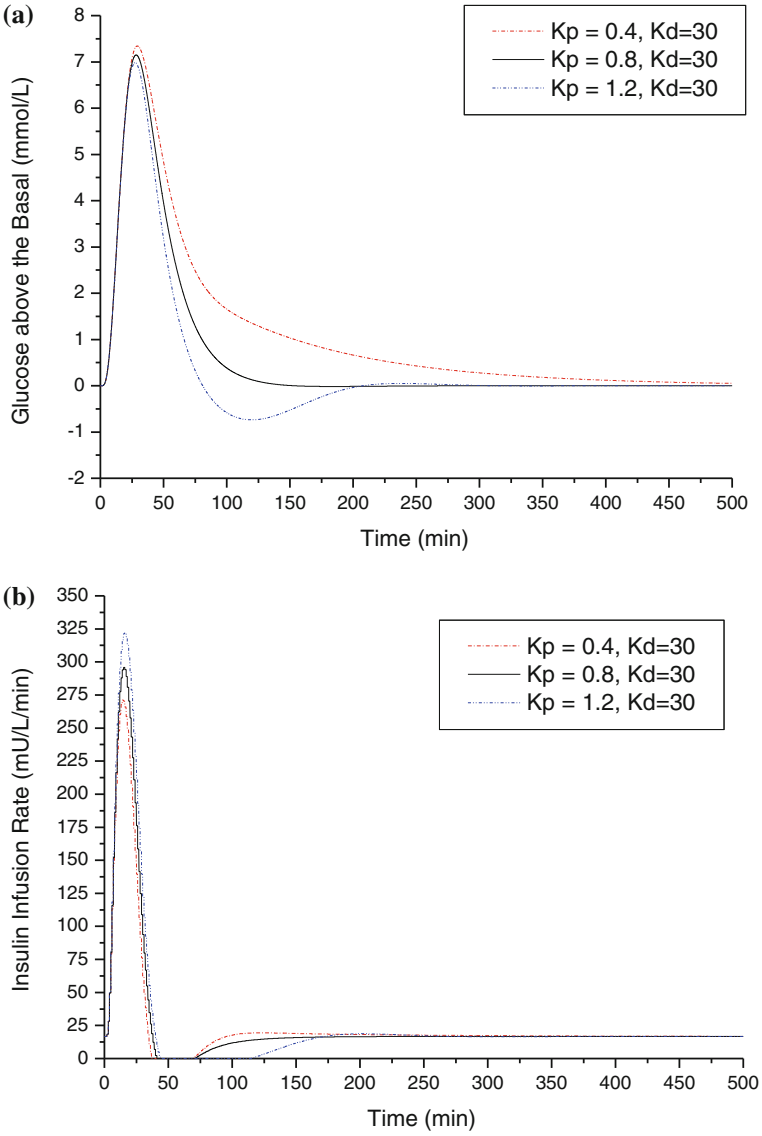


Fig. 3.12 **a** Glucose curve with different k_p. **b** Insulin infusion rates with different k_p

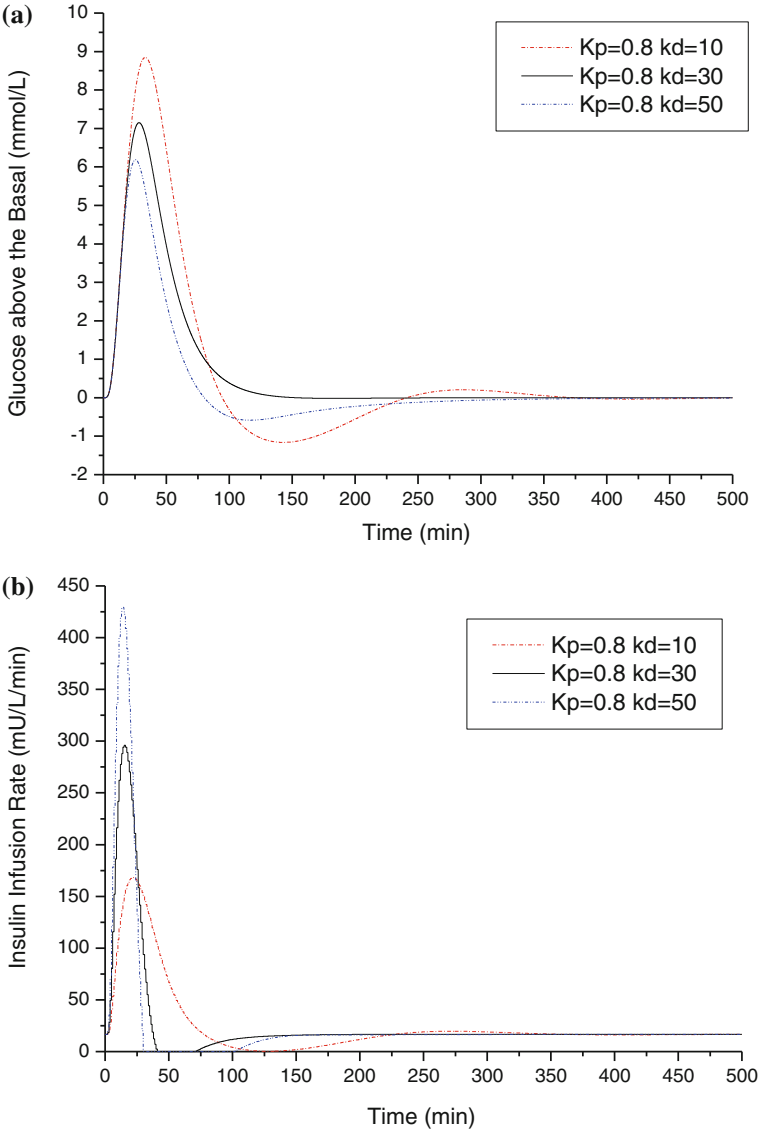


Fig. 3.13 **a** Glucose curve under different k_d. **b** Insulin infusion rate curve under different k_d

3.6 Conclusion

In this chapter, we describe a simulation study which controls glucose levels in diabetics. First, a mathematical model representing the relationship between the human glucose level and insulin is introduced. This model is useful for designing

and simulating a control system. Then, a novel fuzzy logic controller is proposed to regulate the glucose level for diabetes. Fuzzy rules applicable to real patient control are proposed. A simulation study is also carried out, which includes the study of different control parameters settings. The results show that a feedback regulation is feasible if the glucose level measurement is realizable. The fuzzy control system proposed here may also be applicable for open-loop regulation of glucose level. It is quite interesting as in practice it is not easy to monitor and acquire real-time glucose level.

References

1. Van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schietz M et al (2001) Intensive insulin therapy in critically ill patients. N Eng J Med 345(19):1359–1367
2. American Diabetes Association (1998) Economic consequences of diabetes mellitus in the US in 1997. Diabetes Care 21(2):296–309
3. Thomson S, Beaven D, Jamieson M, Snively S, Howl A, Christophersen A (2001) Diabetes New Zealand Inc. Type II diabetes: managing for better health outcomes. PriceWaterHouseCoopers report
4. Chase JG, Lam ZH, Lee JY, Hwang KS (2002) Active insulin infusion control of the blood glucose derivative. In: Seventh international conference on control, automation, robotics and vision (ICARCV'02), Singapore, 2–5 Dec 2002
5. Kienitz K, Yoneyama T (1993) Robust controller for insulin pumps based on H-infinity theory. IEEE Trans Biomed Eng 38(1):57–61
6. Lam ZH, Hwang KS, Lee JY, Chase JG, Wake GC (2002) Active insulin infusion using optimal and derivative-weighted control. Med Eng Phy 24(10):663–672
7. Bode B et al (2004) Alarms based on real-time sensor glucose values alert patients to hypo- and hyperglycemia: the guardian continuous glucose monitoring system. Diabetes Technol Ther 6(2):105–113
8. Bode BW et al (1999) Continuous glucose monitoring used to adjust diabetes therapy improves glycosylated hemoglobin: a pilot study. Diabetes Res Clin Pract 46:183–190
9. Kaufman FR et al (2001) A pilot study of the continuous glucose monitoring system: clinical decisions and glycemic control after its use in pediatric type 1 diabetic subjects. Diabetes Care 24(12):2030
10. Ludvigsson J, Hanas R (2003) Continuous subcutaneous glucose monitoring improved metabolic control in pediatric patients with type 1 diabetes: a controlled crossover study. Pediatrics 5(1):933–938
11. Derr R, Garrett E, Stacy G et al (2003) Is HbA1c affected by glycemic instability. Diabetes Care 26(10):2728–2733
12. Lenhard MJ, Reeves GD (2001) Continuous subcutaneous insulin infusion: a comprehensive review of insulin pump therapy. Arch Intern Med 161(19):2293–2300
13. Davies AG, Baun JD (1988) A decade of insulin infusion pumps. Arch Dis Child 63(3):329–332
14. Mecklenburg R, Benson J, Becker N, Brazel P, Fredlund P, Metz R, Nielson R, Sanner C, Steenrod W (1982) Clinical use of the insulin infusion pump in 100 patients with type 1 diabetes. N Engl J Med 307(9):513–518
15. Plotnick LP, Brancati FL, Clark LM, Erlinger T (2003) Safety and effectiveness of insulin pump therapy in children and adolescents with type I diabetes. Diabetes Care 26(4):1142–1146

16. Doran CV, Chase JG, Shaw GM, Moorhead KT, Hudson NH (2005) Derivative weighted active insulin control algorithms and intensive care unit trials. Control Eng Pract 13(9):1129–1137
17. Bergman RN, Finegood DT, Ader M (1985) Assessment of insulin sensitivity in vivo. Endocr Rev 6(1):45–86
18. Kienitz K, Yoneyama T (1993) Robust controller for insulin pumps based on H-infinity theory. IEEE Trans Biomed Eng 40(11):1133–1137
19. Cobelli C, Nucci G, DelPrato S (1999) Physiological simulation model of the glucose insulin system. In: IEEE conference on engineering in medicine and biology, Atlanta, Georgia, 13–16 Oct 1999

Chapter 4
The Application of Genetic Algorithm for Unsupervised Classification of ECG

Roshan Joy Martis, Hari Prasad, Chandan Chakraborty and Ajoy Kumar Ray

Abstract In this chapter,we have proposed an integrated methodology for electrocardiogram (ECG) based differentiation of arrhythmia and normal sinus rhythm using genetic algorithm optimized *k*-means clustering. Open source databases consisting of the MIT BIH arrhythmia and MIT BIH normal sinus rhythm data are used. The methodology consists of QRS-complex detection using the Pan-Tompkins algorithm, principal component analysis (PCA), and subsequent pattern classification using the *k*-means classifier, error back propagation neural network (EBPNN) classifier, and genetic algorithm optimized *k*-means clustering. The *m*-fold cross-validation scheme is used in choosing the training and testing sets for classification. The *k*-means classifier provides an average accuracy of 91.21 % over all folds, whereas EBPNN provides a greater average accuracy of 95.79 %. In the proposed method, the *k*-means classifier is optimized using the genetic algorithm (GA), and the accuracy of this classifier is 95.79 %, which is equal to that of EBPNN. In conclusion, the classification accuracy of simple unsupervised classifiers can be increased to near that of supervised classifiers by optimization using GA. The application of GA to other unsupervised algorithms to yield higher accuracy as a future direction is also observed.

Keywords Electrocardiogram · Principal component analysis · Neural network · Genetic algorithm · MIT-BIH database

R. J. Martis (✉) · H. Prasad · C. Chakraborty
School of Medical Science and Technology, IIT, Kharagpur, India
e-mail: roshaniitsmst@gmail.com

A. K. Ray
Department of Electronics and Electrical Communication Engineering,
IIT, Kharagpur, India

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_4,

4.1 Introduction

Cardiovascular diseases (CVD) comprise a group of diseases of the heart and blood vessels. Globally, CVD accounts for 16.7 million deaths (29.2 % of total deaths). Around 7.2 million deaths are due to coronary artery disease (CAD) or ischemic heart disease (IHD). Approximately 80 % of all CVD deaths worldwide occur in developing, low-and middle-income countries [1]. A primary concern is that in many countries people of younger generations and those from rural societies are increasingly affected, due to demographic changes and sedentary lifestyles [2, 3]. It has been predicted that between 1990 and 2020 there will be 111 % increase in CVD deaths in India alone. Since the cost of treatment has considerable effects on a country's economy, the development of effective approaches for the early detection and prevention of CAD is important for reducing the burden of heart disease [4].

Arrhythmia occurs due to the anomaly of heart rhythm. Arrhythmias are generally caused by the abnormalities in impulse generation or its conduction or in both. Cardiovascular diseases are the most common etiology for the development of arrhythmias [5]. Many arrhythmias may be life-threatening and require early diagnosis and proper treatment. Arrhythmias like ventricular fibrillation and ventricular flutter are life-threatening medical emergencies.

Electrocardiogram (ECG) is a noninvasive tool for the diagnosis of heart-related abnormalities. It provides both anatomical (i.e., structural) and physiological (i.e., functional) causes of these abnormalities. In normal circumstances, the physician observes the pattern of evolving ECG, understands the disease process, and comes to a diagnoses of the underlying disease. ECG thus has an important role in screening heart abnormalities. Early diagnosis and treatment of heart diseases is crucial; however, in many counties, because of the huge population and limited healthcare resources, it is expensive for medical experts to screen every person. There is, therefore, a need to develop automated screening tools that will make use of some feature extractors and machine-learning algorithms. The work presented in this chapter provides a mass screening method, by classifying arrhythmia and normal sinus rhythm.

Feature extraction techniques, such as principal component analysis (PCA) [6] and linear discriminant analysis (LDA) [6] are used before classification. After feature extraction, pattern classification is to be performed [1, 7–9]. One of the traditional classification algorithms is the *k*-means clustering algorithm was first proposed by MacQueen [10]. The important disadvantage of the *k*-means algorithm is that it will always converge to the local optimum of the objective function. Another supervised classification algorithm is the error back propagation neural network (EBPNN), which has the ability to separate complex data patterns. Again, the EBPNN is also a local optimization of the objective function. A class of classification methods, called evolutionary algorithms, are population-based methods rather than sample-based methods and have heuristically adapted structures. These algorithms always converge to the global optimum of the objective

function. Genetic algorithms are evolutionary algorithms, which borrow principles from natural genetics. There are many works in the literature [11, 12] for a comparative study of the GA.

As with GA, there are many methods of QRS or R-point detection. In this study, we have used the Pan-Tompkins algorithm (1985) for R-point detection. In the past, many automated methods have used the R-point for registration, including some of our earlier works [13–15].

This chapter introduces the application of GA to the ECG classification problem of arrhythmia and normal sinus rhythm. Both normal sinus rhythm and arrhythmia signals are subjected to QRS extraction, PCA, and subsequent pattern classification. Different classifiers used are *k*-means clustering, EBPNN, and GA optimized *k*-means clustering. The *m*-fold cross validation is used to select training and testing patterns for the classifier. The results are compared and discussed below.

The contribution of this chapter is the proposal of a new methodology for ECG classification between arrhythmia and normal sinus rhythm and the use of GA in optimizing the simple unsupervised classifiers like *k*-means clustering so as to improve their classification accuracy. The extension of the application of GA to other classifiers like fuzzy *c*-means clustering and Gaussian mixture model is also observed.

Section 4.2 outlines materials, Sect. 4.3 contains the methodology, Sect. 4.4 includes results and a discussion, and Sect. 4.5 concludes the chapter.

4.2 Materials

In the proposed methodology, the MIT BIH normal sinus rhythm database and MIT BIH arrhythmia database, which are available as open source from www.physionet.org, are used. The MIT BIH normal sinus rhythm database consists of 18 long term ECG recordings of subjects referred to the Arrhythmia Laboratory at Boston's Beth Israel Hospital. Subjects included in this database were found to have had no significant arrhythmias; they include five men, aged 26–45, and 13 women, aged 20–50. The ECG data is digitized at 128 Hz.

The MIT BIH arrhythmia database consisted of 48 half-hour excerpts of two channel ambulatory ECG data obtained from 47 subjects studied by the BIH arrhythmia laboratory between 1975 and 1979. Twenty-three recordings were randomly taken from a set of 4,000 24 h ambulatory ECG data collected from a mixed population including both inpatients (approximately 60 %) and outpatients (approximately 40 %) at Boston's Beth Israel Hospital. The remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias. The ECG recordings are sampled at 360 Hz per channel with 11-bit resolution over a 10 mV range.

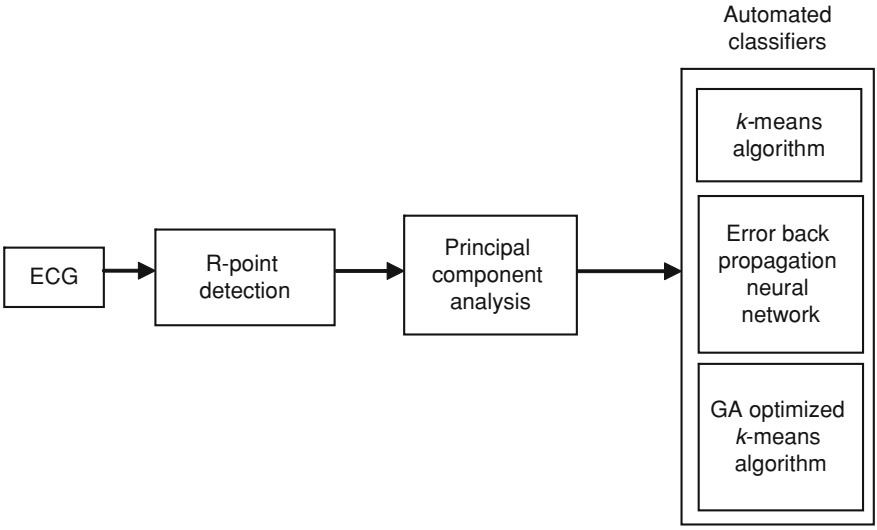


Fig. 4.1 System approach of the proposed methodology

4.3 Methodology

The methodology presented in this chapter consists of preprocessing the ECG, R-point detection using the Pan-Tompkins algorithm, feature compression using principal component analysis (PCA), and subsequent pattern classification using three classifiers, *k*-means clustering, error back propagation neural network (EBNN), and genetic algorithm (GA) optimized *k*-means clustering. The methodology presented in this chapter as a system approach is depicted in Fig. 4.1.

4.3.1 Preprocessing

The signals considered for our analysis are sampled at different rates; hence, it is necessary to choose a common sampling frequency such that equal time spacing is maintained for both signals. A common sampling rate of 250 Hz is chosen for both signals, which are re-sampled using standard techniques [16]. Alternatively, they can also be re-sampled using the fast Fourier transform-based method presented in our previous work [13]. In addition, the open source data may contain muscle artifacts due to movements, powerline interference, and external noise. These unwanted components are removed from the signal by using standard filtering techniques [17].

4.3.2 R-point Detection

The R-point in the ECG has maximum amplitude, which is easily detectable using signal processing methods. Hence, we have used R-point for registration. Other samples are subsequently chosen with respect to the detected R-point. The Pan-Tompkins algorithm is used for the detection of the R-point in our study because of its computational simplicity and better accuracy. In addition, many other methods for R-point detection have been described in the literature, including Fourier transform-based methods [18], wavelet-based methods [19, 20], and Hilbert transform-based methods [21]. The original Pan-Tompkins algorithm consists of taking a derivative using multiple samples, squaring, multiple sample averaging, and thresholding operations.

In this study, an extended version of the Pan-Tompkins algorithm is used. This version uses all simpler operators consisting of computation of the first derivative, rectification, smoothing using a moving average filter, followed by the computation of the second derivative, rectification, smoothing using a moving average filter, summing the two smoothened signals, and thresholding. The derivative provides the slope information, whereas rectification converts all negative magnitudes into positive magnitudes, and smoothing enhances the pulse at the R-point and removes or suppresses the noise components. Once the algorithm provides the location, the R-point is detected by advancing by the number of samples equal to the group delay of all the involved filters.

Based on the detected R-point, 99 samples are chosen to the left of R-point and 100 samples are chosen to the right of R-point, so that a segment of 200 samples is obtained for every subject.

4.3.3 PCA

After segmentation, there is a segment of 200 samples for every subject. Each segment has large dimensionality that imposes a large burden on computation for subsequent classification using automated classifiers. If the information contained in these 200 samples is represented in an efficient manner using fewer components, the computation involved in subsequent classification is reduced due to fewer features. In this study, therefore, PCA is used to reduce the dimensionality of the input data. PCA projects the input data into a new coordinate system, which has axes in the directions of maximum variability. This projection provides new components in which the first component consists of the maximum variations, and the other components consist of variations in decreasing order. Computation of these components consists of computing a data covariance matrix after mean subtraction, decomposing the covariance matrix using Eigen value decomposition, sorting the Eigen vectors in the decreasing order of Eigen values, finally projecting

the data onto the new axes defined by the sorted Eigen vectors. A criterion of containment of 98 % of the total energy of the signal is used to choose the number of components after PCA.

4.3.4 *The* k-*means Algorithm*

The *k*-means clustering algorithm was first proposed by MacQueen [10]. The algorithm is an unsupervised classification method, which assumes a fixed number of clusters. It belongs to the central clustering category which uses Euclidean distance as a distance metric. The algorithm minimizes the total mean squared error between the cluster centroids and the data points. The algorithm then implements the minimization of the following objective function

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2, \tag{4.1}$$

where x_j and μ_i represent the jth pattern and ith cluster center, respectively.

The *k*-means algorithm is given step-by-step in the following.

Step 1: Initialization
Step 2: Data assignment

$$\text{For a data vector, } x_n, \text{ set } y_n = \arg min_k \|x_n - \mu_k\|^2. \tag{4.2}$$

Step 3: Centroid calculation
For each cluster k, let $X_k = \{x_n | y_n = k\}$, the centroid is estimated as

$$\mu_k = \frac{1}{|X_k|} \sum x \in X_k x. \tag{4.3}$$

Step 4: Stop the algorithm if $y_n | n = 1, 2, \ldots N$ does not change; otherwise go back to Step 2.

The *k*-means algorithm can be initialized by choosing a set of k seed points. Seed points can be the first k patterns chosen randomly from the pattern matrix. The first seed point can also be chosen as the centroid of all the patterns, and successive seed points are chosen such that they are a certain distance away from the previously chosen seed points. Each pattern is assigned to a class based on minimum Euclidean distance criterion. Different initial partitions can lead to different clustering results because the *k*-means clustering approach based on the square error criterion can converge to the local minima, rather than the global minima. Therefore, sometimes the *k*-means algorithm must be run many times with different initializations, such that if most of the runs lead to the same results, then we will have some confidence that a global minimum is achieved.

In the data assignment step, the data are partitioned into a class based on the minimum distance between each pattern and the respective class centroid. In the centroid computation step, the average pattern of all the patterns assigned to a given class is computed and is replaced with the previous centroid. The k-means algorithm terminates when the criterion function cannot be improved. The algorithm terminates when the cluster labels for all the patterns do not change between two successive iterations. A maximum number of iterations can be specified to prevent endless oscillations. The computational complexity of the k-means algorithm is of the order $O(NdkT)$, where N is the total number of patterns, d is the number of features, k is the number of clusters, and T is the number of iterations.

4.3.5 EBPNN

An error back propagation neural network [22] is used in this study. It consists of an interconnection of many neurons. The neural network that we have used consists of three layers: the input layer, the hidden layer and the output layer. Initially, random weights are assumed for these interconnections, and the input patterns are fed to the neural network, the output is noted and is compared with the desired output, i.e., class label and, accordingly, the error is back propagated to update the weights. The method is also an optimization which minimizes the following objective function:

$$J = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left\{y_k(x_n, w) - t_k^n\right\}^2, \tag{4.4}$$

where $y_k(x_n, w)$ is the network response for the kth class neuron in the output layer, and t_k^n is the target for the kth class of nth observation feature vector.

The error back propagation algorithm is as follows

1. Begin initialize: η, w, criterion, θ, $m \leftarrow 0$
2. do $m \leftarrow m + 1$
3. $x^m \leftarrow$ training pattern selected sequentially
4. $w_{ji} \leftarrow w_{ji} + \eta\delta_j x_i$; $w_{kj} \leftarrow w_{kj} + \eta\delta_k y_j$
5. until $\|\nabla J(w)\| < \theta\|$
6. return w
7. end

Each pattern is selected sequentially. The network weights are updated using the gradient descent method. If the gradient falls below the threshold θ, the algorithm is stopped.

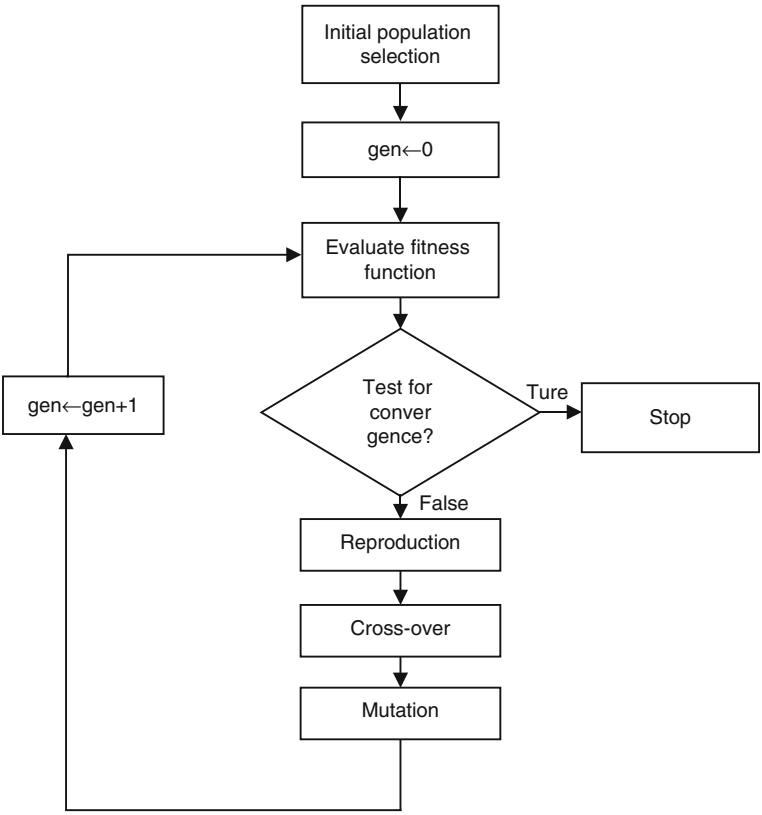


Fig. 4.2 Genetic algorithm-based optimization of cluster centers

4.3.6 GA Optimized k*-means Algorithm*

The *k*-means algorithm discussed in Sect. 4.2 implements the objective function in Eq. 4.1 as a local optimization problem. The *k*-means algorithm is a sample-based optimization strategy. If population-based strategies are used, they may provide the global minimum of the objective function. The genetic algorithm (GA) is an evolutionary algorithm which is a population-based optimization method. We employ GA to optimize the centroids of *k*-means clustering. This GA uses three operators called selection (or reproduction), crossover, and mutation. The principles from natural genetics are used in the algorithm. In contrast to conventional optimization procedures, the GA starts its search from a random set of solutions. The metric used to represent the distance is called a fitness function in the context of GA, which provides relative importance for every population. The GA is depicted in Fig. 4.2 and is explained as follows.

4.3.6.1 Coding and Decoding of Populations

The other traditional algorithms operate on a objective function of real values. However the GA operates in a binary coded string space. Therefore, the centroids which are the optimization variables are to be encoded into strings in binary. Coding the decision variables in a binary string is used to achieve a pseudo chromosomal representation of a solution. Each of the centroid value is represented with a given number of bits. A variable is coded to match each feature present in each subject. Since our optimization variables are in floating point format with x_i^{max} and x_i^{min} as the maximum and minimum value of the ith feature, the decoded value of the centroid is given by,

$$x_i = x_i^{\min} + \frac{x_i^{\max} - x_i^{\min}}{2^{l_i} - 1} DV(s_i), \tag{4.5}$$

Here l_i is the length of the string in order to encode the ith centroid value and $DV(s_i)$ is the fitness value after decoding from binary string s_i. Different variables of decision can be encoded with different precision and number of bits.

4.3.6.2 Fitness Function Evaluation

Each binary string, is to be evaluated for its importance or merit called as fitness value, considering the constraint and objective functions in view. In the current problem, there is no constraint function and hence the fitness function is made equal to the objective function. From the available solutions in the populations, the objective function of Eq. 4.1 is computed by decoding the strings. The current problem is minimization, the actual fitness function which is to be maximized is

$$M = \frac{1}{J} \tag{4.6}$$

4.3.6.3 Reproduction

During reproduction, the strings which are fit are made multiple copies by keeping strings with higher fitness value. The less important strings in terms of their fitness value are discarded. Hence population size is maintained same. This process is called as reproduction. There are many ways to implement the reproduction operation. In the current study the proportionate selection method is used. In this method the strings are multiplied based on the fraction of the total fitness value of a given string.

4.3.6.4 Crossover

The reproduction operation copies the solutions but cannot create any new solution. In both crossover and mutation new populations are generated. In this step, randomly two strings and a crossover site are chosen, and substrings are exchanged between the two strings. In this study single point crossover operation is used. Here in random a crossover bit position is chosen. The chromosome or string is broken into two pieces at the crossover bit position. The two sub strings (or pieces) belonging to two different strings are combined together and the population in the next generation is created.

4.3.6.5 Mutation

Mutation is needed to keep diversity in the population. In this operation, some random bits are chosen and the bit is flipped. Generally, mutation probability is kept small, as is the case with natural genetics. The mutation operation alters the bits in the string in order to create a better string so as to reach the global maxima of the fitness function.

4.3.6.6 Termination

Reproduction, crossover, and mutation are repeated iteratively until the fitness function becomes steady and its value does not change with newer iterations. The GA is said to be converged, and the global maxima of the objective function is attained.

4.3.7 The m-fold Cross Validation

In order to choose the training and testing partition while classification the m-fold cross validation [23] with $m = 3$ is used in this study. Here the total number of observations are disjointly divided into three sets. The first set is used for testing and rest two sets are used for training the classifier in the first fold. The process is repeated in other two sub sets as well to obtain three sub-classification performances which are averaged to estimate final performance of the classifier.

4.4 Results and Discussion

The proposed methodology is implemented as a two-class pattern classification problem using ECG features on the MIT BIH normal sinus rhythm and MIT BIH arrhythmia databases (described in Sect. 4.2). Using the Pan-Tompkins algorithm,

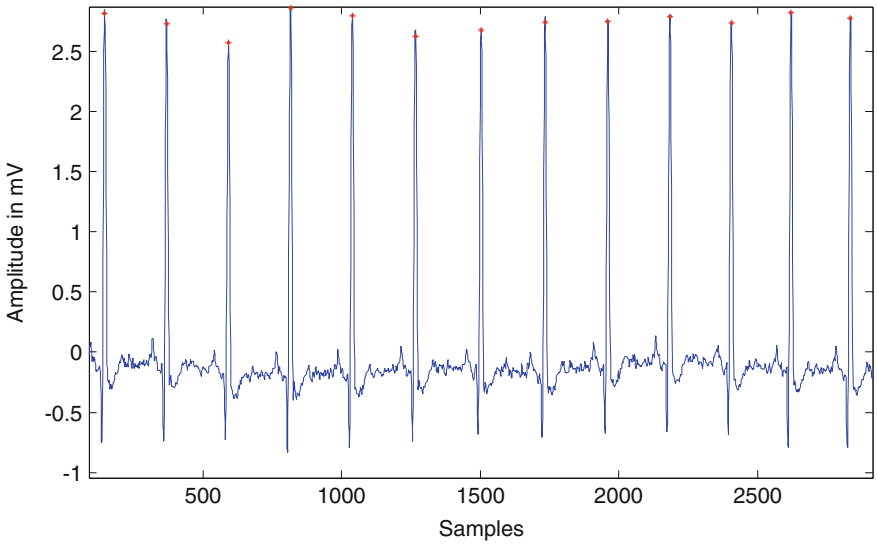


Fig. 4.3 Detection of the R-point in normal sinus rhythm ECG, the R-point is shown as a *red asterisk*

the QRS in the ECG is detected. The exact position of the R-point is obtained by fine tuning, computing the group delays of all the involved filters in the algorithm, and advancing in time by that number of samples. The Pan-Tompkins algorithm is chosen in this study for R-point detection due to its simplicity and the efficient detection of the R-point. Detection of the R-point is shown in Fig. 4.3, in which the detected R-point is shown using a red asterisk. The algorithm consists of multistage filtering (difference, smoothing, etc.) and a nonlinear element (rectification) between the algorithmic steps.

Once the R-point is detected, the ECG signal is segmented into a window of 200 samples such that 99 samples are chosen from the left of the R point and 100 samples are chosen from the right of the R point, including the R-point itself. The 200 samples in every pattern are reduced by the PCA technique. PCA is an orthogonal transformation which reduces the samples by projecting the data into the directions of maximum variability. Eigen value decomposition is used in the PCA to find the variability in each principal component direction. The first principal component (PC) consists of the highest variability; the other PCs consist of the variability in decreasing order. The variability (or the energy or the respective Eigen value) is plotted with respect to the PC dimension in Fig. 4.4. It is observed that the energy contained in these PCs reduces with respect to the dimension of PC. Also, the Eigen values and the percentage energy contained in each dimension are listed in Table 4.1. It is inferred from Fig. 4.4 and Table 4.1 that the first 13 PCs will contain a variability of more than 99.7 %. Therefore, these 13 PCs are used as features for subsequent pattern classification.

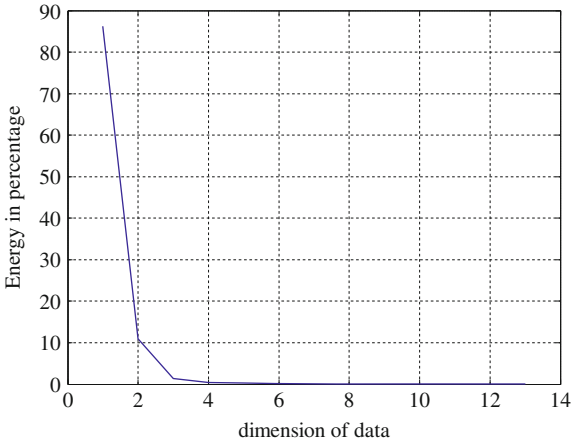


Fig. 4.4 The energy profile of PCs with respect to the dimension

Table 4.1 Energy profile of principal components

PC index	Eigen values	Percentage of energy contained
1	13.1985	86.2342
2	1.6750	10.9438
3	0.2021	1.3204
4	0.0591	0.3864
5	0.0403	0.2636
6	0.0235	0.1533
7	0.0163	0.1067
8	0.0118	0.0773
9	0.0102	0.0663
10	0.0084	0.0546
11	0.0054	0.0350
12	0.0052	0.0342
13	0.0042	0.0272

Table 4.2 Classification accuracy for *k*-means, EBPNN, and GA optimized *k*-means classifiers

Classifier	Accuracy (%)			
	Fold 1	Fold 2	Fold 3	Average
k-means algorithm	91.2088	92.3077	90.1099	91.2088
EBPNN	95.6044	94.5055	97.2527	95.7875
GA optimized *k*-means algorithm	95.6044	95.0549	96.7033	95.7875

The 13 features obtained from PCA are used for subsequent pattern identification using the *k*-means algorithm, EBPNN, and the GA optimized *k*-means algorithm. The *k*-means algorithm is a local optimization algorithm, and its use does not guarantee a global optimum. We can see from Table 4.2 that the *k*-means algorithm provides an average accuracy of 91.2088 and a maximum accuracy of 92.3077.

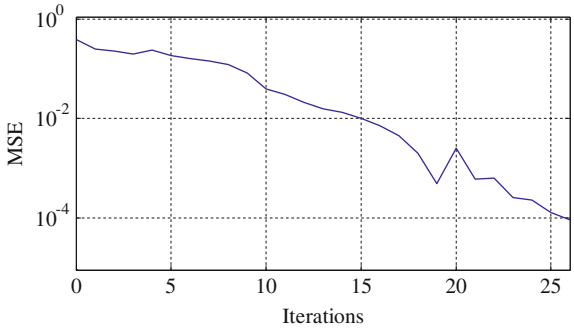


Fig. 4.5 Training of EBPNN: The MSE is decreasing with iterations

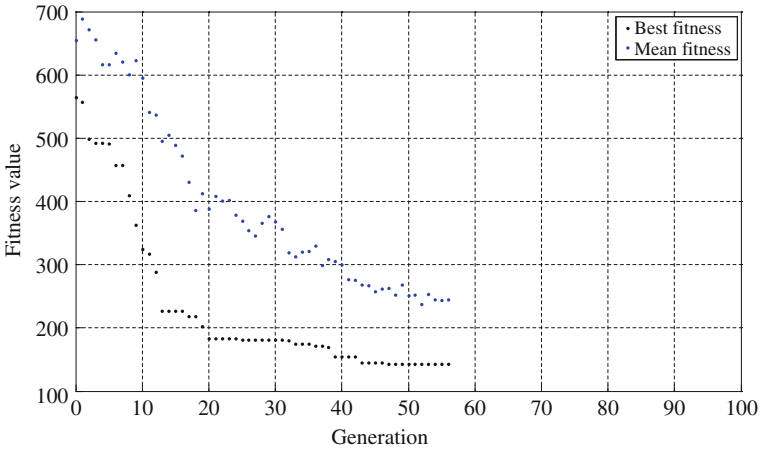


Fig. 4.6 GA classification: The fitness value decreases with the generations

The EBPNN provides better clustering than the *k*-means algorithm does. The decreasing mean-squared error (MSE) used for training the neural network is shown in Fig. 4.5. In our study, we have used neural networks in serial mode. As the epochs iteratively progress, the error is back propagated to update the network weights. A predefined threshold on MSE is defined, and if the MSE reduces to below this threshold, then the algorithm is said to be converged and the iterations are stopped. In our study, the neural network converges in 26 epochs and the threshold chosen is 10^{-04}. We can see from Table 4.2 that EBPNN provides a maximum accuracy of 97.2527 % and an average accuracy of 95.7875 % over the three folds.

The GA optimized *k*-means clustering algorithm is used on the ECG data, in our analysis, and the results are shown in Figs. 4.6, 4.7, 4.8. Figure 4.6 shows best and mean fitness in each generation. Since our optimization is a minimization of the

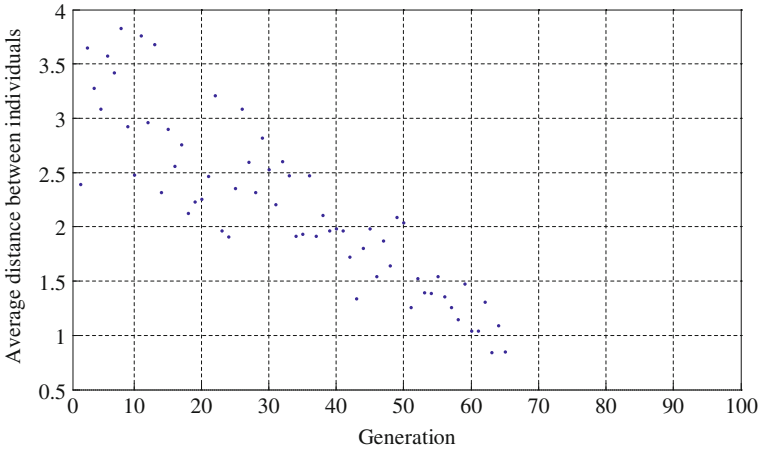


Fig. 4.7 GA classification: average distance between individuals is shown decreasing with iterations

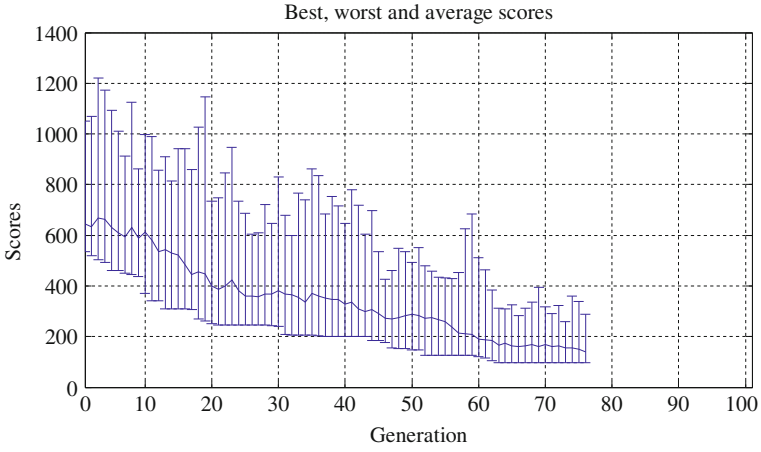


Fig. 4.8 The best, worst, and average scores in every generation in GA classification

objective function, it is expected that both the best and mean fitness should decrease with generations. In Fig. 4.6, the mean and best fitness values decrease with the generations as we have expected. Also, as the generations progress in GA, the new individuals are fitter than the initial ones. Therefore, the average distance between the individuals is expected to decrease with generations, as are the results in Figs. 4.7, 4.8 shows that the best, worst, and average scores decrease with generations in the GA.

The classification accuracy of the GA optimized *k*-means algorithm provides a maximum accuracy of 96.7033 % and average accuracy of 95.7875 % over all three folds (as shown in Table 4.2).

4.5 Conclusion

In this chapter, a new methodology for the classification of ECG belonging to normal sinus rhythm and arrhythmia classes is presented. The *k*-means algorithm, EBPNN, and GA are used for classification. We have seen that the *k*-means algorithm yields a lower accuracy than other supervised classifiers, i.e., EBPNN. We have also seen that if a simple algorithm like *k*-means is optimized using GA, the accuracy increases to that of supervised classifier, EBPNN. In the future, one can optimize other unsupervised algorithms such as the fuzzy *c*-means and the Gaussian mixture model algorithms. In addition, newer variants of the GA are available for faster implementation. These new variants of GAs will be faster and converge in fewer iterations. Also, there is a scope to identify novel operators in the GA and thus to catalyze the algorithm. The methodology used will be of immense utility in machine-learning applications for healthcare informatics.

References

1. Cardiovascular Disease: Prevention and Control (2010) WHO report on global strategy on diet, physical activity and health. World Health Organization, Geneva. http://www.who.int/dietphysicalactivity/publications/facts/cvd/en/. Accessed 5 Sept 2010
2. Park K (2005) Park's textbook of preventive and social medicine, 18th edn. Banarsidas Bhanot publishers, Jabalpur
3. Fauci AS, Braunwald E, Kesper DL, Hauser SL, Longo DL, Jamesonn JL, Loscalzo J (2008) Harrison's principles of internal medicine, 17th edn. Mc-Graw Hill, New York
4. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A (2010) Growing epidemic of coronary heart disease in low-and middle-income countries. Curr Probl Cardiol 35(2):72–115
5. Guyton AC, Hall JE (2006) Textbook of medical physiology, 11th edn. W. B Saunders Co, Philadelphia
6. Duda R, Hart P, Stork D (2001) Pattern classification, 2nd edn. Wiley, New York
7. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, Eagle-wood Cliffs
8. Jain AK, Murthy MN, Flynn PJ (1999) Data Clustering: a review. ACM Comput Surv 31:264–323
9. Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Networks 16(3):645–678
10. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. Paper presented at the proceedings of 5th Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, vol 1, pp 281–297
11. Naldi MC, de Carvalho ACPLF, Campell RJGB, Hruschka ER (2007) Genetic Clustering for data Mining. In: Maimon O, Rokach L (eds) Soft Computing for Knowledge Discovery and Data Mining. Springer, New York, pp 113–132

12. Deb K (2001) Multi-objective optimization using evolutionary algorithms. Wiley, New York
13. Martis RJ, Chakraborty C, Ray AK (2009) A two stage mechanism for registration and classification of ECG using Gaussian mixture model. Pattern Recogn 42(11):2979–2988
14. Martis RJ, Krishnan MM, Chakraborty C, Pal S, Sarkar D, Mandana KM, Ray AK (2012) Automated screening of arrhythmia using wavelet based machine learning techniques. J Med Syst 36(2):677–688
15. Martis RJ, Chakraborty C (2011) Arrhythmia disease diagnosis using genetic algorithm optimized *k*-means clustering. J Mech Med Biol 11(4):897–915
16. Vaidyanathan PP (2003) Multirate systems and filter banks. Pearson education (Asia) Pte, Taiwan
17. Oppenheim AO, Schaffer RA (2003) Discrete time signal processing, Mc-Graw Hill edition, New York
18. Murthy IS, Niranjan UC (1992) Component wave delineation of ECG by filtering in the fourier domain. Med Biol Eng Comput 30(2):169–176
19. Li C, Zheng C, Tai C (1995) Detection of ECG characteristic points using wavelet transforms. IEEE Trans Biomed Eng 42(1):21–29
20. Martinez JP, Almeida R, Olmos S, Rocha AP, Laguna P (2004) A wavelet based ECG delineator: Evaluation on standard databases. IEEE Trans Biomed Eng 51(4):570–581
21. Benitez D, Gaydecki PA, Zaidi A, Fitzpatrick AP (2001) The use of the Hilbert transform in ECG signal analysis. Comput Biol Med 31(5):399–406
22. Bishop C (1995) Neural Networks for pattern recognition, Oxford University press, Oxford
23. Schneider J (1997) Cross validation. http://www.cs.cmu.edu/∼schneide/tu5/node42.html. Accessed on 5 Sept 2010

Chapter 5
Pixel-based Machine Learning in Computer-Aided Diagnosis of Lung and Colon Cancer

Kenji Suzuki

Abstract Computer-aided diagnosis (CAD) for detection of lesions in medical images has been an active area of research. Machine learning plays an essential role in CAD, because representing lesions and organs requires a complex model that has a number of parameters to determine; thus, medical pattern recognition essentially requires "learning from examples" to determine the parameters of the model. Machine learning has been used to classify lesions into certain classes (e.g., abnormal or normal, lesions or non-lesions, and malignant or benign) in CAD. Recently, as available computational power increased dramatically, pixel/voxel-based machine learning (PML) has emerged in medical image processing/analysis, which uses pixel/voxel values in local regions (or patches) in images instead of features calculated from segmented regions as input information; thus, feature calculation or segmentation is not required. Because PML can avoid errors caused by inaccurate feature calculation and segmentation, the performance of PML can potentially be higher than that of common classifiers. In this chapter, MTANNs (a class of PML) in CAD schemes for detection of lung nodules in CT and for detection of polyps in CTC are presented.

5.1 Introduction

Computer-aided diagnosis (CAD) for detection of lesions in medical images [1, 2] has been an active area of research, because evidence indicates that CAD can help improve the diagnostic performance of radiologists or physicians in their image reading and interpretations [3–5]. A lot of investigators have participated in and developed CAD schemes such as those for detection of lung nodules in chest

K. Suzuki (✉)
Department of Radiology, Division of Biological Sciences,
The University of Chicago, Chicago, IL, USA
e-mail: suzuki@uchicago.edu

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_5,

radiographs [6–8] and in thoracic CT [9–12], for detection of microcalcifications/masses in mammography [13], breast MRI [14], and breast US [15, 16], and for detection of polyps in CT colonography (CTC) (also known as virtual colonoscopy) [17–19].

Machine learning plays an essential role in CAD because objects such as lesions and organs in medical images may not be represented accurately by a simple equation. For example, a lung nodule is generally modeled as a solid sphere, but there are nodules of various shapes and nodules with internal inhomogeneities, such as spiculated ones and ground-glass (or non-solid) nodules. A polyp in the colon is modeled as a bulbous object, but there are also polyps which exhibit a flat morphology [20]. Thus, diagnostic tasks in medical imaging essentially require learning from examples (or data).

Machine learning has been used to classify lesions into certain classes (e.g., abnormal or normal, lesions or non-lesions, and malignant or benign) in CAD. Machine-learning algorithms for classification include linear discriminant analysis [21], quadratic discriminant analysis [21], multilayer perceptrons [22], and support vector machines [23]. Such machine-learning algorithms have been applied to lung nodule detection in chest radiography [24] and thoracic CT [10, 25], classification of lung nodules into benign or malignant categories in chest radiography [26] and thoracic CT [27], and polyp detection in CTC [17, 28].

Recently, as available computational power increased dramatically, pixel/voxel-based machine learning (PML), which uses pixel/voxel values in images instead of features calculated from segmented regions as input information, has emerged in medical image processing/analysis; thus, feature calculation or segmentation is not required. Because PML can avoid errors caused by inaccurate feature calculation and segmentation, the performance of PML can potentially be higher than that of common classifiers. By extension of "neural filters" [29] and "neural edge enhancers" [30, 31], which are ANN-based [22] supervised nonlinear image-processing techniques, and an MTANN framework [9], which is a class of PML, have been developed for accommodating the task of distinguishing a specific opacity from other opacities in medical images.

In this chapter, MTANNs (a class of PML) in CAD schemes for detection of lung nodules in CT and for detection of polyps in CTC are introduced. MTANNs have been applied to removal of false-positive detections (FPs) in the computerized detection of lung nodules in low-dose CT [9, 10] and chest radiography [6], for distinction between benign and malignant lung nodules in CT [32], for suppression of ribs and clavicles (i.e., bones) in chest radiographs [33], and for reduction of FPs in computerized detection of polyps in CTC [18, 19, 34–36].

5.2 Pixel-based Machine Learning (PML) in CAD

5.2.1 PML Overview

PML techniques have been developed for tasks in medical image processing/analysis and computer vision. There are three classes of PML techniques: neural filters [29, 37] (including neural edge enhancers [30, 31]), convolution neural networks (NNs) [38–44] (including shift-invariant NNs [45–47]), and massive-training ANNs (MTANNs) [18, 33, 48–50] (including multiple MTANNs [6, 10, 29, 32, 37, 48], a mixture of expert MTANNs [19, 34], a multi-resolution MTANN [33], a Laplacian eigenfunction MTANN (LAP-MTANN), and a massive-training support vector regression (MTSVR) [36, 51]). The class of neural filters has been used for image-processing tasks such as edge-preserving noise reduction in radiographs and other digital pictures [29, 37], edge enhancement in noisy images [30], and enhancement of subjective edges traced by a physician in left ventriculograms [31]. The class of convolution NNs has been applied to classification tasks such as FP reduction in CAD schemes for detection of lung nodules in chest radiographs (CXRs) [38–40], FP reduction in CAD schemes for detection of microcalcifications [41] and masses [42] in mammography, face recognition [43], and character recognition [44]. The class of MTANNs has been used for classification, such as FP reduction in CAD schemes for detection of lung nodules in CXR [6] and CT [4, 10, 48], distinction between benign and malignant lung nodules in CT [32], and FP reduction in a CAD scheme for polyp detection in CTC [18, 19, 34, 51]. The MTANNs have also been applied to pattern enhancement and suppression such as separation of bone from soft tissue in CXR [33, 49] and enhancement of lung nodules in CT [50].

5.2.2 MTANN Filter for Lesion Enhancement

5.2.2.1 Architecture of an MTANN Filter

The architecture of a PML technique which consists of a machine-learning model is shown in Fig. 5.1. In order to enhance actual lesions in medical images, we developed an MTANN supervised filter, which is a class of PML. An MTANN filter consists of a machine-learning regression model such as a linear-output artificial neural network (ANN) model [30] which is a regression-type ANN capable of operating on pixel/voxel data directly, and a support vector regression model [36] which is a regression-type support vector machine. The MTANN filter is trained with input CT images and the corresponding "teaching" images that contain a map for the "likelihood of being lesions." The input to the MTANN filter consists of pixel values in a sub-region (or sub-volume), R_S, extracted from

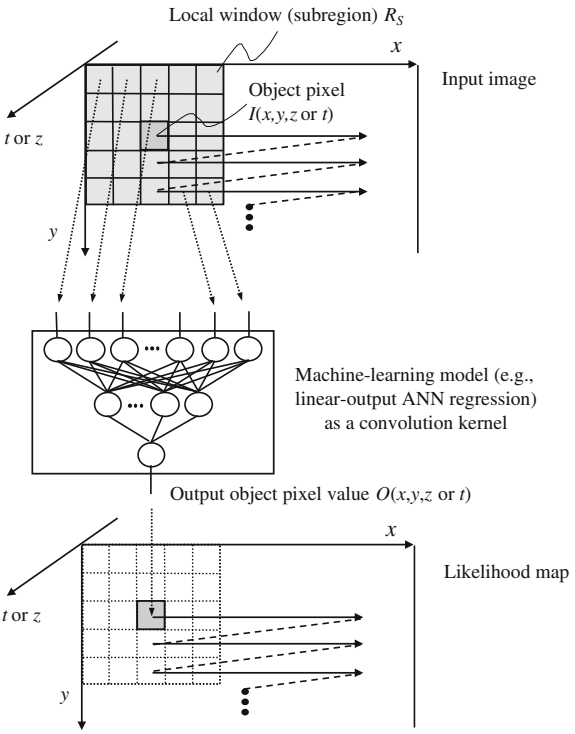


Fig. 5.1 Architecture of a PML (e.g., MTANN) technique consisting of a machine-learning model (e.g., a linear-output ANN regression model or support vector regression) with sub-region (or sub-volume) input and single-pixel output. All pixel values in a sub-region extracted from an input image are entered as input to the machine-learning model. The machine-learning model outputs a single pixel value for each sub-region, the location of which corresponds to the center pixel in the sub-region. The output pixel value is mapped back to the corresponding pixel in the output image

an input image. The output of the MTANN filter is a continuous scalar value, which is associated with the center pixel in the sub-region and is represented by

$$O(x, y, z \text{ or } t) = ML\{I(x - i, y - j, z - k \text{ or } t - k) | (i, y, k) \in R_S\}, \tag{5.1}$$

where x and y are the coordinate indices, $ML(\cdot)$ is the output of the machine-learning regression model, and $I(x, y, z \text{ or } t)$ is a pixel value in the input image. The linear-output ANN model [30] used as the machine-learning regression model in the MTANN employs a linear function, $f_L(u) = a \cdot u + 0.5$, instead of a sigmoid function, $f_S(u) = 1/\{1 + \exp(-u)\}$, as the activation function of the output layer unit because the characteristics and performance of an ANN were improved significantly with a linear function when it was applied to the continuous mapping of values in image processing [30]. Note that the activation function in the hidden layers is still a sigmoid function. The input vector can be rewritten as

$$\overrightarrow{I}_{x,y,z \text{ or } t} = \{I_1, I_2, \ldots, I_m, \ldots, I_{N_I}\}, \tag{5.2}$$

where m is an input unit number and N_I is the number of input units. The output of the nth unit in the hidden layer is represented by

$$O_n^H = f_S\left\{\sum_{m=1}^{N_I} w_{mn}^H \cdot I_m - w_{0n}^H\right\}, \tag{5.3}$$

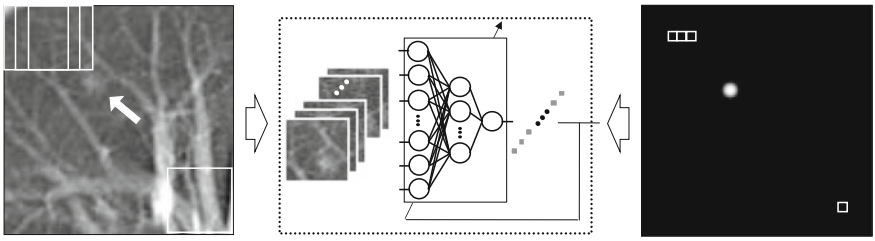


Fig. 5.2 Training of an MTANN filter for enhancement of lesions. The input CT image is divided pixel-by-pixel into a large number of overlapping sub-regions. The corresponding pixels are extracted from the teaching image containing a map for the "likelihood of being lesions." The MTANN filter is trained with pairs of the input sub-regions and the corresponding teaching pixels

where w_{mn}^{H} is a weight between the mth unit in the input layer and the nth unit in the hidden layer, and w_{0n}^{H} is an offset of the nth unit in the hidden layer. The output of the output layer unit is represented by

$$O(x, y, z \text{ or } t) = f_L\left\{\sum_{m=1}^{N_H} w_m^O \cdot O_m^H - w_0^O\right\}, \tag{5.4}$$

where w_m^O is a weight between the mth unit in the hidden layer and the unit in the output layer, N_H is the number of units in the hidden layer, and w_0^O is an offset of the unit in the output layer. For processing the entire image, the scanning of an input CT image with the MTANN is performed pixel-by-pixel, as illustrated in Fig. 5.1.

5.2.2.2 Training of an MTANN Filter

For enhancement of lesions and suppression of non-lesions in CT images, the teaching image $T(x, y, z)$ contains a map for the "likelihood of being lesions," as illustrated in Fig. 5.2. In order to create the teaching image, lesions are first segmented manually to obtain a binary image with 1 being lesion pixels and 0 being non-lesion pixels. Then, Gaussian smoothing is applied to the binary image to smooth the edges of the segmented lesions, because the likelihood of a pixel being a lesion should gradually diminish as the distance from the boundary of the lesion decreases. Note that the ANN was not able to be trained when binary teaching images were used.

The MTANN filter involves training with a large number of pairs of sub-regions and pixels; we call it a massive-sub-region training scheme. In order to enrich the

training samples, a training image, R_T, extracted from the input CT image is divided pixel by pixel into a large number of sub-regions. Note that close sub-regions overlap each other. Single pixels are extracted from the corresponding teaching image as teaching values. The MTANN filter is massively trained by using each of a large number of input sub-regions together with each of the corresponding teaching single pixels; hence, the term massive-training ANN. The error to be minimized by training of the MTANN filter is given by

$$E = \frac{1}{P}\sum_{c}\sum_{(x,y,z\text{ or }t)\in R_T}\{T_c(x,y,z\text{ or }t) - O_c(x,y,z\text{ or }t)\}^2, \tag{5.5}$$

where c is a training case number, O_c is the output of the MTANN for the cth case, T_c is the teaching value for the MTANN for the cth case, and P is the number of training pixels in the training images, R_T. The MTANN filter is trained by a linear-output back-propagation algorithm, in which the generalized delta rule [22] is applied to the linear-output ANN architecture [30]. After training, the MTANN filter is expected to output the highest value when a lesion is located at the center of the sub-region of the MTANN filter, a lower value as the distance from the sub-region center increases, and zero when the input sub-region contains a non-lesion.

Once the trained MTANN enhances lesions in medical images, lesion candidates can be detected by application of a segmentation technique. One of the simplest ways to perform this technique is thresholding. Another segmentation technique can be used for this purpose as well such as multiple thresholding, region growing, level-set segmentation, and active contour segmentation. We used a simple thresholding technique in this study because the MTANN enhanced lesions effectively, i.e., the contrast of lesions was substantially high compared to that of normal structures (see the results in the next section).

5.2.3 MTANN for Classification

5.2.3.1 Training Method of an MTANN for Classification

Once lesion candidates are detected, the next step in a CAD scheme is classification of the candidates into lesions or non-lesions. We can use a PML such as an MTANN for this task as well. For distinction between lesions and non-lesions, the teaching image contains a Gaussian distribution with standard deviation σ_T for a lesion and zero (i.e., completely dark) for a non-lesion, as shown in Fig. 5.3. This distribution represents a map for the "likelihood of being a lesion:"

$$T(x,y,z\text{ or }t) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_r}\exp\left\{-\frac{(x^2+y^2+z^2\text{ or }t^2)}{2\sigma_r^2}\right\} & \text{for a lesion} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

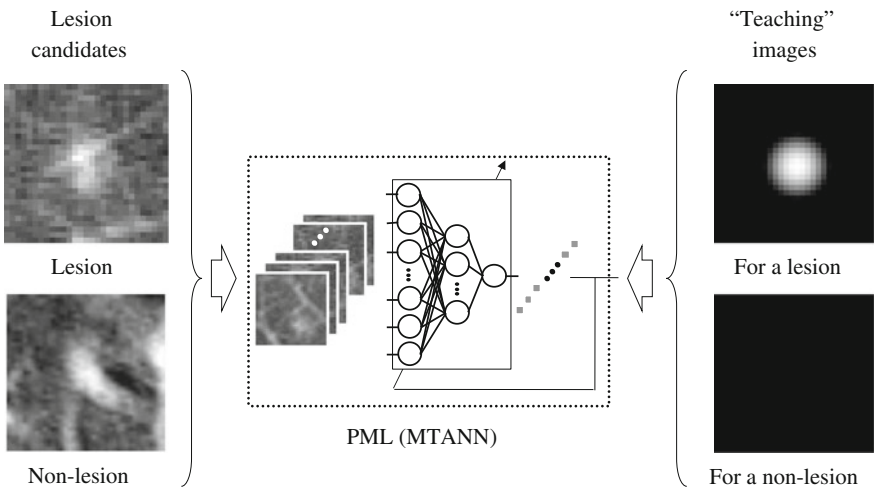


Fig. 5.3 Training of a PML technique (i.e., an MTANN) for classification of candidates into a lesion (e.g., a nodule) or a non-lesion (e.g., a non-nodule). A teaching image for a lesion contains a Gaussian distribution at the center of the image, whereas that for a non-lesion contains zero (i.e., is completely dark)

For enrichment of training samples, a training region (or volume), R_T, extracted from the input image is divided pixel-by-pixel into a large number of overlapping sub-regions. Single pixels are extracted from the corresponding teaching region as teaching values. The MTANN is massively trained by use of each of a large number of the input sub-regions together with each of the corresponding teaching single pixels. After training, the MTANN is expected to output the highest value when a lesion is located at the center of the sub-region of the MTANN, a lower value as the distance from the sub-region center increases, and zero when the input sub-region contains a non-lesion.

5.2.3.2 Scoring Method for Combining Output Pixels

For combining output pixels from a trained MTANN, we developed a scoring method. A score for a given lesion candidate from the trained MTANN is defined as

$$S = \sum_{(x,y,z\,\text{or}\,t)\in R_E} f_G(x,y,z\,\text{or}\,t) \times O(x,y,z\,\text{or}\,t), \tag{5.7}$$

where

$$f_G(x,y,z\,\text{or}\,t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x^2+y^2+z^2\,\text{or}\,t^2)}{2\sigma^2}\right\} \tag{5.8}$$

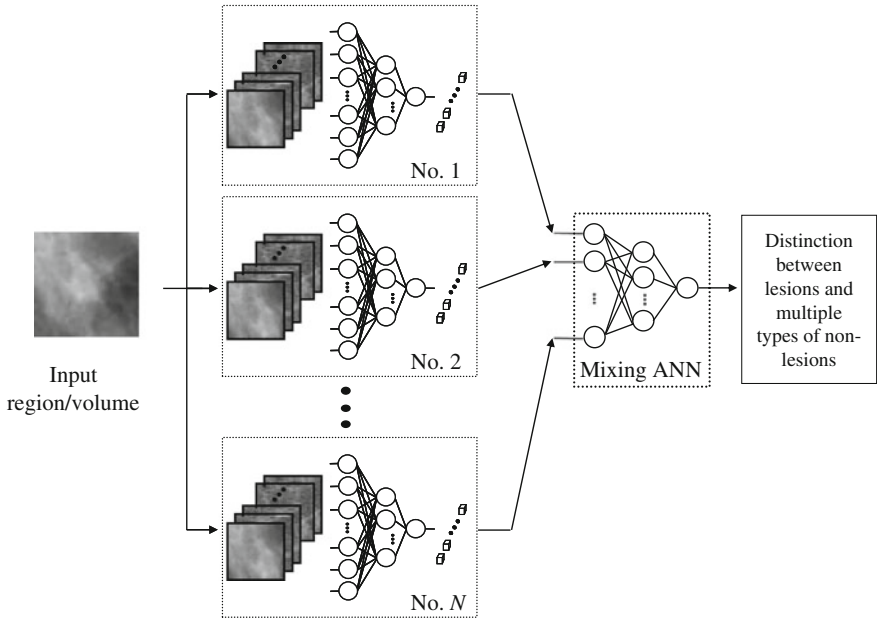


Fig. 5.4 Architecture of a mixture of expert MTANNs for classification of lesion candidates into lesions or multiple types of non-lesions

is a Gaussian weighting function with standard deviation σ, and its center corresponds to the center of the region for evaluation, R_E, and $O(x, y, z \text{ or } t)$ is the output region of the n-th trained MTANN, where its center corresponds to the center of R_E. The use of the Gaussian weighting function allows us to combine the responses (outputs) of a trained MTANN as a distribution. A Gaussian function is used for scoring, because the output of a trained MTANN is expected to be similar to the Gaussian distribution used in the teaching region. This score represents the weighted sum of the estimates for the likelihood that the region (lesion candidate) contains a lesion near the center, i.e., a higher score indicates a lesion, and a lower score indicates a non-lesion.

5.2.3.3 Mixture of Expert MTANNs

In order to distinguish lesions from various types of non-lesions (i.e., FPs), we have extended the capability of a single MTANN, and we have developed a mixture of expert MTANNs. The architecture of the mixture of expert MTANNs is shown in Fig. 5.4. The mixture of experts consists of several MTANNs that are arranged in parallel. Each MTANN is trained independently by using the same

lesions and a different set of non-lesions. Each MTANN acts as an expert for distinction between lesions (e.g., nodules) and non-lesions (e.g., non-nodules) representing a specific non-lesion type. The scores from the expert MTANNs are combined by using a mixing ANN such that different types of non-lesions can be distinguished from lesions. The mixing ANN consists of a linear-output ANN model with a linear-output back-propagation training algorithm [30] for processing continuous output/teaching values; the activation functions of the units in the input, hidden, and output layers are an identity, a sigmoid, and a linear function, respectively. One unit is employed in the output layer for distinction between a lesion and a non-lesion. The scores of each expert MTANN are used for each input unit in the mixing ANN; thus, the number of input units equals the number of expert MTANNs, N. The scores of each expert MTANN act as the features for distinguishing lesions from a specific type of non-lesion for which the expert MTANN is trained. The output of the mixing ANN for the cth lesion candidate is represented by

$$M_c = NN\left[\left\{S_{n,c}\right\} | 1 \le n \le N\right] \tag{5.9}$$

where $NN(\cdot)$ is the output of the linear-output ANN model and n is an MTANN number. The teaching values for lesions are assigned the value one, and those for non-lesions are assigned the value zero. Training of the mixing ANN may be performed by using a leave-one-lesion-out cross-validation scheme [52]. After training, the mixing ANN is expected to output a higher value for a lesion and a lower value for a non-lesion. Thus, the output can be considered as the likelihood of being a lesion. By thresholding the output, a distinction between lesions and non-lesions can be made. The balance between a true-positive (TP) rate and a false-positive (FP) rate is determined by the selected threshold value. If the scores of each expert MTANN properly characterize the type of non-lesion for which the expert MTANN is trained, the mixing ANN combining several expert MTANNs will distinguish lesions from various types of non-lesions.

5.3 A CAD Scheme for Detection of Lung Nodules on CT Images

5.3.1 Lung Cancer Detection in CT

Lung cancer continues to rank as the leading cause of cancer deaths among American men and women [53, 54]; the number of lung cancer deaths each year is greater than the combined number of breast, colon, and prostate cancer deaths. Evidence suggests that early detection of lung cancer may allow more timely therapeutic intervention and thus a more favorable prognosis for the patient [55–58]. Therefore, in the 1970s, screening programs for the early detection of lung cancer were carried out with chest radiography and cytologic examination of sputum in the

United States [59–61] and in Europe [62]. As the CT imaging techniques have advanced, screening with low-dose helical CT has been performed for early detection of lung cancer [63–70] since early 1990.

Because CT is more sensitive than chest radiography in detection of small nodules due to lung carcinoma at an early stage [64, 68], lung cancer screening programs are being conducted in the United States [63–66] and Japan [67–70] with a low-dose single-detector CT as the screening modality. Recently, multi-detector-row CT (MDCT) has been used for lung cancer screening. Helical CT and MDCT, however, generate a large number of images that must be read by radiologists. Such readings may lead to "information overload" for radiologists. Furthermore, radiologists may fail to detect some cancers, which are visible in retrospect, during the interpretation of CT images [71, 72]. Therefore, a CAD scheme for detecting lung nodules in CT has been investigated as a tool for lung cancer screening, because the CAD scheme may detect some cancers that are missed by radiologists [72], and it provides quantitative detection results as a second opinion to assist radiologists in improving their detection accuracy [73].

5.3.2 Database of Lung Nodules in Thick-Slice CT

In order to test the performance of our CAD scheme that utilizes MTANN filters, we created a CT database consisting of 69 lung cancer images obtained from 69 patients [72]. The scans used for this study were acquired with a low-dose protocol of 120 kVp, 25 or 50 mA, 10-mm collimation, and a 10-mm reconstruction interval at a helical pitch of two. The reconstructed CT images were 512 × 512 pixels in size with a section thickness of 10 mm. The 69 CT scans consisted of 2,052 sections. All cancers were confirmed either by biopsy or surgically.

5.3.3 Detection of Nodule Candidates on Thick-Slice CT Images

The flowchart of our CAD scheme utilizing an MTANN supervised lesion enhancement filter and a mixture of expert MTANNs for classification is shown in Fig. 5.5. In order to limit the processing area to the lungs, we segmented the lung regions in a CT image by using thresholding based on Otsu's threshold value determination [74]. Then, we applied a rolling-ball technique along the outlines of the extracted lung regions to include a nodule attached to the pleura in the segmented lung regions [25].

In order to enhance lung nodules in CT images, we trained an MTANN filter with 13 lung nodules in a training database and the corresponding teaching images that contained maps for the "likelihood of being nodules," as illustrated in

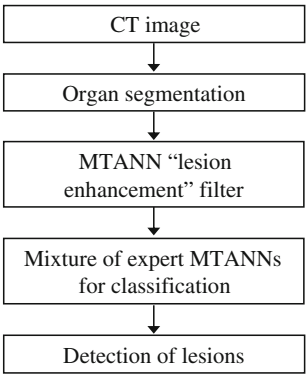


Fig. 5.5 Flowchart of our CAD scheme utilizing an MTANN supervised lesion enhancement filter and a mixture of expert MTANNs for classification

Fig. 5.2. In order to obtain the training regions, R_T, we applied a mathematical morphology dilation filter to the manually segmented lung nodules (i.e., binary regions) such that the training regions sufficiently covered nodules and surrounding normal structures (i.e., a nine times larger area than the nodule region, on average). A three-layer structure was employed for the MTANN filter, because any continuous mapping can be approximated by a three-layer ANN [75]. The number of hidden units was selected to be 20 by using a method for designing the structure of an ANN [76, 77]. The size of the input sub-region, R_S, was 9×9 pixels, which was determined experimentally in our previous studies [9, 10, 78]. The slope of the linear function, a, was 0.01. With the parameters above, the training of the MTANN filter was performed by 1,000,000 iterations. In order to test the performance, we applied the trained MTANN filter to the images of the entire lung. We applied thresholding to the output images of the trained MTANN filter to detect nodule candidates. We then compared the results of nodule-candidate detection with and without the MTANN filter.

We applied the trained MTANN filter to the original CT images. The result of the enhancement of nodules in CT images by the trained MTANN filter is shown in Fig. 5.6. The MTANN filter enhances a nodule and suppresses most normal structures in a CT image. Although some medium and large vessels in the hilum remain in the output image, the nodule with spiculation is enhanced well. We applied thresholding to the output images of the trained MTANN filter. There are a smaller number of candidates in the MTANN-based image, whereas there are many nodule candidates in the binary image obtained by using simple thresholding without the MTANN filter. Note that the large vessels in the hilum can easily be separated from nodules by using their areas.

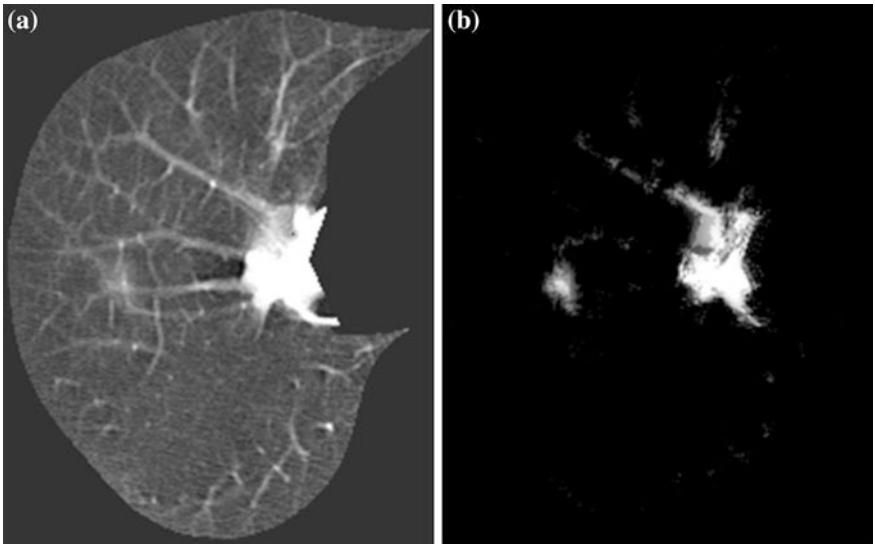


Fig. 5.6 Enhancement of a lesion by using the trained lesion-enhancement MTANN filter for a non-training case. **a** The original image of the segmented lung with a nodule (indicated by an *arrow*). **b** Output image of the trained lesion-enhancement MTANN filter. The nodule is enhanced in the output image, whereas most normal structures are suppressed

5.3.4 Classification of Nodule Candidates in Thick-Slice CT Images

Nodule candidates generally include more non-nodules (FPs) than nodules (TPs). For reduction of FPs, we trained an MTANN filter for classification of nodule candidates into nodules or non-nodules [9, 19]. We used 10 different-sized nodules with various contrasts and 10 non-nodule images including medium-sized and small vessels as training cases for the MTANN. Parameters such as the size of the subregion of the MTANN, the standard deviation of the 2D Gaussian function in the teaching image, and the size of the teaching image were determined by experimental analysis [48] to be 9×9 pixels, 5.0 pixels, and 19×19 pixels, respectively. We employed a three-layer structure for the MTANN, because it has been proven theoretically that a three-layer ANN can approximate any continuous mapping [75]. The number of hidden units in the MTANN was determined to be 20 by using a method for determining the structure of an ANN [76, 77]. Thus, the numbers of input, hidden, and output units were 81, 20, and 1, respectively. With the parameters above, the training of the MTANN was performed 500,000 times, and it converged with a mean absolute error of 0.112.

To investigate the performance of the classification of MTANN, we applied the trained MTANN to non-training cases. Figure 5.7 shows the output images of the trained MTANN, where various-sized actual nodules with different contrasts are represented by bright nodular distributions, whereas medium and small actual

vessels with different orientations are almost eliminated. In order to distinguish nodules from various types of non-nodules, we trained 6 expert MTANNs with 10 typical nodules and 6 types of 10 non-nodules, medium vessels, small vessels, large vessels, soft-tissue opacities, and abnormal opacities from a training database. We applied the trained expert MTANNs to various types of nodules and non-nodules. The trained expert MTANNs enhance nodules and suppress most normal structures including various-sized lung vessels in CT images, as shown in Fig. 5.7. The scores indicating the likelihood of being a nodule from the 6 expert MTANNs were combined with a mixing ANN to form a mixture of expert classification-MTANNs. We used a leave-one-out cross-validation test for testing the mixing ANN in the mixture of expert MTANNs. We evaluated the performance by using free-response receiver-operating-characteristic (FROC) analysis [79].

In order to test the performance of our CAD scheme utilizing the MTANN lesion enhancement filter and the classification MTANNs, we applied it to the test database containing 69 lung cancers. The MTANN lesion enhancement filter followed by thresholding identified 97 % (67/69) of cancers with 6.7 FPs per section. In contrast, the difference-image technique followed by multiple thresholding in a previously reported CAD scheme [10] detected 96 % (66/69) of cancers with 19.3 FPs per section. Thus, the MTANN lesion-enhancement filter was effective in improving the sensitivity and specificity of the CAD scheme. The classification-MTANNs were applied to the nodule candidates for classification of the candidates into nodules or non-nodules. The mixture of expert MTANNs was able to remove 60 % (8,172/13,688) or 93 % (12,667/13,688) of non-nodules with a loss of 1 true positive or 10 true positives, respectively. Thus, our MTANN-based CAD scheme achieved a 96 % (66/69) or 84 % (57/69) sensitivity with 2.7 (5,516/2,052) or 0.5 (1,021/2,052) FPs per section. In contrast, feature analysis and a rule-based scheme in the previously reported CAD scheme [10] removed FPs and achieved 9.3 FPs per section. Finally, with linear-discriminant analysis (LDA), the previously reported CAD scheme yielded a sensitivity of 84 % (57/69) with 1.4 (2,873/2,052) FPs per section. Thus, our CAD scheme utilizing MTANNs achieved a three times lower FP rate at the same sensitivity level. Therefore, MTANNs were effective for improving the sensitivity and specificity of our CAD scheme.

5.3.5 CAD Scheme for Thin-Slice CT

5.3.5.1 Database of Lung Nodules in Thin-Slice CT

Our database for thin slice CT contained 62 nodules in 32 scans acquired from 32 patients with a multi-detector-row CT (MDCT) system with a four-detector scanner. The MDCT scan consisted of an average of 186 thin-slice CT images (the slice thickness ranged from 1.0 to 2.5 mm). Each CT slice had an image matrix size of 512 $\times$ 512 pixels. Nodule sizes ranged from 5 to 30 mm. All nodules were confirmed by consensus of two chest radiologists.

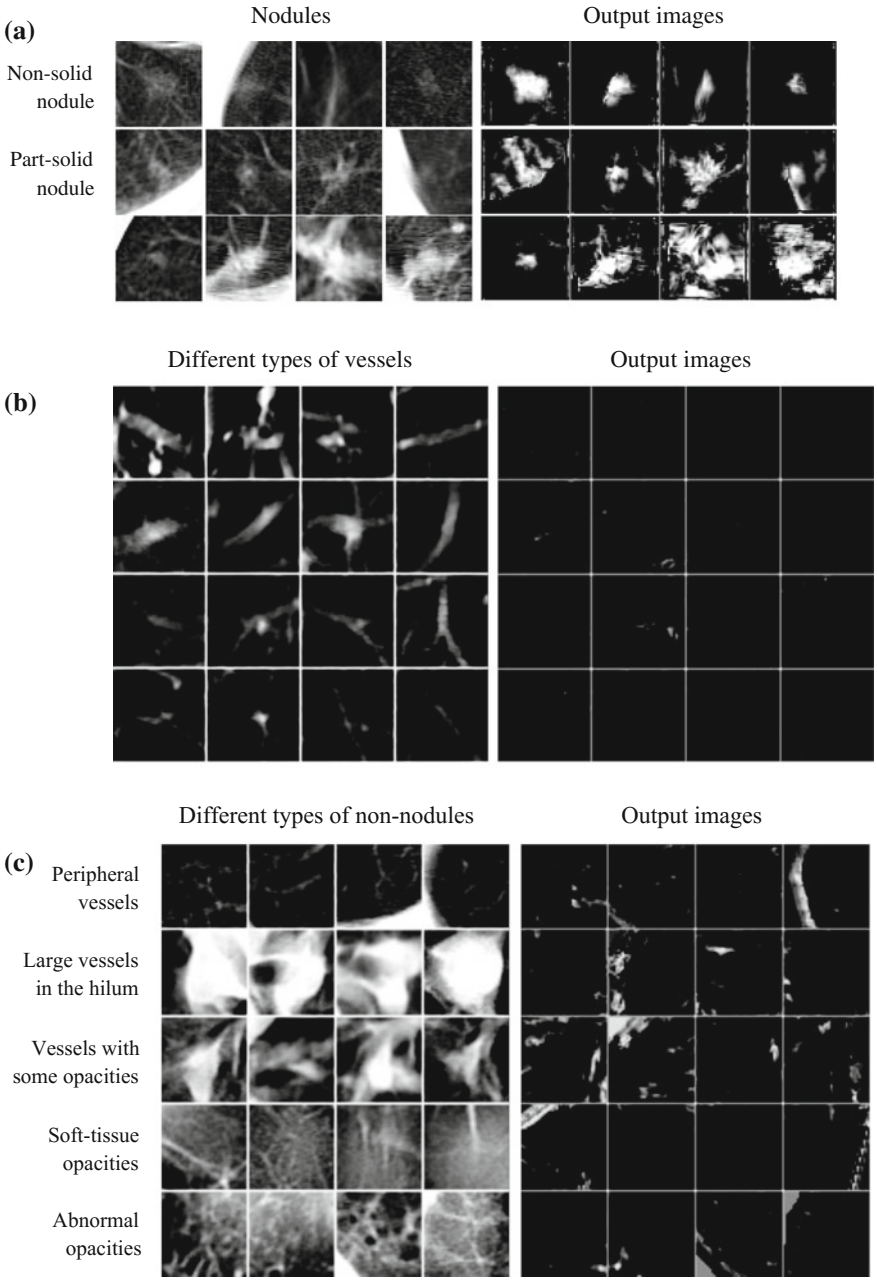


Fig. 5.7 Illustrations of (**a**) various types of nodules and the corresponding output images of the trained MTANN for non-training cases, (**b**) various-sized lung vessels and the corresponding output images, and (**c**) other types of non-nodules and the corresponding output images

5.3.5.2 Detection of Nodule Candidates on Thin-Slice CT Images

We developed an initial nodule detection scheme based on a selective enhancement filter [80] and a rule-based scheme with image features. For handling MDCT slices with different slice thickness, we converted original CT data to isotropic volumes. We applied the selective enhancement filter to the isotropic volumes for enhancing nodules and suppressing vessels. Thresholding followed by the rule-based scheme was applied to the filtered volumes for classification of candidates into nodules and non-nodules.

5.3.5.3 Overall Performance of a CAD Scheme for Thin-Slice CT

With our initial scheme, a sensitivity of 97 % (60/62 nodules) together with an average of 15 (476/32) FPs per patient was achieved. In order to remove eight types of non-nodules (FPs) reported by our initial scheme, we developed a mixture of eight expert 3D MTANNs. The size of the sub-volume and the size of the training volume in the teaching volume were determined to be $7 \times 7 \times 7$ voxels and $15 \times 15 \times 15$ voxels, respectively. Each 3D MTANN was trained 500,000 times with 10 representative nodules and 10 non-nodules in each of the 8 types. For distinction between nodules and each type of non-nodule, a scoring method was applied to the output volume of each trained 3D MTANN. The score was defined by the product of the output volume and a 3D Gaussian weighting function. A higher score indicated a nodule, and a lower score indicated a non-nodule. Eight expert 3D MTANNs were combined with a mixing ANN such that 8 types of non-nodules could be eliminated.

The trained mixture of expert 3D MTANNs was applied for the reduction of the FPs. Each 3D MTANN enhanced nodules and suppressed non-nodules representing the particular non-nodule type with which the 3D MTANN was trained, namely, various nodules in the output volumes of the 3D MTANN were represented by bright distributions, whereas the eight types of non-nodules were almost dark, as shown in Fig. 5.8. Although the distribution of scores for nodules and non-nodules obtained by using the scoring method overlapped, each 3D MTANN distinguished nodules from each type of non-nodule; therefore, the mixture of expert 3D MTANNs removed many non-nodules. The performance of the mixture of expert 3D MTANNs was evaluated by FROC analysis [79]. Results indicated that 57 % (273/476) of FPs were removed without a loss of any TP by the mixture of expert 3D MTANNs, as shown in Fig. 5.9. Thus, the FP rate of our CAD scheme was improved to 6.3 (203/32) FPs per patient at an overall sensitivity of 97 % (60/62 nodules).

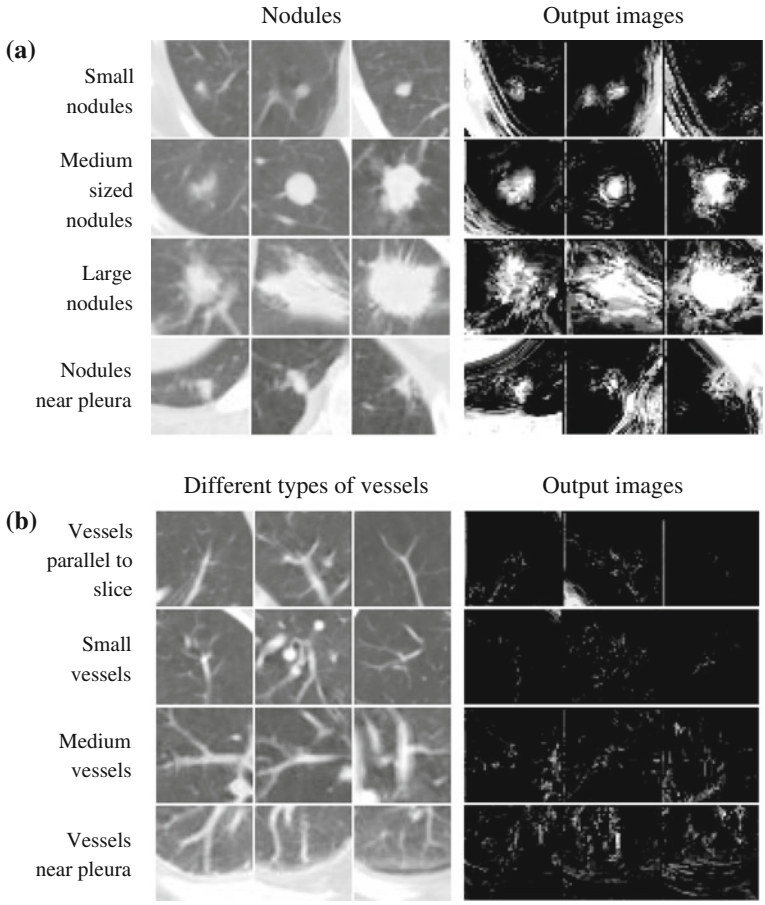


Fig. 5.8 Illustrations of (**a**) various types of nodules and the corresponding output images of the trained MTANN for non-training thin-slice CT images, and (**b**) various types of lung vessels and the corresponding output images

5.4 CAD Scheme for Detection of Polyps in CTC

5.4.1 Colorectal Cancer Detection in CTC

Colorectal cancer is the second leading cause of cancer deaths in the United States [53]. Evidence suggests that early detection of polyps (i.e., precursors of colorectal cancer) can reduce the incidence of colorectal cancer [81, 82]. CT colonography (CTC), also known as virtual colonoscopy, is a technique for detecting colorectal neoplasms by using a CT scan of the colon [83]. The diagnostic performance of CTC in detecting polyps, however, remains uncertain because of a propensity for

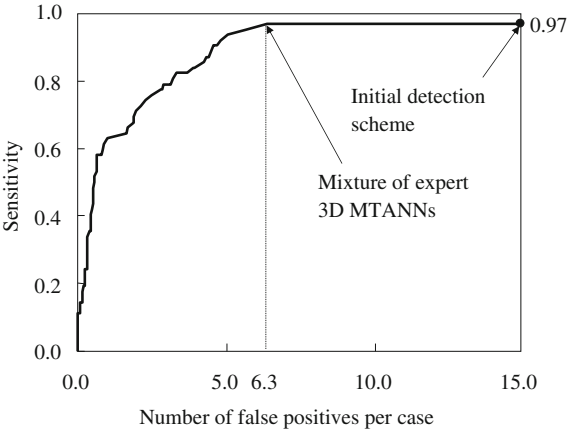


Fig. 5.9 Performance of the mixture of expert 3D MTANNs for classification between 60 nodules and 476 non-nodules on thin-slice CT images. The FROC curve indicates that the mixture of expert 3D MTANNs yielded a reduction of 57 % of non-nodules (FPs) without loss of any true positive, i.e., it achieved a 97 % sensitivity with 6.3 FPs per patient

perceptual errors [84]. Computer-aided detection (CAD) of polyps has been investigated in an effort to overcome the difficulty of CTC. CAD has the potential to improve radiologists' diagnostic performance in the detection of polyps.

Although CAD schemes are useful for improving radiologists' sensitivity in the detection of polyps in CTC, a major challenge for CAD schemes is reducing the number of FPs, while maintaining a high sensitivity. Major sources of FPs generated by CAD schemes include haustral folds, residual stool, rectal tubes, the ileocecal valve, and extra-colonic structures such as the small bowel and stomach [85]. Our purpose in this study was to develop a mixture of expert 3D MTANNs for further reduction of FPs in a polyp-detection CAD scheme while a high sensitivity is maintained.

5.4.2 CTC Database

CTC examinations were performed on 73 patients at The University of Chicago Medical Center. The patients' colons were prepared by standard pre-colonoscopy cleansing with administration of cathartics following a water diet or low-fiber diet, and they were insufflated with room air or carbon dioxide. Each patient was scanned in both supine and prone positions. Our database thus contained 146 CTC datasets. The CT scans were performed with either a single- or a multi-detector-row CT scanner (HiSpeed CTi or LightSpeed QX/i, GE Medical Systems, Milwaukee, WI). The CT scanning parameters included collimations between 2.5 and 5.0 mm, reconstruction intervals of 1.0–5.0 mm [1.0 mm (n = 2, 1 % of the CTC datasets), 1.25 mm (n = 3, 2 %), 1.5 mm (n = 59, 41 %), 2.5 mm (n = 79, 54 %), and 5.0 mm (n = 3, 2 %)], and tube currents of 60–120 mA with 120 kVp. Each reconstructed CT section had a matrix size of 512 × 512 pixels, with an in-plane pixel size of 0.5–0.7 mm. The CT sections were interpolated in isotropic resolution

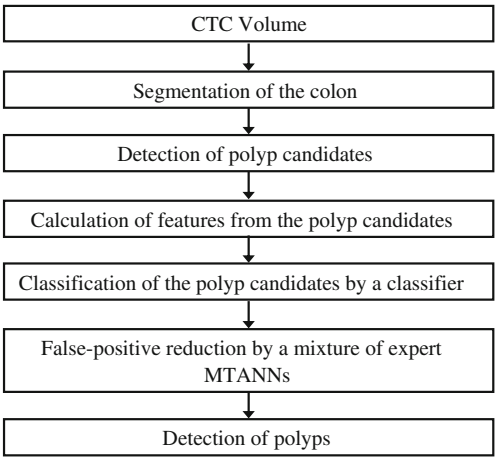


Fig. 5.10 Flowchart of our CAD scheme utilizing the mixture of expert MTANNs for detection of polyps in CTC

by using linear interpolation in the transverse direction. All patients underwent "reference-standard" optical colonoscopy. Radiologists established the locations of polyps in the CTC datasets by using the colonoscopy and pathology reports, as well as multiplanar reformatted views of the CTC on a viewing workstation (GE Advantage Windows Workstation v.4.2, GE Medical Systems, Milwaukee, WI). In this study, we used 5 mm as the threshold for clinically significant polyps [86]. Fifteen patients had 28 polyps, 15 of which were 5–9 mm in diameter and 13 were 10–25 mm. No polyp was submerged in fluid. Fluid was minimized by using a saline cathartic preparation as the standard preparation, not a colon gavage. We also created a training database of non-polyps by manual extraction of volumes containing non-polyps from 27 "normal" (non-polyp) CTC cases.

5.4.3 Performance of Our Initial CAD Scheme

Figure 5.10 is a block diagram of our CAD scheme for the detection of polyps in CTC. We applied our previously reported CAD scheme [17, 87] to the 73 CTC cases. The scheme included a centerline-based extraction of the colon [88], shape-based detection of polyps [17, 89], and an initial reduction of FPs by using a Bayesian ANN [90] based on geometric and texture features [87, 91]. The shape index used in the initial polyp detection step is calculated by using the Hessian matrix. This index determines to which of the following five topologic shapes an object belongs: cup, rut, saddle, ridge, or cap, as shown in Fig. 5.11. Polypoid polyps can be identified with the shape index as a cap shape. A haustral fold can be identified as a saddle or ridge. The colonic wall can be identified as rut or cup. We evaluated supine and prone CTC volumes independently. This CAD scheme achieved a 96.4 % (27/28 polyps) by-polyp sensitivity with an average of 3.1 (224/73) FPs per patient. Forty-eight

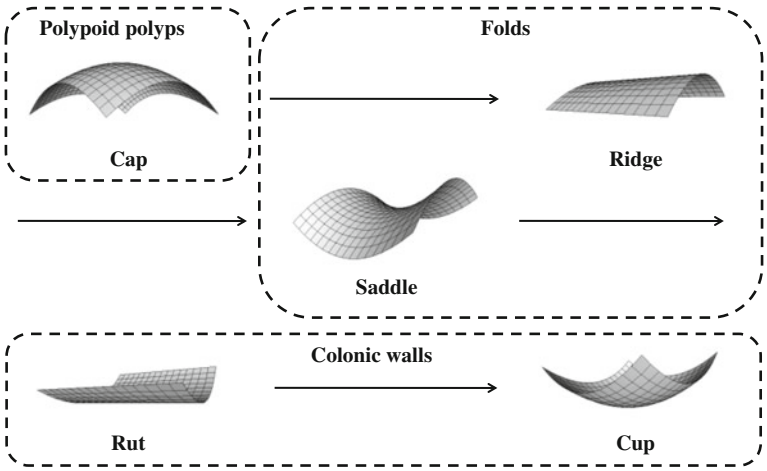


Fig. 5.11 Shape index for characterizing five shapes. Polypoid polyps can be identified with the shape index as a cap. Haustral folds can be identified as a saddle or ridge. Colonic walls can be identified as rut or cup

true-positive polyp detections in both supine and prone CTC volumes represented 27 polyps. We combined our previously reported CAD scheme with the mixture of expert 3D MTANNs for further reduction of FPs.

5.4.4 Training of Expert 3D MTANNs

We manually selected 10 representative polyp volumes (10 polyps) from the 48 true-positive volumes (containing 27 polyps) in our CTC database as the training polyp cases for expert 3D MTANNs. We classified CAD-generated FP sources into eight categories, i.e., rectal tubes, small bulbous folds, solid stool, stool with bubbles, colonic walls with haustral folds, elongated folds, strip-shaped folds, and the ileocecal valve. We manually selected 10 non-polyps in each of the 8 categories from the training non-polyp database (which was not used for testing). The 10 polyps and the 10 rectal tubes were the same as those used in our previous study [18]. The number of sample volumes for each category was 10, because the performance of an expert 3D MTANN was found to be highest when the number of training sample volumes was 20 (i.e., 10 polyps and 10 non-polyps) in our previous study [18], and the performance of 2D/3D MTANNs was not sensitive to the number of sample regions/volumes over different types of non-lesions in our previous studies [9, 18, 32, 78, 92].

The architecture of a mixture of expert 3D MTANNs is shown in Fig. 5.12. We trained 8 expert 3D MTANNs with the 10 polyps and 10 non-polyps in each category. A three-layer structure was employed for the expert 3D MTANNs [75].

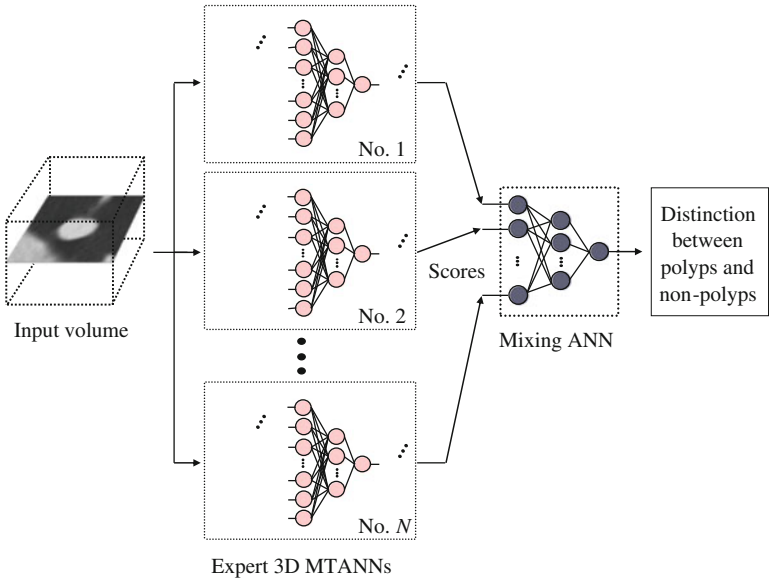


Fig. 5.12 A mixture of expert 3D MTANNs for distinguishing lesions (polypoid and flat lesions) from various types of non-lesions. Each expert 3D MTANN consists of a linear-output ANN regression model. Each MTANN is an expert for distinguishing lesions from a specific type of non-nodule. The outputs of the expert 3D MTANNs are combined with a mixing ANN so that the mixture of expert 3D MTANNs can remove various types of non-lesions

The size of the training volume and the standard deviation of the 3D Gaussian distribution in the teaching volume were 15 × 15 × 15 voxels (i.e., cubic shape) and 4.5 voxels, respectively, which were determined empirically based on our previous studies [9, 10, 18, 78]. The number of hidden units was selected to be 25 by using a method for designing the structure of an ANN [76, 77]. With the parameters above, training of the expert 3D MTANNs was performed by 500,000 iterations. We selected four among the eight expert 3D MTANNs for the mixture of expert 3D MTANNs by experimental analysis, because the mixture of these four expert 3D MTANNs [(1) rectal tubes, (2) stool with bubbles, (3) colonic walls with haustral folds, and (4) solid stool] demonstrated the highest performance.

5.4.5 Evaluation of the Performance for False-Positive Reduction

We applied the trained expert 3D MTANNs to the 27 polyps (48 true-positive volumes) and all 224 non-training FPs identified by our previously reported CAD scheme. The output volumes for these testing cases are shown in Fig. 5.13. The centers of the input volumes corresponded to the detection results provided by the

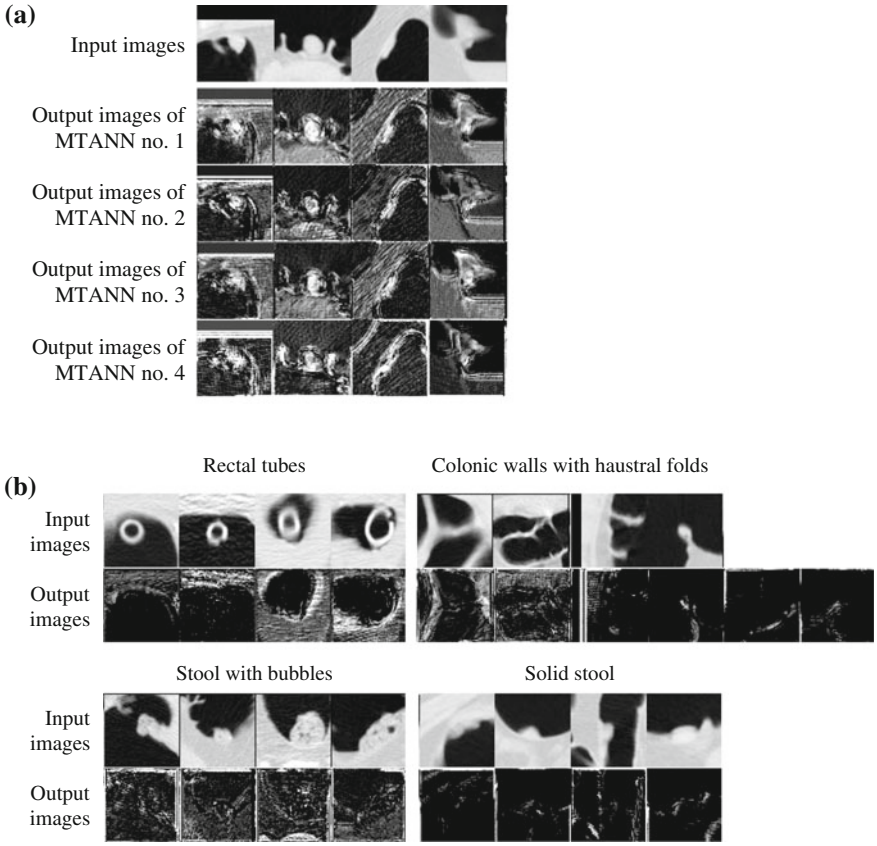


Fig. 5.13 Illustrations of (**a**) various testing polyps and the corresponding output volumes of four trained expert 3D MTANNs and (**b**) four different categories of testing FPs and the output volumes from the corresponding expert 3D MTANNs. In the output volumes, polyps appear as distributions of bright voxels (i.e., they are enhanced), whereas different types of FPs appear as dark voxels (i.e., they are suppressed)

CAD scheme (including both true positives and FPs); thus, this experiment included the effect of actual off-centering of polyp candidates produced by the initial CAD scheme. Various types of polyps, including a sessile polyp (the third image from the left in Fig. 5.13a), are represented in the output by distributions of bright voxels, whereas various types of non-polyps appear as darker voxels, indicating the ability of the expert 3D MTANNs to enhance polyps and suppress different types of non-polyps. We applied the 3D scoring method to the output volumes for polyps and non-polyps. The 3D Gaussian weighting function used the same standard deviation as that for the 3D Gaussian distribution in the polyp teaching volume. Although two distributions of scores in each graph overlapped, a substantial fraction of FPs was eliminated by using the expert 3D MTANNs.

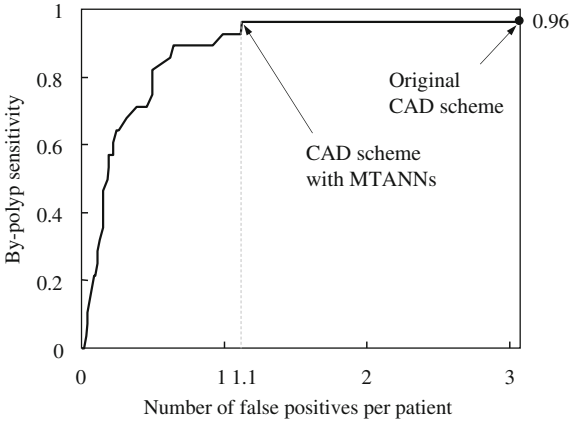


Fig. 5.14 FROC curve that shows the overall performance of the mixture of expert 3D MTANNs when it was applied to the entire database of 27 polyps (48 TPs volumes) and 224 FPs. The FROC curve indicates that the mixture of expert 3D MTANNs yielded a reduction of 63 % (142/224) of non-polyps (FPs) without removal of any true positives, i.e., it achieved a 100 % (27/27 or 17/17) classification performance

We evaluated the overall performance of the mixture of expert 3D MTANNs for FP reduction by using FROC analysis [79]. The FROC curve of the trained mixture of expert 3D MTANNs is shown in Fig. 5.14. The FROC curve was obtained by a change in the threshold value for the output of the mixing ANN. This FROC curve indicates that the mixture of expert 3D MTANNs eliminated 63 % (142/224) of non-polyps (FPs) without removal of any of the 27 polyps, i.e., a 96.4 % (27/28) overall by-polyp sensitivity was achieved at an FP rate of 1.1 (82/73) per patient.

5.4.6 Evaluation of a CAD Scheme with False-Negative CTC Cases

One of the limitations of current CAD research is a lack of evaluation of "difficult" polyps, particularly those which radiologists failed to detect by using standard techniques. Most previously reported studies used polyps detected by radiologists in CTC (i.e., human true-positive (TP)polyps). CAD benefits cannot be fully evaluated based on such TP polyps because these polyps are likely to be detected by radiologists without CAD.

5.4.6.1 Database of False-Negative Polyps

In order to evaluate the performance of a CAD scheme with false-negative (FN) polyps, we collected a database consisting of CTC scans obtained from a previous multicenter clinical trial [93] that included an air-contrast barium enema, as well as same-day CTC and colonoscopy. Six-hundred and fourteen high-risk subjects participating in the original trial were scanned in both supine and prone positions with a MDCT system. The reference standard was a final reconciliation of the

unblinded lesions identified on all of the three examinations. In the original trial, 155 patients had 234 clinically significant polyps of 6 mm or larger. Among them, 69 patients had FN interpretations (i.e., the by-patient sensitivity was 55 %). These patients had 114 "missed" polyps/masses which were not detected by reporting radiologists during their initial clinical reading. Causes of errors included observer errors, i.e., perceptual and measurement errors (51 %), technical errors (23 %), and non-reconcilable cases (26 %) [94]. The perceptual errors were associated with polyps that failed to be detected by observers. The measurement errors refer to the errors associated with the undermeasurement of polyp size as compared to colonoscopy findings as the "reference standard." In our study, we focused on FN cases with observer errors, because the aim of CAD is to prevent observer errors.

We used the inclusion criterion that each case had at least one "missed" polyp due to the perceptual error. As a result, we obtained 24 FN cases with 23 polyps and one mass. An experienced radiologist reviewed CTC cases carefully and determined the locations of polyps with reference to colonoscopy reports. Polyp sizes ranged from 6 to 15 mm, with an average of 8.3 mm. The mass size was 35 mm. Among them, 14 lesions were adenomas. The radiologist determined the difficulty of detection for each polyp/mass as difficult, moderate, or easy, as well as the morphology of each polyp.

5.4.6.2 CAD Performance for False-Negative Cases

Our initial polyp-detection scheme yielded a sensitivity of 63 % with 21.0 FPs per patient. The 3D MTANNs [18, 19] removed many FPs, and our CAD scheme achieved a sensitivity of 58 % (14/24) with 8.6 (207/24) FPs per patient for the 24 missed lesion cases, whereas the conventional CAD scheme with LDA instead of the MTANNs achieved a sensitivity of 25 % at the same FP rate. There were statistically significant differences [95] between the sensitivity of the MTANN CAD scheme and that of the conventional LDA CAD scheme. Therefore, our MTANN CAD scheme has the potential to detect 58 % of missed polyp/mass cases with a reasonable number of FPs [34].

Among the 24 lesions, 17 polyps, 6 polyps, and 1 mass were classified as difficult, moderate, and easy, respectively. Among the 23 polyps, 12, 9, and 2 were categorized as sessile, sessile on a fold, and pedunculated, respectively. Figure 5.15 illustrates FN polyps detected by our MTANN CAD scheme. All examples were graded as difficult to detect. We expect our CAD scheme to be helpful in the detection of difficult polyps.

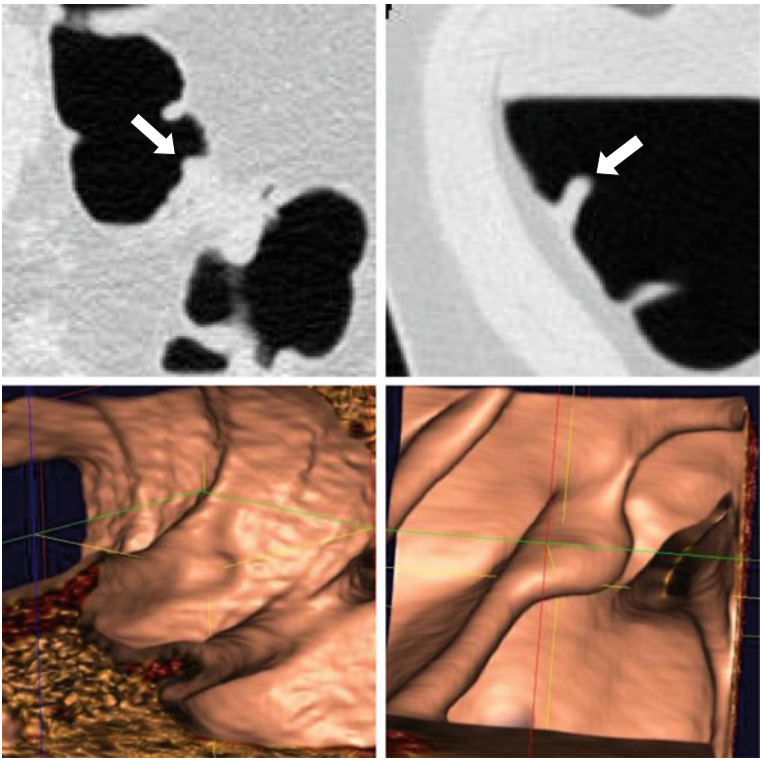


Fig. 5.15 Illustrations of polyps missed by reporting radiologists during their initial reading in the original trial in 2D axial views (*upper images*) and 3D endoluminal views (*lower images*), which were detected by our MTANN CAD scheme. **a** A small polyp (6 mm; hyperplastic) in the sigmoid colon was detected correctly by our CAD scheme (indicated by an *arrow*). This polyp was missed in both CTC and reference-standard optical colonoscopy in the original trial. **b** A sessile polyp on a fold (10 mm; adenoma) in the ascending colon

5.4.7 Detection of Flat Neoplasms by CAD

5.4.7.1 Morphologically Flat Neoplasms (Flat Lesions) in CTC

Current efforts to prevent colorectal cancer focus on the detection and removal of polypoid polyps (i.e., polypoid neoplasms). Recent studies, however, have shown that colorectal cancer can also arise from flat colorectal neoplasms (also known as flat lesions, non-polypoid lesions, superficial elevated lesions, or depressed lesions) [96]. Flat lesions are more likely than polypoid polyps to contain in situ or submucosal carcinoma. One study has shown that flat lesions contributed to 54 % of superficial carcinomas [97]. Flat lesions are also a major challenge for current gold-standard optical colonoscopy, because the subtle findings for these lesions

can be difficult to distinguish from those for the normal mucosa [98]. As compared to the surrounding normal mucosa, flat lesions appear to be slightly elevated, completely flat, or slightly depressed. Although flat lesions were believed to exist primarily in Asian countries such as Japan [99, 100], recent studies have shown their significance in other parts of the world [101] such as the European countries [96] and the United States [97]. Flat lesions in the Western population, thus, may have been missed in current gold-standard optical colonoscopy [102]. Although the detection sensitivity of polyps in CTC is comparable to that in optical colonoscopy [103], flat lesions are a potentially major source of FN CTC interpretations in view of their uncommon morphology [104, 105]. Thus, detection of flat lesions in CTC is essential in colorectal cancer screening.

5.4.7.2 Limitations of Current CAD Schemes for Flat-Lesion Detection

Although current CAD schemes could be useful for detection of polypoid polyps, the detection of flat lesions is a major challenge [106], because existing CAD schemes have focused on the detection of pedunculated and sessile polyps; thus, they are designed for detecting the common polypoid shape. Existing CAD schemes use geometric, morphologic, and textural characteristics to distinguish polyps from normal structures in the colon (e.g., haustral folds, stool, the air/liquid boundary, the ileocecal valve, and a rectal catheter). One of the most promising methods for distinguishing these polyps is to use the mathematical descriptor called the shape index to characterize the shape of a polyp [89]. A polyp is characterized by the shape index as a cap-like structure. Haustral folds and the colonic wall are characterized as saddle-like structures and cup-like structures, respectively. Thus, existing CAD schemes are not likely to detect flat lesions which exhibit a non-polypoid shape.

5.4.7.3 Flat-Lesion Database

To create a flat-lesion database, an expert radiologist measured lesions on CTC images on a CTC viewing workstation (Vitrea 2 software, version 3.9, Vital Images, Minnetonka, MN) [20, 107]. 2D images were viewed with three tailored window/level settings: “lung,” “soft tissue,” and “flat.” Magnified axial, coronal, and sagittal planes were reviewed in 2D for detection of the longest axis and the maximal height of the lesion as seen on each dataset (supine and prone). On a close-angle 3D endoluminal view, the lesion was viewed from various angles for first deciding on its borders. The longest axis and maximal height were measured on each dataset. Comparison of 2D and 3D images before measurements were made were permitted for assessment of the lesion shape and borders in the same session, because this approach corresponds to the method that would be used in clinical practice when lesions are measured. Measurements of maximal thickness

on the 3D-volume-rendered views required the observer to make a subjective best estimate as to where to place the cursor.

We analyzed data from the 3D endoluminal view and the 2D view in each of the three window/level settings to determine which measurements fit the definitions of flat lesions as determined by a height <3 mm or a ratio of height <1/2 of the long axis. Based on the measurements of 50 CTC cases by a radiologist, we found 28 flat lesions in 25 patients (i.e., the prevalence of flat lesions was about 30 %). Eleven flat lesions among the 28 lesions were not detected by reporting radiologists at their initial clinical reading in the original trial; i.e., these were missed lesions; therefore, they can be considered "very difficult" lesions to detect. Lesion sizes ranged from 6 to 18 mm with an average of 9 mm based on optical colonoscopy measurements.

5.4.7.4 Development of a 3D MTANN for Flat Lesions

In order to investigate the feasibility of a 3D MTANN in the detection of flat lesions, we applied a 3D MTANN to flat lesions in the flat-lesion database containing 28 flat lesions in 25 patients. We trained the 3D MTANN with sessile polyps (which are not flat lesions, but appear relatively flat compared to common bulbous polyps) in a different database and with various non-polyps such as a rectal tube, haustral folds, the ileocecal valve, and stool, which are major sources of FPs. We applied the trained 3D MTANN to the 28 flat lesions in the flat-lesion database.

5.4.7.5 Evaluation of the Performance of the CAD Scheme

Our initial polyp-detection scheme without LDA yielded a 71 % by-polyp sensitivity with 25 FPs per patient for the 28 flat lesions, including 11 lesions missed by the reporting radiologists in the original clinical trial. With LDA, 105 FPs were removed with the loss of one TP, thus yielding a 68 % by-polyp sensitivity with 16.3 FPs per patient. We applied the trained expert 3D MTANNs for further reduction of the FPs. The 3D MTANNs removed 39 % of the FPs without removal of any TPs. Thus, our CAD scheme achieved a by-polyp sensitivity of 68 % with 10 FPs per patient, including 6 of the 11 flat lesions missed by the reporting radiologists in the original trial. Our MTANN CAD scheme detected 67 % and 70 % of flat lesions ranging from 6-9 mm and those 10 mm or larger, respectively, including six lesions missed by the reporting radiologists in the original trial with 10 FPs per patient.

Figure 5.16 shows an example of a flat lesion which is very small. Some flat lesions are known to be histologically aggressive; therefore, the detection of such lesions is critical clinically, but they are difficult to detect because of their uncommon morphology. Our CAD scheme detected such difficult flat lesions correctly. It should be noted that this case was missed by the reporting radiologists in the original trial; thus, the detection of the lesion may be considered "very difficult."

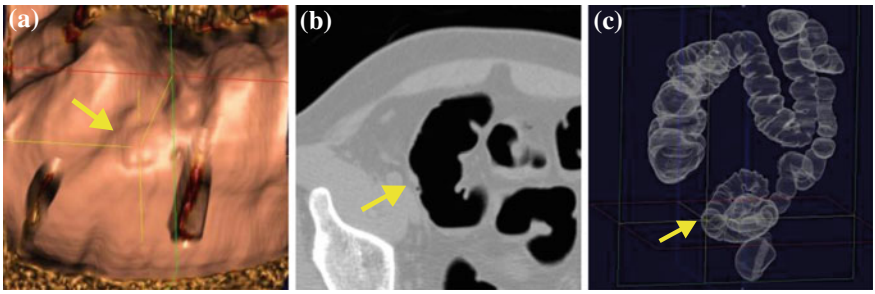


Fig. 5.16 Illustration of a flat lesion which was detected by our MTANN CAD scheme: **a** 3D endoluminal view, **b** 2D axial view, **c** 3D transparent colon view. A small flat lesion (6 mm; adenoma) in the sigmoid colon was detected correctly by our CAD scheme (indicated by an *arrow*). This polyp was missed in CTC in the original trial

5.5 Conclusion

PML is a powerful tool in CAD schemes for detection of lesions in medical images. MTANNs, which are a class of PML, were useful for improving the performance (i.e., both sensitivity and specificity) of CAD schemes for detection of lung nodules in CT and the detection of polyps in CT colonography. The MTANN supervised filter was effective for enhancement of lesions including lung nodules and colorectal polyps and suppression of non-lesions in medical images, which contributed to the improvement of the sensitivity as well as specificity in the initial lesion detection stage in CAD schemes, whereas the classification MTANNs contributed to the improvement of specificity in the FP reduction stage in CAD schemes.

Acknowledgments This work would not have been possible without the help of countless individuals. The author acknowledges the invaluable assistance of all colleagues and support staff. The author is grateful to all members in the Suzuki Laboratory in the Department of Radiology at the University of Chicago, especially Ivan Sheu, Mark L Epstein, Jianwu Xu, and Sheng Chen, for their contributions to the studies, to colleagues and collaborators, especially Abraham H Dachman, Heber MacMahon, Kunio Doi, Samuel G Armato III, Feng Li, Shusuke Sone, Hiroyuki Abe, Qiang Li, Junji Shiraishi, Don C Rockey, Hiroyuki Yoshida, and Janne Nappi for their valuable suggestions, and to Ms. E. F. Lanzl for improving the chapter. The author is also grateful to his wife, Harumi Suzuki, for her assistance with the chapter and studies, and his daughters, Mineru Suzuki and Juno Suzuki, for cheering him up. This work was partly supported by Grant Number R01CA120549 from the National Cancer Institute/National Institutes of Health and by NIH S10 RR021039 and P30 CA14599.

References

1. Giger ML, Suzuki K (2007) Computer-aided diagnosis (CAD). In: Feng DD (ed) Biomedical information technology. Academic Press, New York, pp 359–374
2. Doi K (2005) Current status and future potential of computer-aided diagnosis in medical imaging. Br J Radiol 78(1):S3–S19
3. Li F, Aoyama M, Shiraishi J, Abe H, Li Q et al (2004) Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. Am J Roentgenol 183(5):1209–1215
4. Li F, Arimura H, Suzuki K, Shiraishi J, Li Q et al (2005) Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. Radiology 237(2):684–690
5. Dean JC, Ilvento CC (2006) Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. Am J Roentgenol 187(1):20–28
6. Suzuki K, Shiraishi J, Abe H, MacMahon H, Doi K (2005) False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network. Acad Radiol 12(2):191–201
7. van Ginneken B, ter Haar Romeny BM, Viergever MA (2001) Computer-aided diagnosis in chest radiography: a survey. IEEE Trans Med Imaging 20(12):1228–1241
8. Giger ML, Doi K, MacMahon H (1988) Image feature analysis and computer-aided diagnosis in digital radiography. 3. Automated detection of nodules in peripheral lung fields. Med Phys 15(2):158–166
9. Suzuki K, Armato SG, Li F, Sone S, Doi K (2003) Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose CT. Med Phys 30(7):1602–1617
10. Arimura H, Katsuragawa S, Suzuki K, Li F, Shiraishi J et al (2004) Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening. Acad Radiol 11(6):617–629
11. Armato SG 3rd, Giger ML, Moran CJ, Blackburn JT, Doi K et al (1999) Computerized detection of pulmonary nodules on CT scans. Radiographics 19(5):1303–1311
12. Armato SG 3rd, Li F, Giger ML, MacMahon H, Sone S et al (2002) Lung cancer: performance of automated lung nodule detection applied to cancers missed in a CT screening program. Radiology 225(3):685–692
13. Chan HP, Doi K, Galhotra S, Vyborny CJ, MacMahon H et al (1987) Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. Med Phys 14(4):538–548
14. Gilhuijs KG, Giger ML, Bick U (1998) Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging. Med Phys 25(9):1647–1654
15. Horsch K, Giger ML, Vyborny CJ, Venta LA (2004) Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 11(3):272–280
16. Drukker K, Giger ML, Metz CE (2005) Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 237(3):834–840
17. Yoshida H, Nappi J (2001) Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. IEEE Trans Med Imaging 20(12):1261–1274
18. Suzuki K, Yoshida H, Nappi J, Dachman AH (2006) Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: suppression of rectal tubes. Med Phys 33(10):3814–3824
19. Suzuki K, Yoshida H, Nappi J, Armato SG 3rd, Dachman AH (2008) Mixture of expert 3D massive-training ANNs for reduction of multiple types of false positives in CAD for detection of polyps in CT colonography. Med Phys 35(2):694–703
20. Lostumbo A, Wanamaker C, Tsai J, Suzuki K, Dachman AH (2010) Comparison of 2D and 3D views for evaluation of flat lesions in CT colonography. Acad Radiol 17(1):39–47

21. Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic Press, San Diego
22. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
23. Vapnik N (1995) The nature of statistical learning theory. Springer, Berlin
24. Shiraishi J, Li Q, Suzuki K, Engelmann R, Doi K (2006) Computer-aided diagnostic scheme for the detection of lung nodules on chest radiographs: localized search method based on anatomical classification. Med Phys 33(7):2642–2653
25. Armato SG 3rd, Giger ML, MacMahon H (2001) Automated detection of lung nodules in CT scans: preliminary results. Med Phys 28(8):1552–1561
26. Aoyama M, Li Q, Katsuragawa S, MacMahon H, Doi K (2002) Automated computerized scheme for distinction between benign and malignant solitary pulmonary nodules on chest images. Med Phys 29(5):701–708
27. Aoyama M, Li Q, Katsuragawa S, Li F, Sone S et al (2003) Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images. Med Phys 30(3):387–394
28. Jerebko K, Summers RM, Malley JD, Franaszek M, Johnson CD (2003) Computer-assisted detection of colonic polyps with CT colonography using neural networks and binary classification trees. Med Phys 30(1):52–60
29. Suzuki K, Horiba I, Sugie N (2002) Efficient approximation of neural filters for removing quantum noise from images. IEEE Trans Signal Process 50(7):1787–1799
30. Suzuki K, Horiba I, Sugie N (2003) Neural edge enhancer for supervised edge enhancement from noisy images. IEEE Trans Pattern Anal Mach Intell 25(12):1582–1596
31. Suzuki K, Horiba I, Sugie N, Nanki M (2004) Extraction of left ventricular contours from left ventriculograms by means of a neural edge detector. IEEE Trans Med Imaging 23(3):330–339
32. Suzuki K, Li F, Sone S, Doi K (2005) Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network. IEEE Trans Med Imaging 24(9):1138–1150
33. Suzuki K, Abe H, MacMahon H, Doi K (2006) Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN). IEEE Trans Med Imaging 25(4):406–416
34. Suzuki K, Rockey DC, Dachman AH (2010) CT colonography: Advanced computer-aided detection scheme utilizing MTANNs for detection of "missed" polyps in a multicenter clinical trial. Med Phys 30:2–21
35. Suzuki K, Zhang J, Xu J (2010) Massive-training artificial neural network coupled with Laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in CT colonography. IEEE Trans Med Imaging 29(11):1907–1917
36. Xu J, Suzuki K (2011) Massive-training support vector regression and Gaussian process for false-positive reduction in computer-aided detection of polyps in CT colonography. Med Phys 38(4):1888–1902
37. Suzuki K, Horiba I, Sugie N, Nanki M (2002) Neural filter with selection of input features and its application to image quality improvement of medical image sequences. IEICE Trans Inf Syst E85-D(10):1710–1718
38. Lo SB, Lou SA, Lin JS, Freedman MT, Chien MV et al (1995) Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans Med Imaging 14(4):711–718
39. Lo SCB, Chan HP, Lin JS, Li H, Freedman MT et al (1995) Artificial convolution neural network for medical image pattern recognition. Neural Netw 8(7–8):1201–1214
40. Lin JS, Lo SB, Hasegawa A, Freedman MT, Mun SK (1996) Reduction of false positives in lung nodule detection using a two-level neural classification. IEEE Trans Med Imaging 15(2):206–217

41. Lo SC, Li H, Wang Y, Kinnard L, Freedman MT (2002) A multiple circular path convolution neural network system for detection of mammographic masses. IEEE Trans Med Imaging 21(2):150–158
42. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA et al (1996) Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 15(5):598–610
43. Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw 8(1):98–113
44. Neubauer C (1998) Evaluation of convolutional neural networks for visual recognition. IEEE Trans Neural Netw 9(4):685–696
45. Wei D, Nishikawa RM, Doi K (1996) Application of texture analysis and shift-invariant artificial neural network to microcalcification cluster detection. Radiology 201:696–696
46. Zhang W, Doi K, Giger ML, Nishikawa RM, Schmidt RA (1996) An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. Med Phys 23(4):595–601
47. Zhang W, Doi K, Giger ML, Wu Y, Nishikawa RM et al (1994) Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. Med Phys 21(4):517–524
48. Suzuki K, Armato SG 3rd, Li F, Sone S, Doi K (2003) Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. Med Phys 30(7):1602–1617
49. Oda S, Awai K, Suzuki K, Yanaga Y, Funama Y et al (2009) Performance of radiologists in detection of small pulmonary nodules on chest radiographs: effect of rib suppression with a massive-training artificial neural network. Am J Roentgenol 193(5):W397–402
50. Suzuki K (2009) A supervised 'lesion-enhancement' filter by use of a massive-training artificial neural network (MTANN) in computer-aided diagnosis (CAD). Phys Med Biol 54(18):S31–45
51. Xu JW, Suzuki K (2010) False-positive reduction in computer-aided detection of polyps in CT colonography: a massive-training support vector regression approach. In: MICCAI workshop on computational challenges and clinical opportunities in virtual colonoscopy and abdominal imaging, Beijing, China pp. 55–60
52. Mosier CI (1951) Problems and designs of cross-validation. Educ Psychol Meas 11:5–11
53. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC et al (2005) Cancer statistics, 2005. CA Cancer J Clin 55(1):10–30
54. Cancer Facts and Figures (2005) American Cancer Society, Atlanta
55. Flehinger BJ, Kimmel M, Melamed MR (1992) The effect of surgical treatment on survival from early lung cancer. Implications for screening. Chest 101(4):1013–1018
56. Sobue T, Suzuki T, Matsuda M, Kuroishi T, Ikeda S et al (1992) Survival for clinical stage I lung cancer not surgically treated. Comparison between screen-detected and symptom-detected cases. The Japanese Lung Cancer Screening Research Group. Cancer 69(3):685–692
57. Miettinen S (2000) Screening for lung cancer. Radiol Clin North Am 38(3):479–486
58. Heelan RT, Flehinger BJ, Melamed MR, Zaman MB, Perchick WB et al (1984) Non-small-cell lung cancer: results of the New York screening program. Radiology 151(2):289–293
59. Frost JK, Ball WC Jr, Levin ML, Tockman MS, Baker RR et al (1984) Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Johns Hopkins study. Am Rev Respir Dis 130(4):549–554
60. Flehinger BJ, Melamed MR, Zaman MB, Heelan RT, Perchick WB et al (1984) Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Memorial Sloan-Kettering study. Am Rev Respir Dis 130(4):555–560
61. Fontana RS, Sanderson DR, Taylor WF, Woolner LB, Miller WE et al (1984) Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the Mayo Clinic study. Am Rev Respir Dis 130(4):561–565

62. Kubik A, Polak J (1986) Lung cancer detection. Results of a randomized prospective study in Czechoslovakia. Cancer 57(12):2427–2437
63. Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G et al (1999) Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 354(9173):99–105
64. Miettinen S, Henschke CI (2001) CT screening for lung cancer: coping with nihilistic recommendations. Radiology 221(3):592–596
65. Henschke CI, Naidich DP, Yankelevitz DF, McGuinness G, McCauley DI et al (2001) Early lung cancer action project: initial findings on repeat screenings. Cancer 92(1):153–159
66. Swensen SJ, Jett JR, Hartman TE, Midthun DE, Sloan JA et al (2003) Lung cancer screening with CT: Mayo Clinic experience. Radiology 226(3):756–761
67. Kaneko M, Eguchi K, Ohmatsu H, Kakinuma R, Naruke T et al (1996) Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology 201(3):798–802
68. Sone S, Takashima S, Li F, Yang Z, Honda T et al (1998) Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet 351(9111):1242–1245
69. Sone S, Li F, Yang ZG, Honda T, Maruyama Y et al (2001) Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner. Br J Cancer 84(1):25–32
70. Nawa T, Nakagawa T, Kusano S, Kawasaki Y, Sugawara Y et al (1996) Lung cancer screening using low-dose spiral CT: results of baseline and 1-year follow-up studies. Chest 122(1):15–20
71. Gurney JW (1996) Missed lung cancer at CT: imaging findings in nine patients. Radiology 199(1):117–122
72. Li F, Sone S, Abe H, MacMahon H, Armato SG 3rd et al (2002) Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings. Radiology 225(3):673–683
73. Kobayashi T, Xu XW, MacMahon H, Metz CE, Doi K (1996) Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 199(3):843–848
74. Otsu N (1979) A threshold selection method from gray level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
75. Funahashi K (1989) On the approximate realization of continuous mappings by neural networks. Neural Netw 2:183–192
76. Suzuki K, Horiba I, Sugie N (2001) A simple neural network pruning algorithm with application to filter synthesis. Neural Process Lett 13(1):43–53
77. Suzuki K (2004) Determining the receptive field of a neural filter. J Neural Eng 1(4):228–237
78. Suzuki K, Doi K (2005) How can a massive training artificial neural network (MTANN) be trained with a small number of cases in the distinction between nodules and vessels in thoracic CT? Acad Radiol 12(10):1333–1341
79. Egan JP, Greenberg GZ, Schulman AI (1961) Operating characteristics, signal detectability, and the method of free response. J Acoust Soc Am 33:993–1007
80. Li Q, Sone S, Doi K (2003) Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. Med Phys 30(8):2040–2051
81. Winawer SJ, Fletcher RH, Miller L, Godlee F, Stolar MH et al (1997) Colorectal cancer screening: clinical guidelines and rationale. Gastroenterology 112(2):594–642
82. Dachman H (2003) Atlas of virtual colonoscopy. Springer, New York
83. Macari M, Bini EJ (2005) CT colonography: where have we been and where are we going? Radiology 237(3):819–833
84. Fletcher JG, Booya F, Johnson CD, Ahlquist D (2005) CT colonography: unraveling the twists and turns. Curr Opin Gastroenterol 21(1):90–98
85. Yoshida H, Dachman AH (2005) CAD techniques, challenges, and controversies in computed tomographic colonography. Abdom Imaging 30(1):26–41

86. Johnson CD, Dachman AH (2000) CT colonography: the next colon screening examination? Radiology 216(2):331–341
87. Nappi J, Yoshida H (2003) Feature-guided analysis for reduction of false positives in CAD of polyps for computed tomographic colonography. Med Phys 30(7):1592–1601
88. Frimmel H, Nappi J, Yoshida H (2004) Fast and robust computation of colon centerline in CT colonography. Med Phys 31(11):3046–3056
89. Yoshida H, Masutani Y, MacEneaney P, Rubin DT, Dachman AH (2002) Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study. Radiology 222(2):327–336
90. Kupinski MA, Edwards DC, Giger ML, Metz CE (2001) Ideal observer approximation using Bayesian classification neural networks. IEEE Trans Med Imaging 20(9):886–899
91. Nappi J, Yoshida H (2002) Automated detection of polyps with CT colonography: evaluation of volumetric features for reduction of false-positive findings. Acad Radiol 9(4):386–397
92. Suzuki K, Armato SG, Li F, Sone S, Doi K (2003) Effect of a small number of training cases on the performance of massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose CT. In: Proceedings of SPIE Medical Imaging (SPIE MI), San Diego, CA, pp 1355–1366
93. Rockey DC, Paulson E, Niedzwiecki D, Davis W, Bosworth HB et al (2005) Analysis of air contrast barium enema, computed tomographic colonography, and colonoscopy: prospective comparison. Lancet 365(9456):305–311
94. Doshi T, Rusinak D, Halvorsen RA, Rockey DC, Suzuki K et al (2007) CT colonography: false-negative interpretations. Radiology 244(1):165–173
95. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM (2002) Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 29(12):2861–2870
96. Rembacken BJ, Fujii T, Cairns A, Dixon MF, Yoshida S et al (2000) Flat and depressed colonic neoplasms: a prospective study of 1000 colonoscopies in the UK. Lancet 355(9211):1211–1214
97. Soetikno RM, Kaltenbach T, Rouse RV, Park W, Maheshwari A et al (2008) Prevalence of nonpolypoid (flat and depressed) colorectal neoplasms in asymptomatic and symptomatic adults. JAMA 299(9):1027–1035
98. Soetikno R, Friedland S, Kaltenbach T, Chayama K, Tanaka S (2006) Nonpolypoid (flat and depressed) colorectal neoplasms. Gastroenterology 130(2):566–576; quiz 588–589
99. Kudo S, Kashida H, Tamura T (2000) Early colorectal cancer: flat or depressed type. J Gastroenterol Hepatol 15(Suppl):D66–70
100. Kudo S, Kashida H, Tamura T, Kogure E, Imai Y et al (2000) Colonoscopic diagnosis and management of nonpolypoid early colorectal cancer. World J Surg 24(9):1081–1090
101. Ross S, Waxman I (2006) Flat and depressed neoplasms of the colon in Western populations. Am J Gastroenterol 101(1):172–180
102. Fujii T, Rembacken BJ, Dixon MF, Yoshida S, Axon AT (1998) Flat adenomas in the United Kingdom: are treatable cancers being missed? Endoscopy 30(5):437–443
103. Johnson CD, Chen MH, Toledano AY, Heiken JP, Dachman A et al (2008) Accuracy of CT colonography for detection of large adenomas and cancers. N Engl J Med 359(12):1207–1217
104. Fidler JL, Johnson CD, MacCarty RL, Welch TJ, Hara AK et al (2002) Detection of flat lesions in the colon with CT colonography. Abdom Imaging 27(3):292–300
105. Fidler J, Johnson C (2008) Flat polyps of the colon: accuracy of detection by CT colonography and histologic significance. Abdom Imaging
106. Taylor SA, Suzuki N, Beddoe G, Halligan S (2008) Flat neoplasia of the colon: CT colonography with CAD. Abdom Imaging 34(2):173–181
107. Lostumbo A, Suzuki K, Dachman AH (2009) Flat lesions in CT colonography. Abdom Imaging 34(2):173–181

Chapter 6
Understanding Foot Function During Stance Phase by Bayesian Network Based Causal Inference

Myagmarbayar Nergui, Jun Inoue, Murai Chieko, Wenwei Yu and U. Rajendra Acharya

Abstract Understanding the biomechanics of the human foot during each stage of walking is important for the objective evaluation of movement dysfunction, accuracy of diagnosis, and prediction of foot impairment. Extracting causal relations from amongst the muscle activities, toe trajectories, and plantar pressures during walking assists in recognizing several disease conditions, and understanding the hidden complexity of human foot functions, thus, facilitating appropriate therapy and treatment. To extract these relations, we applied the Bayesian Network (BN) model to data collected in the stance phase of walking. For a better understanding of foot function, the experimental data were divided into three stages (initial contact, loading response to mid-stance, and terminal stance to pre-swing). BNs were constructed for these three stages of data for normal walking and simulated hemiplegic walking, then compared and analyzed. Results showed that BNs extracted could express the underlying mechanism of foot function.

Keywords Bayesian network · Stance phase · Muscle activities · Foot function

M. Nergui (✉)
Graduate School of Engineering, Mechanical Engineering and Medical System Engineering Department, Chiba University, Chiba, Japan
e-mail: myagaa@chiba-u.jp

U. R. Acharya
Department of ECE, Ngee Ann Polytechnic, Clementi, Singapore

U. R. Acharya
Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Lumpur, Malaysia

J. Inoue · M. Chieko · W. Yu
Graduate School of Engineering, Medical System Engineering Department, Chiba University, Chiba, Japan

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_6,

6.1 Introduction

A difference between an actual gait and a normal gait indicates a foot abnormality, which may be caused by dysfunction of the neural or musculoskeletal systems. Currently, most abnormalities of the foot are diagnosed and predicted empirically after subjective assessment, by which correct therapy and rehabilitation for foot impairment cannot be guaranteed. A better understanding of foot function could make an objective assessment possible, and thus is of great significance to not only research in the therapy and rehabilitation area, but also to research in the motor control research area.

The main goal of this research is to develop new tools for identifying the nature and cause of foot function impairment, and assist in treatment and rehabilitation of foot function. Foot function during walking is a result of interaction among the muscular, neural, and skeletal systems, and the walking environment.

In this study, we measured and recorded lower limb, major muscle activities (corresponding to the cause of the motion), the trajectories of toe and ankle joints (reflecting the effect of the motion), and plantar pressure distributions (representing the interaction between the human and environment) during the stance phase. Then, Bayesian network (BN) was applied as the theoretical account of probabilistic illation to extract the causal construction for foot function. Two kinds of walking, normal walking and simulated hemiplegic walking, were measured and analyzed to verify the BN's ability to express and distinguish the significant gait-dependent causal relations.

Our research differs from existing work in the literature in the following aspects.

1. The trajectory and pressure of the tiptoe stance were also recorded and modeled.
   Although the role of the toes in walking has been studied in robotics and gait research [1–3], it has not been studied in a foot function model.
2. The information structure of foot function, expressed by the causal relation among muscle activities, joint trajectories, and plantar pressure recordings were inferred.
   Several studies have been done for estimating foot abnormalities while evaluating normal foot function during walking [4–7]. For example, [5] have reported that the study of using plantar pressures with muscle activities for analyzing and estimating abnormalities of human gait. Also, some scientists have shown that lower-leg muscle impuissance is influenced by upper-leg muscles [8, 9]. Thus, empirical diagnosis cannot bring suitable treatment and rehabilitation methods. Moreover, the probabilistic causal inference could synthesize the information from several types of measurement, without any a priori, for example, the physical connection and relationship between functional components.
3. The BN is used to describe knowledge about an uncertainty. The BN principle consists of probability theory, graph representation, and statistics, and is employed to describe the probabilistic dependencies among random variables.

To better describe knowledge about an uncertainty and represent the probabilistic dependencies among random variables, some researchers explored and determined that the BN can be used to analyze biomedical signals and medical applications. Some examples of this research can be found in [11–15]. In these examples, BNs were applied as many kinds of tools, specifically, to extract the causal relations amongst symptoms and diseases from medical databases, to construct database from incomplete and partially correct statistics for multi-disease diagnosis, and to handle uncertainties in a decision support system.

However, the BNs have been rarely applied to the real continuous sequence of motion-related biomedical data. In [16], BN was used for the upper limb motion categorization. A BN model was used to categorize the healthcare procedure for wheelchair users with spinal injury [17].

This chapter is arranged in the following sections. In Sect. 6.2, the gait measurement experiment for gathering data of normal walking and simulated hemiplegic walking is described. In Sect. 6.3, preprocessing experiment data during the stance phase of walking is shown. In Sect. 6.4, BN, its search algorithm, and node assignments for its construction are briefly outlined. In Sect. 6.5, the results of the analysis are shown. In Sect. 6.6, we discussed and concluded.

6.2 Experiment Data Recording

In our experiment, we collected data for human normal walking and simulated hemiplegic walking. We used an electromyogram (EMG) to record muscle activities, a motion capture system to track foot motions during walking, and plantar pressure and force measurement to measure foot forces (pressures) for further analysis.

6.2.1 Subject

One healthy person, without previous foot abnormalities, took part in the experiment. The subject was required to walk on a normal floor at his normal speed. Table 6.1 shows the subject's weight, health state, and walking speed. To measure the artificial impairment walking (simulated hemiplegic walking), we asked the healthy subject to wear a simulation set, which is a product of Tokushuiryo Co. Ltd. The simulation set contributes constraints to the right-side ankle and knee joint of the subject. Figure 6.1 shows the special lower extremity orthosis constraining the ankle joint by a plantar flexed to 105°.

Before each experiment, informed consent was required from subject.

Table 6.1 Subject data information

Subject	Weight	Health state	Walking state	Speed
Male	65 kgs	Healthy	Normal walking	4 [km/h]
			Artificial impairment walking (Simulated hemiplegic)	1 [km/h]

Fig. 6.1 Lower extremity orthosis used for artificial impairment walking (simulated hemiplegic gait)

6.2.2 Making Records of Muscle Activities Using EMG Sensors

In biomedical research, EMG signals are used as the primary control signal sources to build interfaces for prosthetic applications [18]. Moreover, EMG signals are used as the primary diagnostic tools for clinical neurophysiology, i.e., for distinguishing neuromuscular diseases, and evaluating lower-back pain, kinesiology, and disorders of motor control, etc. [19].

Three EMG sensors (TYK-2007, II Version, Sikikou Engineering) were attached to the muscles shown in Fig. 6.2. The sampling frequency was 1600 Hz.

6.2.3 Making Records of Foot Trajectories by a Motion Capture System

We used a motion capture system (CaptureEx, Library-Inc), containing three cameras (Himawari GE60, 60 fps, Library-Inc) to make records of the trajectories of reflected light markers tied at the thumb, II toe, phalange (heel) bone, cuneiform bone, and ankle joint of right leg. Move-tr/3D software, also a product of Library-Inc., was used to construct a prototype from the recorded reflective marker trajectories and to calculate the toe angles from the prototype. Figure 6.3 shows the procedures used to compute the toe angles from foot trajectories.

The motion tracking system was synchronized with the EMG measurement, through the triggering function of the CaptureEx (Library-Inc).

Fig. 6.2 Muscles used for gait measurement, *EDL* Extensor digitorim longus muscle, *PL* peroneus longus muscle, *TA* Tibialis anterior muscle

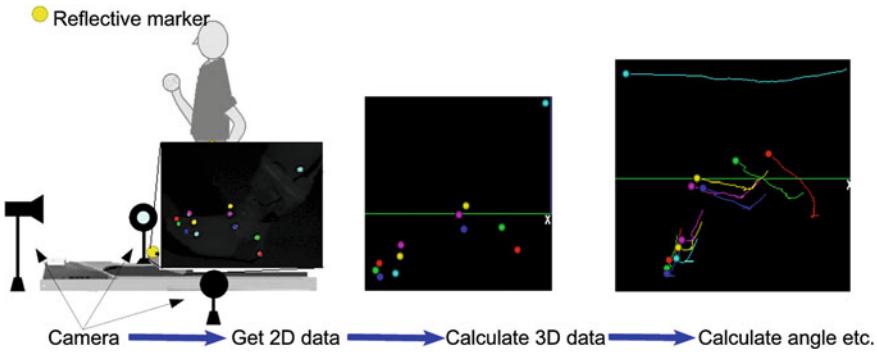


Fig. 6.3 Procedures to compute toe angles from foot trajectories

6.2.4 Making Records of Foot Pressures by Plantar Pressure and Force Measurement

We used the Plantar Pressure and Force Measurement system as an F-scan (Tekscan® technology) system. This system measures dynamic foot pressure and force data and shows interaction between the foot and ground. Conventional visual observation of gait and foot function, in contrast to an F-scan, measures foot force, contacted pressure distribution, and timing. The system consists of sensors, scanning electronic devices, and software. Figure 6.4 shows the sequence of steps for the plantar pressure experiment. This system is used in many applications: in

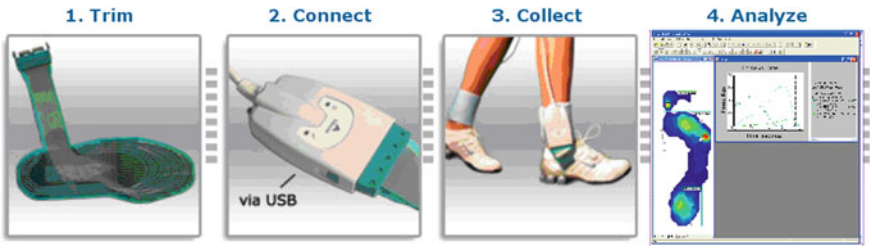


Fig. 6.4 Sequence of steps in the plantar pressure and force measurement experiment

shoe analysis (footwear design), gait analysis, diagnosis of diabetes, and so on. It supports biomechanical parameters, advanced analysis, and confirmation of the effectiveness of treatments.

6.3 Preprocessing Experiment Data During Stance Phase of Walking

6.3.1 Preprocessing EMG Signals in Order for Analysis

In order to analyze the data obtained from the experiment, we did following procedures on measurement data. (1) rectified raw (measured) EMG signals using full waves, (2) performed moving average on the signals, (3) down-sampled the signals to 60 Hz (the sampling rate of the motion capture system), and (4) standardized them. Desired signals (obtained during the stance phase) were extracted from the standardized data. Then, we discretizated the desired signals to three values: Upper, Middle, and Lower. Table 6.2 shows the method used for discretization, which we implemented in MATLAB® 7.1 (MathWorks®, Inc).

6.3.2 Preprocessing Toe Angle Data for Analysis

The thumb and II toe angles were extracted from the trajectory data by the motion capture system. These angles data were filtered and moving averaged, then discretized to three values: Upper, Middle and Lower, as shown in Table 6.2. Figure 6.5 shows an illustration of the angle of thumb and II toe.

Table 6.2 Discretization method

Discretizated value	Threshold value
Upper	0.66*max + 0.33*min < value
Middle	0.66*max + 0.33*min > value > 0.33*max + 0.66*min
Lower	0.33*max + 0.66*min > value

Note Max is maximum value; min is minimum value

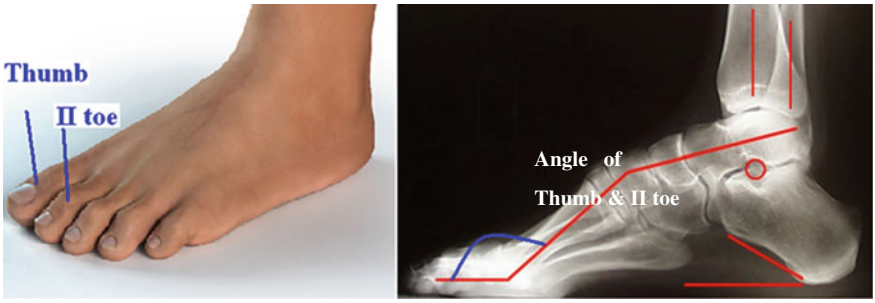


Fig. 6.5 An illustration of toe angle

6.3.3 Preprocessing Plantar Pressure Data for Analysis

Plantar pressure data were divided into three sections for further analysis. Figure 6.6 shows these sections and the corresponding plantar pressures. After the experiment, we calibrated experiment data according to the health condition of our experiment subject. Here, we used 65 kgs as the calibration point because of the weight of the subject. The desired plantar pressure data (obtained during stance phase) were extracted from the overall experimental data. Then, the desired plantar pressure data were discretized by three values: Upper, Middle, and Lower, shown in Table 6.2.

6.4 Outline of Bayesian Network

6.4.1 Concept of Bayesian Network

Bayesian networks (BNs), also known as belief networks or directed acyclic graphic models (DAG), are graphical representations of the probabilistic dependencies among random variables and estimated probabilistic inference obtained by using statistical and computational methods within those variables [20]. Bayesian networks treat random variables, express these variables as a set of nodes, draw arcs expressing probabilistic causal relations and conditional dependency between

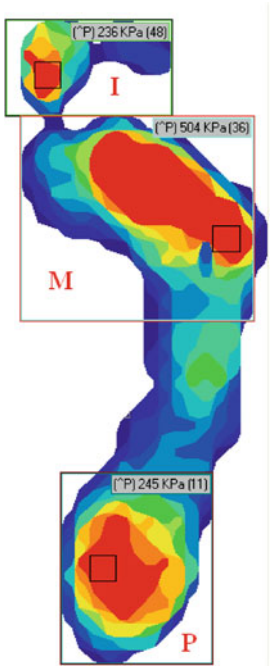


Fig. 6.6 Divided section of plantar pressures

a set of nodes, and extract these causal relations based on conditional probabilities of a set of nodes.

Recently, the Bayesian networks have been employed for medical diagnostics and predictions because they can be used to analyze biomedical signals and handle uncertainty in decision making systems. Bayesian networks are also applied for diagnosing faults in systems, body skill modeling, and so on.

6.4.2 Search Algorithms of BN Structure

There are two kinds of BN learning algorithms: parameter learning and structure learning. The structure learning algorithm is divided by two categories.

1. Constraint-based algorithms learn the network structure by analyzing the probabilistic relations entailed by the Markov property with conditional independence tests and then construct a graph which satisfies the corresponding d-separation statements. The resulting models are often interpreted as causal models even when learned from observational data [20].
2. Score-based algorithms assign a score to each candidate in the Bayesian network and try to maximize it with a heuristic search algorithm. Greedy search algorithms are a common choice, but almost any kind of search procedure can be used.

Table 6.3 Node appointment

Node	Meaning	Node	Meaning
PL	Peroneus longus muscle	M	Fore foot pressure
TA	Tibialis anterior muscle	I	Toes pressure
EDL	Extensor digitorim longus muscle	Thumb	Angle of thumb
P	Rear foot pressure	II toe	Angle of II toe

Table 6.4 Proscribed arcs

Node1	Arc	Node2
EDL	→	Thumb
EDL	←	Thumb
TA	→	II toe
TA	←	II toe
PL	→	Thumb
PL	←	Thumb
PL	→	II toe
PL	←	II toe

We used a score-based algorithm of BNs as a greedy search algorithm and an evaluation function of Bayes factors [9, 10, 20], implemented by a deal-package of an *R* package [21].

6.4.3 Node Assignment for BN Construction

Each EMG sensor attached muscle, each toe angle, and each plantar pressure section was designated a node in the graph.

In every experiment, five BNs were constructed. Each of these BNs expressed causal relations among three nodes of muscles, three nodes of foot pressure sections, and two nodes of angle data. The node appointment is given in Table 6.3. For all the nodes, three values, Upper, Middle and Lower, can be assumed. For further analysis, based on prior knowledge of human walking, some arcs were proscribed for the simplification of calculation. Table 6.4 shows the proscribed arcs.

6.5 Results

In this study, for a better understanding of foot function during the stance phase of different walking patterns, we divided the experiment data into three stages: initial contact (the strike of the heel on the ground), loading response to mid-stance, and terminal stance to pre-swing (ends with the lift of the toe at the beginning of the swing phase of gait). Figure 6.7 shows the three stages of a stance phase of walking.

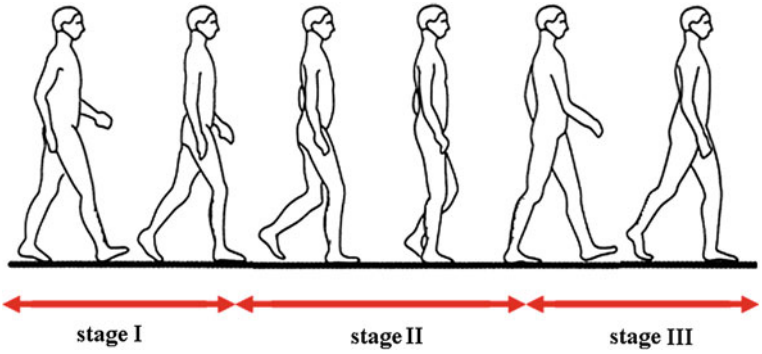


Fig. 6.7 Three stages of the stance phase of the foot

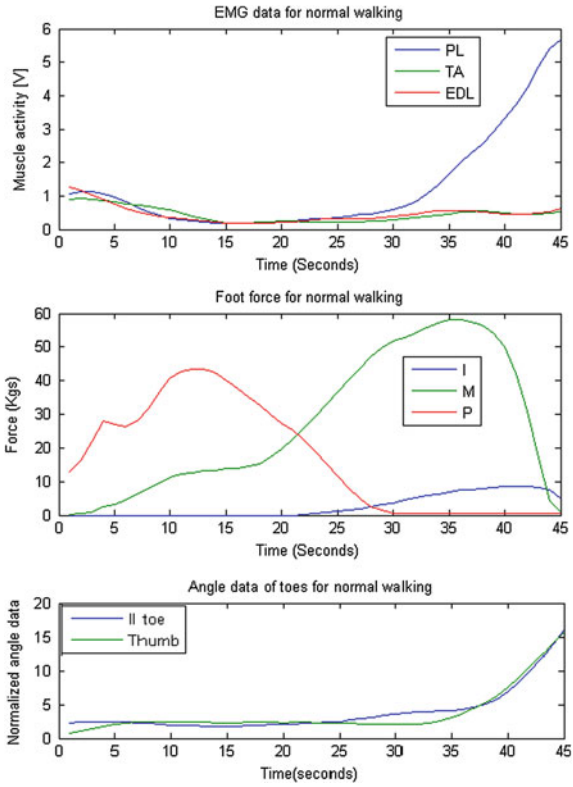


Fig. 6.8 Representation of EMG data, plantar pressure data, and toe angle data before discretization (for normal walking of healthy subject)

Figures 6.8 and 6.9 show a representation of EMG data, plantar pressure data, and toe angle data before the discretization process for normal walking and simulated gait walking, respectively.

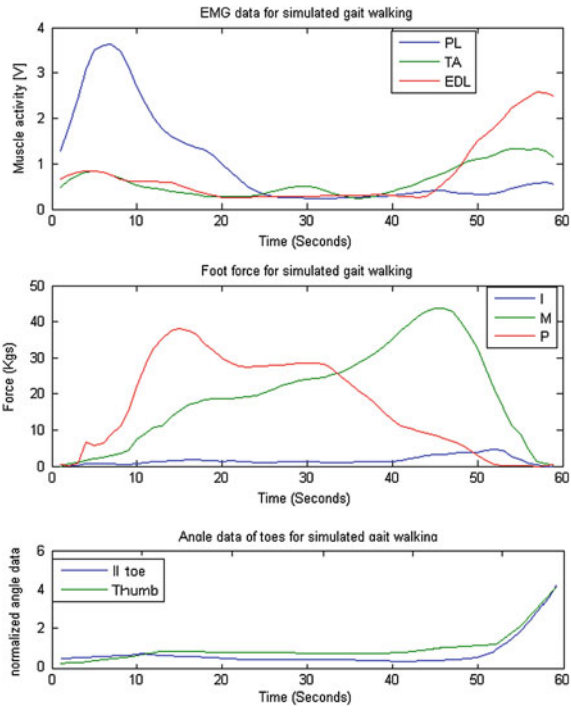


Fig. 6.9 Representation of EMG data, plantar pressure data, and toe angle data before discretization (for simulated gait walking of healthy subject)

Figures 6.10, 6.11 and 6.12 show BN structures of muscle activities and foot pressures for normal walking and simulated hemiplegic walking in three stages, respectively. From Fig. 6.8, we see that the PL muscle is more active than other muscles at Stage III of normal walking. From Fig. 6.11, we see that there are more incoming arcs to the PL node than to other nodes, and it seems that the PL muscle is more active than others. Comparing these two graphical representations, we can tell that extracting the BN-based causal inference among muscle activities and foot pressures is reasonable. From Fig. 6.9, we see that the PL muscle is more active than other muscles at Stage I, and the EDL muscle is more active than other muscles at Stage III of simulated gait walking.

From Figs. 6.11 and 6.12, in case of simulated gait walking, we see that there are more incoming arcs to the PL node than to other nodes and more outgoing arcs from the EDL node than from other nodes. From here, we can see that the PL muscle is more active at Stage I and the EDL muscle is more active at Stage III. From these results, the BN model-based causal inference among muscle activities and foot pressures is reasonable. We also see that BN structures of Stage II are the same in the case of normal walking and artificial impaired walking.

Figures 6.13 and 6.14 show the BN structures of muscle activities and toe angles for normal walking and simulated gait walking with Stages I and III, respectively.

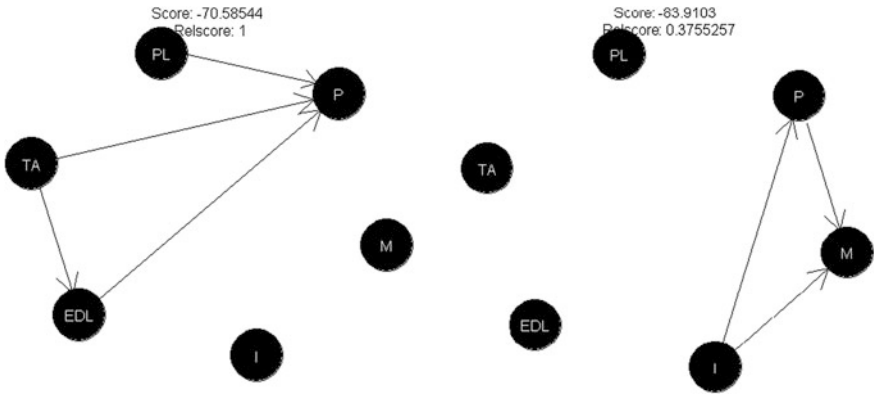


Fig. 6.10 BN structure of muscle activity and foot pressures for normal walking (*left side* is Stage I, *right side* is Stage II)

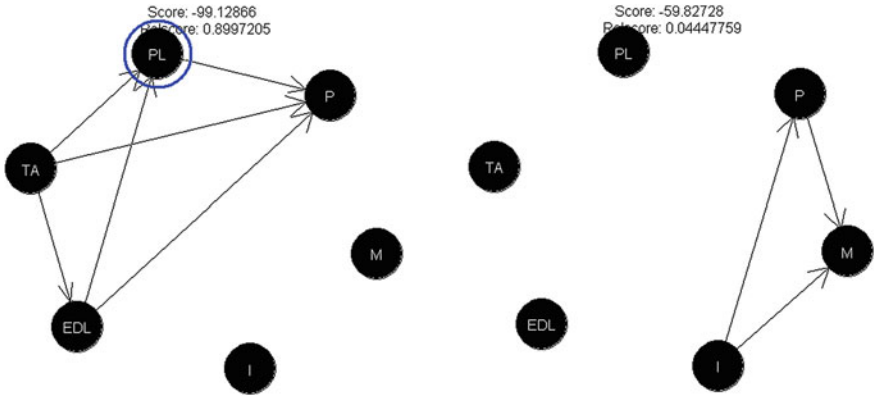


Fig. 6.11 BN structures of muscle activity and foot pressures for simulated gait walking (*left side* is Stage I, *right side* is Stage II)

For Stage II, there are no BN structures extracted from the data because this stage does not have muscle activities and toe motions. Figure 6.13 shows that the BN structures of Stage I for normal walking and simulated gait walking are the same. From Fig. 6.14, we see that there are more incoming arcs to the PL node than to others, and the PL node is more active than others at Stage III of normal walking (see Fig. 6.8).

As shown in this figure, the EDL muscle is more active than others in the case of simulated gait walking (right side). This activity is also shown to be the same as at EMG data representation in Fig. 6.9.

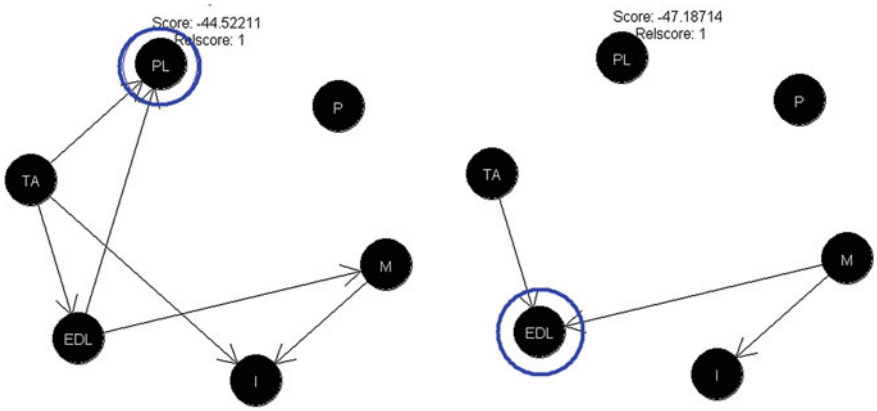


Fig. 6.12 BN structures of muscle activity and foot pressures (*left side* is Stage III for normal walking, *right side* is Stage III for simulated gait walking)

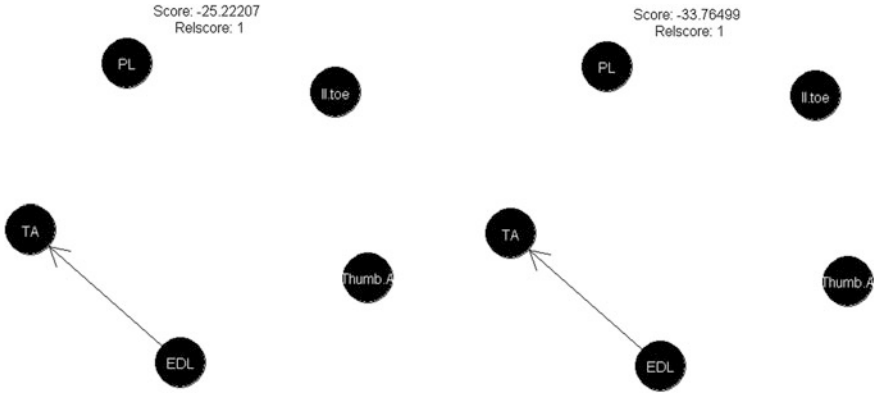


Fig. 6.13 BN structures of muscle activity and toe angles (*left side* is Stage I for normal walking, *right side* is Stage I for simulated gait walking)

In the case of Stage II or loading response to mid-stance, BN structure was not constructed amongst muscle activities and toe angles because all these values are on one discretized value. From here, we can see that in this stage there is no relation amongst muscle activities and toe angles, and it is a stable stage.

Figure 6.15 shows BN structures of muscle activities for the stance phase of walking (the left side is normal walking and the right side is simulated gait walking). From these two graphical representations, we can tell that causal relations amongst muscle activities during stance phase are the same in both cases (normal walking and simulated gait walking). To construct more precise causal relations and conditional dependence amongst foot functions during the stance

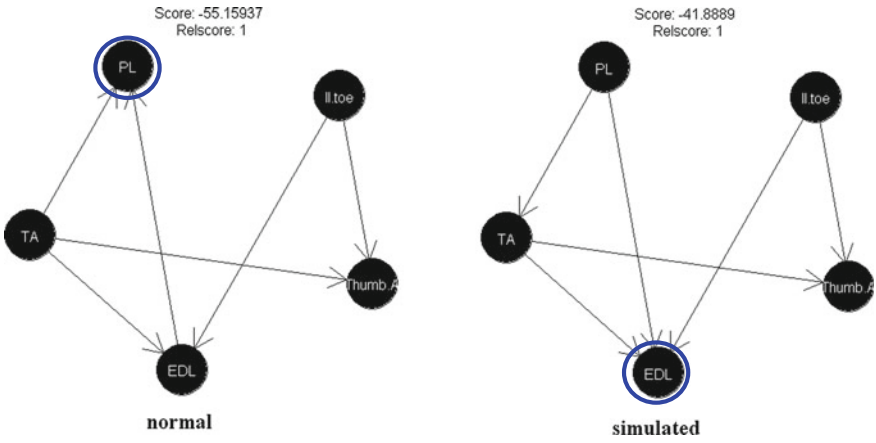


Fig. 6.14 The BN structures of muscle activities and toe angles (*left side* is Stage III for normal walking, *right side* is Stage III for simulated gait walking)

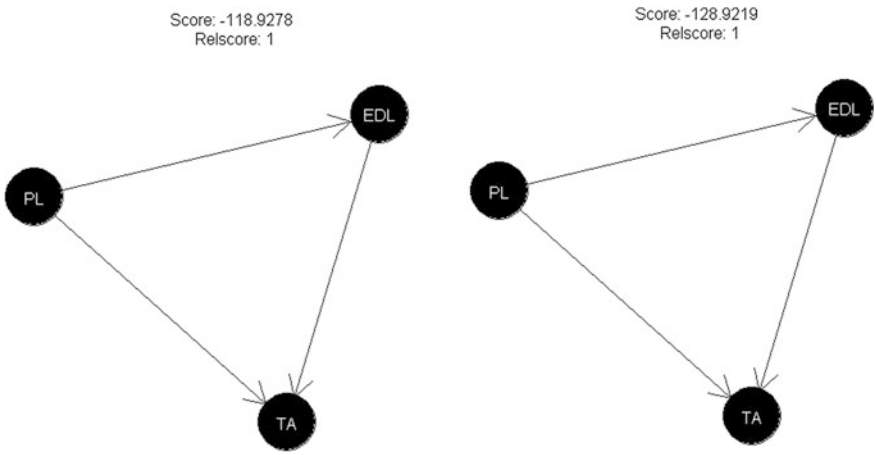


Fig. 6.15 BN structures of muscle activities for stance phase of walking (*left side* is normal walking, *right side* is simulated gait walking)

phase of walking, we divided our experiment data into three stages and discretized three values.

In this study, we analyzed four trial data from measurement experiment data. The resulting probability values are shown in Table 6.5. From this table, we see that two trial data of normal walking and simulated gait walking are similar to each other.

We see from our results that the BNs of normal walking and simulated gait walking are reasonable and good graphical representations of muscle activities, plantar pressure sections, and toe angles.

Table 6.5 Probability value of trial data

Node name	Discretized value	Normal walking (Probability)						Simulated gait walking (Prob)					
		I trial data			II trial data			I trial data			II trial data		
		I	II	III	I	II	III	I	II	III	I	II	III
PL	Upper	0	0	0.42	0	0	0.44	0.42	0	0	0.44	0	0
	Middle	0	0	0.33	0	0	0.35	0.33	0	0	0.34	0	0
	Lower	1	1	0.25	1	1	0.21	0.25	1	1	0.22	1	1
TA	Upper	0.58	0	0	0.62	0	0	0.17	0	0.82	0.15	0	0.8
	Middle	0.42	0	0.67	0.38	0	0.6	0.75	0	0.18	0.72	0	0.2
	Lower	0	1	0.33	0	1	0.4	0.08	1	0	0.13	1	0
EDL	Upper	0.25	0	0	0.22	0	0	0	0	0.55	0	0	0.58
	Middle	0.25	0	0.42	0.23	0	0.4	0.25	0	0.27	0.25	0	0.28
	Lower	0.5	1	0.58	0.55	1	0.6	0.75	1	0.18	0.75	1	0.14
P	Upper	0.5	0.39	0	0.6	0.4	0	0.28	0.63	0	0.33	0.63	0
	Middle	0.5	0.19	0	0.4	0.21	0	0.22	0.2	0	0.23	0.2	0
	Lower	0	0.42	1	0	0.39	1	0.5	0.17	1	0.44	0.17	1
M	Upper	0	0.35	0.66	0	0.4	0.6	0	0.62	0.55	0	0.64	0.56
	Middle	0	0.26	0	0	0.25	0.2	0	0.38	0.18	0	0.36	0.19
	Lower	1	0.39	0.33	1	0.35	0.2	1	0	0.27	1	0	0.25
I	Upper	0	0.16	0.92	0	0.18	0.90	0	0.26	0.91	0	0.3	0.92
	Middle	0	0.26	0.08	0	0.22	0.1	0	0.74	0.09	0	0.7	0.08
	Lower	1	0.58	0	1	0.6	0	1	0	0	1	0	0
Thumb	Upper	0	0	0.25	0	0	0.27	0	0	0.27	0	0	0.28
	Middle	0	0	0.25	0	0	0.28	0	0	0.27	0	0	0.27
	Lower	1	1	0.5	1	1	0.45	1	1	0.46	1	1	0.45
II toe	Upper	0	0	0.25	0	0	0.26	0	0	0.27	0	0	0.27
	Middle	0	0	0.25	0	0	0.27	0	0	0.27	0	0	0.26
	Lower	1	1	0.5	1	1	0.47	1	1	046	1	1	047

6.6 Discussion and Conclusions

The results presented in this chapter show that the BN structure is useful for a better understanding of foot function during the stance phase of human normal and simulated hemiplegic walking.

Biomedical signals are corrupted by external noise during the experiment. The external noise can be environment noise (sound and light), experiment equipment noise, and communication channel noise. To avoid such noise, we filtered and standardized our experimental data.

Foot function during walking is a result of interactions among the muscular, neural, and skeletal systems, and the walking environment.

In this study, we extracted causal structures for foot function by measuring and recording lower-limb major muscle activities, the trajectories of toe and ankle joints, and plantar pressure distributions during the stance phase, and then applied BN as the theoretical account for probabilistic causal inference.

Two different styles of walking, normal and simulated hemiplegic walking of one healthy subject who did not have previous foot abnormalities and weighed 65 kgs, were measured and analyzed to verify the BN's ability to express and distinguish the significant gait-dependent causal relations.

In this work, we assigned as the nodes of BNs to each muscle activity, each plantar pressures section, and each toe angle trajectory, each of which represent commonly cited modes of muscle control analysis and motion analysis. Recently, there have been many Bayesian network models used to diagnose different diseases [11–15, 17] and classify motion [16, 22].

But, our study is revealed to combine different measurements for experimental data, plantar pressure data, muscle activity data, and toe motion data during the stance phase of walking.

We standardized and discretized experiment data into three values and then divided it into three stages during the stance phase of walking: initial contact, loading response to mid-stance, and terminal stance to pre swing, for constructing reasonable relation and dependency amongst muscle activities, plantar pressures, and toe motions. Our preliminary results show that the BNs of normal walking and artificial impairment (simulated hemiplegic walking) are reasonable, and there is no difference between them. We have analyzed four sets of trial data for normal walking and artificial impairment (simulated hemiplegic walking); there are no differences between trial data from probability table.

In future studies, we will try to increase the number of subjects, particularly those for impaired walking cases. We will also conduct experiments for several walking conditions (climbing upstairs, different walking speeds, on gradient walkways, etc.) Moreover, three muscles, three sections of plantar pressures, and two toe angles were investigated in this study, though the other assignment schemes, i.e., multiple muscles, multiple separation to plantar pressures, angle of ankle, inside arch and outside arch of foot should also be studied.

In this study, we used the BN to extract the probabilistic causal information of foot function data, such as muscle activities, plantar pressures, and toe trajectories, from different types of data of human walking phases. The graphical networks extracted from the three stages of the stance phase of gait measurement data are useful for understanding the foot function of the normal walking and simulated hemiplegic walking. Thus, understanding the foot function during walking is important for further analysis of diagnostic, therapy, and training programs for foot impairment.

References

1. Takemura H, Iwama H, Ueda J, Matumoto Y, Ogasawara T (2003) A study of the toe function for human walking. JSME Symp Welfare Eng 3:97–100
2. Kaapel-Bargas A, Woolf R, Cornwall M, McPoil T (1998) The windlass mechanism during normal walking and passive first metatarsalphalangeal joint extension. Clin Biomech 13(3):190–194

3. Nishiwaki K, Kagami S, Kuniyoshi Y, Inaba M, Inoue H (2002) Toe joints that enhance bipedal and fullbody motion of humanoid robots. In: Paper presented at the proceedings of the 2002 IEEE international conference on robotics & automation (ICRA02), Washington, DC, pp. 3105—3110, 11–15 May 2002
4. Hutton WC, Dhanendran M (1979) A Study of the distribution of load under the normal foot during walking. Int Orthop (SICOT) 3(2):153–157
5. Kong K, Tomizuka M, (2008) Estimation of abnormalities in a human gait using sensor-embedded shoes. In: Paper presented at the proceedings of the 2008 IEEE/ASME international conference on advanced intelligent mechatronics, 2–5 July 2008
6. Warren GL, Maher RM, Higbie EJ (2004) Temporal patterns of plantar pressures and lower-leg muscle activity during walking: effect of speed. Gait Posture 19:91–100
7. Nergui M, Murai C, Koike Y, Yu W, Acharya R (2011) Probabilistic information structure of human walking. J Med Syst 35(5):835–844, Springer
8. Goldberg EJ, Nepture RR (2007) Compensatory strategies during normal walking in response to muscle weakness and increased hip joint stiffness. Gait Posture 25:360–367
9. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, CA
10. Jensen FV (2001) Bayesian networks and decision graphs. Springer, New York
11. Nikovski D (2000) Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. IEEE Trans Knowl Data Eng 12(4):509–516
12. Suojanen M, Andreassen S, Olesen KG (2001) A method for diagnosing multiple diseases in MUNIN. IEEE Trans Biomed Eng 48(5):522–532
13. Meloni A, Ripoli A, Positano V, Landini L (2009) Mutual information preconditioning improves structure learning of bayesian networks from medical databases. IEEE Trans Inf Technol Biomed 13(6):984–989
14. Meloni A, Landini L, Ripoli A, Positano V (2009) Improved learning of bayesian networks in biomedicine. Paper presented at the Ninth international conference on intelligent systems design and applications, 30 Nov–2 Dec 2009
15. Rose C, Smaili C, Charpillet F (2005) A dynamic bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis. Paper presented at the proceedings of the 17th IEEE international conference on tools with artificial intelligence (ICTAI'05), 14–16 Nov 2005
16. Nan B, Okamoto M, Tsuji T (2009) A hybrid motion classification approach for EMG-based human—robot interfaces using bayesian and neural networks. IEEE Trans Rob 25(3):502–511
17. Athanasiou M, Clark JY (2007) A bayesian network model for the diagnosis of the caring procedure for wheelchair users with spinal injury. Paper presented at the twentieth IEEE international symposium on computer-based medical systems (CBMS'07), 20–22 June 2007
18. Yu W, Yamaguchi H, Maruishi M, Yokoi H, Mano Y, Kakazu Y (2002) EMG automatic switch for FES control for hemiplegics using artificial neural network. Elsevier. Robot Auton Syst 40(2):213–224
19. Ebara K, Ohashi M, Kubota T (1999) Rehabilitation program for walking impairment. Ishiyaku Publishers Inc, Tokyo
20. Neapolitan RE (2004) Learning bayesian networks. Prentice Hall, New Jersey
21. Bøttcher SG, Dethlefsen C (2003) Learning bayesian networks with R third Edition. In: Paper presented at the proceedings of the international workshop on distributed statistical computing, 20–22 Mar 2003
22. Cheng C, Ansari R, Khokhar A (2004) Cyclic articulated human motion tracking by sequential ancestral simulation. Paper presented at the IEEE computer society conference on computer vision and pattern recognition (CVPR'04), 27 June–2 July 2004

Chapter 7
Rule Learning in Healthcare and Health Services Research

Janusz Wojtusiak

Abstract Successful application of machine learning in healthcare requires accuracy, transparency, acceptability, ability to deal with complex data, ability to deal with background knowledge, efficiency, and exportability. Rule learning is known to satisfy the above criteria. This chapter introduces rule learning in healthcare, presents very expressive attributional rules, briefly describes the AQ21 rule learning system, and discusses three application areas in healthcare and health services research.

Keywords Rule learning · Attributional calculus · AQ21 system · Health services research · Aggregated data · Healthcare billing data

7.1 Introduction

Healthcare requires modern computational tools to handle the complexity of data and workflows. The healthcare environment is dynamic and frequently changing: New knowledge is published on a daily basis, new drugs are constantly available, and the best practice guidelines change. Moreover, healthcare is a critical area in which success is measured by patient survival and wellbeing. Unfortunately, many existing treatment and reimbursement systems used in healthcare treat individual patients as "average" cases without tailoring to patient characteristics.

The above reasons call for machine learning methods to manage the complexity and automatically adapt to frequent changes. This chapter focuses on one of the best known and most important methods in machine learning in healthcare: rule learning. It briefly describes rule learning methods, discusses their use in

J. Wojtusiak (✉)
Machine Learning and Inference Laboratory, Department of Health Administration and Policy, George Mason University, Fairfax, VA, USA
e-mail: jwojtusi@gmu.edu

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_7,

healthcare delivery, research, administration and management, and presents advantages of using rule learning rather than traditional computational approaches and other machine learning methods.

In order to fully justify the use of rule learning in healthcare, the following sections briefly outline aspects of machine learning that are particularly important in this application area.

7.1.1 What is Needed in Healthcare and Health Services Research?

Machine learning methods have a wide range of applications in healthcare delivery, research, administration, and management. Many of these applications are slowly emerging as the healthcare community becomes more familiar with machine learning and its immense potential. On the other hand, most machine learning researchers are not familiar with healthcare settings and over-trivialize them. This mutual lack of understanding between healthcare and machine learning communities results in the lack of advanced machine learning methods adoption.

Among the healthcare areas that benefit the most from machine learning are those that rely on automated processes or that can be automated. The ability of machine learning methods to adapt to dynamically changing environments, previously unseen situations, and new challenges make them ideal for these types of applications. Two of the most common applications of machine learning in healthcare are: decision support systems and knowledge discovery. Decision support systems rely on computational models that aid decision makers in a variety of situations. These models can be constructed and maintained using machine learning. In addition, knowledge discovery, which primarily derives from medical datasets, can be used to study patterns of healthcare delivery systems, management, billing, etc. Machine learning has, thus, great potential when correctly applied to hard problems that cannot be solved with more traditional computational methods or manually without the use of computers.

However, for machine learning to be adopted in healthcare, methods need to fulfill several requirements. These requirements are eminent and applicable to virtually all domains in which machine learning is or can be used. However, some of these requirements are particularly important in healthcare when the adoption of new technologies and results are exceptionally challenging.

- **Accuracy**. Models have to provide reliable predictions and/or reliably describe data, which is, in most cases, their main function. Multiple measures of accuracy are available, all of which perform some form of counting/scoring of correct and incorrect predictions and combinations thereof. Some commonly used measures of accuracy include precision, recall, sensitivity, specificity, F-score, and others.
- **Transparency**. Medical and healthcare studies require models to be easily understood by people not trained in machine learning, statistics, and other

advanced data analysis methods. In this sense, providing just the reliable predictions is not sufficient, as models should also "explain" why a specific prediction is made and what the model actually does. This corresponds not only to methods that lead to creation of new knowledge, but also to autonomous systems that because of their critical role need to leave an "audit trail" and be analyzed/verified periodically.

The concept of understandability and interpretability has been well known since early work on expert systems and artificial intelligence, but has been largely ignored by many modern machine learning methods. One reason for this is that it is very hard to measure the complexity of created models and hypotheses, and use that measurement as one of knowledge representation selection criteria. It is virtually impossible to consistently measure and compare the transparency of models learned in different representations. (How do we compare transparency of specific SVM-based, NN-based, and rule-based models for diagnosing liver diseases? How do we generalize the measure?) Moreover, compound knowledge representations, which are natural to people, tend to be difficult to learn through machine learning methods. One such representation, called attributional calculus consists of attributional rules, which are briefly outlined in Sect. 7.2.1.

- **Acceptability**. Models need to be accepted by their potential users. While partially related to transparency, acceptability requires that the models that do not contradict the knowledge of existing experts are otherwise "reasonably" congruent with what is currently being done, and correspond to existing workflows. Acceptability is a key issue in healthcare, more than in any other industry. Clinicians, administrators, and supporting staff do not want to change the way they work, even if the developed models being used are accurate and superior to methods currently being used. The use of ML algorithms should immediately lead to improved work and provide incentives to participants; otherwise results may not be adopted.
- **Ability to handle complex types of data**. Healthcare data are complex. Even relatively simple applications of machine learning to healthcare data require making numerous conversions, data pre-processing, encoding of variables, and so on. In order to have widespread acceptance in healthcare, machine learning methods should be able to operate directly with healthcare data without the need to artificially encode. Healthcare data are not, and should not be, treated by ML tools as a collection of numbers without meaning. Although more advanced ML methods recognize a wide range of data types (nominal, structured, ordinal, interval, ratio, absolute, compound, etc.), prevalent standards such as ICD-9, ICD-10, CPT, SNOMED, and HL7 are currently not directly supported by ML tools.
- **Ability to handle background knowledge**. Computers require massive amounts of data to make simple decisions or discover simple facts. Humans do exactly the opposite—we are able to make important decisions and discover important facts based on minimal information. Although there are many differences in human and computer inference/learning processes, one of the most important is the ability to use background knowledge to place problems into the appropriate context. Similarly, machine learning algorithms that are provided with large knowledge bases and a wealth of background knowledge need not

have access to huge amounts of data. This allows machine learning algorithms to focus on the discovery of novel facts and not what is already known to experts. Extremely large repositories of medical and healthcare knowledge (is often not coded and in many cases only available as text of published manuscripts) can be incorporated into the machine learning process.

- **Efficiency**. Both model induction and model application algorithms need to be efficient. Machine learning algorithms applied in healthcare should be able to cope with very large amounts of data. The data may have many examples (sometimes called records or datapoints), attributes (sometimes called variables or features), or both. The theoretical estimates of algorithm complexity are often available for many methods. More importantly users want the methods to be executed in a specific period of time, even if it means that results are only approximate or "good enough."
- **Exportability**. Results of machine learning should be directly transferable to decision support and other systems where they can be immediately applied. It is not unusual that the learned models will work along with already existing models and thus need to be compatible. For example, learned models can be translated or directly learned in the form of rules in Arden Syntax, a popular representation language in clinical decision support systems. If models are learned in completely different representations, they need to be translated (usually approximately) to the target form.

This chapter focuses on the use of rules and rule learning methods in different healthcare areas. Rules are known to be one of the most transparent knowledge representations that also conform to other criteria outlined above.

7.2 Rule Learning

Over the past few decades multiple rule learning algorithms and software have been developed. Multiple types of rules are considered in machine learning research depending on their use and form, including: association rules (which are used to represent regularities in data), decision rules (which are used to support decisions) and their subtype classification rules (used to classify examples into concepts), rules with exceptions (that include part describing when the rule does not apply), *m*-of-*n* rules (used to count true values or statements), and attributional rules (the most expressive form of rules considered here).

The AQ21 system is particularly suitable for problematic healthcare situations because of its flexibility, ability to deal with multiple types of attributes, handle both large and small datasets, use background knowledge in different forms, learn from individual and aggregated data, manage meta-values, cope with noise, perform constructive induction, generate alternative hypotheses, and many other features. AQ21 uses attributional rules as the main form of knowledge representation. The following subsections briefly introduce attributional rules, and outlines AQ21 main algorithms.

7.2.1 Attributional Rules

Healthcare applications require rules that are more expressive than typically used

$$CLASS\ IF\ CONDITION \tag{7.1}$$

Most software creates rules in which *CONDITION* is a conjunction of simple conditions in the form *ATTRIBUTE* = *VALUE*. Many such rules are needed to describe even simple concepts. Attributional rules are currently the most expressive form of rules induced by machine learning algorithms. They are the main knowledge representation in a formal language called *attributional calculus*, AC [9]. AC has been created to support *natural induction*, an inductive learning process which has results that are natural to people because of their form and content.

Natural induction requires that knowledge be equivalent to statements in natural language (i.e. English), so those who are not experts in machine learning or knowledge mining, or do not have a technical background may understand it. Thus, medical doctors, healthcare administrators, nurses, and researchers should be able to understand, interpret, modify, and apply knowledge learned by computer systems. Such a goal requires that knowledge discovery programs use a language that can either be automatically translated to natural language or easily understood on its own.

Learned knowledge is represented in attributional calculus in the form of *attributional rules,* which consist of *attributional conditions*. An attributional condition takes the form:

$$[L\ rel\ R : A], \tag{7.2}$$

where L is an attribute, an internal conjunction or disjunction of attributes, a compound attribute, a counting attribute, or an expression. *rel* is one of $=, >, <, \leq, \geq, :$, or $\neq$. R is an attribute value, an internal disjunction of attribute values, an attribute, an internal conjunction of values of attributes that are constituents of a compound attribute, or an expression. A is an optional annotation that may list statistical information describing the condition. The annotation often includes $|p|$ and $|n|$ values for the condition, defined as the numbers of positive and negative examples, respectively, that satisfy the condition, and the condition's consistency defined as $|p|/(|p| + |n|)$.

There are several forms of attributional rules allowed by attributional calculus. Three important forms of attributional rules are presented below:

$$CONSEQUENT <== PREMISE \tag{7.3}$$

$$CONSEQUENT <== PREMISE \lfloor EXCEPTION \tag{7.4}$$

$$CONSEQUENT <== PREMISE \lceil PRECONDITION \tag{7.5}$$

Table 7.1 Table with example conditions and rules

[Length > 7.3]
The length of an entity is greater than 7.3 units (*as defined in the attribute's domain*).
[Color = red v blue: 40, 2]
The color of an entity is red or blue. The condition is satisfied by forty positive and two negative examples.
[Length & Height ≤ 12]
An entity's length and height are both smaller or equal to 12 units. The units are defined in the attributes' domains.
[Weather: sunny & windy]
The weather is sunny and windy. This is an example of a condition that includes a compound attribute Weather.
[Part = acceptable] <== [Width = 7.12] & [Length <3] & [Material = steel v plastic]
A part is acceptable if its width is between 7 and 12, its length is less than 3 and its material is steel or plastic.
[Activity = play] <== [Condition = cloudy v sunny: 7, 8] & [Temp = medium v high] ⌊[Condition = cloudy] & [Wind = yes] & [Temp = high] : p = 7, n = 0, q = 1
An activity is play if the condition is cloudy or sunny and temperature is medium or high, except for when the condition is cloudy, there is wind and temperature is high. The rule covers 7 positive and no negative examples. Its quality of the rule is 1.

where *PREMISE*, *CONSEQUENT*, *EXCEPTION*, and *PRECONDITION* are complexes, that is, conjunctions of attributional conditions. An *EXCEPTION* can also be an explicit list of examples that constitute exceptions to the rule. The rules without exception or preconditions are interpreted as the *CONSEQUENT* is true whenever the PREMISE is true. The rules with exceptions are interpreted that the *CONSEQUENT* is true whenever the PREMISE is true, except for when the *EXCEPTION* is true. The rules with preconditions are interpreted that the *CONSEQUENT* is true whenever the *PREMISE* is true, provided that the *PRECONDITION* is true. The symbols ⌊ and ⌈ are used to denote exception and precondition, respectively. Each rule may be optionally annotated with several parameters such as numbers of covered examples (positive and negative), the rule complexity, etc.

One class of the data is usually described using several rules, called a *ruleset*. Rules considered here are independent, i.e., the truth status of one rule does not affect interpretation of other rules. This is in contrast to many other rule learning programs that learn sequential rules that need to be evaluated in a specific order. A set of rulesets that describe all considered classes in the data (often defined by possible values of an output/dependent attribute) is called a *ruleset family*, a.k.a. classifier. Depending on the problem at hand, the goal may be learn a complete classifier, a ruleset for one class of interest, or individual rules representing regularities/patterns in the data. Selected example attributional conditions and rules along with explanations are presented in Table 7.1.

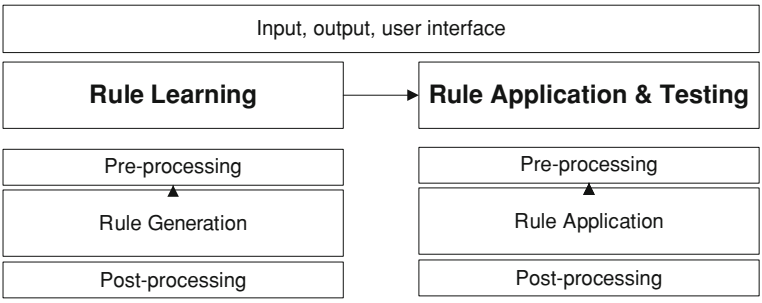


Fig. 7.1 AQ21 system architecture

7.2.2 AQ21

The well-known family of AQ programs originated with the simple version of the A^q algorithm for solving the general covering problem used at the core of rule learning [5]. Numerous implementations and extensions of the method were developed over the years. Among the best known AQ implementations are AQ7 [6], AQ11 [7], AQ15c [11], AQ17 [1], AQ19 [8], and most recently AQ21 [12–14].

The AQ21 system consists of two main modules for learning attributional rules, and for their application (Fig. 7.1). The learning module consists of data and background knowledge, a pre-processing module, a rule generation module, and a post-processing module. Similarly, the testing module consists of a pre-processing module which converts data and rules to common representation, a rule application module which matches examples against rules, and a post-processing which calculates summaries and statistics.

Rule learning starts with the pre-processing of data and background knowledge which both need to be converted into the right representation and then prepared for rule generation. The process may involve simple steps such as encoding of attribute values, and/or more complex ones including constructive induction. The goal for the latter is to automatically determine the representation space (a set of attributes, their types, and domains). This method is best suitable for the learning problem at hand. AQ21 implements two of three known classes of constructive induction (data-driven (DCI) [3], knowledge-driven (KCI), hypothesis-driven (HCI) [10], and multi-strategy [2]), DCI and KCI. The methods include operators such as attribute selection, attribute generation, and attribute modification.

At the core of the AQ learning is its rule generation module. The method pioneered the separate-and-conquer approach to rule learning, in which data representing a target class being learned are sequentially covered in a way that avoids negative examples. The AQ21 rule generation module starts by focusing on a single example and generates possible generalizations of that example that are consistent or partially consistent with the data and background knowledge. This process, called *star generation*, results in a rule or set of rules that describe part of the data. Multiple stars may be generated in parallel, in order to prevent erroneous

generalizations due to noise in the data. The process of star generation is repeated until all data or a significant portion of data are covered (explained) by generated rules. The quality of rules in AQ21 is evaluated using lexicographical evaluation functional (LEF), a method which sequentially evaluates rules through multiple criteria. Numerous variants of the AQ rule generation algorithm have been investigated over the years and are widely described in literature.

The rule post-processing method includes rule optimization, selection of the final rules to be used in a hypothesis or a set of alternative hypotheses, and calculation of statistical parameters describing these rules. The final rules are presented to the user or transferred to the testing and application module.

The rule testing and application module starts with the pre-processing of hypotheses and examples in order to match their representation and prepare for the actual application process. Each considered example (application case) is evaluated against rules. In the case of application of rules in decision support, only one example is usually considered. Rules can be evaluated strictly (when an example either matches a rule or not) and flexibly (when a degree of match, DM, ranging from zero to one is calculated). Multiple schemas [9] are available on how to flexibly evaluate individual conditions, rules, and entire rule sets.

Unlike most classifiers that always give one definitive answer, the AQ21 application module may either provide multiple possible answers, or simply answer "don't know." In this philosophy, it is better to provide users with more than one plausible answer with high confidence, or not answer at all, than give a likely incorrect definitive answer.

7.3 From Rule Learning to Decision Support

Decision support systems are broadly defined as computer systems that aid decision makers. This definition can include everything from simple spreadsheet applications, through simulation models, to rule-based expert systems. In this chapter, we focus on knowledge-based decision support systems in which computers provide support to their users based on the content of their knowledge bases.

Traditionally, decision support systems are static in the sense that their knowledge does not change over time without explicit intervention by the user. Machine learning-based decision support systems can, however, evolve and adapt to dynamically changing environments in which they operate. Adaptability is, thus, one of two important areas in which machine learning can help in decision support.

Consider an alert system which provides clinicians with messages informing them about important events related to a specific patient, i.e., allergies, drug–drug interactions, abnormal results. An oversensitive alert system that displays too many messages causes a well-known phenomenon called alert fatigue. In such a case, physicians no longer read alerts, but rather ignore all of them. A typical approach to the problem is to create a system-wide policy/threshold so that alerts

do not overwhelm users. This one-size-fits-all approach ignores all the differences between physicians and the way they practice. A machine learning-based solution is able to adapt to specific users (physicians) and show only alerts that have the lowest chance of not being overwritten.

The second important area in which machine learning can be used in healthcare is knowledge generation. The majority of decision support systems are based on rules. These rules, sometimes called Medical Logic Modules (MLMs), are prepared by panels of experts based on the best practice and known evidence. Their creation is a long and difficult process. One of the important applications of machine learning is knowledge generation—the knowledge if present in the right forms can help in preparation of MLMs.

Because rules created by the AQ21 system are independent (i.e. unordered), they can be easily incorporated into decision support systems. For example, attributional rules described above can be directly written in ARDEN syntax [4]. The actual rules are written in the "logic" slot of MLMs while the "data" slot is used to derive attribute values and translate then into the required format. Because one MLM corresponds to a complete decision, it includes multiple rules forming a complete ruleset family. Attributional rules can be also manually inspected by experts and modified as rules and compliance requirements change.

7.4 Review of Selected Applications

This section describes three recent studies that applied rule learning in diverse areas of healthcare. They span over medical, comparative effectiveness, and managerial datasets.

7.4.1 Hospital Bills Classification

The purpose of the described study is to improve billing by advancing healthcare provider operations and performance through the use of machine learning methods [16]. Across the country, healthcare providers are experiencing ongoing pressure from declining revenues. Payers are under increasing pressure to contain costs. The implementation of healthcare reform through the Patient Protection and Affordable Care Act (Public Law 111–148) will further exacerbate this issue. These and additional demands to combat waste, fraud, and abuse are creating mounting pressures to achieve 'perfection' in all phases of healthcare billing and reimbursement authorization for hospitals and independent healthcare providers (e.g. physicians and medical group practices). In order to ensure that payments are appropriate, payers must ascertain that there is proper documentation of care prior to reimbursement. Providers must be diligent in maintaining proper documentation to receive the correct payment and avoid loss of revenue.

The opposing pressures from payers and providers call for the use of decision support/screening methods, to better manage the billing and revenue cycle and detect inconsistencies in coverage, care/service documentation and payments, and to guide financial and clinical personnel through this process. Specifically, we are using machine learning to create models for screening billing information for inconsistency. The initial, proof-of-concept, study presented here is based on the batch processing of obstetrics data collected from a one year period in 2008.

In the first step, the data are pre-processed to match requirements of the machine learning application used. Data available in multiple tables in the hospital information system need to be converted into a flat file. Additional processing of variables needs to be done. In the second step, the AQ21 machine learning system [13], which creates predictive models in the form of highly transparent attributional rules, is used. In order to apply the method to create models, the data is classified as "normal payment" and "abnormal payment" which correspond to payments consistent and not consistent with contractual agreements, respectively. Finally, after the rule learning phase, the models are used to predict whether a specific bill is likely to receive normal payment in advance to its submission to the payer.

Initial application of the method in analyzing billing information for obstetrics patients covered by Medicaid achieved promising results. The presented method provides two strong benefits in analyzing billing information. First, the use of machine learning allows one to automatically create models for predicting bill payments before their submission. The models allow screening of billing information before the bill is sent to payees, therefore maximizing the chance of receiving full payments, and reducing unnecessary denials. Second, the use of highly transparent representations of models in the form of attributional rules, allows for the detection of regularities in bill denials which may lead to potential workflow improvement.

7.4.2 Comparative Effectiveness Research

The gold standard for biomedical research is randomized clinical trials (RCT). In many cases, RCTs are impossible or unethical to perform, and only secondary analysis of existing data form clinical records is possible. Rule learning is an attractive approach to comparative effectiveness research of alternative treatments or medications. The latter are often prescribed based on trial and error.

The problem considered in comparative effectiveness research is substantially different from one considered in typical concept learning in which examples are labeled with classes. Here, the data are in the form of rows including C_i, T_i, and O_i, where C_i are the ith patient case characteristics, T_i is the applied treatment or combination of treatments, and O_i indicates outcomes [15]. Models are created and tested using the following three steps, also illustrated in Fig. 7.2.

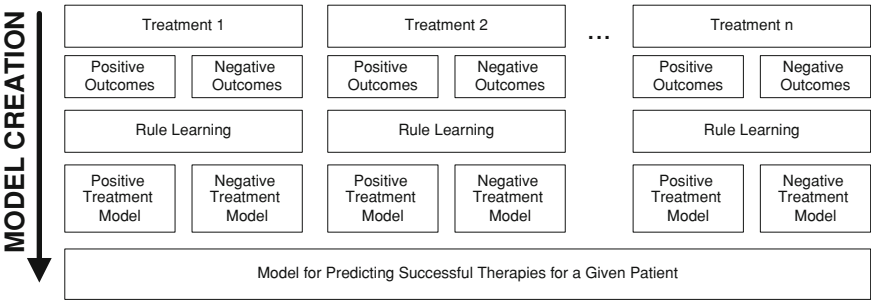


Fig. 7.2 Creation of models for comparative effectiveness research

1. For each treatment or combination of treatments, T, select P_T cases from the database for which therapy T was successful and N_T cases for which therapy T was unsuccessful.
2. Apply rule learning to induce general models, M_T, based on P_T and N_T to predict whether therapy T will be successful given a patient's characteristics. A collection of such models for all considered combinations of treatments will be the final model M. Similarly, create models M_{NT} to predict that a given therapy will not be successful. The reason for creating both positive and negative models is that using both models allows for better control of the level of generalization, and thus increases the confidence in the final models.
3. Given a set of patient characteristics $< c_1, \ldots c_k >$, model M will return a set of possible combinations of treatments $\{T_1, \ldots T_n\}$that are likely to be successful, $M(< c_1, \ldots c_k >) = \{T_1, \ldots T_n\}$. It is possible that for a given case more than one combination of treatments is returned, i.e. $n > 1$, or no considered combination of treatments is returned, i.e. $n = 0$. Similarly, models M_{NT} are applied, to create a list of potentially improper combinations of treatments.
4. Test model M on a subset of "unused" data consisting of P "successful" cases and N "unsuccessful" cases. Results of the testing are reported in terms of specificity, selectivity, and statistical significance of individual models and all models together.

The created models define groups (or clusters) of patient characteristics that are likely to have positive or negative outcomes. Note that the groups may be intersecting i.e., more than one combination of treatments may appropriate in a specific case, and not exhaustive, i.e., there may be cases for which none of the examined combinations of treatments is predicted to be successful. In the latter case, a *flexible interpretation* of rules may be used to select the closest potentially successful combination of therapies. Within groups of patients selected by machine learning, traditional comparative effectiveness can be performed.

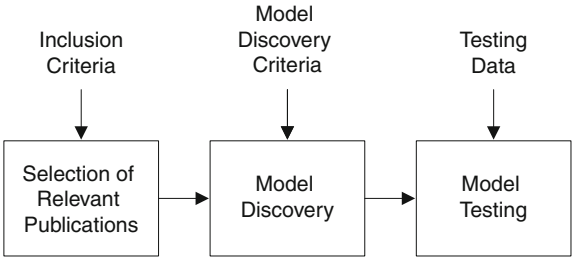


Fig. 7.3 Steps in rule learning from published aggregated data

7.4.3 Aggregated Data

There is a growing need to combine data originated from multiple clinical studies. A majority of published studies describe relatively small cohorts and produce platform-dependent results that often lack consistency. Individual measurements of the clinical parameters are protected by The Health Insurance Portability and Accountability Act (HIPAA), thus precluding a combination of multiple cohorts into the large database to perform secondary analyses. A combination of multiple studies, which is the goal of systematic reviews, relies on meta-analysis methods to statistically combine results of the studies. Traditional meta-analysis, however, does not perform knowledge discovery or build predictive/classification models from aggregated data [14].

The problem addressed here is how to learn rules from aggregated data published from multiple studies, rather than from individual examples (subjects). The goal of the method is to discover a model M for diagnosing diseases D, from published results in which data satisfy a set of criteria C. One important characteristic of the method is that the studies do not need to describe diagnostic methods for diseases D, but to only include relevant data summaries. Common inclusion criteria that are prerequisites for the traditional meta-analysis methods are not required either. It is sufficient that the criteria are disclosed, so they can serve as inputs to the model along with the aggregated data. The process of the model development is depicted in Fig. 7.3.

The rule learning problem considered here induces a rule-based classifier $M(X) \rightarrow D$ that can be used to diagnose X patients into diseases from D. The model is induced using aggregated data describing groups of patients, not individual datapoints as typically handled by machine learning algorithms. Specifically, the method uses aggregated data A, inclusion criteria C, and other groups' information G to create model M. This process extends learning from aggregated data that deals with multiple cohorts of patients described as mean $\pm$ standard deviation of each clinical parameter.

The method has been applied to deriving diagnostic models for metabolic-syndrome related liver complications from summarized (aggregated) descriptions of the small cohorts of patients available from published manuscripts. The significance of this topic is large because approximately 47 million people in the

United States have metabolic syndrome (MS) and this number is on the rise. The aggregated clinical data were retrieved from articles published in leading peer-reviewed journals. By applying the developed rule learning methodology, we arrived at several different possible rulesets (sets of rules that together form a model to make a specific diagnosis) that can be used to predict three considered complications of MS, namely non-alcoholic fatty liver disease (NAFLD), simple steatosis (SS), and nonalcoholic steatohepatitis (NASH). It should be noted that the NAFLD group comprises both SS and NASH cases, which means that values of the output attribute form a hierarchy.

Seven NAFLD or NASH predicting rulesets were generated using the AQ21 system executed with different parameters. Resultant rulesets predicting NAFLD or NASH were blindly validated using a well-defined NAFLD database containing 489 patients with biopsy-proven NAFLD, NASH or SS with extensive clinical and laboratory data.

An example of typical automatically learned rule states that patients with BMI >26.85 are likely to have NAFLD, except for when AST is at most 27.2 and adiponectin level are at least 7.25 [14]. The rule is formally shown as:

$$[Group = NAFLD] <== [BMI > 26.85] \lfloor [AST \leq 27.2] \,\&\, [Adiponectin \geq 7.25]. \tag{7.6}$$

Validation of this rule for predicting NAFLD resulted in a positive predictive value (PPV) of 85–87 %, reflecting relatively high "rule-in" characteristic of the algorithm. The best rule for the prediction of NASH relied on combination of fasting insulin, HOMA and adiponectin values with an accuracy of 78 %, with PPV of 71 % and negative predictive value (NPV) of 37 %.

7.5 Summary

This chapter briefly presented rule learning and its uses in healthcare and health services research. The focus of this paper was on the AQ21 rule learning and testing system because of the system's applicability to healthcare problems. AQ21 can be viewed more like a laboratory for experimentation with healthcare data rather that a single computer program, which can be executed on data and produce rules. Rule learning performed by AQ21 is particularly suitable for healthcare applications because its high transparency increases the chance that models will be accepted by users.

Acceptability of machine learning methods is a central criterion among those listed in Sect. 7.2. Other criteria (accuracy, transparency, etc.) lead to the acceptability of models, which in healthcare community is very hard to achieve. While other types of models, such as decision trees and Bayesian networks, are known to be highly transparent, attributional rules follow most of the criteria listed in Sect. 7.2.

Among the numerous current and potential applications of rule learning in healthcare and health services research, three diverse applications were briefly presented in this chapter. Each application demonstrates that rule learning has great potential and can give good results. The application of rule learning is, however, always straight-forward, and significant work and preparations need to be done before rule learning can be effectively/efficiently used.

Future work on rule learning should focus on four directions. (1) Richer and more natural (to people) rule-based knowledge representations can be created by extending attributional calculus to capture concepts that are natural to healthcare practitioners and researchers. (2) Easy to use tools that deal directly with healthcare data can be developed. One attempt to make computational intelligence and machine learning (CIML) tools accessible to the healthcare community was through CIML Virtual Organization [17]. The VO's goal is to provide the healthcare community with access to CIML tools, advice, educational materials, and networking. (3) Efficiency of rule learning methods can be improved. High complexity or rule based representations require long computation times, particularly when advanced methods, such as constructive induction, are used. (4) Machine learning, in particular rule learning, can be popularized as an attractive approach to data analysis and systems' adaptability, to healthcare community.

References

1. Bloedorn E, Wnek J, Michalski RS, Kaufman K (1993) AQ17 A multistrategy learning system the method and users guide. reports of the machine learning and inference laboratory: MLI 93-12. School of Information Technology and Engineering. George Mason University, Fairfax
2. Bloedorn E (1996) Multistrategy constructive induction. Ph.D. Dissertation, Reports of the machine learning and inference laboratory : MLI 96-7. School of Information Technology and Engineering. George Mason University, Fairfax
3. Bloedorn E, Michalski RS (1998) Data-driven constructive induction. IEEE intelligent systems special issue on feature transformation and subset selection: 30–37
4. Hripcsak G (1994) Writing arden syntax medical logic modules. Comput Biol Med 5(24):331–363
5. Michalski RS (1969) On the quasi-minimal solution of the general covering problem. In: Proceedings of the 5th international symposium on information processing (FCIP 69) (Switching Circuits). A3: 125–128. Yugoslavia, Bled
6. Michalski RS, Larson J (1975) AQVAL/1 (AQ7) User's guide and program description. Department of Computer Science. University of Illinois, Urbana
7. Michalski RS, Larson J (1983) Incremental generation of VL1 hypotheses: the underlying methodology and the description of program AQ11. Department of computer science: reports of the intelligent systems group, ISG 83-5, UIUCDCS-F-83-905. University of Illinois, Urbana
8. Michalski RS, Kaufman K (2001) The AQ19 system for machine learning and pattern discovery: a general description and user's guide. Reports of the machine learning and inference laboratory: MLI 01-2. George Mason University, Fairfax

9. Michalski RS (2004) Attributional calculus: a logic and representation language for natural induction. Reports of the machine learning and inference laboratory: MLI 04-2 George Mason University, Fairfax
10. Wnek J, Michalski RS (1994) Hypothesis-driven constructive induction in AQ17-HCI: a method and experiments. Mach Learn 14(2):139–168
11. Wnek J, Kaufman K, Bloedorn E, Michalski RS (1996) Inductive learning system AQ15c: The method and user's guide. Reports of the machine learning and inference laboratory: MLI 96-6 George Mason University, Fairfax
12. Wojtusiak J (2004) AQ21 User's guide. Reports of the machine learning and inference laboratory: MLI 04-3 George Mason University, Fairfax
13. Wojtusiak J, Michalski RS, Kaufman K, Pietrzykowski J (2006) The AQ21 natural induction program for pattern discovery: initial version and its novel features. In: Proceedings of the 18th IEEE international conference on tools with artificial intelligence. Washington, D.C
14. Wojtusiak J, Michalski RS, Simanivanh T, Baranova AV (2009) Towards application of rule learning to the meta-analysis of clinical data: an example of the metabolic syndrome. Int J Med Informatics 78(12):104–111
15. Wojtusiak J, Alemi F (2010) Analyzing decisions using datasets with multiple attributes: a machine learning approach. In: Yuehwern Y (ed) Handbook of healthcare delivery systems. CRC Press, Boca Raton
16. Wojtusiak J, Shiver J, Ngufor C, Ewald R (2011) Machine learning in hospital billing management. presentation. In: HIMSS 2011 acedemic forum. AUPHA, Orlando
17. Zurada JM, Mazurowski MA, Abdullin A, Ragade R, Wojtusiak J, Gentle JE (2009) Building virtual community in computational intelligence and machine learning. Comput Intell Mag 4(1):43–54

Chapter 8
Machine Learning Techniques for AD/MCI Diagnosis and Prognosis

Dinggang Shen, Chong-Yaw Wee, Daoqiang Zhang, Luping Zhou and Pew-Thian Yap

Abstract In the past two decades, machine learning techniques have been extensively applied for the detection of neurologic or neuropsychiatric disorders, especially Alzheimer's disease (AD) and its prodrome, mild cognitive impairment (MCI). This chapter presents some of the latest developments in the application of machine learning techniques to AD and MCI diagnosis and prognosis. We will divide our discussion into two parts: single modality and multimodality approaches. We will discuss how various biomarkers as well as connectivity networks can be extracted from the various modalities, such as structural T1-weighted imaging, diffusion-tensor imaging (DTI) and resting-state functional magnetic resonance imaging (fMRI), for effective diagnosis and prognosis. We will further demonstrate how these modalities can be fused for further performance improvement.

Chong-Yaw Wee, Daoqiang Zhang and Luping Zhou contributed equally to this book chapter.The work was performed when Luping Zhou was with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill.

D. Shen (✉) · C.-Y. Wee · D. Zhang · P.-T. Yap
Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
e-mail: dgshen@med.unc.edu

C.-Y. Wee
e-mail: cywee@med.unc.edu

D. Zhang
e-mail: dqzhang@nuaa.edu.cnzhangd@med.unc.edu

P.-T. Yap
e-mail: ptyap@med.unc.edu

D. Zhang
Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

L. Zhou
University of Wollongong, Wollongong, NSW 2522, Australia
e-mail: lupingz@uow.edu.au

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_8,

Keywords Alzheimer's disease · Mild cognitive impairment · Machine learning · Diagnosis · Prognosis · Connectivity networks · Multimodality

8.1 Background

Alzheimer's disease (AD) is the most common form of dementia, characterized by cognitive and intellectual deficits that interfere with daily life if effective treatment is not available. AD gets worse over time by gradually destroying brain cells, causing loss in memory and the ability to reason, make judgments, and communicate. In 2006 the worldwide prevalence of AD was 26.6 million, and it is projected that 1 in 85 persons will be affected by 2050 [1]. The number of people who develop AD is expected to continue to increase as life expectancy increases. With the aging of the world population, AD has become a serious problem and a huge burden to the healthcare system. Recognizing the urgent need to slow down or completely prevent the occurrence of a worldwide healthcare crisis, effort has been under way to develop and administer effective pharmacological and behavioral interventions for delaying the onset and progression of the disease.

A significant body of literature [2–4] suggests that pathological manifestation of AD begins many years before it can be diagnosed using cognitive tests. At the stage where symptoms can be observed, significant neurodegeneration has already occurred. Studies suggest that individuals with mild cognitive impairment (MCI), a prodrome of AD, are expected to convert to probable AD at an annual rate of 10–15 % [5], whereas healthy controls develop dementia at an annual rate of 1–2 % [6]. Compared to AD, MCI is more difficult to diagnose due to its very mild cognitive impairment symptoms. At the present time, AD-related neurodegeneration such as structural atrophy [7], pathological amyloid depositions [8], and metabolic alterations [9] have been identified as potential biomarkers.

Advanced statistical machine learning and pattern recognition techniques have been actively applied to map neurodegenerative patterns during the early stage [10–13]. Examples of machine learning techniques that are widely used in medical imaging analysis include support vector machines (SVMs) [14], boosting-based learning [15], artificial neural networks [16], *k*-nearest neighbor classifier [17], and linear discriminant analysis [18]. In addition to determining group differences, pattern classification methods can be trained to identify individuals who are at risk for AD [11, 12, 19–23]. A recent study demonstrated that classification methods are capable of identifying AD patients via their MRI scans and achieved accuracy comparable to that obtained by experienced neuroradiologists [19]. Efforts have also been undertaken to develop regression techniques for relating clinical scores to imaging data [24–26], facilitating continuous monitoring of AD progression. In this chapter, we will focus on machine learning based diagnosis and prognosis of AD/MCI using information obtained from single and multiple modalities.

8.1.1 Single-Modality-based Diagnosis and Prognosis

Single-modality-based methods are clinically more feasible due to simpler scanning protocols and lesser image acquisition effort. For example, many methods use only structural MRI brain images for classification between AD/MCI patients and normal controls [12, 27, 28]. Popular neuroimaging measurements include: regional brain volumes [29, 30], cortical thickness [31–33], and hippocampal volume and shape [34, 35], etc.

The understanding of brain anatomical circuitry has been experiencing remarkable progression due to the development of diffusion tensor imaging (DTI), where white matter (WM) fiber bundles can be delineated through characterization of water diffusion [36]. WM tracts connecting brain regions can be reconstructed in vivo using diffusion tractography (or fiber tracking) to characterize brain circuitry [36]. Diffusion measures such as fractional anisotropy (FA) and mean diffusivity (MD) are commonly utilized as features in statistical analysis to localize WM changes related to AD and MCI [37, 38].

Functional connectivity is defined as the temporal correlation between regional neurophysiological signal fluctuations [39, 40]. Blood oxygenation level dependent (BOLD) signal, which extracted from functional magnetic resonance imaging (fMRI) data, exhibits low-frequency spontaneous fluctuations in the resting brain and shows a high degree of temporal correlation across different brain regions. Since the seminal work of Biswal et al. [41], resting-state fMRI (rs-fMRI) has been widely applied to the analysis of various neuropsychological diseases including MCI [42] and AD [43]. One apparent advantage of resting-state fMRI over task-activation fMRI is that no complicated experimental design is required. Experiments can be performed easily with patients who may have difficulties performing specific task inside the scanner, especially those with disorders that exhibits prominent cognitive degeneration, such as AD [40].

Another important imaging modality for AD/MCI detection is fluorodeoxyglucose positron emission tomography (FDG-PET) [29]. With FDG-PET, reduction of glucose metabolism was found in the parietal, posterior cingulate and temporal brain regions of AD patients [44]. Besides neuroimaging techniques, biological or genetic biomarkers are effective alternatives for AD/MACI diagnosis. Researchers found that (1) the increase of cerebrospinal fluid (CSF) total tau (t-tau) and tau hyperphosphorylated at threonine 181 (p-tau) are related to neurofibrillary tangle, (2) the decrease of amyloid β ($A\beta_{42}$) indicates amyloid plaque deposit, and (3) the presence of the apolipoprotein E (APOE) $\varepsilon 4$ allele can predict cognitive decline or conversion to AD [45].

8.1.2 Multimodality-based Diagnosis and Prognosis

It has been demonstrated that different imaging modalities can provide complementary information to enhance AD/MCI diagnosis [45–47]. For example, it was reported that FDG-PET and MRI measures may be complementarily and differentially sensitive to memory in health and disease, with metabolism being the stronger predictor in normal controls, and morphometry most related to memory function in AD [47]. Also, it is shown that morphometric changes in AD and MCI, although are related to CSF biomarkers, can provide complementary information [45]. A more recent study, which compared the respective prognostic ability of genetic, CSF, neuroimaging, and cognitive measures obtained from the same participants, demonstrated that complementary information provided by these different modalities can be used for enhanced AD/MCI diagnosis [46]. Inspired by these findings, a number of studies used two or more biomarkers simultaneously to detect AD and MCI: MRI and CSF [10, 13], MRI and cognitive testing [48], MRI and PET [22, 49], MRI and APOE biomarkers [50], FDG-PET and CSF [51], FDG-PET and cognitive testing [52], and MRI, CSF, and FDG-PET [53].

8.2 Single-Modality-based Diagnosis and Prognosis

AD and other similar progressive degenerative neurological diseases exhibit spatially and temporally pathology, where the brain is damaged on a large-scale, highly connected network, rather than in a single isolated region. In view of this, a sensitive description of interregional connections is required to better delineate the pathology of disease for accurate diagnosis. Models of whole-brain connectivity, which comprise networks of brain regions connected either by anatomical tracts or functional associations, have drawn a great deal of interest recently due to the increasing reliability of network characterization through neurobiologically meaningful and computationally efficient measures [54, 55]. In this section, we will discuss some recently proposed network-based techniques using biomarkers from single imaging modality for AD and MCI diagnosis and prognosis.

8.2.1 Structural Analysis via Enriched White Matter Connectivity Networks

Recently, an enriched description of WM connections via diffusion tractography [56] was proposed to convey topological and biophysical information of the connections. This description is achieved by using a collection of diffusion parameters that are derived during whole-brain streamline fiber tractography and is aimed to effectively describe small variations on WM regions caused by

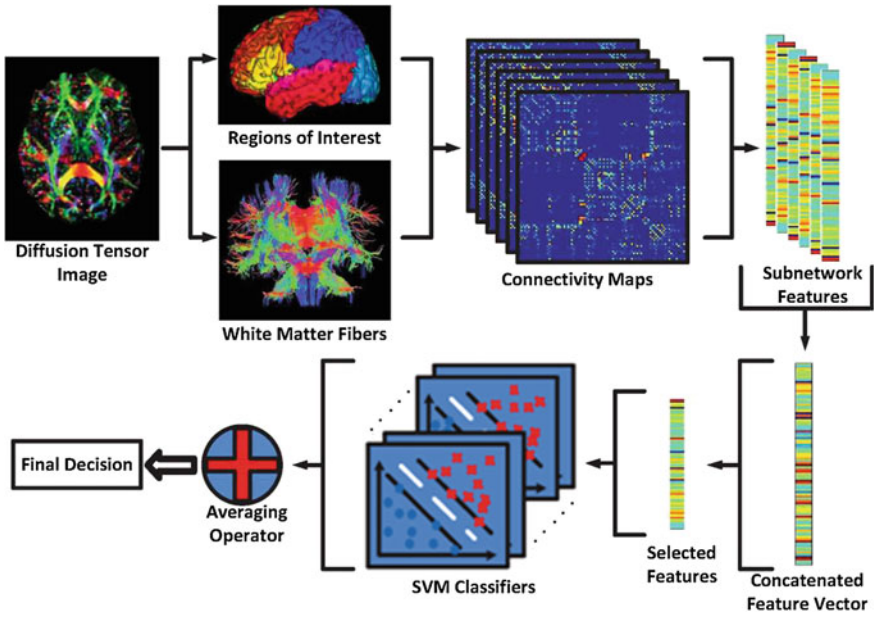


Fig. 8.1 Classification based on enriched description of WM connections

pathological attacks. The MCI classification framework using this enriched description is shown in Fig. 8.1.

Six diffusion parameters are included in the enriched description, i.e., fiber count, fractional anisotropy (FA), mean diffusivity (MD), and principal diffusivities (λ_1, λ_2, λ_3). During tractography, the number of fibers passing through each pair of regions is counted. Two regions are considered anatomically connected if there are fibers passing through their respective masks. Counting the number of connecting fibers between every possible pair of regions provides us the connection topology of the network. Connectivity networks of FA, MD, and principal diffusivity can also be derived by taking the average values along the connecting fibers. These five networks share identical connection topology as the fiber count network, but conveying different biophysical properties. An example of the six connectivity networks for one subject is provided in Fig. 8.2.

Network measures typically quantify connectivity profiles associated with the nodes and reflect the way how these nodes are embedded in the network. Clustering coefficients [57, 58], which measures the cliquishness of a network, is commonly used to extract information from the constructed brain connectivity networks for group analysis. The original clustering coefficient is formulated to work only with unweighted graphs and is intended to provide a summary statistics of the whole network. To increase sensitivity to pathology induced network changes, a weighted version of local clustering coefficient [57] can be used instead. However, the use of local fine-grained features will produce a high-dimensional

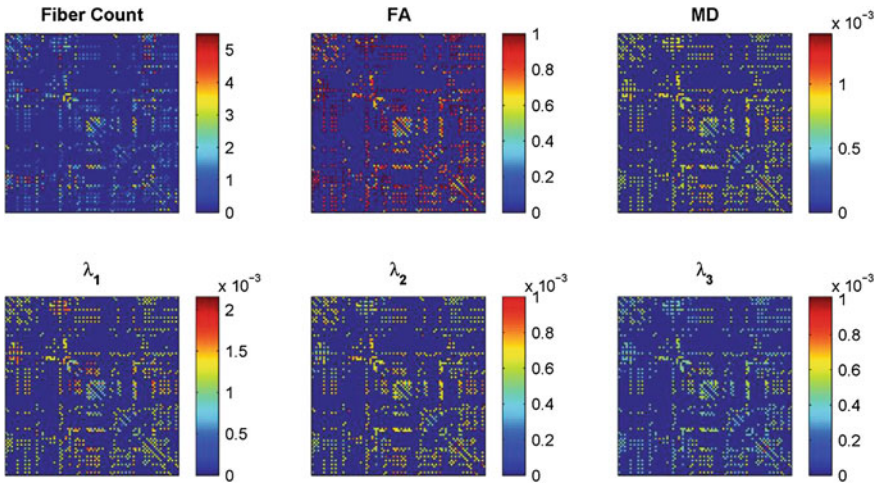


Fig. 8.2 Connectivity networks constructed with different diffusion parameters

feature pool which may cause the problem of curse of dimensionality, particularly in the graph theoretic approach. Good classification performance is normally difficult to achieve if all extracted features are directly used indiscriminately. This difficulty arises because not all the features are equally important for classification. A proper feature selection procedure needs to be employed to select an optimal subset of features with the most discriminative power to improve generalization performance.

The discriminative power of a feature can be quantitatively evaluated by its relevance to classification as well as its generalizability. Relevancy of a feature to classification is measured through its correlation with clinical labels [21]. Pearson correlation coefficient is commonly used to rank features based on this relevancy. Features with larger absolute value of the Pearson correlation coefficient are considered to be more relevant to classification.

The generalizability of a feature is evaluated via leave-one-out cross-validation (LOOCV) when measuring the correlation of the feature with respect to the clinical labels [21]. Specifically, for n training samples, the worst absolute Pearson correlation coefficient resulting from the n leave-one-out correlation measurement is conservatively selected as the effective correlation coefficient. This approach is particularly important for minimizing the effect of outliers when evaluating a huge number of features.

Nevertheless, the ranking score is computed independently for each feature, without considering the correlation with other features. This method inevitably causes some redundant features to be selected, thus affecting classification performance. To minimize this effect, a wrapper-based feature selection method called an SVM-RFE algorithm [59, 60] is used to select the final optimal subset based on feature ranking.

Classification performance of the enriched WM connectivity description method is evaluated using a nested LOOCV strategy [56] to ensure a relatively unbiased estimate of the generalization power of the classifiers to new subjects. In each LOO case, one subject is first left out as the testing subject, and the remaining subjects are used for feature extraction, feature selection and classifier training. A second or inner LOO loop is applied to the training set to construct and optimize an ensemble classifier. Specifically, for n total number of subjects involved in the study, one is left out for testing, and the remaining $n - 1$ are used for training. From the remaining $n - 1$ samples, $n - 1$ different training subsets are formed by each time leaving one more sample out, giving us $n - 2$ subjects in each training subset. For each subset, an SVM classifier is construct with its performance is evaluated using the second left out subject. This procedure is repeated $n - 1$ times, once for each training subset. This procedure ensures that the selected diffusion parameters maximize the area under the receiver operating characteristic (ROC) curve. When the unseen (omitted during the training and parameter optimization process) test sample is to be classified, all n – 1 classifiers are used, and their outcomes are averaged to provide the final classification decision. This process is repeated n times, each time leaving out a different subject, finally leading to overall cross-validation classification accuracy.

8.2.2 Functional Analysis via Multi-Spectral Connectivity Networks

Over the past several years, rs-fMRI has emerged as a novel informative method for investigating the development of large-scale functional networks in the human brain. This method, first used to demonstrate coherent spontaneous low-frequency fluctuations in BOLD signal within the adult somatomotor system [41], involves measuring the hemodynamic response related to neural activity in the brain or spinal cord from participants as they lay in the MRI scanner in the "resting condition". This system was recently employed to identify individuals with MCI from healthy controls and performed well in the tests [61].

Wee et al. [61] suggested an efficient characterization of rs-fMRI time series via: (1) Multi-spectral characterization to quantify relatively small changes of BOLD signal by decomposing the mean time series of each ROI into five distinct frequency sub-bands, and (2) Graph theoretic analysis to characterize topological properties and strengths of brain functional connectivity networks through neurobiologically meaningful and computationally efficient measures [54, 55, 62].

In vivo neuroimaging studies suggest that normal aging [63] and AD [64, 65] are associated with GM volume loss. There is an emerging body of evidence that MRI can observe deterioration, including progressive loss of GM in the brain, from MCI to full-blown AD [66]. It has been shown that the GM volume of the human brain decreases linearly by approximately 5.0 % per decade throughout

lifetime after 9 years of age [63]. It has been reported that local GM loss rates are approximately 5.3 and 0.9 % per annum in AD and healthy aging, respectively, with an asymmetric trend where a faster loss rate is observed in the left hemisphere than in the right [65]. Furthermore, removal of signals from the ventricles and WM can reduce the noise caused by the cardiac and respiratory cycles [67]. Based on these observations, only the BOLD signal extracted from the GM is used. To achieve this, tissue segmentation is performed on T1-weighted image of each subject to label the GM, WM, and CSF. Then, the segmented GM image is used to mask the fMRI images. This procedure eliminates signal contamination originating from WM and CSF in the fMRI time series. Anatomical parcellation is used to divide the brain into different regions-of-interest (ROIs).

For each subject, the mean time series of each individual ROI is obtained by averaging the GM-masked fMRI time series over all voxels in the ROI. Temporal band-pass filtering with a frequency interval $(0.025 \leq f \leq 0.100\,\mathrm{Hz})$ is then applied to the mean time series of each individual ROI, trading-off between avoiding the physiological noise associated with higher-frequency oscillations [68] and the measurement error associated with estimating very low-frequency correlations from truncated time series [69]. In conventional approaches, the regional mean time series of entire spectrum is directly employed to construct functional connectivity networks. However, this whole-spectrum approach might not be sensitive enough to describe complex yet subtle pathological patterns of the neurological diseases.

In [61], a multi-spectral characterization of the regional mean time series is proposed to construct functional connectivity networks. The band-pass filtered GM-masked mean time series of each region is decomposed into five equally divided frequency subbands using the fast Fourier transform (FFT). Using this multi-spectral approach, small BOLD signal variations can be better preserved.

Functional connectivity, which indicates interregional correlations in neuronal variability [39], can be measured using a pairwise Pearson correlation coefficient between a given pair of ROIs. Given a set of N random variables, the Pearson correlation matrix is a symmetric matrix in which each off-diagonal element is the correlation coefficient between a pair of variables. The brain regions can be considered a set of nodes and the correlation coefficients can be considered signed weights on the set of edges. The normality of Pearson correlation coefficients is improved by applying a Fisher's r–to–z transformation. The feature extraction, feature selection and high-dimensional multivariate classification steps used for MCI diagnosis in [56] are similarly applied to the case of rs-fMRI.

Examples of the functional connectivity maps constructed using the multi-spectral characterization for one normal control (NC) and one MCI patient are shown in Fig. 8.3.

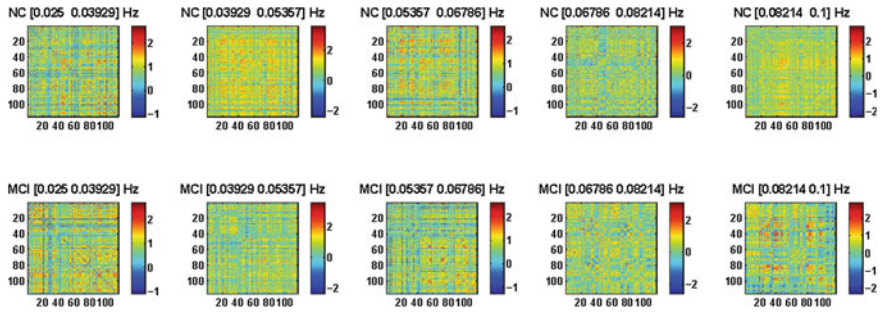


Fig. 8.3 Multi-spectral functional connectivity maps for a normal control (NC) and an MCI patient

8.2.3 Hierarchical Brain Networks from T1-Weighted MRI

Because of its clinical accessibility, T1-weighted MRI has been widely utilized for the diagnosis and prognosis of MCI and AD. Conventionally, the mean tissue volumes of GM, WM, and CSF are calculated locally within ROI, and used as features for classification. Nevertheless, it is realized that disease-induced brain structural changes may happen in several inter-related regions instead of isolated spots. Therefore, it is proposed in [70] that compared with the traditional local isolated measures, representing the brain as a system of interconnected regions may be a more effective way to characterize subtle brain changes. For this purpose, this approach constructs a hierarchical brain network to directly model the pairwise ROI relationships within a subject, with each node denoting a ROI and each edge characterizing the pairwise connection. The node of ROI is represented by a volumetric vector that consists of the mean volumes of GM, WM, and CSF in this ROI. The relationship between two ROIs within the same subject is computed by the Pearson correlation between the two corresponding volumetric vectors. The correlation value indicates the similarity of the tissue compositions between a pair of brain regions. When a patient is affected by MCI, the correlation values of some brain regions with other regions will be affected, due possibly to factors such as tissue atrophy.

By computing the pairwise correlation between ROIs, the approach in [70] provides a second order measure of the ROI volume, while the conventional approaches only employ the first order measure of the volume. As higher order measures, the proposed new features may be more descriptive, but also more sensitive to noise, such as registration errors. Therefore, a four-layer hierarchy of multi-resolution ROIs (Fig. 8.4a) is introduced to increase the robustness of classification. Effectively, the correlations are considered at different scales of regions to provide different levels of noise suppression and discriminant information, which can be further selected by the proposed classification scheme. This approach considers the correlations both within and between different resolution

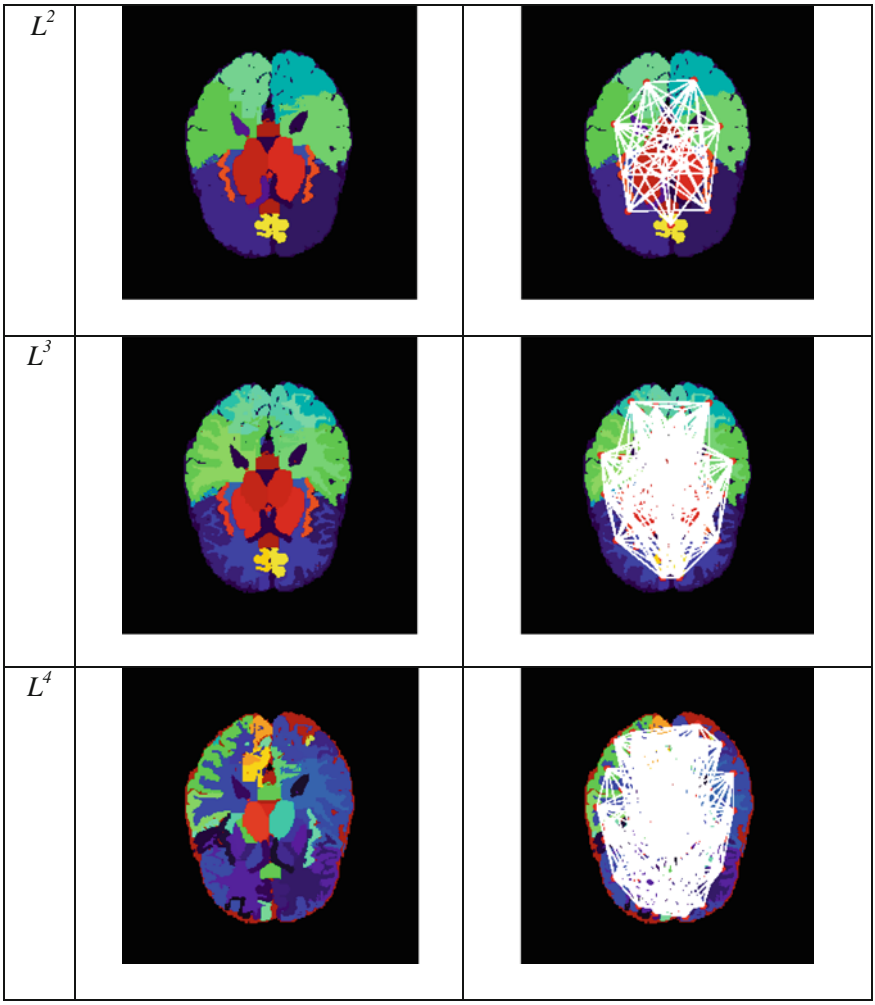


Fig. 8.4 **a** Hierarchical ROIs in three layers (the *top layer* is a whole brain which is not shown), **b** Network connections between ROIs within different layers

scales (Fig. 8.5), because a certain "optimal" scale often cannot be known a priori. The brain network could be very complicated as partially shown in Fig. 8.4b. To efficiently construct the informative network features, a membership matrix is created to indicate the relationship of ROIs from different layers. The membership matrix is computed offline: it is fixed once the hierarchical structure has been determined. For a new brain image, this approach only needs to compute the ROI interactions on the bottommost layer that has the highest resolution of ROIs, and then use the membership matrix to propagate the correlations to other layers effectively.

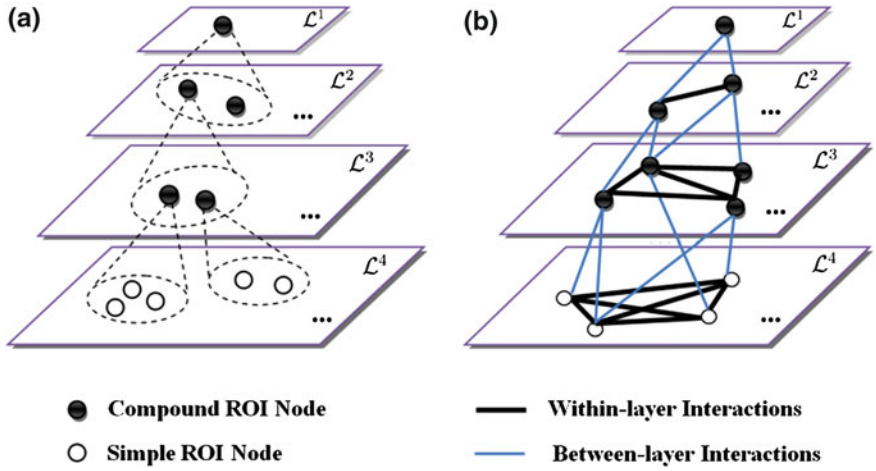


Fig. 8.5 Schematics of the network model. **a** Two types of nodes are included in the hierarchical network: the simple node in the bottommost layer and the compound node in other layers. Each compound node is obtained by grouping several simple nodes in an agglomerative fashion. **b** Two types of edges are included in the hierarchical network, modeling the within-layer and between-layer interactions, respectively

Note that the proposed brain network may not be sparse, as shown in the DTI and fMRI networks [62], because the connections in this case are not based on functional or real neuron-connections. The dense adjacency matrix resulting from the correlation of tissue compositions implies that WM, GM, and CSF fractions of brain regions are consistently similar. Note that the far-away region pairs can have meaningful tissue composition similarity since distance information is not included in this approach. Because the network is fully connected, some commonly used network features, such as local clustering coefficients, do not work as efficiently as they do for the sparse networks in DTI and fMRI. Therefore, the weights of edges are directly used as features, that is, the elements in the upper triangle matrices of correlation matrices are concatenated to form the feature vectors.

This approach produces significantly larger number of features than conventional methods. If improperly handled, classifier training may become intractable due to this large number of features. Conventionally, there are usually two ways to deal with the high dimensionality of features: (1) select the most discriminative subset of features from the original features, known as feature selection, or (2) combine the original features linearly or non-linearly to obtain a lower dimensional new feature space, known as feature embedding. Zhou et al. proposed a dimensionality reduction process to efficiently reduce the feature dimensionality to a manageable level while preserving as much discriminative information as possible [70]. This method combines feature selection and feature embedding via partial least square (PLS) analysis [71] in an integrated optimization process. PLS is a supervised learning method, which makes use of classification labels for data

embedding. Therefore, it achieves a better discrimination than many of the popular unsupervised methods, such as principal components analysis (PCA) and Laplacian Eigenmap, and even than some advanced supervised methods, such as kernel Fisher discriminant analysis (KFDA).

Taking advantage of PLS analysis, the approach presented in [70] employed four steps to achieve good classification and generalization: rough feature selection, refined feature selection, feature embedding and linear classification. They are elaborated as follows.

In *Step 1*, a rough feature selection is performed to filter out a large amount of features that have little relevance to the classification. The relevance is computed by the Pearson correlation between each original feature and the classification label. Features with absolute correlation values lower than a threshold are treated as irrelevant features and filtered out.

In *Step 2*, a refined feature selection is performed to pick out the candidate features for the PLS feature embedding. For this purpose, the selected features in Step 1 are used to train a PLS model, and then ranked by variable importance on projection (VIP) score [72] to estimate their discriminative power for the PLS model. After this step, about 60–80 discriminative features with the top VIP scores are reserved for feature embedding in the next step.

In *Step 3*, a PLS embedding is performed to further reduce the dimensionality of the network features. Using the refined selected features in Step 2, a new PLS model is constructed to seek an embedding space that best preserves the discrimination of features. Then the selected features in Step 2 are projected onto a much lower dimensional space learned by PLS analysis in Step 3.

In *Step 4*, after PLS embedding, a small number of features in the new space have been able to capture the major class discrimination. This greatly reduces the complexity of relationships between data. Therefore, in Step 4, using the features in the embedded space, a linear SVM has been sufficient for an accurate prediction of MCI patients.

Note that the number of selected features in each step is determined by cross-validation on the training data.

The merits of the proposed method are as follows. First, the proposed method uses a second-order volumetric measure that is more descriptive than the conventional first-order volumetric measure. Second, while the conventional approaches only consider local volume changes, the proposed method considers global information by pairing spatially separated ROIs. Third, at the top of the hierarchy the proposed method introduces a whole-brain ROI, with which, each ROI can provide a first-order measurement of local volume. In this way, the proposed method seamlessly incorporates both the local volume features and the proposed global network features into the classification. Fourth, the proposed method involves only linear methods, leading to interpretability of classification results, which is equally important as classification accuracy in neuroimaging analysis. Finally, the proposed method investigates the *relative* speeds of disease progression in different brain regions, providing a complementary perspective of the spatial atrophy patterns to conventional methods.

8.3 Multimodality-based Diagnosis and Prognosis

A number of studies have shown that biomarkers from different modalities may contain complementary information useful for diagnosis of AD/MCI, and several works on combining different modalities have been reported [10, 13, 27, 47, 50, 51, 73]. A common trait of these methods is that they concatenate all the features from different modalities into a long feature vector. However, approaches as such do not distinguish between modalities and are hence not the best way to combine information from multiple sources. In this section, we provide an alternative method that uses a multiple kernel combination to integrate biomarkers. Compared with the direct feature concatenation method, the kernel combination method has the following advantages: (1) It can combine heterogeneous data that cannot be directly concatenated; (2) it provides more flexibility by using different weights on the biomarkers of different modalities. Furthermore, to overcome the small sample size problem in training multimodality classifier, we adopt a semi-supervised learning technique that can learn from both labeled and unlabeled data.

8.3.1 Multimodality Data Fusion via Multi-Kernel SVM

Zhang et al. [74] recently proposed a general framework based on kernel methods developed by Scholkopf and Smola [75] to combine multiple biomarkers (i.e., MRI, PET, and CSF) for discriminating between AD (or MCI) and normal controls. The proposed method is based on kernel combination and can be easily embedded into the conventional SVM classifier for high-dimensional pattern classification. Moreover, unlike other kernel combination methods which can only process one data type, i.e. numbers, this method can combine numeric data, strings, and graphs, etc. The framework proposed by Zhang et al. [74] is explained as below.

In SVM, by using a kernel-induced implicit mapping function, linearly non-separable samples are first mapped to a higher or infinite dimensional space, where they are more likely to be linearly separable than in the original space. A maximum margin hyperplane is then sought in the higher-dimensional space. Multiple-kernel learning (MKL), which is pioneered by Lanckriet et al. [76] and Bach et al. [77], is an additive extension of the single kernel SVM by incorporating multiple kernels. Suppose that we are given n training samples and each of them contains M modalities. Let $\mathbf{X}_i^{(m)}$ denote a feature vector of the mth modality of the ith sample, and its corresponding class label be $y_i \in \{-1, 1\}$.

Multiple-kernel based SVM solves the following primal problem:

$$\begin{aligned} &min_{w^{(m)},b,\xi_i} \frac{1}{2}\sum_{m=1}^{M}\beta_m\left|\left|w^{(m)}\right|\right|^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t. \qquad y_i\left(\sum_{m=1}^{M}\beta_m\left(\left(w^{(m)}\right)^T\phi^{(m)}\left(x_i^{(m)}\right)+b\right)\right) \geq 1-\xi_i \\ &\qquad \xi_i \geq 0, \quad i=1,\ldots,n, \end{aligned} \tag{8.1}$$

where $\mathbf{W}^{(m)}$, $\phi^{(m)}$ and $\beta_m \geq 0$ denote the normal vector of hyperplane, the kernel-induced mapping function and the weighting factor of the mth modality, respectively.

Similar to the conventional SVM, the dual form of multiple-kernel SVM can be formulated as below:

$$\begin{aligned} &max_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \sum_{m=1}^{M}\beta_m k^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right) \\ &s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \\ &\qquad 0 \leq \alpha_i \leq C, i=1,\ldots,n \end{aligned} \tag{8.2}$$

where $k^{(m)}\left(\boldsymbol{x}_i^{(m)}, \boldsymbol{x}_j^{(m)}\right) = \phi^{(m)}\left(\boldsymbol{x}_i^{(m)}\right)^T\phi^{(m)}\left(\boldsymbol{x}_j^{(m)}\right)$ is the kernel function for the two training samples on the mth modality.

For a new test sample $\boldsymbol{x} = \left\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(\boldsymbol{M})}\right\}$, we first denote $k^{(m)}\left(\boldsymbol{x}_i^{(m)}, \boldsymbol{x}^{(m)}\right) = \phi^{(m)}\left(\boldsymbol{x}_i^{(m)}\right)^T\phi^{(m)}\left(\boldsymbol{x}^{(m)}\right)$ as the kernel between the test sample and each training sample on the mth modality. Then, the decision function for the predicted label can be obtained as below:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{n} y_i\alpha_i\sum_{m=1}^{M}\beta_m k^{(m)}(x_i^{(m)}, x^{(m)}) + 1\right). \tag{8.3}$$

Multiple-kernel based SVM can be naturally embedded into the conventional single-kernel SVM if we denote $k(x_i, x_j) = \sum_m \beta_m k^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right)$ as a mixed kernel between the multimodality training samples $\mathbf{X}_i$ and $\mathbf{X}_j$, and $k(\boldsymbol{x}_i, \boldsymbol{x}) = \sum_m \beta_m k^{(m)}\left(\boldsymbol{x}_i^{(m)}, \boldsymbol{x}^{(m)}\right)$ as a mixed kernel between the multimodality training sample $\mathbf{X}_i$ and the test sample $\mathbf{X}$.

It is worth noting that the multiple-kernel SVM proposed by Zhang et al. [74] is different from previous multi-kernel learning methods [78, 79]. One key difference is that the weights β_ms are not jointly optimized with other SVM parameters (such as α). Instead, Zhang et al. enforce the constraint $\sum_m \beta_m = 1$ and use a coarse-grid search through cross-validation on the training samples to select the optimal values. The obtained β_m values are used to combine kernels into a single mixed

kernel, which can be incorporated into the standard SVM to be solved using conventional SVM solvers, e.g., LIBSVM [80].

8.3.2 Semi-Supervised Learning Using Unlabeled Data

One challenge in AD patient identification is that the number of AD patients and normal controls (NCs) is generally very small, thus making it difficult to train an effective classifier. As a remedy, we note that MCI subjects, although their cognitive status is uncertain, can be helpful for improving classifier construction. To exploit the potential of using MCI subjects to aid classification between AD and NC subjects, Zhang et al. [81] treat MCI subjects as unlabeled data (i.e., not classified either as AD or NC), and then employ a semi-supervised learning technique [82, 83] to solve the classification problem. In the following, we will first introduce the semi-supervised learning technique, called Laplacian regularized least squares (LapRLS) method [84], and then derive its multimodality extension (mLapRLS).

8.3.2.1 Laplacian Regularized Least Squares

Assume we have l labeled data (from AD and NC samples), (x_i, y_i), $i = 1, \ldots, l$, and u unlabeled data (from MCI samples), (x_j, y_j), $j = l+1, \ldots, l+u$. Suppose $k(.,.)$ is a Mercer kernel function, and let H be the associated reproducing kernel Hilbert space (RKHS) and $||.||$ be the corresponding norm. The LapRLS algorithm solves the following least-squared loss function [84]:

$$min_{f \in H} \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \gamma_A \|f\|^2 + \frac{\gamma_B}{(u+l)^2} \boldsymbol{f}^T L \boldsymbol{f} \tag{8.4}$$

where $\boldsymbol{f} = [f(x1), \ldots, f(xl+u)]^T$. L is the graph Laplacian given as $L = D - W$, where W_{ij}s are the edge weights in the adjacency graph defined on both labeled and unlabeled data and the diagonal matrix $\boldsymbol{D}$ is given by $D_{ii} = \sum_j W_{ij}$. Symbols γ_A and γ_B are the two regularization parameters. Intuitively, the first two terms in Eq. (8.4) are for the supervised learning on only the labeled data (AD and NC samples), while the last term in Eq. (8.4) involves both labeled and unlabeled data (AD, NC and MCI samples) for unsupervised estimation of the intrinsic geometric structure of the whole data. According to the Representer Theorem [84], the solution to Eq. (8.4) is an expansion of kernel functions over both labeled and unlabeled data:

$$f(x) = \sum_{i=1}^{l+u} \alpha_i k(x, x_i). \tag{8.5}$$

Substituting Eqs. (8.5) into (8.4), we arrive at the dual form of Eq. (8.4) with respect to the $(l+u)$-dimensional variable vector $\alpha = [\alpha_1, \ldots, \alpha_l + u]^T$:

$$\min_{\alpha \in R^{l+u}} \frac{1}{l}(Y - JK\alpha)^T(Y - JK\alpha) + \gamma_A \alpha^T K\alpha + \frac{\gamma_B}{(u+l)^2}\alpha^T KLK\alpha, \tag{8.6}$$

where $K = \{k(x_i, x_j)\}$ is an $(l + u) \times (l + u)$ kernel matrix over all labeled and unlabeled data; $Y = [y_1, \ldots, y_l, 0, \ldots 0]$ is an $(l + u)$-dimensional label vector, and $J = diag\ (1, \ldots, 1, 0, \ldots, 0)$ is an $(l + u) \times (l + u)$ diagonal matrix with the first l diagonal entries as 1 and the rest as 0. By computing the derivative of Eq. (8.6) with respect to α as zero, we obtain the following solution:

$$\alpha = \left(JK + \gamma_A l\mathbf{I} + \frac{\gamma_B l}{(u+l)^2}\right)^{-1} Y, \tag{8.7}$$

where $\mathbf{I}$ is the identity matrix. It is worth noting that, when γ_B, Eq. (8.7) gives zero coefficients over the unlabeled data, and the coefficients over the labeled data are exactly those given by the standard regularized least squares (RLS) method, i.e., LapRLS degenerates to RLS.

8.3.2.2 Multimodality LapRLS

Now, we derive the multimodality extension of LapRLS, called mLapRLS, for classification between AD and NC. Given l labeled data (from AD and NC samples), (x_i, y_i), $i = 1,\ldots,l$, and u unlabeled data (from MCI samples), (x_j, y_j), $j = l+1,\ldots,l + u$, we assume each data x_i is composed of M modalities, i.e., $x_j = \{x_i^{(1)},\ldots, x_i^{(M)}\}$, $i = 1,\ldots,l + u$.

Define the distance function between two multimodality data x_i and x_j as

$$d(x_i, x_j) = \sum_{m=1}^{M} \beta_m d^{(m)}\left(x_i^{(m)}, x_j^{(m)}\right), \tag{8.8}$$

where $d^{(m)}(.,.)$ denotes the distance function on the mth modality, and β_ms are the nonnegative weighting parameters used to balance the contributions of different modalities. All β_m s are constrained by $\sum_m \beta_m = 1$. According to Eq. (8.8), we can compute the adjacency graph for the multimodality data, and then obtain the corresponding edge weights matrix W and graph Laplacian L on the multimodality data. Next, we can define the kernel function on two multimodality data x and x_i as

$$k(x, x_i) = \sum_{m=1}^{M} \beta_m k^{(m)}\left(x^{(m)}, x_i^{(m)}\right), \tag{8.9}$$

where $k^{(m)}$ denotes the kernel matrix over the mth modality, similar to the definition given above for the single modality case. With the definition of $k(x, x_i)$, the $(l + u) \times (l + u)$ kernel matrix K on the multimodality data can be straightforwardly obtained as $K = k(x_i, x_j)$. Once we have the graph Laplacian L, the

definition of the kernel function $k(x, x_i)$ on the multimodality data, and the kernel matrix K, the mLapRLS solution to the multimodality data can be obtained in the same way as LapRLS is obtained in Eq. (8.7). Similar to LapRLS, mLapRLS will degenerate to the corresponding multimodality RLS (mRLS) when $\gamma_B = 0$. In this case, mRLS uses only AD and NC samples for training.

8.4 AD/MCI Diagnosis and Prognosis

In this section, we will evaluate the machine learning based classification techniques that are discussed in the previous sections for AD/MCI diagnosis and prognosis using single and multiple modality data.

8.4.1 Single-Modality-based Diagnosis and Prognosis

8.4.1.1 MCI Diagnosis Using Enriched White Matter Connectivity Description

The dataset contains images of 27 participants (10 MCI patients and 17 socio-demographically matched NCs) who were recruited by the Duke-UNC Brain Imaging and Analysis Center, North Carolina, USA. Informed consent was obtained from all participants, and the experimental protocols were approved by the institutional ethics board. Confirmation of diagnosis for all subjects was made via expert consensus panels at the Joseph and Kathleen Bryan Alzheimer's Disease Research Center (Bryan ADRC) and the Department of Psychiatry at Duke University Medical Center. Diagnosis was based upon available data from a general neurological examination, neuropsychological assessment evaluation, collateral and subject symptom and functional capacity reports. Demographic information of the participants is shown in Table 8.1.

A priori knowledge of the number of features that should be used for classification is not available and this number is automatically determined as part of inner loop of the nested LOOCV. Although it generally yields slightly lower classification performance, the nested LOOCV provides a better indicator of the generalizability of a classifier. The classification accuracy by the enriched description of WM connections (with six parameters) is 88.9 %, which is at least an 14.8 % increment from that using any single physiological parameter. The area under receiver operating characteristic (ROC) curve (AUC) of the enriched description method is 0.929, indicating its excellent diagnostic power. It is found that simple connectivity description, which uses only a single diffusion parameter, is unable to provide good generalization power, as indicated by the much smaller AUC values. The classification performance of the enriched and simple connectivity descriptions is provided in Table 8.2.

Table 8.1 Demographic information of the participant involved in the study

–	MCI	NC
No. of subjects	10	17
No. of males	5	8
Age (mean ± SD)	74.2 ± 8.6	72.1 ± 8.2
Years of education (mean ± SD)	17.7 ± 4.2	16.3 ± 2.4
MMSE score (mean ± SD)	28.4 ± 1.5	29.4 ± 0.9

Table 8.2 Classification performance and AUC values for enriched and simple connectivity descriptions

Description	Accuracy (%)	AUC
Enriched	88.89	0.929
Fiber count	70.37	0.653
FA	74.07	0.859
MD	59.26	0.647
λ_1	59.26	0.629
λ_2	55.56	0.594
λ_3	59.26	0.612

A subset of most discriminant features is selected using the SVM-RFE algorithm [59] in a backward sequential way to remove one feature at a time. The selected subset is a group of features that yields the best classification performance based on the training set. Since the selected subset of features might be different for each LOO case, the most significant ROIs are determined as the regions (features) with the highest selected frequency in all LOO cases. The most discriminant regions that are selected during training stage are: (1) rectus gyrus, which is located on the orbital surface of the frontal lobe; (2) insula, which is located within lateral fissure between the temporal lobe and the frontal lobe; and (3) precuneus, which is a part of the superior parietal lobe hidden in the medial longitudinal fissure between the two cerebral hemispheres.

Note the classification framework is a data-driven approach where the assumption of the set of brain measurements that optimally differentiate MCI patients from cognitively normal individuals are not known a priori, but can only be determined from the data. The LOOCV used guards against data overfitting, a persistent problem in high dimensionality analyses of datasets with relatively small sample size.

8.4.1.2 MCI Diagnosis Using Multi-Spectral Connectivity Characterization

Thirty-seven participants (12 MCI patients and 25 socio-demographically matched NCs) were recruited by the Duke-UNC Brain Imaging and Analysis Center, North

Table 8.3 Demographic information of the participant involved in the rs-fMRI study

–	MCI	NC
No. of subjects	12	25
No. of males	6	9
Age (mean ± SD)	75.0 ± 8.0	72.9 ± 7.9
Years of education (mean ± SD)	18.0 ± 4.1	15.8 ± 2.4
MMSE score (mean ± SD)	28.5 ± 1.5[a]	29.3 ± 1.1

[a] One of the patients does not have a MMSE score

Table 8.4 Classification accuracies and AUC values of whole- and multi-spectral network characterization methods

Approach	Accuracy (%)	AUC
Unmasked + whole-spectrum	56.76	0.530
GM-masked + whole-spectrum	59.46	0.543
Unmasked + multi-spectral	67.57	0.620
GM-Masked + multi-spectral	86.49	0.863

Carolina, USA. Informed consent was obtained from all participants, and the experimental protocols were approved by the institutional ethics board. Demographic information of the participants is provided in Table 8.3.

Classification performance for the multi-spectral characterization of rs-fMRI regional mean time series was compared with the conventional whole-spectrum characterization. The nested LOOCV procedures described in the Sect. 8.2.1 were applied for performance evaluation.

The effectiveness of GM-masked and unmasked BOLD signals was evaluated in relation to the whole- and multi-spectral characterization methods. The comparison results are shown in Table 8.4.

In agreement with the hypothesis, GM-masked BOLD signal with multi-spectral characterization outperforms the unmasked and whole-spectrum characterization methods. GM-masking, when used with the conventional whole-spectral characterization, only shows slightly improvement in terms of classification accuracy and AUC value. However, when combined with the multi-spectral characterization, the classification accuracy increases by more than 18.9 % while the AUC value increases by more than 0.24, indicating significant improvement in diagnostic power. This marked improvement in performance demonstrates the effectiveness and robustness of the GM-masked multi-spectral characterization in providing relatively fine and localized analysis.

The most discriminant regions that are selected for classification are mainly located in prefrontal cortex areas and temporal lobes. The selected regions involved parts of frontal lobe such as rectus gyrus, orbitofrontal cortex and frontal gyrus, parts of temporal lobe such as temporal poles, amygdala and parahippocampal gyrus, superior occipital gyrus of occipital lobe and precuneus of parietal lobe.

Table 8.5 Demographic information of the subjects involved in the study

–	P-MCI	NC
No. of subjects	100	125
No. and percentage of males (%)	57 (57.0)	61 (48.8)
Baseline age (mean ± SD)	75.0 ± 7.1	76.1 ± 6.1
Baseline MMSE score (mean ± SD)	26.5 ± 1.7	29.1 ± 1.0

8.4.1.3 MCI Diagnosis Using Hierarchical Brain Networks

A set of 125 normal control subjects and 100 progressive MCI (P-MCI) subjects were involved in this study. This dataset was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The ADNI database contains approximately 200 cognitively normal elderly subjects to be followed for 3 years, 400 subjects with MCI to be followed for 3 years, and 200 subjects with early AD to be followed for 2 years. The P-MCI subjects refer to converters who developed probable AD after the baseline scanning. The diagnosis of AD is made according to the NINCDS/ADRDA criteria [85] for probable AD. The demographic and clinical information of all the selected subjects are summarized in Table 8.5.

The effectiveness of constructing hierarchical brain network from T1-weighted MRI for MCI prediction was evaluated by the comparison of the discrimination power of the network and the volumetric features, and the comparison of the performance of different classifiers for the network features.

Comparison of Features

The 125 subjects were randomly partitioned into 20 training and test groups, each with 150 subjects for training and 75 subjects for test. Five methods were tested in the experiment: (1) FN: the proposed method, using the four-layer hierarchical network features; (2) SN: using the network features from only the bottommost layer with the highest resolution of ROIs; (3) FN-NC: using the network features from all the four layers, but removing the edges across different layers; (4) SV: using the volumetric features from only the bottommost layer with the highest resolution of ROIs; (5) FV: using volumetric measures from all four layers.

Table 8.6 summarized the results. The classification accuracy is averaged across all the training and test groups. In order to demonstrate the advantage of the proposed network features, a paired t test is conducted between the proposed method (FN) and the other four methods, respectively. The p-value of the paired t-test is reported in Table 8.6. It can be seen that the proposed method (FN) is always statistically better (at the significance level of 0.05) than any of the other four methods.

Table 8.6 Comparison of discrimination efficacy of features

–	Mean test accuracy (%)	Paired *t*-test *p*-value
FN	85.07	–
SN	83.00	0.00272
FN-NC	83.13	0.00367
SV	81.93	0.00166
FV	81.47	0.00015

From Table 8.6, the following results are observed:

- The proposed hierarchical network features in FN outperform the conventional volumetric features in SV. The advantage may come from using both regional correlations and the hierarchical structure.
- The better performance of network features over volumetric features with the same hierarchical structure (SN vs. SV, and FN vs. FV) exhibit the benefits purely from using the regional correlations.
- The better performance of the four-layer network features FN over the single layer network features SN demonstrates the statistically significant benefit purely from the hierarchy. Moreover, the result that the full hierarchy FN statistically outperforms the hierarchy without cross-layer correlations FN-NC indicates the necessity of using the cross-layer edges in the network.

Comparison of Classifiers

The classification performance of the proposed classification scheme was compared with other six possible schemes shown in Table 8.7. To facilitate the description, the proposed scheme was denoted as P1, while the other six schemes in comparison were denoted as P2–P7. In order for comparison, each of the six schemes P2–P7 was also partitioned into four steps: rough feature selection, refined feature selection, feature embedding and classification. Note that all schemes P1–P7 employ the same rough feature selection as their first step.

The classification results are given in Fig. 8.6 and Table 8.7. In Table 8.7, the overall classification accuracy is an average accuracy over different numbers of training samples in Fig. 8.6. The results reveal that, among all the classification schemes, the proposed scheme P1 (VIP selection + PLS embedding + a linear SVM) achieves the best overall classification accuracy of 84.35 %. This is slightly better than that of P2, where a nonlinear SVM is employed. As shown in Table 8.7, the classification schemes with PLS embedding (P1–P4) outperform those without PLS embedding (P5–P7), achieving an overall accuracy above 84.0 %. In addition, the supervised embedding methods, i.e., PLS (P1–P4) and KFDA (P7), perform better than the unsupervised Laplacian Eigenmap embedding

Table 8.7 Configuration of classification schemes

Scheme	Configuration	Accuracy (%)
P1	VIP selection + PLS embedding + linear SVM	84.35
P2	VIP selection + PLS embedding + nonlinear SVM	84.03
P3	No selection + PLS embedding + linear SVM	84.11
P4	No selection + PLS embedding + nonlinear SVM	84.10
P5	SVM-RFE selection + no embedding + nonlinear SVM	80.07
P6	No selection + Laplacian Eigenmap embedding + nonlinear SVM	79.16
P7	No selection + KFDA embedding + linear SVM	81.08

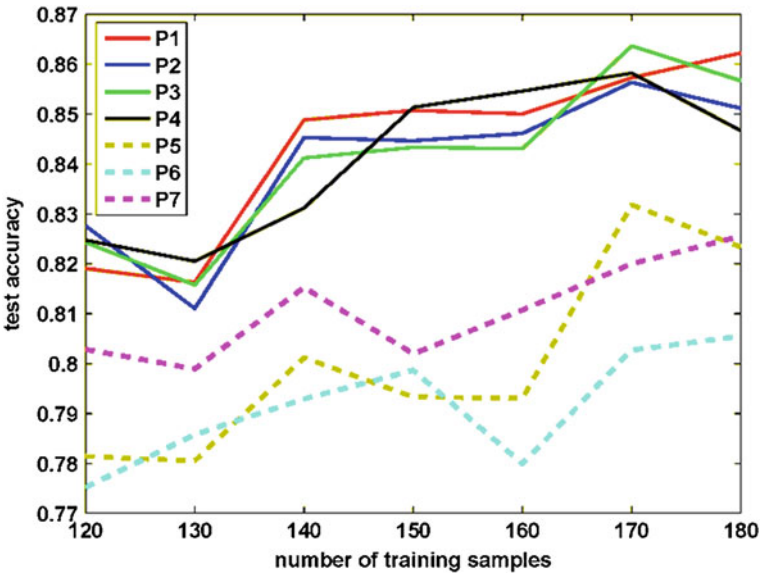


Fig. 8.6 Comparison of seven classification schemes on network features. The classification accuracy at each number of training samples is averaged over 20 randomly partitioned training and test groups. The scheme configurations are shown in Table 8.7

(P6). Moreover, although it is a linear method, PLS embedding (P1–P4) even beats the nonlinear supervised embedding of KFDA (P7).

Spatial Patterns

Some discriminative features resulting from the proposed two-step feature selection method are shown in Table 8.8. These features are consistently selected by more than half of the 20 training and test groups. Note that each network feature encodes the pairwise ROI relationship, instead of referring to only a single ROI. There are two parts in Table 8.8. On the upper portion of the table, both ROIs

Table 8.8 Selected discriminative features

Hippocampus—amygdala
Hippocampus—lingual gyrus
Hippocampus—uncus
Hippocampus—prefrontal/superolateral frontal lobe
Hippocampus—globus palladus
Hippocampus—entorhinal cortex
Hippocampus—cingulate region
Hippocampus—ventricle
Hippocampus and amygdala and fornix—ventricle
Uncus—fornix
Hippocampus—posterior limb of internal capsule
Globus palladus—anterior limb of internal capsule
Hippocampus—occipital lobe WM

associated with a network feature may be related to MCI diagnosis, such as hippocampus, entorhinal cortex, fornix, cingulate, etc., as reported in the literature [86, 87]. But the change speeds of tissue volumes are different over the two clinic groups. Take the correlation between hippocampus and ventricle as an example. It is known that the enlargement of ventricle is a biomarker for the diagnosis of the AD [88]. However, different from the hippocampus volume loss that often occurs at the very early stage of the dementia, the ventricle enlargement often appears in the middle and late stages. On the lower portion of the table, the first ROI may be affected by the disease, while the second ROI may remain constant to the disease. For example, it has been reported in a DTI study [89] that the anterior and posterior limbs of internal capsule and the occipital lobe WM may not significantly differ between MCI and NCs. Table 8.8 may suggest that, it is the different progression pattern that makes the correlation between the two regions the discriminative feature.

8.4.2 Multimodality-based Diagnosis and Prognosis

A series of experiments were performed on the multimodality data using the ADNI database. Here, ADNI subjects with all corresponding MRI, PET, and CSF data at baseline were used, leading to a total of 202 subjects, including 51 AD patients, 99 MCI patients, and 52 NCs. Table 8.9 lists the subject characteristics.

Standard image pre-processing was performed for all MRI and PET images. Specifically, anterior commissure (AC)—posterior commissure (PC) correction is first performed, followed by skull-stripping, removal of cerebellum, and segmentation of structural MR images into three different tissues: GM, WM, and CSF. Through atlas warping, we partitioned each subject image into 93 ROIs. For each ROI, we calculated the GM tissue volume from the subject's MRI image. For each

Table 8.9 Demographic information of the subjects involved in the study

–	AD (n = 51, 18F/33M)			MCI (n = 99, 32F/67 M)			NC (n = 52, 18F/34 M)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Age	75.2	7.4	59–88	75.3	7.0	55–89	75.3	5.2	62–85
Education	14.7	3.6	4–20	15.9	2.9	8–20	15.8	3.2	8–20
MMSE	23.8	2.0	20–26	27.1	1.7	24–30	29.0	1.2	25–30
CDR	0.7	0.3	0.5–1	0.5	0.0	0.5–0.5	0.0	0.0	0–0

PET image, we first rigidly aligned it with its corresponding MRI image, and then calculated the average value of PET signals in each ROI. Therefore, for each subject, we got totally 93 features from its MRI image, 93 features from its PET image, and 3 features ($A\beta_{42}$, *t*-tau and *p*-tau) from the CSF biomarkers.

8.4.2.1 AD/MCI Diagnosis Using Multi-Kernel SVM

We used standard 10-fold cross-validation to measure the classification accuracy, as well as the sensitivity and the specificity. Specifically, the whole set of subjects were equally partitioned into 10 subsets, and each time the subjects within one subset were selected as the testing set and all remaining subjects in the other 9 subsets were used for training the multiple-kernel classifier. This process was repeated for 10 independent times. The SVM classifier was implemented using LIBSVM toolbox [80], using a linear kernel and a default value for the parameter C (i.e., $C = 1$). The weights in the multiple-kernel classification method were selected from the training samples through a grid search in the range of 0–1 with a step size of 0.1. For each feature f_i in the training samples, feature normalization was performed, i.e., $f_i = (f_i - \bar{f}_i)/\sigma_i$, where $\bar{f}_i$ and σ_i are respectively the mean and standard deviation of the ith feature across all training samples. The estimated $\bar{f}_i$ and σ_i will be used to normalize the corresponding feature of each test sample.

Multimodality Classification Based on MRI, PET, and CSF

Table 8.10 shows the classification result of the multimodality classification method, compared with the methods based on each individual modality only. It's worth noting that Table 8.10 reports the averaged results of 10 experiments, with the minimal and maximal values given in brackets. As can be seen from Table 8.10, the combined use of MRI, PET, and CSF consistently achieve more accurate discrimination between AD (or MCI) patients and normal controls. Specifically, for AD versus NC classification, the multimodality classification method achieves a classification accuracy of 93.2 %, a sensitivity of 93.0 %, and a specificity of 93.3 %, while the best accuracy on individual modality is only 86.5 % (when using PET). On the other hand, for MCI versus NC classification,

Table 8.10 Comparison of performance of single-modal and multimodal classification methods

Method	AD versus NC			MCI versus NC		
	ACC (%)	SEN (%)	SPE (%)	ACC (%)	SEN (%)	SPE (%)
MRI	86.2	86.0	86.3	72.0	78.5	59.6
	(82.9–89.0)	(82.7–88.7)	(83.1–89.1)	(68.4–74.7)	(75.6–80.6)	(55.1–63.7)
CSF	82.1	81.9	82.3	71.4	78.0	58.8
	(80.o–84.9)	(80.0–84.7)	(80.0–85.1)	(68.2–73.3)	(75.6–79.4)	(54.3–61.7)
PET	86.5	86.3	86.6	71.6	78.2	59.3
	(82.9–90.5)	(82.7–90.3)	(83.1–90.6)	(67.4–74.7)	(75.0–80.6)	(52.9–63.7)
Combined	93.2	93.0	93.3	76.4	81.8	66.0
	(89.0–96.5)	(88.7–96.3)	(89.1–96.6)	(73.5–79.7)	(79.4–84.4)	(62.6–70.3)
Baseline	91.5	91.4	91.6	74.5	80.4	63.3
	(88.5–96.5)	(88.3–96.3)	(88.6–96.6)	(71.9–78.2)	(78.3–83.3)	(59.7–68.3)

The numbers in each bracket denote the minimal and maximal classification rate in 10 independent experiments

AD Alzheimer's disease, *MCI* mild cognitive impairment, *NC* normal control, *ACC* classification ACCuracy, *SEN* SENsitivity, *SPE* SPEcificity

the multimodality classification method achieves a classification accuracy of 76.4 %, a sensitivity of 81.8 %, and a specificity of 66.0 %, while the best accuracy on individual modality is only 72.0 % (when using MRI).

Furthermore, to compare with other multimodality classification methods, we also use direct feature concatenation as a baseline method for multimodality AD (or MCI) classification. Specifically, for each subject, we first concatenated 93 features from MRI, 93 features from PET, and 3 features from CSF, into a 189 dimensional vector. Remember that each feature has been normalized to have zero mean and unit standard deviation. Then, we performed SVM-based classification on all samples, with corresponding results shown in the bottom row of Table 8.10. As can be observed from Table 8.10, our kernel combination method consistently outperforms the baseline method for each performance measure.

Comparison of Different Combination Schemes

To study the effect of different combining weights, i.e., β_{MRI}, β_{CSF}, β_{PET}, on the performance of the multimodality classification method, all the possible values, ranging from 0 to 1 at a step size of 0.1, were tested under the constraint of ($\beta_{\mathrm{MRI}} + \beta_{\mathrm{CSF}} + \beta_{\mathrm{PET}} = 1$). Figures 8.7 and 8.8 show the corresponding classification results, including accuracy (top row), sensitivity (bottom left), and specificity (bottom right), with respect to different combining weights of MRI, PET, and CSF. Note that, in each subplot, only the squares in the upper triangular part have valid values due to the constraint ($\beta_{\mathrm{MRI}} + \beta_{\mathrm{CSF}} + \beta_{\mathrm{PET}} = 1$). For each plot, the three vertices of the upper triangle, i.e., the top left, top right, and bottom left squares, denote individual-modality based classification results using only PET ($\beta_{\mathrm{PET}} = 1$), CSF ($\beta_{\mathrm{CSF}} = 1$), and MRI ($\beta_{\mathrm{MRI}} = 1$), respectively.

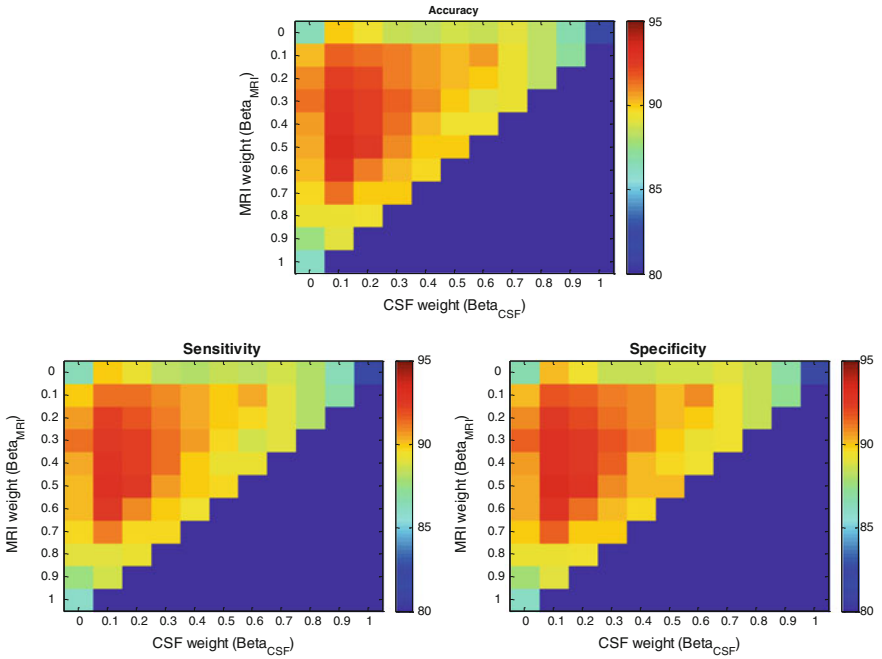


Fig. 8.7 AD Classification results with respect to different combining weights of MRI, PET and CSF. Only the *squares* in the *upper triangular* part have valid values, due to the constraint: ($\beta_{\mathrm{MRI}} + \beta_{\mathrm{CSF}} + \beta_{\mathrm{PET}} = 1$). Note that for each plot, the *top left*, *top right*, and *bottom left squares* denote the individual-modality based classification results using PET ($\beta_{\mathrm{PET}} = 1$), CSF ($\beta_{\mathrm{CSF}} = 1$), and MRI ($\beta_{\mathrm{MRI}} = 1$), respectively

As can be seen from Figs. 8.7 and 8.8, nearly all inner squares of the upper triangle have larger values (better classification) than the three vertices, demonstrating the effectiveness of the multimodality combination in AD (or MCI) classification. Furthermore, Figs. 8.7 and 8.8 also show that the squares with higher accuracy appear mainly in the inner of each triangle, instead of the boundary. This implies that each modality is indispensable for achieving good classification. Similar to what can be observed from Table 8.10, Figs. 8.7 and 8.8 also show that, for AD classification, the differences among accuracy, sensitivity, and specificity are small, while, for MCI classification, it tends to have a higher sensitivity but lower specificity.

Diagnosis Using Semi-Supervised Multimodality Classification

The mLapRLS was compared with mRLS on the multimodality (MRI, PET, and CSF) data. Specifically, a 10-fold cross-validation was performed on 51 AD patients and 52 NC subjects to get the labeled training data and testing data.

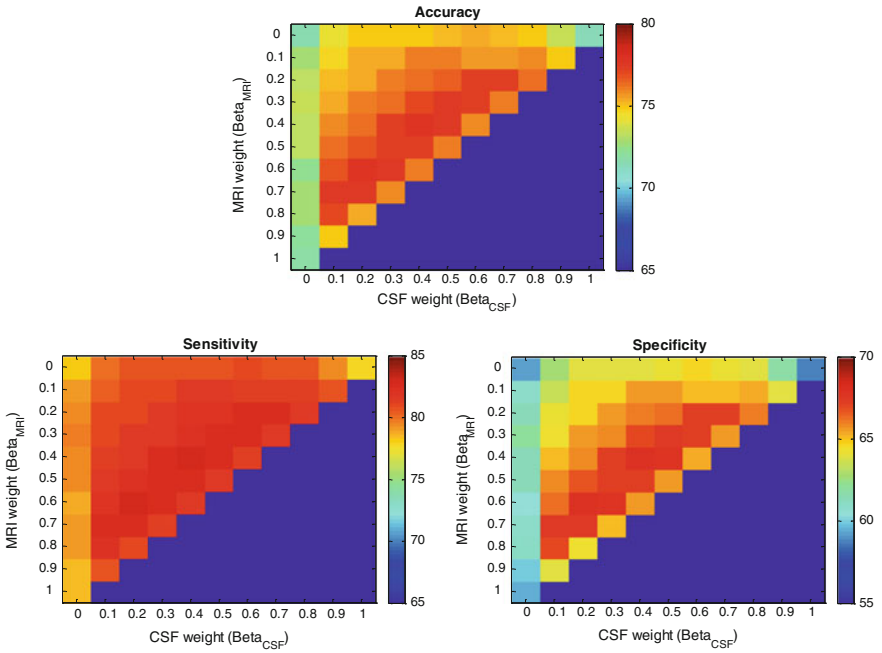


Fig. 8.8 MCI Classification results with respect to different combining weights of MRI, PET and CSF. Only the squares in the upper triangular part have valid values, due to the constraint: ($\beta_{\mathrm{MRI}} + \beta_{\mathrm{CSF}} + \beta_{\mathrm{PET}} = 1$). Note that for each plot, the *top left*, *top right*, and *bottom left squares* denote the individual-modality based classification results using PET ($\beta_{\mathrm{PET}} = 1$), CSF ($\beta_{\mathrm{CSF}} = 1$), and MRI ($\beta_{\mathrm{MRI}} = 1$), respectively

Unlabeled data were obtained from those 99 MCI subjects. A linear kernel was used for both algorithms. Following [84], for mRLS, the parameters were set as $\gamma_A = 0.05/l$ and $\gamma_B = 0$, while for mLapRLS, they were set as $\gamma_A = 0.05/l$ and $\gamma_B = 0.05(l+u)^2/l$. Here, l denotes the number of AD and NC subjects, and u is the number of MCI subjects. The Euclidean distance is used for each modality in Eq. (8.8). For both algorithms, the values of the weighting parameters β_ms were determined through cross-validation using grid search.

Figure 8.9 shows the classification results of both algorithms on the multimodality data, which include classification accuracy, sensitivity, specificity, and AUC. The results in Fig. 8.9 indicate that, by using the MCI subjects as additional unlabeled data, mLapRLS significantly improves the performances of distinguishing AD from NC subjects, compared to those by mRLS that uses only AD and NC subjects as samples for training classifier. Specifically, the AUC values of mLapRLS and mRLS are 0.985 and 0.946, respectively. The results validate the effectiveness of mLapRLS in using additional data (i.e., MCI subjects) to enhance the AD classification.

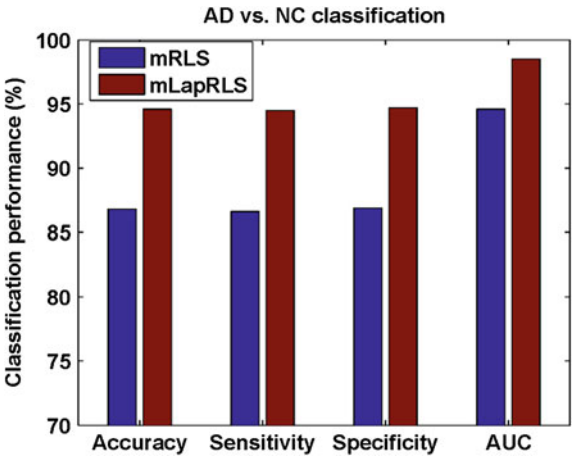


Fig. 8.9 Classification results on multimodality data

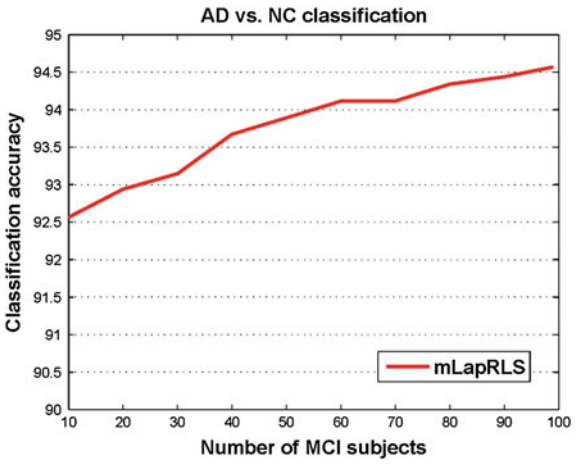


Fig. 8.10 Classification accuracy with respect to the different number of MCI subjects used to help train the multimodality classifier

Finally, in Fig. 8.10, we show the classification accuracy of the mLapRLS algorithm with respect to different number of MCI subjects used for helping training. As we can see from Fig. 8.10, as the number of included MCI subjects increases, the classification accuracy of mLapRLS also steadily increases, which again validates the usefulness of using MCI subjects for helping classification between AD and NC.

8.5 Summary

In the past two decades, machine learning techniques have been proven to be important for effective neurodegenerative disorders diagnosis and prognosis, particularly for AD and MCI. Essentially, machine learning techniques that have been applied for AD and MCI diagnosis and prognosis can be categorized into single modality and multimodality based approaches. Some recent developments in this area have been discussed in this chapter. In single modality based approaches, information on local microstructural characteristics of water diffusion (DTI), hemodynamic response related to neural activity (fMRI) and structural atrophy (T1-weighted imaging) is extracted using connectivity networks to provide a comprehensive representation of brain alterations for improved classification performance. For DTI, a collection of physiological parameters are derived along the tracked fibers for better characterization of brain circuitry. The multispectral characterization provides a localized analysis of BOLD signals by decomposing the frequency interval into several sub-bands. For T1-weighted images, hierarchical brain connectivity networks derived from the structural images provides a more effective way of characterizing subtle changes than by using local isolated measures. It is widely accepted that different modalities can convey complementary information and is useful for AD and MCI diagnosis and prognosis. Based on this observation, many machine learning techniques have been applied to integrate information from multiple modalities. Multi-kernel SVM, when used to integrate complementary information from structural MRI, PET and CSF, demonstrates significant improvements in AD and MCI diagnosis and prognosis. As a remedy to small sample size problem, a semi-supervised learning technique is introduced to derive additional information from MCI data for improving the discriminative power of the constructed classifiers. The increased accuracy, sensitivity, and specificity of these approaches indicate that machine learning techniques are a viable alternative to clinical diagnosis of brain alterations associated with cognitive impairment.

References

1. Brookmeyer R et al (2007) Forecasting the global burden of Alzheimer's disease. Alzheimer's Dementia 3(3):186–191
2. Johnson SC et al (2006) Activation of brain regions vulnerable to Alzheimer's disease: the effect of mild cognitive impairment. Neurobio Aging 27(11):1604–1612
3. Thompson PM, Apostolova LG (2007) Computational anatomical methods as applied to ageing and dementia. Br J Radiol 80:S78–S91
4. Whitwell JL et al (2007) 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 130(7):1777–1786
5. Grundman M et al (2004) Mild cognitive impairment can be distinguished from Alzheimer's disease and normal aging for clinical trials. Arch Neurol 61(1):59–66

6. Bischkopf J, Busse A, Angermeyer MC (2002) Mild cognitive impairment—a review of prevalence, incidence and outcome according to current approaches. Acta Psychiatr Scand 106:403–414
7. Jack CR Jr et al (2005) Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. Neurology 65(8):1227–1231
8. Jack CR Jr et al (2010) Hepothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 9(1):119–128
9. Nestor PJ, Scheltens P, Hodges JR (2004) Advances in the early detection of Alzheimer's disease. Nature 5:S34–S41
10. Davatzikos C et al (2010) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging 32:e19–e27
11. Davatzikos C et al (2008) Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobio Aging 29:514–523
12. Fan Y et al (2008) Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39:1731–1743
13. Vemuri P et al (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73(4):294–301
14. Vapnik VN (1999) The nature of statistical learning theory (Statistics for Engineering and Information Science). Springer, Heidelberg
15. Morra JH et al (2010) Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE Trans Med Imaging 29(1):30–43
16. Jiang J, Trundle P, Ren J (2010) Medical image analysis with artificial neural networks. Comput Med Imaging Graph 34(8):617–631
17. Fitzpatrick JM, Sonka M (2000) Handbook of medical imaging, vol 2. In: Sonka M (ed) Medical image processing and analysis. PM80SC. SPIEthe International Society for Optical Engineering
18. Bankman IN (ed) (2008) Handbook of medical image processing and analysis. Academic Press, New York
19. Kloppel S et al (2008) Automatic classification of MR scans in Alzheimer's disease. Brain 131(3):681–689
20. Fan Y et al (2008) Unaffected family members and Schizophrenia patients share brain structure patterns: A high-dimensional pattern classification study. Biol Psychiatry 63(1):118–124
21. Fan Y et al (2007) COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans Med Imaging 26(1):93–105
22. Davatzikos C et al (2008) Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol Aging 29:514–523
23. Vemuri P et al (2008) Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 39(3):1186–1197
24. Duchesne S et al (2005) Predicting clinical variable from MRI features: application to MMSE in MCI. Med Image Comput Comput Assist Interv 8(1):392–399
25. Chu C et al (2007) Regression analysis for clinical scores of Alzheimer's disease using multivariate machine learning method. In: Human Brain Mapping, Chicago
26. Fan Y, Kaufer D, Shen D (2009) Estimating clinical variables from brain images using Bayesian regression. Alzheimer's Dimentia 5(4):372
27. Westman E et al (2010) Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. Neuroimage 54(2):1178–1187
28. Lao Z et al (2004) Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 21:46–57
29. Chetelat G, Baron JC (2003) Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage 18(2):525–541

30. Jack CR Jr et al (1998) Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. Neurology 51(4):993–999
31. Thompson PM et al (2004) Mapping cortical change in Alzheimer's disease, brain development, and Schizophrenia. J Neurosci 23:S2–S18
32. Dickerson BC et al (2009) The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb Cortex 19(3):497–510
33. Thompson PM et al (2001) Cortical changes in Alzheimer's disease detected with a disease-specific population-based brain atlas. Cereb Cortex 11(1):1–16
34. Chupin M et al (2009) Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19(6):579–587
35. Colliot O et al (2008) Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248(1):194–201
36. Gong G et al (2009) Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography. Cereb Cortex 19:524–536
37. Rose SE, Janke AL, Chalk JB (2007) Gray and white matter changes in Alzheimer's disease: a diffusion tensor imaging study. J Magn Reson Imaging 27(1):20–26
38. Zhang Y et al (2007) Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and Alzheimer disease. Neurology 68(1):13–19
39. Friston KJ et al (1993) Functional connectivity: the principal-component analysis of large (PET) data sets. J Cereb Blood Flow Metab 13:5–14
40. Greicius M (2008) Resting-state functional connectivity in neuropsychiatric disorders. Curr Opin Neurol 21:424–430
41. Biswal B et al (1995) Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34(4):537–541
42. Sorg C et al (2007) Selective changes of resting-state networks in individuals at risk for Alzheimer's disease. PNAS 104(47):18760–18765
43. Greicius MD et al (2004) Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. PNAS 101(13):4637–4642
44. Diehl J et al (2004) Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. A PET study. Neurobio Aging 25(8):1051–1056
45. Fjell AM et al (2010) CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. J Neurosci 30(6):2088–2101
46. Landau SM et al (2010) Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75(3):230–238
47. Walhovd KB et al (2010) Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobio Aging 31(7):1107–1121
48. Geroldi C et al (2006) Medial temporal atrophy but not memory deficit predicts progression to dementia in patients with mild cognitive impairment. J Neurol Neurosurg Psychiatry 77:1219–1222
49. Hinrichs C et al (2009) Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. Neuroimage 48(1):138–149
50. Ye J et al (2008) Heterogeneous data fusion for Alzheimer's disease study. In: Paper presented at the proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, 2008
51. Fellgiebel A et al (2007) FDG-PET and CSF phospho-tau for prediction of cognitive decline in mild cognitive impairment. Psychiatry Res Neuroimag 155(2):167–171
52. Chetelat G et al (2005) FDG-PET measurement is more accurate than neuropsychological assessments to predict global cognitive deterioration in patients with mild cognitive impairment. Neurocase 11(1):14–25

53. Walhovd KB et al (2010) Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. Am J Neuroradiol 31(2):347–354
54. Hagmann P et al (2008) Mapping the structural core of human cerebral cortex. PLoS Comput Biol 6:e159
55. Sporns O, Zwi JD (2004) The small world of the cerebral cortex. Neuroinformatics 2:145–161
56. Wee CY et al (2011) Enriched white matter connectivity networks for accurate identification of MCI patients. Neuroimage 54(3):1812–1822
57. Rubinov M, Sporns O (2010) Complex networks measures of brain connectivity: uses and interpretations. Neuroimage 52(3):1059–1069
58. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
59. Guyon I et al (2004) Gene selection for cancer classification using support vector machines. Machine Learning 46(1–3):389–422
60. Rakotomamonjy A (2003) Variable selection using SVM based criteria. J Mach Learn Res: Special issue on special feature 3:1357–1370
61. Wee CY et al (2011) Classification of MCI patients via functional connectivity networks. In: ISMRM' 2011 Québec, Canada
62. Bassett DS, Bullmore E (2006) Small-world brain networks. The Neuroscientist 12(6):512–523
63. Courchesne E et al (2000) Normal brain development and aging: quantitative analysis at in vivo MR imaging in healthy volunteers. Radiology 216:672–682
64. Karas GB et al (2003) A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 18(4):895–907
65. Thompson PM et al (2003) Dynamics of gray matter loss in Alzheimer's disease. J Neuroscience 23(3):994–1005
66. Whitwell JL et al (2008) MRI patterns of atrophy associated with progression to AD in amnestic mild cognitive impairment. Neurology 70(7):512–520
67. Van Dijk KRA et al (2010) Intrinsic functional connectivity as a tool for human connectomics: theory, properties and optimization. J Neurophysiol 103:297–321
68. Cordes D et al (2001) Frequencies contributing to functional connectivity in the cerebral cortex in "resting-state" data. Am J Neuroradiol 22:1326–1333
69. Achard S et al (2008) Fractal connectivity of long-memory networks. Phys Rev E Stat Nonlin Soft Matter Phys 77(3 Pt 2):036104
70. Zhou L et al (2011) Hierarchical anatomical brain networks for MCI prediction by partial least square analysis. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR)
71. Rosipal R, Kramer N (2006) Overview and recent advances in partial least squares. Lect Notes Comput Sci 3940:34–51
72. Wold S et al (1993) PLS—partial least-squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design: theory methods and applications, vol 1. ESCOM, Leiden, pp 523–550
73. Hinrichs C et al (2009) MKL for robust multi-modality AD classification. Med Image Comput Comput Assist Interv Part II:786–794
74. Zhang D et al (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55(3):856–867
75. Scholkopf B, Smola AJ (2002) Learning with Kernels. MIT Press, Massachusetts
76. Lanckriet GRG et al (2004) Learning the Kernel matrix with semidefinite programming. J Mach Learn Res 5:27–72
77. Bach FR, Lanckriet GRG, Jordan MI (2004) Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the twenty-first international conference on Machine learning (ICLM'04), p 6

78. Wang Z, Chen S, Sun T (2008) MultiK-MHKS: a novel multiple kernel learning algorithm. IEEE Trans Pattern Analysis Mach Intell 30(2):348–353
79. Lanckriet GR et al (2004) Kernel-based data fusion and its application to protein function prediction in yeast. Pac Symp Biocomput, In, pp 300–311
80. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines
81. Zhang D, Shen D (2011) Semi-supervised multimodal classification of Alzheimer's disease. In: IEEE international symposium on biomedical imaging (ISBI'11)
82. Tiwari P et al (2010) Semi supervised multi kernel (SeSMiK) graph embedding: identifying aggressive prostate cancer via magnetic resonance imaging and spectroscopy. Med Image Comput Comput Assist Interv 2010:666–673
83. Chapelle O, Scholkopf B, Zien A (eds) (2006) Semi-supervised learning. MIT Press, Cambridge
84. Belkin M et al (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7:2399–2434
85. McKhann G et al (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34(7):939–944
86. Cuingnet R et al (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage 56(2):766–781
87. Pengas G et al (2010) Focal posterior cingulate atrophy in incipient Alzheimer's disease. Neurobio Aging 31(1):25–33
88. Nestor SM et al (2008) Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. Brain 131(9):2443–2454
89. Bozzali M et al (2002) White matter damage in Alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging. J Neurol Neurosurg Psychiatry 72(6):742–746

Chapter 9
Using Machine Learning to Plan Rehabilitation for Home Care Clients: Beyond "Black-Box" Predictions

Mu Zhu, Lu Cheng, Joshua J. Armstrong, Jeff W. Poss, John P. Hirdes and Paul Stolee

Abstract Resistance to adopting machine-learning algorithms in clinical practice may be due to a perception that these are "black-box" techniques and incompatible with decision-making based on evidence and clinical experience. We believe this resistance is unfortunate, given the increasing availability of large databases containing assessment information that could benefit from machine-learning and data-mining techniques, thereby providing a new and important source of evidence upon which to base clinical decisions. We have focused our investigation on the clinical applications of machine-learning algorithms on older persons in a home care rehabilitation setting. Data for this research were obtained from standardized client assessments using the comprehensive RAI-Home Care (RAI-HC) assessment instrument. Our work has shown that machine-learning algorithms can produce better decisions than standard clinical protocols. More importantly, we have shown that machine-learning algorithms can do much more than make "black-box" predictions; they can generate important new clinical and scientific insights. These insights can be used to make better decisions about treatment plans for patients and about resource allocation for healthcare services, resulting in better outcomes for patients, and in a more efficient and effective healthcare system.

M. Zhu (✉) · L. Cheng
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
e-mail: m3zhu@uwaterloo.ca

J. J. Armstrong · J. W. Poss · J. P. Hirdes · P. Stolee
School of Public Health and Health Systems, University of Waterloo, Waterloo, ON N2L 3G1, Canada

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_9,

9.1 Introduction

Machine-learning algorithms (see, e.g., [5, 26]) have been used extensively in biomedical applications, such as in predicting the role of genes and proteins [37]. Their use in clinical decision-making has been comparatively limited, although some applications have been reported, for example, in predicting coronary syndromes [25], in assessing the severity of pancreatitis [52], or in diagnosing breast cancer [56] or melanoma [23]. Some of the resistance to adopting machine-learning algorithms in clinical practice may be due to a perception that such methods are "black-box" techniques [54] that are incompatible with decision-making based on explicit evidence-based care pathways combined with the clinician's own experience and insights. Although understandable, we believe this resistance is unfortunate, given the increasing availability of large databases containing assessment information that could benefit from machine-learning and data-mining techniques, thereby creating a new and important source of evidence upon which to base clinical decisions. As we will illustrate in this chapter, innovative applications of these algorithms can go beyond "black-box" predictions to yield valuable clinical and scientific insights.

Because few studies have investigated the use of machine-learning methods on rehabilitation for the elderly, this is where we have focused our investigation. Several groups have used machine-learning approaches to classify walking conditions [33, 35] or movement patterns [4]. Preliminary applications for predicting rehabilitation outcomes have produced mixed results [51, 61]. Through the work described in this chapter, we found that machine-learning algorithms generated more accurate predictions than established clinical protocols [73, 74].

Our work has been conducted as a component of "InfoRehab" (www.inforehab.uwaterloo.ca), a multidisciplinary research program funded by the Canadian Institutes of Health Research to enhance the rehabilitation of the elderly through more effective use of health information. One of our research objectives is to understand whether improved clinical decision-making and improved client outcomes can be achieved through more sophisticated use of routinely collected health assessment information. While rehabilitation can improve the functional independence and quality of life of older persons, and thus save the healthcare system money, resources for rehabilitation services (primarily physical and occupational therapy) are limited, and many elderly patients who could benefit from rehabilitation do not receive any therapy [31]. Thus, it is critically important that limited rehabilitation resources be targeted to those persons most likely to benefit. Recent reviews and consensus processes have identified a major research priority to improve methods for identifying the patients most likely to benefit from rehabilitation [7, 60, 65].

The elderly may benefit from rehabilitation for a variety of reasons; common reasons include musculoskeletal disorders (e.g., hip fracture and osteoarthritis), stroke, or deconditioning resulting from prolonged hospital stays. The frailty, clinical heterogeneity, medical complexity, and multiple comorbidities which are

common in older patients present significant challenges to clinical decision-making for rehabilitation in this population [65]. Cognitive impairment is an example of these challenges. Cognitive function has been identified as an important factor in predicting the success of rehabilitation for older patients [28] and is often used as a key criterion in assessing rehabilitation potential [24], the rationale being that adequate cognition is necessary to follow instructions for therapy and exercise programs. On the other hand, clinicians are often able to identify patients with lower cognitive function who would be suitable candidates for rehabilitation [7].

Ghisla and colleagues [21] found that, while good cognition was associated with functional gains in older geriatric rehabilitation patients, patients with poor cognition could also improve in physical function as a result of rehabilitation. Colombo and colleagues [11] found that mental status score was not correlated to functional improvement in a geriatric rehabilitation ward. A randomized controlled trial of an interdisciplinary care program with better access to rehabilitation therapy for hip fracture patients [50] produced the surprising finding that, while the intervention produced no overall benefit in patient outcomes, there was a trend toward improvement in patients with cognitive impairment. Using cognitive impairment as a selection criterion for rehabilitation is thus problematic, and it is likely that the potential for successful rehabilitation in persons with cognitive impairment is related to a varying combination of multiple factors such as mood, premorbid and baseline physical function, motivation, comorbidities, presence of a caregiver, and other client characteristics [7, 66].

InfoRehab has focused on rehabilitation in the home care setting. While the growing importance of home care and other community-based healthcare services is widely recognized, funding for home care services is still lower than funding for hospitalization and other institutional services. In Ontario, where our research is currently being conducted, reports suggest that already limited resources for home care rehabilitation are being further constrained [27], adding urgency to research that can support effective planning and allocation of rehabilitation services.

In Ontario, home care services are coordinated by Community Care Access Centres (CCACs). CCAC case managers assess all long-stay home care clients with the RAI-HC (also referred to as the MDS-HC, [36, 48]), a comprehensive assessment system developed by an international research consortium (interRAI; www.interrai.org), and one of a suite of assessment tools developed for use in care planning, outcome measurement, quality improvement, and resource allocation [29]. The RAI-HC, which contains more than 300 items measuring a wide range of client characteristics, including functional status, diagnoses, cognition, communication, mood and behavior, informal supports, and other information, is currently in use in many jurisdictions around the world. The assessment items can be used to derive specific measures for health issues such as cognition, depression or ability to perform activities of daily living (ADLs). They also form the basis of Clinical Assessment Protocols (CAPs), which can be used to guide care planning and decision-making. Clients are assessed on admission and at follow-up intervals of approximately 6 months.

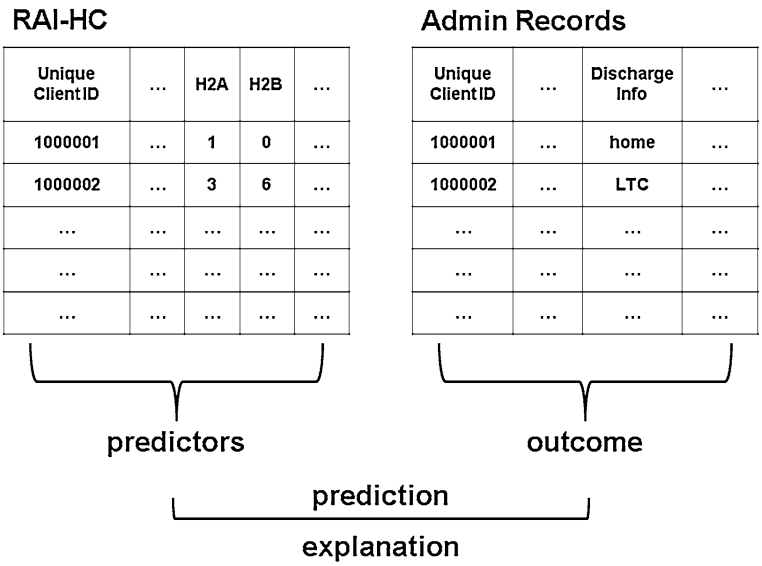


Fig. 9.1 Schematic illustration of data set structure: RAI-HC assessment items are linked to health outcome and/or service utilization data to facilitate our analysis

The current available database of RAI-HC assessment information in Ontario provides a rich resource for machine-learning and data-mining analyses. Recently, these assessment data have been linked with administrative data using detailed information on service utilization, as well as data on mortality and discharge disposition (e.g., to hospital or to a long-term care home).

In this chapter, we will describe several examples of our investigation to apply machine-learning techniques in clinical decision-making for both *predictive* and *explanatory* tasks (see, e.g., [58]). For both types of tasks, we require data on health outcomes. To do so, we combined the RAI-HC with client discharge and/or service utilization information from the CCAC administrative records, often within 6 months or 1 year of the initial assessment (See Fig. 9.1).

In the first example, we illustrate the application of machine learning for predictive purposes, and address an application relevant to the research priority described earlier—how to identify elderly patients most likely to benefit from rehabilitation. Whereas predictive tasks are the forte of machine-learning techniques, clinicians are often concerned with explanatory tasks. In the second example, we show how machine learning can be used to identify key variables (client characteristics) that best explain who receives rehabilitation services. Given resource limitations and the evidence that many home care clients who could benefit do not receive rehabilitation [31], it is instructive to explore what client characteristics may be guiding current clinical practices and decision-making around rehabilitation services. The third example describes the application of an off-the-shelf machine-learning algorithm for both prediction and explanation. We

try to predict institutionalization in a long-term care (LTC) home, as well as to identify key risk factors for LTC placement.

Other than illustrating the use of a number of popular machine-learning algorithms and addressing a number of pressing issues in the rehabilitation of older persons, these examples also convey a number of general messages. In particular, we will see that data-driven algorithms can often make better predictions than expert-driven protocols, that complex algorithms are not necessarily superior to simple ones, and that what appears to be a "black-box" algorithm can sometimes contain useful scientific insights and other times be used to extract scientific insights directly.

9.2 Example 1: Rehabilitation Potential (A Predictive Task)

Clinical Assessment Protocols (CAPs) are triggered when specified combinations of RAI-HC assessment items suggest that specific problems or risks are present and warrant further investigation [14, 36, 48]. The tasks they perform, i.e., predicting whether certain risks exist, are well-suited for machine-learning algorithms. Here, we illustrate the use of two machine-learning algorithms to predict whether a client has rehabilitation potential. Accurate predictions of clients' rehabilitation potential constitute one step toward solving the aforementioned problem that many who could benefit from rehabilitation currently do not receive it [31].

9.2.1 Focal Point: A Tale of Two Algorithms Versus an Existing Protocol

The CAP most relevant to rehabilitation planning and the assessment of rehabilitation potential is the ADLCAP, where "ADL" stands for "Activities of Daily Living." In this section, we compare the ADLCAP with a simple machine-learning algorithm known as the K-nearest neighbors (KNN), and with a mathematically sophisticated, modern algorithm known as the support vector machine (SVM).

9.2.1.1 ADLCAP

The ADLCAP is derived using a number of nested if–then statements that use different combinations of 19 variables (Table 9.1) from the RAI-HC instrument as conditions [49]. The ADLCAP is triggered if the client is unable to perform two or more of the "activities of daily living" items (H2A to H2J in Table 9.1), if the

Table 9.1 The 19 RAI-HC items used by the ADLCAP to assess a client's rehabilitation potential: we recoded the values of these variables so that the machine-learning algorithms would effectively interpret these predictors in exactly the same way as the ADLCAP

Predictor items from RAI-HC and descriptions		Orig. value	Recoded value	% = 1
H2A	Mobility in bed	0, 1, …, 8	0, 1 → 0; else 1	9.5
H2B	Transfer	0, 1, … , 8	0, 1 → 0; else 1	18.0
H2C	Locomotion in home	0, 1, …, 8	0, 1 → 0; else 1	14.8
H2D	Locomotion outside of home	0, 1, … , 8	0, 1 → 0; else 1	38.2
H2E	Dressing upper body	0, 1, … , 8	0, 1 → 0; else 1	32.0
H2F	Dressing lower body	0, 1, … , 8	0, 1 → 0; else 1	37.8
H2G	Eating	0, 1, … , 8	0, 1 → 0; else 1	10.4
H2H	Toilet use	0, 1, … , 8	0, 1 → 0; else 1	19.8
H2I	Personal hygiene	0, 1, … , 8	0, 1 → 0; else 1	25.6
H2J	Bathing	0, 1, … , 8	0, 1 → 0; else 1	77.9
C3	Ability to understand others	0, 1, … , 4	0, 1, 2 → 0; else 1	4.7
P6	Overall change in care need	0, 1, 2	0, 1 → 0; else 1	34.8
H3	ADL decline	0, 1	0, 1	39.8
K8B	Unstable conditions	0, 1	0, 1	29.1
K8C	Flare-up of recurrent/chronic problem	0, 1	0, 1	7.8
K8D	Treatment changed in last 30 days	0, 1	0, 1	16.6
H7A	Client optimistic about improvement	0, 1	0, 1	22.6
H7B	Caregivers optimistic about improvement	0, 1	0, 1	11.4
H7C	Good prospect of recovery	0, 1	0, 1	10.7

Source Zhu et al. [74]

client is able to understand others (C3), and if any one of the conditions described by the other covariates is present (P6 to H7C in Table 9.1). Although there is no explicit weighting of the items, the protocol implies a particular importance for a cognition item, the ability to understand others (C3). Considered necessary for the success of a rehabilitative program, it is the single item that must always be present for the ADLCAP to be triggered.

9.2.1.2 K-Nearest Neighbors

The K-nearest neighbors (KNN) algorithm [12] is a classic and extremely easy-to-describe technique. Suppose we have a database consisting of a total of n observations, (x_i, y_i), for $i = 1, 2, \ldots, n$, where x_i is a vector of covariates and y_i is a binary outcome of either zero or one. The database is called the training set. Given any two observations, x_i and x_j, let $d(x_i, x_j)$ be a distance metric. To predict the response for a new observation x_{new} with the KNN algorithm, we first identify a set, $N(x_{new}, K)$, consisting of K observations from the training set that are closest to x_{new} in terms of the metric $d(\cdot, \cdot)$. We then estimate the probability that $y_{new} = 1$ by

$$\widehat{P}(y_{new=1}|x_{new}) = \frac{1}{|N(x_{new}, K)|} \sum_{x_i \in N(x_{new}, K)} y_i,$$

where the notation $|A|$ means the size of the set, A. That is, we compute the proportion of neighbors that have binary outcomes equal to one, and predict y_{new} to be one if this proportion exceeds a certain threshold, say, the fraction of observations with $y_i = 1$ in the training set. The reason why $|N(x_{new}, K)|$ is not always equal to K is because of the possibility of ties, meaning that two or more observations could sometimes have the same distance to x_{new}.

9.2.1.3 Support Vector Machine

The support vector machine (SVM) is a relatively new algorithm that has received a tremendous amount of attention in the machine-learning community as of late (e.g., [6, 13, 64]). To predict the outcome for x_{new}, the SVM uses quadratic programming (e.g., [22]) to construct a model of the following form:

$$f(x_{new}) = w_0 + \sum_{x_i \in SV} w_i K(x_{new}; x_i),$$

where w_0 and w_i are model coefficients, and $K(u; v)$ is a kernel function specified by the user. For our experiments, we used the radial basis kernel function. A key feature of the SVM is that, once the parameters w_0 and w_i are computed, the final model only depends on a subset of the training data, denoted in the equation above by "SV." These observations are called "support vectors" and are determined by the SVM algorithm automatically. For the SVM, binary outcomes are coded as -1 and +1 rather than zero and one, and one predicts y_{new} to be +1 if $f(x_{new}) > 0$ and -1 if $f(x_{new}) < 0$.

9.2.2 Data and Various Details of the Analysis

For this example, our data consisted of initial RAI-HC assessments of $N = 24{,}724$ clients from eight Ontario CCACs. The outcome variable was whether a client had true rehabilitation potential.

We defined a client as having true potential $(y = 1)$ if (1) there was an improvement in the client's ADL function, assessed using the interRAI ADL long form measure [47], over a follow-up period of approximately 1 year, or if (2) the client remained at home at the end of the treatment program. The rationale for this definition was that, for frail elderly clients for whom the likely course was functional decline, any improvement in ADL function was important. Also, persons discharged from home care who remained in their own homes (i.e., were not admitted to a long-term care home) could also be considered to have had a successful outcome. Other disposition outcomes included discharge to a nursing home, or death, which could be considered indications of rehabilitation failure. In this dataset, 6,567 clients were so defined as having true rehabilitation potential.

In order to make conservative and fair comparisons with the ADLCAP, we allowed KNN and SVM to use only the 19 predictors considered by the ADLCAP. Moreover, we recoded these 19 variables so that both KNN and SVM would effectively interpret these predictors in exactly the same way as the ADLCAP did.

For example, the ADLCAP treats the predictor H2A (mobility in bed) in the following way:

if $H2A = 2, 3, 4, 5, 6,$ *or* 8 (indicating levels of dependence),

then consider as dependent;

else (meaning $H2A = 0$ *or* 1, indicating independence) consider as independent.

Suppose that client A had $H2A = 2$ and client B had $H2A = 6$. The ADLCAP would not distinguish these two clients with regard to this variable. Therefore, the variable $H2A$ could be recoded as a binary variable: (2, 3, 4, 5, 6, and 8) as one, and all other values (0 and 1) as zero. Table 9.1 summarizes how the 19 variables were recoded according to the ADLCAP.

We used KNN and SVM to make predictions on the eight regional CCAC data sets one-by-one. When making predictions for cases from a particular region, we used a random sample of 2,500 clients from the *other seven regions* as the training set. This methodology precluded each algorithm from using one's own data to predict one's own outcome (and thereby creating a bias toward better prediction). Furthermore, tuning parameters of the algorithms were selected by cross-validation on the training set alone, using the overall error rate as the guiding criterion. Cross validation is a standard procedure in machine learning to determine the value of various tuning parameters in any given algorithm (see, e.g., [26]). While the role played by these tuning parameters is extremely important, we will not describe cross validation in this chapter.

9.2.3 Results

Table 9.2 compares how the machine-learning algorithms performed against the ADLCAP. For binary predictions, the false positive (FP) and false negative (FN) rates are intuitive performance measures corresponding, respectively, to the probabilities of the two types of errors one can make, namely, predicting a true zero to be a one (FP) and predicting a true one to be a zero (FN). Clearly, the smaller these values are, the better the results will be. The positive and negative diagnostic likelihood ratios (DLR + and DLR–, respectively) are somewhat less intuitive. More details about these readings are given in the Appendix at the end of the chapter. In short, an informative prediction method should have $DLR+ > 1$ and $DLR- < 1$. Moreover, given two prediction methods, A and B, A can be said to be more informative than B if $DLR+(A) > DLR+(B)$ and if $DLR-(A) < DLR-(B)$. Table 9.2 clearly shows that both KNN and SVM are significantly better than the ADLCAP in predicting a client's rehabilitation potential. Data-driven algorithms can often make better predictions than expert-driven protocols.

Table 9.2 Predicting rehabilitation potential: performance evaluation of ADLCAP (CAP), K-nearest neighbors (KNN) and support vector machine (SVM)

	Overall error			False+ (FP)			False– (FN)			DLR+			DLR–		
Region	CAP	KNN	SVM	CAP	KNN	SVM	CAP	KNN	SVM	CAP	KNN	SVM	CAP	KNN	SVM
1	0.37	0.34	0.35	0.30	0.34	0.35	0.65	0.36	0.35	1.18	1.88	1.86	0.92	0.55	0.54
2	0.37	0.32	0.29	0.31	0.31	0.25	0.62	0.38	0.43	1.24	2.01	2.22	0.89	0.55	0.58
3	0.38	0.31	0.32	0.32	0.27	0.29	0.63	0.50	0.46	1.14	1.84	1.88	0.93	0.68	0.64
4	0.46	0.32	0.31	0.36	0.30	0.28	0.65	0.35	0.36	0.99	2.15	2.25	1.00	0.50	0.51
5	0.31	0.23	0.24	0.27	0.18	0.22	0.67	0.53	0.44	1.25	2.57	2.58	0.91	0.65	0.56
6	0.43	0.28	0.29	0.38	0.24	0.27	0.62	0.41	0.38	1.01	2.40	2.33	1.00	0.55	0.52
7	0.48	0.30	0.32	0.43	0.28	0.31	0.59	0.37	0.34	0.95	2.29	2.14	1.04	0.51	0.50
8	0.42	0.31	0.33	0.37	0.28	0.32	0.62	0.42	0.37	1.03	2.08	1.96	0.98	0.58	0.54
Mean	0.40	0.30	0.31	0.34	0.28	0.29	0.63	0.42	0.39	1.10	2.15	2.15	0.96	0.57	0.55

Source Zhu et al. [74]

However, Table 9.2 also shows that there is no substantive difference between KNN and SVM. If the SVM is mathematically a much more sophisticated algorithm than the KNN, then why did it not produce much better predictions? One must bear in mind that the amount of information contained in the data is not affected by the algorithm used to extract it. It is true that some algorithms have natural limitations. For example, one cannot use a linear algorithm to estimate a nonlinear decision boundary (unless one employs a trick to apply the linear algorithm in a different input space altogether). But if a simple algorithm has already extracted the right amount of information, then one cannot expect a complex algorithm to magically find more information in the same data. Complex algorithms are not necessarily superior to simple ones.

9.2.4 Discussion

The interRAI consortium has recently undertaken a review and revision of the CAPs, including the ADLCAP. This work was informed in part by the results found in our analyses. In this regard, machine learning has had some impact on the planning of rehabilitation services in practice. In particular, machine-learning algorithms can "set the bar" for clinical predictions, and be used to refine clinical protocols in order to achieve improved performance.

While both KNN and SVM produced superior results, these algorithms have not replaced the ADLCAP as the screening tool used in practice. This may relate to the perception, discussed earlier, that these prediction procedures are "black-boxes." Even though the predictions could be validated empirically to be more accurate, it would be hard for clinicians to understand why a particular prediction was made for any given case. We address this challenge next.

First, it is possible to explain the intuition of the KNN algorithm with a clinical analogy. In particular, one can argue that physicians also rely on an implicit KNN algorithm to make clinical decisions. A physician's clinical decision is undoubtedly influenced by his or her past clinical experiences. For example, a physician will likely recommend a particular treatment program to a new patient if the new patient's clinical profile matches those patients who have been successfully treated by the physician in the past with the same program. Hence, a physician's past patients can be regarded as the training set. Matching the clinical profile of a new patient to those of past patients is similar to finding a number of nearest neighbors from the training set. In this sense, we can think of the KNN algorithm as an artificial "super expert" who has had the "experience" of "treating" virtually every patient recorded in the database and can, therefore, use this extensive "clinical experience" to make informed and intelligent decisions.

Next, instead of using the SVM to just make predictions, it is also possible to derive useful scientific insights from its output. In SVM, observations chosen as support vectors are either very close to or on the wrong side of the decision boundary; non-support vectors, on the other hand, are on the correct side of the

Table 9.3 Fraction of clients with covariate = 1: Differences between those who most clearly have and those who most clearly do not have rehabilitation potential, according to SVM

Covariate = 1	Clearly have potential	Clearly no potential	Absolute difference	
H2A	0.01	0.10	0.09	
H2B	0.08	0.16	0.08	
H2C	0.04	0.15	0.11	
H2D	0.22	0.37	0.15	
H2E	0.11	0.32	0.21	
H2F	0.17	0.37	0.20	
H2G	0.01	0.12	0.12	
H2H	0.03	0.24	0.20	
H2I	0.07	0.27	0.20	
H2J	0.29	1.00	0.71	*
C3	1.00	0.94	0.06	
P6	0.49	0.11	0.37	
H3	0.49	0.12	0.37	
K8B	0.17	0.29	0.12	
K8C	0.05	0.05	0.00	
K8D	0.32	0.02	0.30	
H7A	0.65	0.00	0.65	*
H7B	0.35	0.00	0.35	
H7C	0.46	0.00	0.46	*

Source Zhu et al. [74]

boundary and at least a certain distance away from it (see, e.g., [13, 26]). In other words, non-support vectors are the easy-to-classify observations in the dataset. In our context, these observations would be the clients that, according to SVM, either clearly had or clearly did not have any rehabilitation potential. A careful examination of these two groups of clients, therefore, could yield additional insights.

Notice that this is *different* from simply comparing clients with $y = +1$ and those with $y = -1$. Observations near the decision boundary are often noisier than the rest; thus, excluding them can often make it easier for us to understand the exact nature of the class separation.

To illustrate this approach, we built an SVM with a random sample of 10,000 observations from all eight CCAC datasets and examined the resulting two groups of non-support vectors. Table 9.3 shows these observations. In the table, each row shows the fraction of observations with the corresponding covariate equal to 1 in each of these two groups. Recall from Table 9.1 that all covariates had been recoded as binary in our study. It is evident from Table 9.3 that these two groups of clients are most different in terms of H2J (bathing), H7A (client optimistic about functional improvement), and H7C (client rated as having good prospects of recovery), suggesting that these three variables were the most important ones for predicting rehabilitation potential. This indicates that the SVM, despite appearing to be a "black-box" prediction algorithm, can be used nonetheless to derive useful clinical and scientific insights.

9.3 Example 2: Receipt of Rehabilitation Services (An Explanatory Task)

As we have indicated, clinicians may be uncomfortable with the "black-box" type of algorithms despite their superior predictive performance. Here, we illustrate the use of machine-learning techniques to extract scientific insights directly, rather than to make predictions. Our goal is to identify the most important factors for determining whether a client will receive rehabilitation services. As described earlier, this analysis provides important insights into the factors currently guiding clinical decision-making.

9.3.1 Focal Point: Variable Screening by an Ensemble Application of the LASSO

The study objective for this example, identifying the most important items from the RAI-HC for predicting the binary outcome of receiving rehabilitation services, constitutes a classic "variable selection" problem. In addition to more conventional statistical approaches [42], there are many machine-learning algorithms for performing this task, such as stepwise selection (e.g., [43]), SCAD [17], LARS [16], the elastic net [69], PGA [72], and VISA [55], among many others. For this example, we focus on an algorithm known as the LASSO [63].

9.3.1.1 Variable Selection by the LASSO

First proposed by [63], the LASSO algorithm has become the most studied variable-selection tool by statisticians over the last decade. Many variations (e.g., [40, 69]) now exist. We use the following notations to denote the data and the model parameters:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_d \end{bmatrix}.$$

In our case, y_i is, again, a binary outcome of either zero or one. As usual, x_{ij} is the j-th predictor variable for subject i; and β_j is the j-th regression coefficient. Let $l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$ denote the log-likelihood function based on modeling each y_i as a Bernoulli random variable with parameter $p_i \equiv P(y_i = 1)$, and linking p_i to the predictors $x_{i1}, x_{i2}, \dots, x_{id}$ by the logistic equation,

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{id}\beta_d.$$

Classical logistic regression (e.g., [38]) estimates the regression coefficients by maximizing $l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$. The LASSO estimates these coefficients by solving the following optimization problem instead:

$$max_{\beta}\, l(\beta; \boldsymbol{X}, \boldsymbol{y}) - \lambda\Omega(\beta), \tag{9.1}$$

where

$$\Omega(\beta) = \sum_{j=1}^{d} |\beta_j|$$

is a penalty function that is able to shrink the regression coefficients, $\beta_1, \beta_2, \ldots, \beta_d$, and force some of them to become zero. If $\beta_j = 0$, then the j-th predictor has no effect. Hence, only predictors with nonzero regression coefficients are "selected" by the model that the LASSO produces.

9.3.1.2 Variable Ranking by the LASSO

The non-negative parameter, λ, controls the amount of shrinkage; more coefficients will become zero (and fewer predictors will be selected) as λ is increased. To a large extent, the choice of λ controls which predictors end up being selected. This means the value of λ used must be carefully justified.

This "inconvenience" can be circumvented by taking into account not just one solution to the optimization problem (9.1) given by one particular (and perhaps subjective) choice of λ, but the entire solution path (see, e.g., [16]) as λ changes. The idea is as follows. Start from a large λ, so large that all regression coefficients are forced to be zero and, hence, no predictor variable is selected. As λ is gradually decreased, the regression coefficients become nonzero and the predictor variables enter the model sequentially; see Fig. 9.2 for an illustration. The relative importance of the predictor variables can be *ranked* by the order in which they enter the model. As far as we are aware, such an application of the LASSO is novel.

9.3.1.3 Ensemble Approach

Furthermore, instead of ranking all the variables just once using the entire dataset, we adopted the "ensemble approach" (e.g., [68, 71]) to obtain a more stable ranking. We first drew 100 random subsamples from our data set, say, $S_1, S_2, \ldots, S_{100}$, each of size $n = 10,000$. Let $r(b, j)$ denote the rank of variable j based on sample S_b. We then calculated the average rank over the 100 samples for each variable j,

$$\overline{r}(j) = \frac{1}{100} \sum_{b=1}^{100} r(b, j),$$

as well as $\sigma(j)$, the standard deviation of $\overline{r}(j)$.

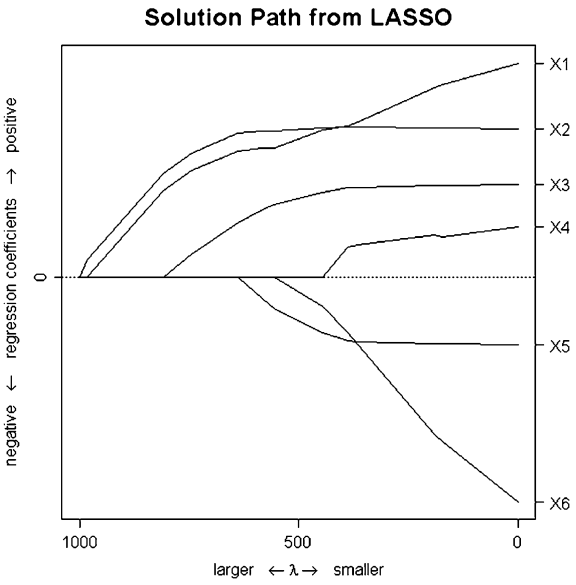


Fig. 9.2 Solution path of the LASSO (illustration): There are six predictors, X_1, X_2, … , and X_6. When λ is very large, all six coefficients are forced to be zero. As λ decreases, the coefficients become nonzero (and the predictors enter the model) in the following order: X_2, X_1, X_3, X_5, X_6, and X_4

9.3.2 Data and Various Details of the Analysis

For this example, our data consisted of initial RAI-HC assessments of $N = 135,184$ home care clients from Ontario. The outcome variable was a binary indicator of whether the client received rehabilitation services, which we defined to mean the receipt of either physical or occupational therapy (PT or OT) within 6 months of a client's initial assessment.

Altogether, a total of $d = 239$ items from the RAI-HC were treated as potential predictors. All items in RAI-HC were included *except* a few "obviously" irrelevant variables, for example, those in the BB8 section, which have to do with responsibility for payment, and obvious "cues" for receiving rehabilitation service such as P2O and P2P, receipt of OT and PT within last seven days, respectively.

In the actual implementation, we used a variation of the original LASSO algorithm, called the "group LASSO" [39]. We used this variation because many of our predictors were categorical. A categorical predictor is often coded by a number of dummy variables in regression analysis, and the group LASSO forces these dummy variables to enter or exit the model together as a group along the solution path.

Table 9.4 Receipt of rehabilitation: top-ranked variables (average rank <20)

Predictor items from RAI-HC (j) and description		$\bar{r}(j)$	$\sigma(j)$	Effect on odds ratio
K6A	Unsteady gait	1.00	(0.00)	+
H3	ADL decline	2.02	(0.01)	+
K6B	Limits going outdoors	3.38	(0.15)	+
J1X	Cancer (last 5 years)	3.58	(0.13)	−
H7A	Client believes can improve	3.68	(0.07)	+
K5	Falls frequency	4.59	(0.16)	+
K4C	Pain intensity disrupts usual activity	5.63	(0.24)	+
P2F	Chemotherapy during last 7 days	6.39	(0.27)	−
P2W	Nurse monitoring < daily	8.75	(0.26)	−
H4B	Mode of locomotion—outdoors	9.49	(0.11)	+
H5	Stair climbing	13.39	(1.34)	+
H4A	Mode of locomotion—indoors	14.30	(0.45)	+
H1FB	Shopping—difficulty	16.57	(0.58)	+
P2V	Daily nurse monitoring	17.04	(0.41)	−
O2B	Better-off if different environment	17.26	(0.29)	−

If a variable has a positive/negative effect on the odds ratio, it means high levels of this variable increase/decrease the probability of receiving PT/OT (within 6 months of initial assessment)

9.3.3 Results

Table 9.4 shows the results that we obtained from this analysis. The last column (effect on odds ratio) was obtained as follows. First, a logistic regression model was fitted using all the top-ranked variables in Table 9.4. For categorical variables (all except K5—falls frequency), level "8" (which, where applicable, meant "activity did not occur") was not considered (standard interRAI practice), and level "0" was treated as the baseline (see RAI-HC instrument).

For each of the other levels, a 95 % confidence interval was then obtained for its odds-ratio relative to the baseline. Let L denote the minimum of lower confidence limits over all levels, and U, the maximum of upper confidence limits over all levels. The interval (L, U) could be regarded as a kind of conservative, meta-confidence interval. For a continuous variable (K5 in this case), this would be the regular confidence interval. Finally, let $M = (L + U)/2$ be the midpoint between L and U. The effect on odds ratio was summarized to be positive (+) if $M > 1$, and negative (−) otherwise. Other than H5 (stair climbing), none of these variables had a meta-interval that included unity, i.e., either $L < U < 1$ or $1 < L < U$. The positive and negative effects summarized in Table 9.4 were, therefore, unambiguous, except for H5.

While many of the variables identified by the LASSO algorithm as strongly associated with the receipt of rehabilitation services were related to impairments in gait (K6A) or ADL (H3) that could be improved through rehabilitation therapy, others were associated with a reduced likelihood of receiving rehabilitation, such as a cancer diagnosis (J1X) or recent receipt of chemotherapy (P2F). These associations may be because rehabilitation is considered inappropriate for some

persons with a terminal illness. However, this is an area that warrants further investigation, as many cancer patients would benefit from, but do not receive, rehabilitation services [62].

Also of note is that cognitive impairment was *not* specifically identified as an important factor associated with not receiving rehabilitation. This disassociation may be because other variables are acting as proxies for adequate cognitive function, for example, H7A (client belief in potential for improvement) and/or O2B (client seen as better off elsewhere). Alternatively, cognitive impairment may *not* be currently used as a criterion to limit rehabilitation services, for reasons suggested in the Introduction section.

We clearly see that machine-learning techniques not only can be used to make good predictions, but also can be used to obtain useful scientific insights.

9.3.4 Discussion

The approach we took above to select/rank/screen predictor variables is an example of the so-called "ensemble approach." The ensemble approach for solving prediction problems was first made popular in the machine-learning community through such algorithms as boosting [18, 20], bagging [8], random forest [9], and the gradient boosting machine [19]. The ensemble approach for variable selection was pioneered by [72] and formalized by [68], while similar ideas have appeared in the literature (e.g., [41]).

To describe the main idea, suppose there are p candidate variables. A variable-selection ensemble (of size B) can be represented by a $B \times p$ matrix, say $\boldsymbol{E}$, whose j-th column contains B repeated measures of how important variable j is [68]. Let $E(b,j)$ denote the (b,j)-th entry of $\boldsymbol{E}$. Using the ensemble $\boldsymbol{E}$ as a whole, one can *rank* the importance of variable j using a majority-vote type of summary, such as

$$R(j) = \frac{1}{B}\sum_{b=1}^{B} E(b,j). \tag{9.2}$$

The key for generating a variable-selection ensemble (VSE), therefore, lies in producing multiple measures of importance for each candidate variable. By contrast, "regular" variable selection procedures typically produce just one such measure, that is, $B = 1$. It is easy to understand why averaging over a number of repeated measures is often statistically beneficial. This benefit is the main reason that VSEs are attractive and more powerful than many classical approaches.

The formal definition of VSEs above makes it clear that VSEs can be generated in many ways. In Example 1, we identified that the most important variables for predicting a client's rehabilitation potential were H2J, H7A, and H7C by studying the non-support vectors. This finding can be verified by a simple VSE. Recall that, in Example 1, we had created eight training datasets. Using each of these datasets, we performed stepwise variable selection on a standard logistic regression model

using the Akaike Information Criterion (AIC, [1]) as the selection criterion. We thus obtained eight slightly different subsets of selected variables. The only variables that appeared in the intersection of *all* eight subsets were H2J, H7A, and H7C.

We believe the ensemble approach we used in this example to screen predictor variables has a good deal of potential in health informatics. We think variable *selection* per se is usually not quite the right scientific objective, whereas variable *ranking* is. Imagine the problem of searching for a few biomarkers that are associated with a certain disease. What type of answer is more useful to a medical doctor? Telling him or her that you think it is biomarkers A, B, and C that are associated? Or giving him or her a ranked list of the biomarkers instead? We think the latter is more useful, and such a list is precisely what the ensemble approach is designed to provide.

Using an ensemble approach, variable selection is performed in two steps. We first rank the variables, e.g., by (9.2), and then use a certain thresholding rule to make the selection. As proponents of the ensemble approach, we believe that the task of ranking is the more fundamental of the two steps. From a decision-theoretic point of view, once the variables are ranked, the choice of the decision threshold has more to do with one's prior belief of how sparse the model is likely to be. Hence, the variable importance measure (9.2) is a particularly nice feature of this approach.

9.4 Example 3: Institutionalization (A Combined Predictive and Explanatory Task)

In our third and final example, we illustrate the use of an off-the-shelf machine-learning algorithm to perform both predictive and explanatory tasks.

Currently, within the province of Ontario, there is an increased emphasis on "aging at home" initiatives [67]. These initiatives are aimed at improving the ability of seniors to remain independent in the community so that they can remain out of institutional care (long-term care homes). As institutionalization is very costly, these "aging at home" initiatives are also designed to ensure the sustainability of the overall health system.

Here, we try to predict whether a home care client would be placed into a long-term care (LTC) facility within a year of initial RAI-HC assessment, and to identify the top risk factors for LTC placement. Home healthcare plays a critical role in managing the transition between community and institutional living for older adults; thus the ability to identify at-risk individuals in the home care population and to recognize factors that predict institutionalization can be most valuable.

Table 9.5 The random forest algorithm (for classification)

1. For each $b = 1$ to B, fit a maximal-depth decision tree, $f_b(\mathbf{x})$, as follows:
   (a) (Bootstrap) Draw a bootstrap sample of the training data; call it D^{*b}. Use D^{*b} to fit f_b.
   (b) (Random Subset) When building f_b, randomly select a subset of $m < d$ predictors before making each split—call it S, and make the best split over the set S rather than over all possible predictors.
   End For.
2. Output an ensemble classifier, i.e., to classify $\mathbf{x}_{new}$, simply take majority vote over all trees, $\{f_b(\mathbf{x}_{new}), b = 1, 2, \ldots, B\}$. Alternatively, rank the likelihood that $\mathbf{x}_{new}$ belongs to class k by the fraction of times $f_b(\mathbf{x}_{new}) = k$, for $b = 1, 2, \ldots, B$.

9.4.1 Focal Point: Using the Random Forest as an Off-the-Shelf Algorithm for Both Prediction and Explanation

Most machine-learning algorithms contain a few tuning parameters that must be selected carefully by the user, for example, the parameter K—the number of neighbors—in KNN. Some algorithms are quite sensitive to the choice of these tuning parameters, e.g., the SVM, which makes them hard to use by a non-specialist. The phrase "off-the-shelf" means the algorithm is *relatively* insensitive to these choices. As a result, it is relatively easy to use and does not require much customizing to produce *reasonably* good results. In this section, we illustrate the use of the random forest [9], one of the best off-the-shelf prediction algorithms available [71], to perform both predictive and explanatory tasks.

9.4.1.1 Using the Random Forest to Make Predictions

The random forest algorithm works by building a collection of decision trees. Predictions are made by taking majority vote over all trees. The decision tree is a deterministic algorithm. Given the same data set, the algorithm will produce exactly the same tree in multiple runs. Naturally, a forest made up of many identical copies of the same tree is not interesting or useful. To generate different decision trees, the random forest algorithm employs two stochastic mechanisms. First, the algorithm draws a bootstrap sample [15] from the data before building each tree. Second, it forces each decision tree to optimize its splits over a randomly selected subset of predictors. Table 9.5 describes the algorithm.

9.4.1.2 Using the Random Forest to Rank Predictors

Not only is the random forest algorithm capable of producing excellent predictions, it is also capable of computing a variable-importance measure, which we simply denote as RF-VIM, for each predictor. The RF-VIM is based on marginal evaluations of the would-be deterioration in the overall model performance, had

the values of a predictor been permuted [10]. The rationale is that, if permuting the value of a predictor does not have much effect on the model performance, it must not be an important predictor, and vice versa. It turns out that these evaluations can be performed efficiently inside the random forest algorithm because the bootstrap step (Table 9.5, step 1a) implies that, on average, each observation is only used by about 2/3 of the trees in the forest. We will not go into the technical details of this aspect in this chapter; refer to [9].

9.4.2 Data and Various Details of the Analysis

For this example, the outcome of interest was a binary indicator of whether a client was placed into an LTC facility within a year of initial assessment, and our data consisted of RAI-HC assessments of $N = 13,006$ clients from eight Ontario CCACs. These clients were from a subset of the data used in Example 1, including only older adults (age 65+) who remained within the home care system for more than 30 days. Many older adults received a RAI-HC assessment but did not receive any formal home care services. For example, some individuals were assessed but did not require any supportive services, whereas others were assessed and placed directly into LTC due to their low level of functioning and/or high care needs. These clients were excluded from this analysis because, for the outcome of LTC placement, it was more meaningful to focus only on the "active" home care clients.

A total of $d = 189$ predictors were used in this analysis, consisting of relevant items in the RAI-HC instrument as well as some other measures that were embedded within the RAI-HC system, such as the CHESS score (Changes in Health, End-stage disease and Symptoms and Signs, [30]), a composite measure of RAI-HC items that is used to predict mortality and to measure health instability in geriatric populations, and the CPS (Cognitive Performance Scale, [46]), another composite measure designed to measure the cognitive assets of individuals that has been shown to correspond closely with the Mini-Mental State Examination (MMSE). We also included a Frailty Index [57], a measure based on the idea that the concept of frailty is a non-specific multi-factorial state that can be characterized by examining an individual's accumulated deficits during his/her life course. This measure was constructed for the RAI-HC data using procedures discussed by [59].

For the predictive task, predicting whether a client would be placed into LTC, we followed the same procedures as in Example 1, and made predictions on the eight CCAC data sets one-by-one. When making predictions for cases from one CCAC, we used cases from the other seven CCACs as the training set. We used a default forest size of 500 trees, and ranked the clients by the fraction of trees (out of 500) that predicted LTC placement within a year.

For the explanatory task, identifying key risk factors for LTC placement, we followed the same procedures as in Example 2. As indicated above, a different

Table 9.6 Predicting LTC placement: performance evaluation of the random forest and the MAPLe algorithm

	Area under the ROC Curve (AUC)	
Region	Random forest	MAPLe
1	0.807	0.727
2	0.765	0.702
3	0.791	0.727
4	0.757	0.693
5	0.709	0.644
6	0.766	0.706
7	0.816	0.726
8	0.777	0.716
Mean	0.773	0.705

training set was used for each of the eight CCAC datasets. Let $s(b,j)$ be the RF-VIM score for variable j based on training set b, for $b = 1, 2, \ldots, 8$. For each predictor j, we computed

$$\overline{s}(j) = \frac{1}{8}\sum_{b=1}^{8} s(b,j),$$

as well as $\sigma(j)$, the standard deviation of $\overline{s}(j)$.

9.4.3 Results

It is well-known that, when a numeric score is available to rank the likelihood of a binary outcome, varying the decision threshold will generate a continuum of false positive and false negative rates. In particular, lowering the threshold will lead to more false positives, whereas raising the threshold will lead to more false negatives. The receiver-operating characteristic (ROC) curve (see, e.g., [53]) is a simple two-dimensional graph measuring a model's false positive rate against its true positive rate over all possible decision thresholds. The area under the ROC curve, or simply "area under the curve" (AUC), is a common metric used in this context for evaluating the effectiveness of the ranking produced [53]. A perfectly ranked set of predictions would give $AUC = 1$, and a randomly ranked set would give $AUC = 0.5$.

Using the AUC, Table 9.6 compares the predictive/ranking performance of the random forest with that of an existing decision-support algorithm, MAPLe (Method for Assigning Priority Levels, [32]). Designed for the Ontario home care context to assist case managers in determining the relative priority of a client's need for support services, the MAPLe algorithm has been shown to be a strong predictor of nursing home placement, caregiver distress, and for being rated as requiring alternative placement to improve outlook [32].

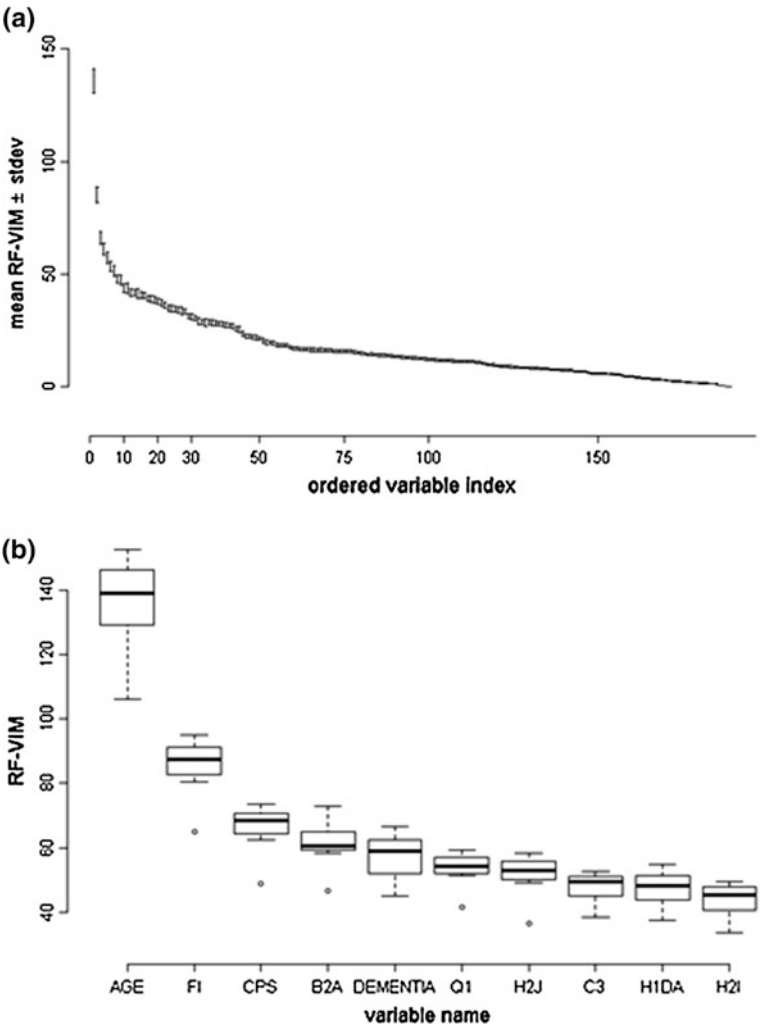


Fig. 9.3 **a** Potential risk factors for LTC placement within 1 year of initial RAI-HC assessment, ranked by the variable-importance measure derived from the random forest algorithm (RF-VIM)—mean and stdev from 8 CCAC datasets. **b** The top 10 risk factors—distribution of RF-VIM from 8 CCAC datasets

Results for the explanatory task are displayed in Fig. 9.3 and Table 9.7. The top two predictors, age and the Frailty Index (FI), clearly stood out from the rest. The other top predictors were mostly related to cognition (e.g., CPS, B2A, C3) and the ability to perform *basic* ADLs (e.g., H2J—bathing, and H2I—personal hygiene).

Table 9.7 Top 10 risk factors for LTC placement within 1 year of initial assessment, as ranked by the variable-importance measure derived from the random forest algorithm (RF-VIM)—mean and stdev from eight CCAC datasets

Risk factors (j) and descriptions		RF-VIM	
		Mean $\bar{s}(j)$	Stdev $\sigma(j)$
Age	Approximate age	135.79	(5.35)
FI	Frailty index	85.28	(3.31)
CPS	Cognitive performance scale	66.09	(2.73)
B2A	Daily decision making	61.17	(2.67)
Dementia	Presence of dementia	57.35	(2.57)
Q1	Number of medications	53.42	(1.95)
H2J	Bathing	51.53	(2.39)
C3	Ability to understand others	47.76	(1.68)
H1DA	Managing medications—performance	47.35	(2.10)
H2I	Personal hygiene	43.87	(1.93)

9.4.4 Discussion

While the MAPLe algorithm was developed using sophisticated methods, the aim was nonetheless to produce an algorithm that was straightforward, comprehensible, defensible, and intuitively appealing, and that had input and buy-in from clinical, health system, and policy experts. Our results suggest that the explicit nature of the MAPLe, understandable given its application in assigning priorities for long-term care placement, may be achieved at a cost to its predictive ability. Although the MAPLe algorithm predicted LTC placement quite well (AUC $\sim$ 0.705), machine-learning approaches are often better able to address the inter-dependencies and non-linear relationships of data that represent the complex characteristics of human subjects and human systems [45].

As for results from the explanatory task, age came out as a key risk factor for LTC placement. Other than age, it is interesting but perhaps not surprising that the Frailty Index emerged as the biggest predictor of LTC placement in our analysis, significantly ahead of other cognition and basic ADL items. Analyses reported by the developers of the Frailty Index [44], as well as our own work [2] and that of others (e.g., [34]) have found the FI to be a strong predictor of death, institutionalization, and other adverse health outcomes. As a composite measure calculated based on an individual's total number of measured deficits, the FI thus reflects multiple risk factors. Therefore, our findings further validate the use of the FI as a quantitative summary measure of vulnerability in older adults.

This example has illustrated that the random forest algorithm is a viable, off-the-shelf method capable of making superior predictions, while also generating explanatory information on factors associated with those predictions, thus generating clinical and scientific insights into what is going on inside the "black-box."

9.5 Summary

In this chapter, we have described several examples of how machine-learning algorithms can be used to guide clinical decision making, and to generate scientific insights about these decisions. Our specific focus has been on rehabilitation in home care, but we believe these techniques have a much wider applicability. Our examples have included a conventional application of a comparatively simple and classic algorithm (KNN) to perform a predictive task, a novel application of a more sophisticated algorithm of increasing importance (LASSO) to perform an explanatory task, and a straight-forward application of an off-the-shelf algorithm (random forest) to perform both.

In clinical applications, our work has shown that machine-learning algorithms can produce better decisions than standard clinical protocols. Our work also suggests that a "simple" algorithm such as the KNN may work just as well as a more complex one such as the SVM. More importantly, we have shown that machine-learning algorithms can do much more than make "black-box" predictions; they can generate important new clinical and scientific insights. These insights can be used to make better decisions about treatment plans for patients and about resource allocation for healthcare services, resulting in better outcomes for patients, and in a more efficient and effective healthcare system.

Acknowledgments The InfoRehab project is supported by the Canadian Institutes of Health Research (CIHR). We thank Chloe Wu for her assistance with the management of data.

A.1 9.6 Appendix: Evaluation of Binary Predictions

Suppose we have a certain procedure (whether an algorithm or a protocol) for predicting binary outcomes of either zero or one. The false positive (FP) and false negative (FN) rates are intuitive measures of the prediction performance. They are the probabilities of the two types of errors the procedure can make, namely, calling a true zero a one (FP) and calling a true one a zero (FN).

The positive diagnostic likelihood ratio $(DLR+)$, and the negative diagnostic likelihood ratio $(DLR-)$ are less intuitive but extremely useful measures; they

> "quantify the change in the odds of [the true outcome] obtained by knowledge of [the prediction]" or "the increase in knowledge about [the true outcome] gained through [the prediction]" [53].

Let

$$prior - odds = \frac{P(outcome = 1)}{P(outcome = 0)},$$

$$posterior - odds(prediction) = \frac{P(outcome = 1|prediction)}{P(outcome = 0|prediction)}.$$

By a simple application of Bayes' theorem [3], it can be shown that

$$posterior - odds(prediction = 1) = (DLR+) \times (prior - odds),$$

$$posterior - odds(prediction = 0) = (DLR-) \times (prior - odds).$$

Therefore, $DLR+$ can be interpreted as the factor by which a prediction of one can increase the prior odds, and $DLR-$ can be interpreted as the factor by which a prediction of zero can decrease the prior odds. Therefore, informative prediction procedures should have $DLR+ > 1$ and $DLR - < 1$. Given two prediction methods, A and B, A can be said to be more informative than B if $DLR+(A) > DLR + (B)$ and if $DLR-(A) < DLR - (B)$. The use of $DLR+$ and $DLR-$ to evaluate procedures for making binary predictions has been gaining popularity in the last two decades [53].

References

1. Akaike H (1974) A new look at the statistical model identification. IEEE T Automat Contr 19:716–723
2. Armstrong JJ, Stolee P, Hirdes JP, Poss JW (2010) Examining three frailty conceptualizations in their ability to predict negative outcomes for home care clients. Age Ageing 39:755–758
3. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Phil Trans Roy Soc 53:370–418
4. Begg R, Kamruzzaman J (2005) A machine learning approach for automated recognition of movement patterns using basic, kinetic and kinematic gait data. J Biomech 38:401–408
5. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
6. Boser B, Guyon I, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceeding 5th annual workshop computer learn theory. ACM Press, New York
7. Borrie MJ, Stolee P, Knoefel FD, Wells JL, Seabrook JA (2005) Current best practices in geriatric rehabilitation in Canada. Geriatr Today Can J Geriatr Med Geriatr Psych 8:148–153
8. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
9. Breiman L (2001) Random forests. Mach Learn 45:5–32
10. Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16:199–231
11. Colombo M, Guaita A, Cottino M, Previderé G, Ferrari D, Vitali S (2004) The impact of cognitive impairment on the rehabilitation process in geriatrics. Arch Gerontol Geriatr Suppl 9:85–92
12. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE T Inform Theory 13:21–27
13. Cristianini N, Shawe-Taylor J (2002) An introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, New York
14. Diwan S, Shugarman LR, Fries BE (2004) Problem identification and care plan responses in a home and community-based services program. Med Care 23:193–211
15. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
16. Efron B, Hastie TJ, Johnstone IM, Tibshirani RJ (2004) Least angle regression (with discussion). Ann Stat 32:407–499

17. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 96:1348–1360
18. Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. Machine Learning Proc 13th Int Conf. Morgan Kauffman, San Francisco, pp 148–156
19. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
20. Friedman JH, Hastie TJ, Tibshirani RJ (2000) Additive logistic regression: A statistical view of boosting. Ann Stat 28:337–407
21. Ghisla MK, Cossi S, Timpini A, Baroni F, Facchi E, Marengoni A (2007) Predictors of successful rehabilitation in geriatric patients: subgroup analysis of patients with cognitive impairment. Aging Clin Exp Res 19:417–423
22. Gill PE, Murray W, Wright MH (1986) Practical optimization. Academic Press, London
23. Gilmore S, Hofmann-Wellenhof R, Soyer HP (2010) A support vector machine for decision support in melanoma recognition. Exp Dermatol 19:830–835
24. Hanks RA, Rapport LJ, Millis SR, Deshpande SA (1999) Measures of executive functioning as predictors of functional ability and social integration in a rehabilitation sample. Arch Phys Med Rehabil 80:1030–1037
25. Harrison RF, Kennedy RL (2005) Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. Ann Emerg Med 46:431–439
26. Hastie TJ, Tibshirani RJ, Friedman JF (2001) The elements of statistical learning: data mining, Inference and Prediction. Springer, New York
27. Hepburn B (2010) Healthcare in Ontario is cracking under stress. Toronto Star, August
28. Hershkovitz A, Kalandariov Z, Hermush V, Weiss R, Brill S (2007) Factors affecting short-term rehabilitation outcomes of disabled elderly patients with proximal hip fracture. Arch Phys Med Rehabil 88:916–921
29. Hirdes JP, Fries BE, Morris JN, Steel K, Mor V, Frijters DH et al (1999) Integrated health information systems based on the RAI/MDS series of instruments. Health Manage Forum 12:30–40
30. Hirdes JP, Frijters D, Teare G (2003) The MDS-CHESS scale: a new measure to predict mortality in institutionalized older people. J Am Geriatr Soc 51:96–100
31. Hirdes JP, Fries BE, Morris JN, Ikegami N, Zimmerman D, Dalby DM et al (2004) Home Care Quality Indicators (HCQIs) based on the MDS-HC. Gerontologist 44:665–679
32. Hirdes JP, Poss JW, Curtin-Telegdi N (2008) The method for assigning priority levels (MAPLe): a new decision-support system for allocating home care resources. BMC Med 6:9
33. Ji SY, Smith R, Huynh T, Najarian K (2009) A comparative analysis of multi-level computer-assisted decision-making systems for traumatic injuries. BMC Med Inform Decis Mak 9:2
34. Kulminski AM, Ukraintseva SV, Kulminskaya IV, Arbeev KG, Land K, Yashin AI (2008) Cumulative deficits better characterize susceptibility to death in elderly people than phenotypic frailty: lessons from the cardiovascular health study. J Am Geriatr Soc 56:898–903
35. Lau HY, Tong KY, Zhu H (2009) Support vector machine for classification of walking conditions of persons after stroke with dropped foot. Hum Mov Sci 28:504–514
36. Landi F, Tua E, Onder G, Carrara B, Sgadari A, Rinaldi C et al (2000) Minimum data set for home care: a valid instrument to assess frail older people living in the community. Med Care 38:1184–1190
37. Lucas P (2004) Bayesian analysis, pattern analysis, and data mining in healthcare. Curr Opin Crit Care 10:399–403
38. McCullagh P, Nelder JA (1989) Generalized Linear Models. Chapman and Hall, Boca Raton
39. Meier L, van de Geer S, Bühlmann P (2008) The group Lasso for logistic regression. J Royal Stat Soc B Met 70:53–71
40. Meinshausen N (2007) Relaxed Lasso. Comput Stat Data An 52:374–393
41. Meinshausen N, Bühlmann P (2010) Stability selection (with discussion). J Royal Stat Soc B Met 72:417–473

42. Melin R, Fugl-Meyer AR (2003) On prediction of vocational rehabilitation outcome at a Swedish employability institute. J Rehabil Med 35:284–289
43. Miller AJ (2002) Subset selection in regression. Chapman and Hall, New York
44. Mitnitski AB, Graham JE, Mogilner AJ, Rockwood K (2002) Frailty, fitness and late-life mortality in relation to chronological and biological age. BMC Geriatr 2:1
45. Mitnitski AB, Mogilner AJ, Graham JE, Rockwood K (2003) Techniques for knowledge discovery in existing biomedical databases: Estimation of individual aging effects in cognition in relation to dementia. J Clinical Epidemiol 56:116–123
46. Morris JN, Fries BE, Mehr DR, Hawes C, Phillips C, Mor V et al (1994) MDS cognitive performance scale. J Gerontol 49:M174–M182
47. Morris JN, Fries BE, Morris SA (1999) Scaling ADLs within the MDS. J Gerontol Med Sci 54:M546–M553
48. Morris JN, Fries BE, Steel K, Ikegami N, Bernabei R, Carpenter GI et al (1997) Comprehensive clinical assessment in community setting: applicability of the MDS-HC. J Am Geriatr Soc 45:1017–1024
49. Morris JN, Fries BE, Steel K, Ikegami N, Bernabei R (1999) Primer on use of the Minimum Data Set-Home Care (MDS-HC) version 2.0© and the Client Assessment Protocols (CAPs). Hebrew Rehabilitation Center for Aged, Boston
50. Naglie G, Tansey C, Kirkland JL, Ogilvie-Harris DJ, Detsky AS, Etchells E, Tomlinson G, O'Rourke K, Goldlist B (2002) Interdisciplinary inpatient care for elderly people with hip fracture: a randomized controlled trial. Can Med Assoc J 167:25–32
51. Ottenbacher KJ, Linn RT, Smith PM, Illig SB, Mancuso M, Granger CV (2004) Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Ann Epidemiol 14:551–559
52. Pearce CB, Gunn SR, Ahmed A, Johnson CD (2006) Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-reactive protein. Pancreatology 6:123–131
53. Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York
54. Price RK, Spitznagel EL, Downey TJ, Meyer DJ, Risk NK, El-Ghassawy OG (2000) Applying artificial neural network models to clinical decision making. Psychol Assess 12:40–51
55. Radchenko P, James G (2008) Variable inclusion and shrinkage algorithms. J Amer Statist Assoc 103:1304–1315
56. Ramos-Pollán R, Guevara-López MA, Suárez-Ortega C, Díaz-Herrero G, Franco-Valiente JM, Rubio-Del-Solar M, González-de-Posada N, Vaz MA, Loureiro J, Ramos I (2012) Discovering mammography-based machine learning classifiers for breast cancer diagnosis. J Med Syst 36(4):2259–2269
57. Rockwood K, Mitnitski AB (2007) Frailty in relation to the accumulation of deficits. J Gerontol A Biol Sci Med Sci 62:722–727
58. Shmueli G (2010) To explain or to predict? Stat Sci 25:289–310
59. Searle S, Mitnitski AB, Gahbauer E, Gill T, Rockwood K (2008) A standard procedure for creating a frailty index. BMC Geriatr 8:24
60. Stolee P, Borrie MJ, Cook S, Hollomby J, the participants of the Canadian Consensus Workshop on Geriatric Rehabilitation (2004) A research agenda for geriatric rehabilitation: the Canadian consensus. Geriatr Today J Can Geriatr Soc 7:38–42
61. Tam SF, Cheing GLY, Hui-Chan SWY (2004) Predicting osteoarthritic knee rehabilitation outcome by using a prediction model using data mining techniques. Int J Rehabil Res 27:65–69
62. Thorsen L, Gjerset GM, Loge JH, Kiserud CE, Skovlund E, Fløtten T, Fosså SD (2011) Cancer patients' needs for rehabilitation services. Acta Oncol 50:212–222
63. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Royal Stat Soc B 58:267–288
64. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York

65. Wells JL, Seabrook JA, Stolee P, Borrie MJ, Knoefel F (2003) State of the art in geriatric rehabilitation. Part I: review of frailty and comprehensive geriatric assessment. Arch Phys Med Rehabil 84:890–897
66. Wells JL, Seabrook JA, Stolee P, Borrie MJ, Knoefel F (2003) State of the art in geriatric rehabilitation. Part II: clinical challenges. Arch Phys Med Rehabil 84:898–903
67. Williams AP, Lum JM, Deber R, Montgomery R, Kuluski K, Peckham A et al (2009) Aging at home: integrating community-based care for older persons. Healthc Pap 10:8–21
68. Xin L, Zhu M (2012) Stochastic stepwise ensembles for variable selection. J Comput Graph Stat 21:275–294
69. Zou H (2006) The adaptive Lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
70. Zou H, Hastie TJ (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc B 67:301–320
71. Zhu M (2008) Kernels and ensembles: perspectives on statistical learning. Am Stat 62:97–109
72. Zhu M, Chipman HA (2006) Darwinian evolution in parallel universes: a parallel genetic algorithm for variable selection. Technometrics 48:491–502
73. Zhu M, Chen W, Hirdes JP, Stolee P (2007) The K-nearest neighbors algorithm predicted rehabilitation potential better than current clinical assessment protocol. J Clin Epidemiol 60:1015–1021
74. Zhu M, Zhang Z, Hirdes JP, Stolee P (2007) Using machine learning algorithms to guide rehabilitation planning for home care clients. BMC Med Inform Dec Mak 7:41

Chapter 10
Clinical Utility of Machine Learning and Longitudinal EHR Data

Walter F. Stewart, Jason Roy, Jimeng Sun and Shahram Ebadollahi

Abstract The widespread adoption of electronic health records in large health systems, combined with recent advances in data mining and machine learning methods, creates opportunities for the rapid acquisition and translation of knowledge for use in clinical practice. One area of great potential is in risk prediction of chronic progressive diseases from longitudinal medical records. In this Chapter, we illustrate this potential using a case study involving prediction of heart failure. Throughout, we discuss challenges and areas in need of further development.

Keywords Electronic health records · Hearth failure · Machine learning · Prediction models · Text mining

10.1 Introduction

The adoption of electronic health records (EHR) in clinical practice is increasing as technology advances, and regulatory pressures from healthcare reform efforts. Large health care delivery systems in the US adopted such systems early and are

W. F. Stewart (✉)
Sutter Health, Concord, CA, US
e-mail: stewarwf@sutterhealth.org

J. Roy
University of Pennsylvania, Philadelphia, PA, US
e-mail: jaroy@upenn.edu

J. Sun · S. Ebadollahi
IBM TJ Watson Research Center, Hawthorne, NY, US
e-mail: jimeng@us.ibm.com

S. Ebadollahi
e-mail: ebad@us.ibm.com

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_10,

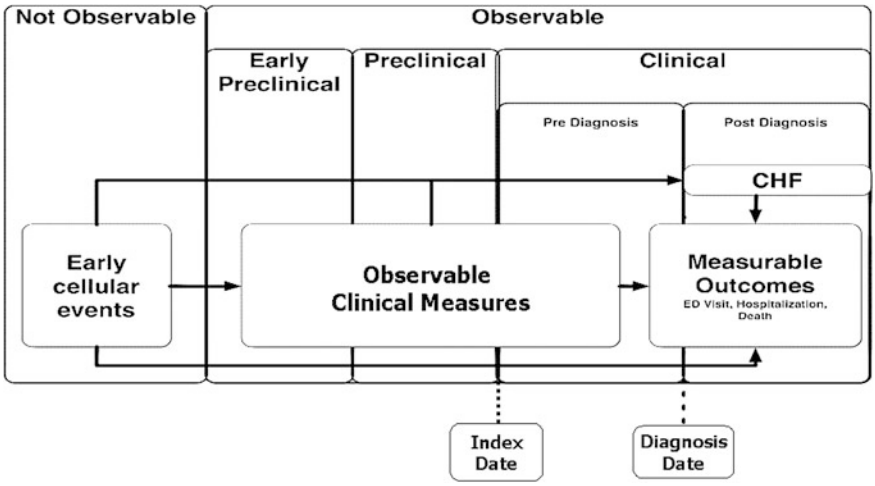


Fig. 10.1 Chronic disease progression from early non-observable events to early preclinical and later stages

rapidly demonstrating the diversity of ways in which EHRs can be used to create value by improving the quality, safety, and accessibility of care at a lower cost. In particular, EHRs are used to accelerate the translation of knowledge for real-time use in clinical practice. In this chapter, we consider how machine learning tools may be used to develop decision aids and diagnostics for real-time use in clinical practice.

The accessibility of EHRs presents significant opportunities for utilizing clinical daata and other information in ways not possible with paper records. In particular, longitudinal clinical care data are accessible from EHRs for disease risk prediction and personalized decision making. The opportunities are particularly compelling given the recent co-evolution of powerful machine learning and data mining techniques. These techniques offer potentially promising means for the rapid extraction of information from EHRs and translation of that information into clinical care decision support.

The disciplines of data mining (DM) and machine learning (ML) are rapidly maturing with great success in many applications such as search engines, image analysis and recommendation systems. The opportunity in healthcare is to use these tools to accurately derive insights from longitudinal patient data that can effectively be used change the clinical pathway of the patient for an optimal outcome.

Longitudinal EHR data are increasingly being used to predict future events or outcomes for a given patient. For example, chronic diseases emerge over time, mediated by early physiologic and pathological changes for which overt or surrogate indicators are documented in a patient's record (Fig. 10.1). The primary goal of predictive modeling in this context is to move detection of the disease from

Table 10.1 Examples of application of data mining and machine learning to EHR data

Examples	Time scale	Value of predictive model	Who benefits?
Risk of chronic progressive disease	12–36 months	Slowing progression preventing onset	Patient and payer
Risk of disease progression	12–60 months	Slowing progression, preventing rapid decline	Patient and payer
Optimizing choice of interventions	Variable		Patient and others
Time to inpatient discharge	Days	Improve discharge preparation, reduce readmission	Hospital and patient
Risk of 30-day readmission	Days to weeks	Reduce risk of re-admission	Hospital and patient
Identifying future costly patients	12–36 months	Prevention and case management to reduce cost of care	Payer and patient

a frank disease state to an earlier clinical or pre-clinical disease state so that the natural history of the diseases itself can be changed. In this framework, population level data can be mined to detect robust signals before an event or to influence the course of events (e.g., optimizing choice of treatment). The time scale for prediction depends on the clinical context. For example, effective use of a signal of future risk of chronic progressive diseases like heart failure would likely require that the disease be detected 1–2 years before usual diagnosis if the use of an intervention at an earlier time is to be used to influence the natural course of disease. In contrast, the time scale for predicting 30-day readmission risk is days to weeks. The formula that is developed to quantify the signal can be applied to individual patients in real time (i.e., algorithms applied to extracts of patient EHR data) or in batches, as needed. Quantitative signals of this type have many potential applications during routine encounters or for population level screening or management (Table 10.1).

The focus of this chapter is on the use of data-mining and machine-learning tools to detect a patient's risk for chronic progressive diseases like diabetes, dementia, kidney disease, and heart failure (HF), among others. These conditions are and will continue to be the dominant drivers of healthcare costs as increasing prevalence in aging populations drives the demand for healthcare and increases per capita healthcare costs. While evidence indicates that medical home and other high touch care models can improve patient outcomes and reduce the cost of care, the options for success are narrowly defined given the deterministic nature of progressive illnesses. An alternative approach is to develop cost effective means of early disease detection and intervention as a means to slow disease progression. This approach is especially sensible when safe and low-cost diagnostic tests and treatments are available and the health problem can be detected early enough to change natural history. In this context, we view longitudinal EHR data as a clinical care asset [1], where patient data can be searched using sophisticated data-mining and machine-learning tools for early signals of disease or disease progression. In

many respects, formalizing the use of such tools in clinical practice is akin to using screening or diagnostic tests to identify patients at high risk of progressive deterioration. In this chapter, we focus on recent work we have completed for the early detection of heart failure.

10.1.1 Why Heart Failure?

Heart disease has long been the leading cause of death in the US. The routine identification of risk factors (e.g., hypertension, hyperlipidemia, use of tobacco products, changes to diet) and growth in the diversity and use of effective treatments has contributed to a substantial decline in the rate of heart attacks and the cause specific mortality rate over the past 60 years. As a consequence, the age specific prevalence of heart failure has increased.

HF is a heterogeneous disease. Two common variants (i.e., diastolic HF and systolic HF) account for 80–85 % of prevalent cases. Onset is subtle and detection is difficult. The more common symptoms (e.g., shortness of breath on exertion, ankle edema) are somewhat non-specific and can be explained away by e.g., poor conditioning, weight, prolonged standing, venous insufficiency, and certain medications. The emerging disease is often missed until it is more serious and has concomitant expression of multiple symptoms (e.g., rales, shortness of breath without exertion, tachycardia, pleural effusion) or because of a more explicit diagnosis (i.e., ejection fraction <50 %) of underlying disease. While HF is usually first detected in primary care, it is often diagnosed at such a late stage that it will continue to progress and the patient will deteriorate over a 5 year period. Early detection of diastolic HF, in particular, is of interest because emerging evidence indicates that low cost benign treatments may be effective in slowing disease progression. Previous efforts at early detection using screening questionnaires have not worked [2]. We have begun to explore how longitudinal patient data can be used in combination with data-mining and machine-learning tools to detect patients at high risk of a future heart failure diagnosis.

In the sections that follow, we first describe issues with the use of electronic health records data for predictive modeling and specific considerations of how to use structured and unstructured data. We then describe approaches to modeling and close with an explanation of how such tools might be used in clinical practice.

10.1.2 Use of EHR Data for Predictive Modeling

The format and accessibility of EHR data vary by manufacturer, clinical setting (e.g., primary care or specialty care), organization, and time. Our focus in this section is not on this aspect of EHR data. Rather, our focus is on the categories of data routinely available in EHRs data and how these data might be used and

represented for predictive modeling. We first consider the source population and how it differs from traditional longitudinal studies, followed by a discussion of features of structured data and the features of unstructured data.

10.1.2.1 Study Context

While a primary care population is similar to a general population sample, the source of EHR data differs substantially from how data are collected in a typical epidemiologic study. A longitudinal epidemiological study is usually initiated to understand causal relations. A representative sample of the population of interest is first identified and recruited. Selection bias can occur at the outset because the option to participate or not is not random. The data and the data collection schedule are fixed. That is, the data to be collected are defined beforehand by protocol and are collected from individuals at fixed or defined time intervals. Typically, subject participation erodes over time and the remaining participants are less and less representative of the original source population. Nonetheless, outcome status of the participants is determined both through routine follow-up and by other means (e.g., death certificates, inpatient treatment).

EHR data arise from a source population defined by the type of care provided. Primary care practices represent the most general population sample, while specialty care patients usually represent a more select population depending on the specialty, location (e.g., major medical center versus community provider), and established referral patterns. The purpose in developing a predictive model will dictate the source population that is most appropriate. Table 10.2 describes features of a longitudinal study using EHR data on a primary care population versus a traditional general population sample selected for longitudinal study (Table 10.2). The nature of selection bias and missing data are somewhat different for these two considerations.

For clinical care, a predictive model is developed to infer something predictive who has sought care. That is, the predictive model is developed so that it can be applied to the same type of individual longitudinal data that were available for the development of the model in the first place. As such, notions of selection bias are more difficult to define and depend on a number of factors. First, the equivalent of patient recruitment and enrollment begins with the first visit to a given primary care practice or a specific primary care provider (PCP). Patients choose their provider rather than being selected for longitudinal follow-up. Second, the act of seeking care and the frequency of encounters are related to health status and care seeking behavior of the patient. In contrast to a traditional epidemiologic study, the visit pattern itself may contain information about the patient. Third, “data collection” in primary care is unscheduled. Data are obtained because a patient is scheduled for a visit. While some data (e.g., weight, blood pressure, pulse, temperature) are routinely collected, most data are related to the reason for visit. That visit could be for routine care (e.g., periodic physical), for a specific need (e.g., acute health problem, disabling pain, etc.), or for other possible reasons

Table 10.2 Comparative methods for a traditional prospective cohort and "study cohort" selected from a primary care practice with an EHR where retrospective data are used

Study population or design feature	Traditional prospective study	Primary care EHR "Cohort"
Identification and recruitment	The sampling frame is defined by the investigator, and individuals are identified to be representative of the source population	Source population is selected by the investigator from among those who have sought care
Enrollment	Active outreach and enrollment: Individual decides to opt in or out	Individual "participates" by choosing a PCP and seeking care as needed
Follow-up	Active scheduling of participant at defined intervals: Drop out or censuring occurs because the patient opts out or moves from area	Individual is defined as a participant as long as they seek care from the PCP
Data collection	Fixed by protocol	Defined by patient need and purpose of visit and decisions made by the physician
Data	standardization	Fixed by protocol
Can vary over time		
Outcome assessment	Fixed by protocol and obtained directly from the individual and from other sources	Passively ascertained through care that is delivered and actively pursued from other sources (e.g., hospitalizations, mortality)

(e.g., prescription refill). The data that are obtained can vary among patients who present with the same health problem. Finally, the extent of follow-up for health problems also varies. Outcomes may or may not be captured depending on the type (e.g., process of care, mediating measures such as blood pressure, or LDL, and serious outcomes like stroke), the visit pattern, and continued use of a primary care practice or network of practices connected to a common data repository (e.g., system level EHR). More generally, the predictive models that are developed must be conditioned by what can be observed with regard to duration of time as a patient, number of visits, and the data that are available.

In contrast to a traditional study framework, the application of machine-learning to longitudinal EHR data poses unique challenges because data collection is unscheduled. In part, the decision to seek care or to schedule a visit is under the control of the patient and usually motivated by health problems. As such, information about the patient is inherent to this unscheduled feature of the EHR data itself.

10.1.2.2 General Principles on Use of EHR Data

EHR data are recorded in a variety of formats. The types of measures and variables can be characterized by the following features: static versus dynamic, the scale of measure, whether the measure is hierarchical, repeated measures over time, and

Table 10.3 Common examples of EHR data and their features

Data source	Feature	Feature type
Demographics	Age, gender, ethnicity	Static
Diagnosis	ICD9, HCC (higher level aggregation of ICD9)	Temporal discrete
Imaging	Imaging test order	Temporal discrete
Medication	Generic name, drug class and subclasses	Temporal discrete
Lab	Component name	Temporal continuous
Vital	Systolic and diastolic blood pressure, pulse, temperature, weight, height	Temporal continuous
Social history	Drinking and smoking history	Temporal discrete
Symptom	Framingham criteria and other related symptoms	Temporal discrete
Temporal patterns	Visitation frequency, frequently co-occurring diagnosis codes and medication	Temporal discrete

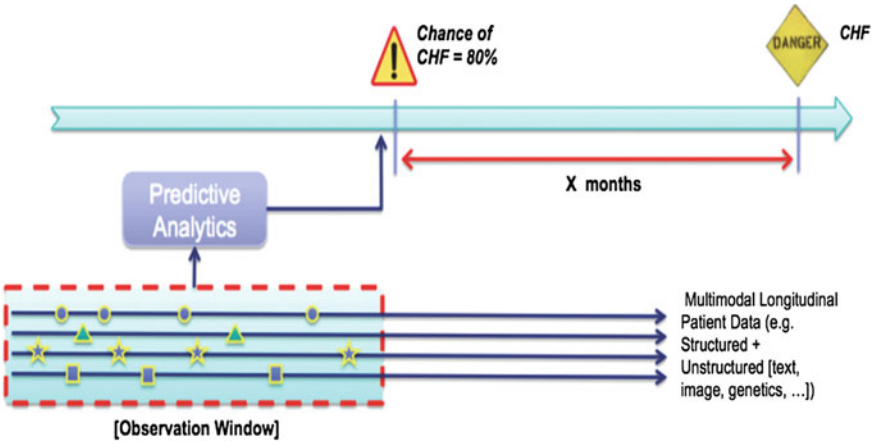


Fig. 10.2 General representation of EHR data for predictive modeling

temporal relation to clinically meaningful anchors (Table 10.3). The approach to representing features for a hypothetical patient is described in Fig. 10.2, which shows that observation time begins with the first primary care visit and ends with the diagnosis of heart failure or with an equivalent index time for a control. The richness of the features for predictive modeling depends on the data domain which is related, in part, to dynamicity. Importantly, every feature may have up to three temporal descriptors.

(1) The measure can occur within or outside the observation window (i.e., the time period for predictive measures obtained in advance of diagnosis).
(2) The time between the date of the measure and the date of diagnosis and
(3) The temporal correlation among measures from different domains in the observation window may be particularly important for detecting more complex but meaningful and robust signals.

The objective of modeling is to determine how each type of measure and combination of measures can be used along with relevant temporal measures to predict a future event with sufficient accuracy to be clinically useful.

A limited set of variables (e.g., demographics, blood pressure, beats per minute, weight, sex, etc.) are routinely available on all or almost all patients. Most other measures, however, are strongly dependent on the healthcare utilization tendencies (e.g., scheduling routine physical exam) of a given patient and, more importantly, on which health problems a patient experiences that motivate whether and when healthcare is sought. Given this unique feature of how EHR data come to be, a "missing value" framework is useful in defining how to represent data in a model, given that any variable may not be available on a subgroup of patients. That is, in contrast to traditional epidemiologic studies, EHR data are captured in an unscheduled manner and the specific value of a variable (e.g., diagnosis of a given disease) for a given patient may not be known. This means that the act of documenting data in the EHR may contain information about the future status of a patient, above and beyond, the actual value of the variable itself at a specific time. For some variables, the initial documentation of the feature itself will suffice and can be represented by a binary 0/1 indicator. This representation is sensible if the disease is almost always brought to the attention of a physician and diagnosed (e.g., type II diabetes). But most health related measures are not routinely ascertained. One simple technnique for this type of "missing data" Problem is to represent a variable by a 0/1 binary variable indicating the availability of the measure and an interaction term between the binary variable and the actual variable value can be used to separately represent the value of the variable among those who have one or more values. By using this type of representation a variable is effectively "observed" for all subjects, and takes a value of zero if the variable was never measured [3].

A diversity of variables can be measured repeatedly on a given patient, for example, repeated measures of common disease mediators (e.g., LDL, blood pressure). The number of repeated measures varies greatly among patients. Moreover, the frequency of measurement is related to health status independent of the variable values. Repeated measures can be represented by measures of central tendency and variability, but these summary measures may be insensitive to predicting future events. The most recent repeated measures within the observation window may be the most relevant, but predictive power may also depend on the value of previous repeated measures. For examples, the first documentation of elevated blood pressure or hypertension indicates the beginning of a disease process, but the impact of this process in mediating vascular pathology will vary depending on the actual systolic and diastolic pressure over time. The temporal effect of blood pressure can also be represented by an area defined by the pressure level above or below a clinically recognized threshold (e.g., diastolic 80 mm Hg) and the time above or below this limit. Finally, The temporal variance of repeated measures of a disease mediator, in particular, may be a useful indicator of instability or change in latent disease state, especially if variability is increasing with time.

10.1.2.3 Relevant Features from Structured Data

While space does not permit a detailed description of how to use different types of EHR variables, we consider the specific measures that are available in the most common domains (Table 10.3). Demographic variables are usually available on all patients in fixed field format and include sex, date of birth, and race or ethnicity.

Health behavior is a dominant mediator of disease risk, and tobacco and alcohol use are among the two most important measures. While, these features are not always available in the EHR, these data are likely to become more common given meaningful use regulations. However, there is no national standard for collecting data on smoking and alcohol use or on other health behaviors. Data obtained on health behavior are likely to vary in specificity and to be represented in a variety of formats. The simplest and most common format for smoking status is likely to be represented by the patient's status (i.e., current smoker, not currently using tobacco, never used tobacco) and possibly some information on level of use. Repeated measures of current status can be used to create a time dependent exposure measure. However, documentation of smoking and alcohol status may be missing or may be coded as "not asked." If the value is missing, it may be best to code the measure as missing or alternatively as a non-smoker and non-user of alcohol. Sensitivity analysis can be used to determine the validity of assumptions about missing data (i.e., whether not asked means not used versus unknown) and can provide useful predictive information.

Prescription medications are represented in the EHR by a time stamped fixed field order along with other descriptors. Each medication order includes information that may be used to calculate or estimate the day's supply. Order spans may be strung together to define a period of potential active medication use. Specifically, the number of days between orders defines the ***time span*** during which medication could have been used. The number of days' supply of medication times the medication dose can be used to calculate the total dosage available. Dividing the total dosage by the time span gives the average daily dose. Dividing the days' supply by the time span gives the proportion of days that a medication was taken or the medication possession ratio (MPR). Medications will be switched from time-to-time. However, it is important to distinguish medication switching (i.e., stop current medication and switch to another) from the addition of a medication (i.e., add new medication to the current therapy).

One general limitation to EHR order data is that it does not indicate of the patient actually picked up the medication and used it. Insurance claims data are required to determine if the medication was obtained by the patient. National efforts have been underway to return data on claims adjudication back to the prescribing physician's EHR. In the absence of the claims data, memorable assumptions can be made regarding use of medications. For example, if a provider submits sequential prescription orders for the same problem (e.g., type II diabetes) over time, it is more likely than not that the patient was using the medication from earlier orders. To make use of prescription order data, the following variables should be considered.

(1) ***Medication order start date***, defined by order date;
(2) ***Medication order end date*** is not explicitly defined. It must be inferred by either the date of the subsequent order for the same health problem or by the number of days' supply in the last order (i.e., add this to the last order date), or the average time span between previous prescription orders.
(3) ***Number of days' supply*** is defined by the number of pills per prescription and the number of pills prescribed for use each day. The number of pills per day is usually found in the medication ***sig,*** a free text field entered by the clinician at the time of prescription. Comprehensive databases of sig field texts do not exist, making it challenging to convert the sig into the number of pills per day. Nonetheless, prescription orders are often for a standard number of days. For chronic progressive conditions like hypertension, the number of days' supply is usually in monthly intervals versus an acute prescription (e.g., a bacterial infection) which would usually be in days. It is common practice to determine if the same patient had orders for other medications of the same class or subclass that overlap the medication span.
(4) ***Medication switch versus add*** is usually dependent on the study aim. For example, if the question of interest is focused on switching medications within a drug sub-class (e.g., from one calcium channel blocker to another) for anti-hypertensives (the major class), a switch will be defined differently than if the question is focused on changing the medication from one sub-class to another (e.g., calcium channel blockers to ACE inhibitors). If a new medication order occurs prior to the end date for the previous medication, a decision needs to be made on whether it was a medication switch or the addition of a second medication. This decision can be based on clinical judgment (i.e., is it common practice to treat with combination therapy for the respective medications) and by querying for re-orders in the future. That is, the new medication is not likely to be a switch if it is ordered in the future along with the original medication, and
(5) ***Medication dose*** is determined by multiplying the dose within each pill (which is found embedded within the medication name) and the number of pills per day (which is identified by translating the medication sig). There may be different combinations that result in the same dose per day. For example, 400 mg twice a day is the same dose as 200 mg four times a day. These medication variables can be used to both create order spans and to link these spans together to define a time period of active medication use. Considerations can be made for the length of allowable gap in coverage (i.e., a new order occurs 100 days after an order that had only 90 day supply). Finally, we note that medication use is also documented in the active medication list that is used by a nurse or provider to record medications that the patient verifies or reports he or she is using at the time of an office visit encounter. An advantage of this list is that it can be used to identify other important medications that are not ordered by the physician, including use of over the counter medications like aspirin. A disadvantage of this list is that there are no standards for how to obtain these data from patients and the reliability and validity of the data has not been established.

Table 10.4 Operational criteria for defining selected co-morbidities

Co-morbidity	Relevant ICD-9 codes	Minimal Operational Criteria to indicate presence of the health problem	Supplementary Data
Hypertension	401.**–405.**	2 encounter diagnosis or 2 medication orders with an ICD-9 diagnostic code for hypertension	Untreated blood pressure indicating presence of hypertension that is not explained by other factors
Diabetes	250.**	2 encounter diagnosis or 2 medication orders with an ICD-9 diagnostic code for diabetes	Untreated Hemoglogin A1c or fasting glucose that is consistent with diabetes
Hypocholesterolemia	272.**	2 medication orders with an ICD-9 diagnostic code for hyperlipidemia	Untreated LDL measures consistent with hyperlipidemia
Asthma	493.**	2 encounter diagnosis or 2 medication orders with an ICD-9 diagnostic code for asthma	Untreated pulmonary function test results consistent with asthma

As previously noted, documentation that a patient has a specific disease depends, in part, on how the patient uses care. Serious diseases are likely to be universally documented in the EHR. For example, most, if not all, patients who have type II diabetes will eventually be diagnosed by a physician, even though patients may vary by the time interval between actual disease onset and diagnosis. For other health problems (e.g., depression) the likelihood of diagnosis will depend on disease severity. Separate from the completeness of documentation, operational criteria must be developed to designate whether a patient has a specific disease. Typically, operational criteria for disease designations require the repeated documentation of selected ICD-9 codes in separate encounters occurring in a limited time period (e.g., 12 months) along with supporting documentation from relevant clinical measures. The appearance of an ICD-9 code as an encounter diagnosis or with a medication order or on the problem list is usually meaningful. In contrast, an ICD-9 code that appears with an image order may not be meaningful as it may simply represent a possible diagnosis that had to be expressed to execute the order. The sensitivity and specificity of operational criteria will vary by the number of required mentions of an ICD-9 code. Operational criteria that are based on fewer mentions (e.g., 2 or more) will be more sensitive, whereas operational criteria that are based on more mentions (e.g., 3 or more) will be more specific. Examples of operational criteria are summarized in Table 10.4. Finally, the duration of time with a disease may be an important predictor of risk of other diseases like HF. The first documentation of the presence of a disease in the EHR does not indicate that the disease was first diagnosed. Distinguishing incident from prevalent conditions requires a consideration of the time interval between the first primary care visit and

the first documentation of a disease. A short time interval (e.g., <6 months) more likely than not indicates that the disease was diagnosed before the patient saw a new primary care physician for the first time.

Vital sign measures are typically stored in fixed fields and are measured repeatedly because they are obtained during most ambulatory encounters. Vital signs include systolic and diastolic blood pressure, pulse, body temperature, height, and weight. Of these, measures of height for adults tend to be incomplete and can vary unexpectedly because of differences in how height was obtained (e.g., with and without shoes). Median height may be the most sensible measure to use when deriving other measures such as body mass index.

Laboratory measures are represented by an order and usually followed by a result report. Some tests are ordered as part of routine care (e.g., basic metabolic panel or lipids). Others are ordered to further study a particular disease (e.g., hemoglobin A1c used to evaluate diabetes). Additionally, the length of time between tests may vary depending on the test that is being conducted. For example, lipids may be tested every 5 years in a young, healthy adult but every 6–12 months for those with high cholesterol. Other tests are may be performed repeatedly over the course of several days, especially if a patient has been hospitalized. The lab result may be numeric (e.g., total cholesterol), text (e.g., blood culture results are recorded as positive or negative), or both (eGFR is numeric if <60 but then grouped into >60 when appropriate). A test that has a numeric result may have a text result which may be very meaningful (e.g., LDL = "UNINTERPRETABLE" when triglycerides are extremely elevated). Regardless of the lab type, the data elements for a given test usually include the resulting value, and date the test was administered.

10.1.3 Relevant Features from Unstructured Data

In addition to structured information, EHR data also contain unstructured information, such as physician notes and radiology reports, which are in text format. In fact, text data comprise a majority of all data available in an EHR. These text notes often contain the medical practitioners' documentation for various types of patient encounters. Common types of encounter notes include "Office Visit," "Case Manager," and "Radiology." The practitioners uses a standard set of section labels to characterize the content, most commonly in SOAP (subjective, objective, assessment, and plan) format. Other types of sections in these notes include Examination, History, and Comment. These notes often contain much more detailed description about signs and symptoms, which are not usually available in the structured data, but that may be quite useful for predictive modeling. For example, progress notes can be used to identify the occurrences of sign and symptoms as well as the context in which they are mentioned.Text mentions of Framingham signs and symptoms may be particularly important. Specifically, text

can indicate whether a symptom was defined as present or absent, severe, persistent, and other important descriptors that indicate whether it was consistent with heart failure or another disease (e.g., COPD).

10.2 Early Detection of Heart Failure: A Case Study in Approaches to Machine Learning

In this section, illustrate a specific example involving heart failure prediction modeling [4]. We begin with a description of initial work, which involved predictive modeling using structured data, and then describe ongoing enhancements to the model by including variables created from text mining.

Data for the early detection of heart failure were obtained from the Geisinger Clinic (GC) EHR. GC is a multispecialty group practice in central and northeastern Pennsylvania that includes 41 outpatient community practice sites, each of which offers primary care. GC primary care patients are similar to the population of the region in age, sex, and race. GC has used EpicCare EHR since 2001 for all practice-based tasks including viewing test results, clinical messaging, order entry, and progress notes and for the storage and exchange of administrative and clinical data (e.g., appointment, admission, financial, clinical results, and dictations). Since 1993, GC has been serviced by a single laboratory company.

10.2.1 Selection of Cases and Controls

We used a nested case-control design to develop apredictive model. Subjects who met one of the following HF diagnosis criteria between January 1, 2003 and December 31, 2006 were identified.

(1) HF diagnosis appeared at least once on the problem list,
(2) HF appeared in the EHR for two outpatient encounters, indicating consistency in clinical assessment,
(3) At least two medications were prescribed with an associated ICD-9 diagnosis of HF, or
(4) HF appeared on one or more outpatient encounter and at least one medication was prescribed with an associated ICD-9 diagnosis for HF.

If a patient met one of these criteria, the date of diagnosis was defined as the first appearance in the EHR of a HF diagnosis with an order, on the problem list, or as a reason for visit or encounter diagnosis. To exclude prevalent cases, a patient was defined as an incident case only if he or she had at least 1 year of care with a GC primary care provider in which there was no previous documentation of a HF diagnosis.

10.2.1.1 Validation of HF Diagnosis Criteria

We completed a chart review of a random sample of 100 individuals who met the operational criteria for HF. A clinician supervised two research staff, each of whom independently reviewed the records of cases. We documented the first appearance and related date of all major or minor Framingham criteria for HF. Of the 100 HF cases, 86 met Framingham criteria for HF. Of the 14 cases that did not meet Framingham criteria, 7 had two minor criteria, 2 had one major criterion, 2 had one minor criterion, and 2 had no Framingham criteria. The documentation of Framingham criteria are expected to be incomplete for three reasons. Some criteria (e.g., circulation time of 25 s) are dated and simply not used. The physician may not be aware of all criteria or may simply adopt a practice of only using some of thesigns and symptoms. Finally, text documentation may take more time when compared to selecting a diagnosis from an EHR menu. For the 86 cases the assigned date of diagnosis was earlier for 52 when using operational criteria, earlier for 28 when using Framingham criteria, and the same for 8 cases when using either criterion.

We only included patients who were between ages 50 and 79 at the date of diagnosis. We also required that subjects had their first GC encounter at least 2 years prior to the diagnosis date, to ensure that there was sufficient prior data to use in the prediction model. We identified a total of 536 incident HF cases.

10.2.1.2 Selection of Controls

We selected up to ten eligible clinic-, sex-, and age-matched (in 5-year age intervals) controls for each incident heart failure case. Primary care patients who had no history of HF diagnosis before December 31, 2006, had their first GC office encounter within 12 months of the first office visit of a matching incident HF care, and had at least one office encounter 30 days before or any time after the HF diagnosis date, were eligible to be in the control group. In situations where ten matches were not available, all available matches were selected. For 81 % of cases, nine or ten controls were identified. A total of 3953 controls were included in the analytic file.

10.2.2 Feature Creation and Missing Value Handling

For this analysis, the focus was on detecting HF diagnosis 6 months or more before the actual diagnosis date. Thus, the **index date** was defined as the date 6 months prior to the diagnosis date. Controls were assigned the same index date as their matched case. Only values in the EHR that occurred on or before the index date were used in prediction modeling.

Variables from all domains listed in Table 10.3 were used in the analysis. In most cases, the most recent value prior to the index date was used as the time-dependent variable. However, more than one type of feature was used for many of the variables. For example, features of comorbidities included both an indicator of diagnosis (e.g., diagnosis of diabetes) and the duration (e.g., time since diabetes diagnosis). Other features were created from available variables, such as pulse pressure and the proportions of widened (i.e., >40 mm Hg) and narrowed (i.e., <30 mm Hg) pulse pressure measurements out of all physician visits. For utilization of health care, we created a variable that is a count of the number of physician visits in each of a sequence of six-month windows prior to the HF diagnosis date (or comparable date for controls). Finally, for each abnormal laboratory measure, we created a variable that is the time between the first abnormal lab measurement date and the 6-month prior diagnosis reference date.

As described in the previous section, whether or not a procedure (e.g., ECHO image) was ordered might be an important signal separate from the findings for the procedure. Further, the predictive model should depend on all information that was available at the time of the prediction. Thus, the hypothetical values of labs that were never ordered are not of practical use. Instead, we simply included an indicator that the test was ordered and the result of the test (if ordered) as features. Operationally, we include an indicator variable (e.g., indicator that hemoglobin a1c was ordered) and the interaction between the value and the order indicator (e.g., hemoglobin a1c value times hemoglobin a1c indicator variable). These features are always observed, as the interaction is equal to 0 if the test was never ordered [3].

A total of 179 unique features were created and considered for inclusion in the analysis.

10.2.3 Machine-Learning Models and Feature Selection

Three machine-learning methods were compared: logistic regression, support vector machine (SVM), and boosting. Logistic regression is a classical approach for predicting a binary outcome using many features. SVM transforms the original data variable space into a "feature space," which is a higher dimensional space. An advantage of this approach is that the search for a linear classification decision boundary may be easier in the higher dimensional feature space than in the lower dimensional input space. One popular machine-learning ensemble method is boosting. Boosting combines the outputs of many "weak classifiers" to produce a strong "committee." At each iteration, the weights are adjusted based on classification errors from the previous run (misclassified cases from the previous iteration get more weight at the next run). Thus, difficult-to-correctly-classify observations receive ever-increasing weight, and are thus most influential.

For logistic regression, variables were selected to minimize AIC or BIC. The L1-norm variable selection technique was used for SVM [5]. For AdaBoost, we used the variable importance scores to select features [6].

We used the generalized linear models step functions in *R* for fitting and variable selection in the logistic regression models. For the SVM models, we used the radial basis kernel with the default values from the kernlab *R* package. Finally, the AdaBoost package in *R* was used for boosting.

Models were compared using the area under the curve (AUC) under a ten-fold cross-validation analysis.

10.3 Results

Logistic regression and boosting performed similarly, with AUCs of about 0.77 and 0.75, respectively. SVM did not perform as well, with an AUC of just below 0.65. For logistic regression and boosting, an AUC of about 0.75 was achieved with between 10–15 variables in the model. Features commonly selected across methods and across the ten subsets of the data include past diuretic medication orders, diagnosis of atrial fibrillation, and presence of respiratory symptoms.

While these results were promising, the predictive power of the models could potentially be enhanced by including information that is not captured in structured fields. Thus, current work is underway to use novel text mining methods to extract key information from health records. We next describe this work.

10.3.1 Extension to Text Mining and Temporal Mining

We develop the text analytics for extracting Framingham symptoms using advanced text analytics tools [7] for basic text processing and for building application-specific dictionaries and grammars. The processing results were then inserted into a comprehensive text analysis pipeline, built within the open source Apache UIMA system (Unstructured Information Management Architecture) [8] which provides for acquisition of the clinical note texts and the following three major steps:

(1) Basic text processing encompassed application-independent analytics, including paragraph and sentence boundary detection, tokenization, dictionary look-up, morphological analysis, and part-of-speech tagging.
(2) Several dictionaries were created in an iterative fashion to recognize feature candidates. These dictionaries included words and phrases used to recognize Framingham risk criteria and other symptoms, grammar for annotating segment header words and phrases (e.g., “Patient History:”), grammar and supporting information for annotating potential mentions of the Framingham

Weight-Loss feature (e.g., "lost 15 pounds over three days"), and a set of grammar for creating NegatedContext annotations (e.g., "negative for pleural effusion or rales").

(3) The development of Text Analysis Engines (TAEs) built with UIMA included the following components: A segment annotator, based on the segment headers found by the dictionaries and grammar; a WeightLoss annotator, which filters WeightLoss mentions found by the dictionaries and grammar by checking for a diuretic treatment, and the proper ranges for the amount of weight lost and the time taken; an annotator which filters all candidate mentions by checking word co-occurrence constraints associated with each feature type and by checking constraints on which features can appear in which segment types; and a TAE which checks whether or not each feature appears in a NegatedContext and outputs the feature to a file of asserted features or denied features, accordingly.

In terms of evaluation, focusing on positive predictive value of asserted features, we selected 5 cases at random and extracted features from the resulting set of 784 text files, yielding a file of 703 asserted features. One cardiologist manually examined the resulting feature file and used the textual feature mentions and their sentential contexts to assign a "correct" or "incorrect" label to each extracted Framingham feature. A total of 93 % of the asserted features were identified as "correct" by manual review. We then processed the entire set of notes for the cases and controls.

The average number of features per case was approximately twice that of the controls, whereas the denial of features was only 9 % more common among cases. Further, the ratio of denied to asserted features was more than twice as high for controls as for cases.

Another extension for predictive modeling of HF is the ability to adopt frequent temporal sequences as features. Specifically, we are interested in determining whether there are important temporal sequences of events. We will leverage previous temporal mining work [9, 10, 11], to mine frequent sequential patterns from the longitudinal variables and use these patterns as features. The core idea here is to develop and apply algorithms that extract frequent sequential patterns of longitudinal events within and across patient cohorts. Once those patterns are identified, we can include them as additional features to build the predictive model and test the predictive power of the temporal features.

10.3.2 Challenges to Developing Scalable Solutions

Deploying predictive models for use in clinical care will eventually require a scalable and well-integrated solution that can be used in existing EHR systems. Typically, there are two major parts to the work :model building and model scoring. In model building, an optimal model is trained based on training data. In

model scoring, the trained model is deployed to score future data. Model building is frequently an offline process, and model scoring should be performed in real time. The challenge for model building is with the comlpex heterogeneous and high-dimensional EHR data and the need for efficient means of processing and preparing these data. To obtain a good model, the model needs to be constructed on a subset of all available data (training set) and tested on the remaining data (test set).

This training and testing process needs to be conducted multiple times to derive statistically valid results. The final deployed model can be the ensemble of all those models trained on different subsets. Or based on the training process, an optimal set of training parameters are obtained and then or used the optimal set to train another model on the entire data. In the context of HF prediction, the model building process typically involves the following steps:

(1) Feature construction: A comprehensive set of variables is derived from longitudinal EHR data (note: we use the terms features, variables, covariates, and characteristics interchangeably).
(2) Feature selection: Features that are most distinctive with respect to their impact on prediction results are identified, given a specific time interval before HF diagnosis and type of HF or other subgroup descriptors.
(3) Classification: Various machine learning methodologies are used for mapping features to labels (HF or not HF in the prediction time interval).
(4) Model evaluation: Statistical validation of the quality of the predictive models is provided. To really speed up the process, we often perform multiple runs of the model building process in parallel. With the help of parallel computation framework such as MapReduce, we can imagine to train a sophisticated model on a large EHR dataset within a short period of time (say in hours).

The resulting model can then be deployed to operational environment for model scoring on future data. Once there, a set of features will be constructed and scored by the model for the risk of HF when a new encounter occurs based on the patient's historical EHR information. Then, based on the score and confidence, different treatment protocols can be invoked. The challenge for model scoring is to conduct all the above in real time. To achieve real-time performance, the appropriate indexing and caching mechanism can be performed *a priori* to speed up the scoring process. The other option is to perform the computation in a scheduled fashion. For example, the system can be configured to conduct the scoring process automatically based on the doctor schedule and to cache the result in preparation for patient visits. During the visit, the pre-computed HF score and corresponding recommendation can be presented to the patient.

10.4 Conclusions

Our preliminary work in applying text analytics, data mining, and machine learning tools to clinical, laboratory, diagnostic, and other data routinely captured in the EHR has shown promising results for the early detection of HF. Future work in this area will include incorporating more features from unstructured data into the model and translating prognostic tools for clinical practice. An important aspect of this research will be in determining the value of these tools for improving patient outcomes and reducing healthcare costs.

References

1. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM (2007) Bridging the inferential gap: the electronic health record and clinical evidence. Health Aff 26:w181–91
2. Fonseca C, Oliveira AG, Mota T, Matias F, Morais H, Costa C, Ceia F (2004) Evaluation of the performance and concordance of clinical questionnaires for the diagnosis of heart failure in primary care. Eur J Heart Fail 6:813–820
3. Roy J, Hennessy S (2011) Bayesian hierarchical pattern mixture models for comparative effectiveness of drugs and drug classes using healthcare data: a case study involving antihypertensive medications. Stat Biosci 3:79–93
4. Wu J, Roy J, Stewart WF (2010) Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care 48(6 Suppl):S106–113
5. Zhu J, Rosset S, Hastie T (2003) 1-norm support vector machines. Neural Inf Proc Sys 2003:16
6. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, Stanford
7. IBM. Text analytics tools and runtime for IBM Language ware. Available at: http://www.alphaworks.ibm.com/tech/lrw2011
8. Apache UIMA. Available at: http://uima.apache.org/
9. Norén G, Hopstadius J, Bate A, Star K, Edwards I (2010) Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Disc 20:361–387
10. Mörchen F, Ultsch A (2007) Efficient mining of understandable patterns from multivariate interval time series. Data Min Knowl Disc 15:181–215
11. Wang W, Yang J (2005) Mining sequential patterns from large data sets, series. Adv Database Sys 28

Chapter 11
Rule-based Computer Aided Decision Making for Traumatic Brain Injuries

Ashwin Belle, Soo-Yeon Ji, Wenan Chen, Toan Huynh and Kayvan Najarian

Abstract This chapter provides an overview of various machine learning algorithms which are typically adopted into many predictive computer-assisted decision making systems for traumatic injuries. The objective here is to compare some existing machine learning methods using an aggregated database of traumatic injuries. These methods are used towards the development of rule-based computer-assisted decision-making systems that provide recommendations to physicians for the course of treatment of the patients. Since physicians in trauma centers are constantly required to make quick yet difficult decisions for patient care using a multitude of patient information, such computer assisted decision support systems are bound to play a vital role in improving healthcare. The content of this chapter also presents a novel image processing method to assess traumatic brain injuries (TBI).

A. Belle (✉) · K. Najarian
Department of Computer Science, School of Engineering, Virginia Commonwealth University, East Hall, Room 4248, 401 West Main Street, 843019Richmond, VA 23284-3019, USA
e-mail: bellea@vcu.edu

K. Najarian
e-mail: knajarian@vcu.edu

S.-Y. Ji
Department of Computer Science, Bowie State University, Building Suite 207, 14000 Jericho Park Road, Bowie, MD 20715, USA
e-mail: sji@bowiestate.edu

W. Chen
Department of Biostatistics, Virginia Commonwealth University, Seventh Floor, 830 East Main Street, 980032Richmond, VA 23298-0032, USA
e-mail: chenw6@vcu.edu

T. Huynh
The Department of General Surgery, Division of Trauma, Surgical Critical Care and Acute Care Surgery, Carolinas Medical Center, 1000 Blythe Blvd, Charlotte, NC 28203, USA
e-mail: toan.huyhn@carolinashealthcare.org

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_11,

11.1 Background

According to a Center for Disease Control (CDC) report in 2010, approximately 1.7 million new cases of traumatic brain injury (TBI) are reported annually [1]. Of which nearly 52,000 of these cases results in death and amongst those who survive many suffer permanent disabilities. TBI is a contributing factor to a third (30.5 %) of all injury-related deaths in the United States. Almost half a million (473,947) emergency department visits for TBI cases are made annually by children aged 0–14 years, of which a significant percentage suffer from neurological impairment [2]. Reports also suggests that traumatic brain injuries are the most expensive affliction in the United States, direct medical costs and indirect costs such as lost productivity of TBI totaled an estimated $60 billion in the United States in 2000 [3].

TBIs usually occur due to specific causes and since their methods of treatment are well established, long-term disabilities and fatal complications can be reduced with the use of computer-aided systems. Decision making and resource allocation for trauma care can also be significantly improved with such systems since they are less subjective and more precise [4]. Furthermore, research suggests that the cost of trauma care can be greatly reduced by utilizing an inclusive trauma care system which focuses on computer-aided resource utilization [5].

The treatment of traumatic brain injuries demands optimal and prompt decisions making which in turn can increase the likelihood of patient survival [6, 7]. Since these injuries are highly time-sensitive, providing the ability to predict the possible duration of stay in the Intensive Care Unit (ICU) can be an important factor when deciding the means to transport a patient (i.e., ambulance or helicopter) from the scene of the accident to the hospital. Patients with critical injuries stand to benefit the most from helicopter transportation since they require immediate medical care are expected to spend more time in intensive care units. In fact Cunningham's [8] comparison of the outcome of treatment given to trauma patients suggests that, patients in critical condition are more likely to survive if transported via helicopter. However, due to the high costs of helicopter transportation, resource allocation for all cases becomes an issue [9, 10].

In trauma medicine, several computer-assisted decision-making systems are currently available. Such systems are typically designed to perform statistical survey based on patient demographics found within trauma databases [11, 12]. However these systems are usually not sufficiently accurate or specific for practical implantation. Some computer based decision making systems also use neural network for prediction [12–15]. However, since the inner workings of such neural network based systems are unknown, the reasoning behind their predictions and recommendation decisions are not transparent. Currently multiple issues inherent with such computer assisted diagnostic systems prevent its wide-spread use in trauma centers. Some of the main reasons being: the use of 'black-box' like methods, such as neural networks; the lack of a comprehensive database integrating all relevant patient information for specific prediction processes; and

exclusion of relevant attributes and the inclusion of irrelevant ones in developing predictions specific to a certain task, resulting in rules that are clinically not meaningful or unnecessarily complicated.

There are several common machine learning algorithms which are utilized towards medical applications. These include support vector machines (SVM), and decision tree algorithms such as C4.5 and Classification and Regression Trees (CART). Boosting is also sometimes utilized for improving classification accuracy. Although these algorithms perform relatively well for medical applications, they seem to suffer with large feature sets containing numerous attributes, as they have limited success in separating and identifying the important variables with respect to the specific application. This suggests that to better recognize and understand patterns in medical data, machine learning concepts needs to be combined with a method to identify the most correlated sets of attributes thereby being able to create more reliable rules for predictions.

The literature of biomedical informatics reinforces the benefits of employment of machine learning. This can be seen in research such as by Andrews et al. [16] where decision tree and logistic regression methods are used to compare and analyze similarities and differences between different medical databases. Kuhnert's research [17] emphasizes that methods, such as multivariate adaptive regression splines and CART which are nonparametric based, can provide more informative models. Signorini et al. [18] designed a simple model containing variables such as age and Glasgow Coma Scale (GCS); however, due to the limited number of variables, the reliability of the generated rules may be questionable. Hasford [19] compares CART and logistic regression, and finds that CART is more successful in outcome prediction than logistic regression alone. Guo [20] on the other hand finds that CART is more effective when it is combined with the logistic model.

Therefore, combining both statistical techniques and machine learning [21, 22] could be a more promising approach towards creating more accurate and reliable rules generating systems for decision making. In this chapter the performance analysis of several combinations of logistic regression and machine learning algorithms is discussed, in particular, the focus is on the extraction of significant variable that aid in the generation of reliable predictions. In the interest of comparison, other methods (such as neural networks) are also compared with the transparent rule-based systems. The research study presented in this chapter is based on a previously published article in BMC Medical Informatics and Decision Making 2009 [23]. The final outcomes, such as home/rehab, alive/dead or predicted length of stay in ICU are predicted based on a developed computational model. In addition, factors and attributes which affect the decision making process the most during the treatment of traumatic injuries are also identified.

Table 11.1 On-site dataset

Variable	Possible values
Gender*	(Male, Female)
Blunt*	Blunt, penetrating
ChiefComp*	MVC, fall, pedestrian, motorcycle crash, etc.
Position*	Passenger, driver, cyclist, motorcycle passenger, etc.
Age	Patient's age
FSBP (Initial Blood Pressure)	$0 \leq \text{FSBP} \leq 300$
GCS (Glasgow Coma Score)	$3 \leq \text{GCS} \leq 15$
ISS (Total Injury Severity Score)	$0 \leq \text{ISS} \leq 75$
Pulse	$0 \leq \text{Pulse} \leq 230$
Respiration rate	$0 \leq \text{Respiration} \leq 68$

Categorical variables are starred (Table Source [23])

11.2 Description of Data Used

For this study, an aggregated database of Traumatic Brain Injury cases was used. This database primarily contains three different sources of datasets, on-site, offsite, and helicopter. The datasets were provided by the Carolinas Healthcare System (CHS) and the National Trauma Data Bank (NTDB).

11.2.1 On-Site Dataset

The on-site dataset as the name suggests contains patients' data captured at the site of the accident. Decision making based on the limited variables available at the scene an accident is particularly difficult and critical, especially due to the incomplete access of important patient information such as pre-existing conditions (comorbidities), demographic information, etc. Therefore decisions under such critical circumstances must be made without such key information and certain physiological measurements, which usually are collected only after the arrival at the hospital. In Table 11.1, four categorical and six numerical attributes collected for this dataset is presented.

11.2.2 Off-Site Dataset

The off-site dataset is a more comprehensive dataset with several variables including additional information on comorbidities and complications. Both categorical and numerical attributes are included as inputs. A total of 1589 cases are included in the database of which 588 fatal and 1001 non-fatal. Here predicted outcomes are defined by either 'alive' or 'dead' which refers to the patient's

Table 11.2 Off-site dataset

Variables	Alive	Dead	Rehab	Home
Cases	1001	588	628	213
Male*	704 (70.3 %)	404 (68.7 %)	443 (70.5 %)	150 (70.4 %)
Female*	297 (29.7 %)	184 (31.3 %)	185 (29.5 %)	63 (29.6 %)
Age	41.2 ± 19.6	49.2 ± 24.1	39.6 ± 19.3	37.2 ± 16.6
FSBP	126 ± 33.4	119.3 ± 45.6	125.3 ± 31.6	124.5 ± 34.1
FURR	15.3 ± 10.9	13.9 ± 11.9	14.4 ± 11.1	18.2 ± 10.5
GCS	8.7 ± 5.3	27.5 ± 5.2	7.9 ± 5.2	10.5 ± 5.1
ISS	30.5 ± 12.8	35.3 ± 14.7	32 ± 13.2	27.1 ± 11.7
EDEYE	2.4 ± 1.4	2.1 ± 1.4	2.2 ± 1.4	2.8 ± 1.4
ED verbal	2.7 ± 1.8	2.3 ± 1.7	2.4 ± 1.8	3.3 ± 1.8
EDRT	4.6 ± 3.2	3.8 ± 3.3	4.1 ± 3.3	5.7 ± 2.89
Head AIS	3.0 ± 1.6	3.6 ± 1.6	3.1 ± 1.8	2.5 ± 1.4
Thorax AIS	2.3 ± 1.7	2.4 ± 1.8	2.3 ± 1.8	2.4 ± 1.7
Abdomen AIS	1.1 ± 1.5	1.1 ± 1.6	1.0 ± 1.5	1.5 ± 1.7
Intubation*	Yes/No			
Prexcomor*	Values: Acquired Coagulopathy, Chronic Alcohol Abuse, Chronic Obstructive Pulmonary Disease, Congestive Heart Failure, Coronary Artery Disease, Coumadin Therapy, Documented history of Cirrhosis, Gastric or Esophageal Varices, Hypertension, Insulin Dependent, Myocardial infarction, non-insulin Dependent, obesity, Pre-existing Anemia, Routine Steroid Use, serum Creatinine >2 mg % (on Admission), Spinal Cord Injury			
Complications*	Acute Respiratory Distress Syndrome (ARDS), Aspiration Pneumonia, Bacteremia, Coagulopathy, Intra-Abdominal Abscess, Pneumonia, Pulmonary Embolus			
Safety*	Seat belt, none used, air bag deployed, helmet, other, infant/child car seat, protective clothing			

First, the number of Cases in each group (Alive, Dead—and within the surviving patients, Rehab and Home) is listed. For the numerical attributes the table provides Mean ± standard deviation. Finally, the categorical variables are listed with their possible values. ISS provides the overall injury severity score (ISS) for patients with multiple injuries, and GCS is the Glasgow Coma Score. Many studies make heavy use of GCS and ISS, as these measures are considered standard metrics in assessing patient condition and degree of injury. Note that surviving patients who were transported to other hospitals are not included in the rehab + home total (Table Source [23])

survival outcome. Amongst those predicted to survive, a more exact outcome is further classified into categories of either 'rehab' or 'home'. Table 11.2 presents the attributes of the off-site dataset. Here "Prexcomor" is a categorical dataset which represents any comorbidities that could possibly have negative impact on the patient's chances of recovery from injury and any complications. Other terms used are defined in the table description.

Table 11.3 Helicopter dataset

Variable	Severe (ICU stay > 2 days)	Non-severe (ICU stay ≤ 2 days)
Cases	301	196
Male	201 (66.8 %)	132 (67.3 %)
Female	100 (33.2 %)	64 (32.7 %)
Age	30.6 ± 16.6	32.9 ± 17.2
FSBP	137.7 ± 23.2	127.6 ± 28.0
ISS	14.2 ± 8.1	23.7 ± 9.47
Pulse	101.4 ± 22.3	108.2 ± 26.6
Resp rate	15.6 ± 9.44	6.45 ± 10.6
ISS-HN	2.83 ± 0.86	3.46 ± 0.91

The number of severe and non-severe cases is listed, along with the percentages of each that are male and female. Mean ± standard deviation is given for each numerical attribute (Table Source [23])

11.2.3 Helicopter Dataset

This dataset comprises of data accumulated from those patients who were transported to a hospital in a helicopter. The variables are cheifcomp (the type of injury), age, gender, blood pressure, prefluids (the amount of blood provided to the patients), airway (the type of device used to assist patients with breathing), GCS, heart rate, respiration rate, ISS (Injury Severity Score) and ISS-Head and Neck. In particular Age, blood pressure, Glasgow Coma Scale (GCS), heart rate, Injury Severity Score (ISS), ISS-Head and Neck, and respiration rate are classified as numerical variables. The number of days spent in ICU is considered as the most informative measure when deciding the means of transport to the hospital, hence this is used as the final outcomes measure. In this dataset, duration of ICU stay attribute ranges between 0 and 49 days. When relatively small dataset with multiple outcomes are used for such predictions the resultant model generated may become unnecessarily complex and hard to understand. Hence, to eliminate complexity, the dataset is classified into just two groups, non-severe and severe, as done so by Pfahringer [24]. The non-severe group contains patients who stayed in the ICU less than 2 days while the severe group consists of patients who stayed in the ICU more than 2 days. This threshold was developed after obtaining feedback and consultation from trauma experts. In total, the dataset contains 497 cases: 196 severe and 301 non-severe. Table 11.3 provides the details of the helicopter dataset.

11.3 Special Topic: A Novel Method for Assessment of Traumatic Brain Injuries

Typically, traumatic brain injuries (TBI) often cause changes in the size and position of the ventricular system inside the brain. However the severity of TBIs can be characterized by the shift in the midline of the brain. Being able to identify

the midline shift may allow us to predict the intracranial pressure (ICP) to some extent. Having an estimate of ICP is important when treating TBI patients, since elevated ICP often results in secondary injuries in TBI causing potentially deadly consequences such as ischemia or herniation, which can be fatal if unrecognized or untreated.

A standard and accurate method of monitoring ICP is inserting pressure sensors inside ventricles through a brain surgery, which may cause infection and both short-term and long-term damages to the brain. Hence non-invasive measurement of ICP can be vital and useful for TBI cases.

The method described here for assessment of TBI, is based on a novel framework for automated midline shift measurement using CT (Computer Tomography) images [25]. A brief description is presented here, further details of this method can be found in [25].

11.3.1 Dataset

The testing CT dataset was provided by the Carolinas Healthcare System (CHS). All subjects in this dataset were diagnosed to have either mild or severe cases of TBI upon admission into the hospital. The dataset has 40 patients, comprising of 391 axial CT scan slices which show ventricles or region that should have contained ventricles.

11.3.2 Method

The midline shift measurement can be divided into three steps as seen in Fig. 11.1. The patients CT scans are taken and first the ideal midline of the brain, i.e., the expected midline before injury, is found via a hierarchical search based on skull symmetry and tissue features. Second, the ventricular system is segmented from the brain CT slices. Third, the actual midline is estimated from the deformed ventricles by shape matching method.

The horizontal shift in the ventricles is then calculated based on the estimation of ideal midline and the actual midline in TBI CT images. After the midline shift is successively estimated, features including midline shift, texture information of CT images, as well as other demographic information are used to predict ICP.

11.3.3 Detection of Ideal Midline

Increased ICP can cause the actual midline of the brain to shift from its original position. There are two important purposes for computing the ideal midline—i.e.,

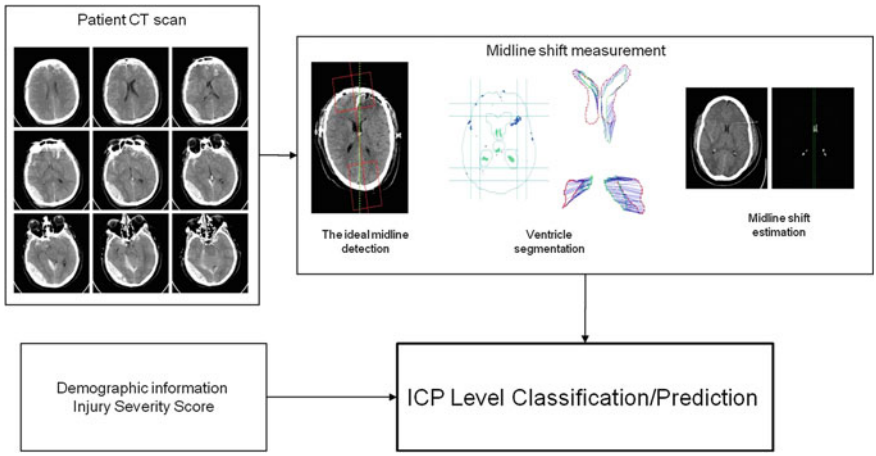


Fig. 11.1 Methodology overview

the midline as expected in a healthy brain with ICP under normal levels. The first purpose is that the ideal midline can be utilized as a reference line to measure shifts in the brain tissue from its ideal position. The second purpose for computing the ideal midline is to utilize it for calibrating each scan based on head orientation and rotation, since depending on the patient's position there can be variations in how the head rotation is captured between different CT scans.

Usually the symmetry of the brain can be used as a feature to roughly approximate the ideal midline. However, for a more accurate detection of the ideal midline some anatomical features must be considered in the computation. Certain anatomical features of the skull do not get affected by shifts in the midline of the brain. Therefore these features such as the falx cerebri fold present in the lower part of the skull and the bone protrusion in the upper part of the skull can be used effectively in locating or computing the ideal midline. Although the ideal midline can be roughly approximated using the symmetry of the brain, the anatomical features must be considered for more accurate detection. The ideal midline detection method has three steps:

1. Detect the approximate midline using symmetry.
2. Detect the falx-cerebri and anterior bone protrusion.
3. Use these features to refine the midline position.

At first, each slice is processed independently to detect the ideal midline. To compensate the certain inaccuracies in computing individual slices an adjustment is applied across the identified midlines of all scans.

11.3.4 Ventricular System Segmentation

In CT scans, the ventricle system is typically seen as a darker color as compared to other tissue matter. However, CT scans usually also contain noise (which spans across different grayscale values) in tissue areas which could be misrepresented as ventricles. Another challenge usually faced during CT analysis is the blurriness between different tissue structures. Hence to combat these issues the ventricular system segmentation process is separated into two parts. First an initial low level segmentation method is applied to group pixels into different parts. Specifically, Iterated Conditional Modes (ICM) [26] and Maximum A posteriori Spatial Probability (MASP) [27] algorithms are adopted for low level CT brain segmentation. Next a high level template matching is used to identify ventricles from segmented result. The template matching is employed to further identify ventricle areas.

11.3.5 Midline Shift Estimation

Using the segmentation information of the ventricles, the actual midline can be computed as the line that lies between the left and the right lateral ventricles. The results of the ventricle segmentation is a in the form of a binary image (with ventricle regions considered as object and non-ventricle regions as background). With the binary format, the only information which can be obtained is usually the shape information. In order to identify separate regions of the ventricles and estimate the actual midline, a mapping is built between the segmented shapes and the standard ventricle template with its respective annotated information. Using the slice of the bilateral ventricle, the point that lies in the middle of the edges of the left and right lateral ventricles is computed. This mid-point is then used to estimate the midline which separates the left and the right sides. The midline shift is then calculated using the estimated actual midline and the ideal midline computed from the first step.

Finally once the midline shift has been estimated, other features are extracted from the CT images such as texture analysis, estimation of blood amount, etc. Using these extracted features as well as demographic information of the patient, the ICP levels are predicted. Further details on feature extraction and ICP calculation can be found in [25].

11.4 Comparative Analysis

In trauma cases, the treatment outcomes for patients with similar conditions may turn out to be significantly different. Hence recognizing patterns in trauma cases is not very straight forward. Linear methods sometimes have proven insufficient for

pattern analysis even for trivial cases. Therefore because of the under-performance of linear regression methods for computer-aided trauma systems, non-linear techniques are encouraged for such applications [28]. Amongst non-linear techniques, neural network has been a popular choice, however due to their non-transparent nature the learning structure and weights of the trained network model remain unknown [29]. Some methods do exist which are capable of extracting approximate rules to represent this hidden knowledge; however, they fail in truly representing the trained networks [30]. Machine learning techniques such as AdaBoost and Support Vector Machines (SVM's) also hide the content of the trained network. The knowledge of the developed model for these techniques is also not visible, which is an extremely important requirement in medical applications.

Hence the use of certain rule-based methods such as C4.5 and CART can be useful for this purpose. Since these rule-based machine learning methods utilize some nonlinear capabilities while still providing transparency within the decision making process.

A brief description of each of the machine learning technique used in this comparative study has been given below.

11.4.1 Learning Algorithms

11.4.1.1 Neural Networks

An artificial neuron is an information processing and learning paradigm which was largely inspired by the biological processes of the human brain. A neural network is a network of artificial neurons. It is basically composed multiple processing elements or neurons which are highly interconnected to form a dense network. These neurons within the network work in alliance to solve specific problems.

A neural network learns by examples, it processes training examples individually which it then compares its initial classification (usual arbitrary) of the input with the true classes of the input examples. In particular, neural networks based on Radial Basis Function (RBF), are ideal for solving pattern classification problems. This is due to their capability for faster learning and their simple topological structure.

A standard RBF network comprises of a feed-forward back propagation neural network which is supervised. This network has an input layer, a hidden layer and an output layer. The families of Gaussian functions are usually most popular basis functions used within the hidden layer of the network. The outputs of these Gaussian functions are inversely proportional to the distance from the neuron center.

Given a finite set of training data $\{(x_j, y_j)|j = 1, \ldots, m\}$, where c_i is the center vector of the basis function, the equation for a standard output is as follows:

$$y_i = \varphi(x) = \sum_{i=1}^{N} \alpha_i \rho(||x_{j-} c_i||). \tag{11.1}$$

where N is the number of neurons within the hidden layer, α_i is the set of weights which minimize the least square between approximate output and the real output.

Typically the basis function used is given by the Gaussian activation function which produces the radial function representing the distance between each pattern vector and each hidden unit weight vector. This function is given as:

$$\rho(||x_i - c_i||) = \exp\left(-\frac{||x_i - c_i||}{\sigma^2}\right) \tag{11.2}$$

Here the neuron radius is given by σ [31, 32]. The weight of each neuron in RBFs is calculated using their distance in the feature space.

11.4.1.2 Support Vector Machine

Support Vector Machines (SVMs) [33] are methods based on supervised learning and are primarily used for classification. A SVM splits its input data into two sets of vectors positive and negative examples, in an n-dimensional space. Using this space the SVM algorithm computes an optimal hyperplane that maximizes the distance between the two vector set [34]. SVM's are typically used in solving problems such as image classification, text categorization, cancer data classification, protein analysis, and hand writing recognition. SVM works well for such applications due to its ability to handle large feature spaces [35].

Consider a labeled training example set of size N. $D = (x1, y1), \ldots, (xn, yn)$ with $yi \in \{+1, -1\}$ and $x \in R^d$, where d is the dimensionality of the input. Let $\emptyset : R^d \rightarrow F$ represent a function that maps the input space to the feature space. Here the SVM algorithm finds a hyperplane (w, b) that maximizes the margin between the two classes given that the classes are linearly separable.

$$\gamma = \min_i\{y_i < w, \emptyset(x_i) > -b\} \tag{11.3}$$

where b is a real number or the bias term and w and F have the same dimensions. For an unknown input vector x_j, classification would be defined by:

$$f(x_i) = sgn(y_i < w, \emptyset(x_i) > -b) \tag{11.4}$$

This minimization occurs when $w = \sum_i \alpha_i \gamma_i \emptyset(x_i)$, where α_i represents the strength of training point x_i and is a positive real number in the final classification decision. The set of points closest to the hyperplane is the set of points with non-zero α_i. These non-zero points represented the actual support vectors. Despite SVM being computationally expensive, its ability to handle large feature spaces makes it popular for application in many real world problems [35].

11.4.1.3 Adaptive Boost (AdaBoost)

AdaBoost, which stands for Adaptive Boosting, was introduced by Yoav Freund and Rovbert Schapire [36]. It is a machine learning algorithm which uses a linear combination of weak classifiers to construct a robust classifier. The AdaBoost method works by repeatedly calling a particular learning algorithm which is perceived to be weak in a set of rounds $t = 1, \ldots, T$. In this method for the training set a weights distribution is maintained, where $D_t(k)$ is the distribution's weight for a particular training example k in round t. The aim here is to find a good weak hypothesis for a weak learner such as $h_t : X \rightarrow \{-1, +1\}$ for the distribution D_t, where the goodness measure is computed by the error of the hypothesis with respect to D_t. Following which D_t is updated by increasing the weights of the examples that were incorrectly classified. In contrast those examples which were classified correctly are given lesser weights. This forces the weak classifier to focus on more difficult training examples.

To highlight the importance of h_t, AdaBoost selects some parameter α_t. After the completion of all the rounds t, the final hypothesis H is denoted by a weighted majority vote of all weak hypothesis T. Similar to other boosting algorithms it has been proven that if each computed hypothesis is at least slightly better than random, then the training error reduces at an exponential rate. However, Adaboost is also able to adapt to the error rates of individual weak hypotheses, so each subsequent classifier is adjusted in favor of examples mislabeled by previous classifiers [37].

11.4.1.4 Classification and Regression Tree

Classification and Regression Tree (CART) has become a popular and fundamental method in building statistical model from simple feature set. Binary decision trees were invented by Leo Breiman and his colleagues who named them classification and regression trees or in short CART [38]. CART is powerful due to its ability to deal with multiple types of features (enumerated sets, floats, etc.) both in terms of input features and predicted features. It can also deals with incomplete data (i.e., missing value) and the output generated trees produces rules which are humanly readable. CART applies information theoretic concepts to create a decision tree. These decision trees contain binary questions at each node of the tree (e.g. "Is this patient's ISS $>$ 15" with Yes/No answers). This allows rather complex patterns to be observed within a given data, and generate expression in the form of transparent grammatical rules [39]. CART's nonlinear extensions makes it highly suitable for application in machine learning and data mining, due to the efficiency of the algorithm in dealing with a variety of data types [40] and missing data. In cases of missing data, CART simply uses a substitution value which is a pattern similar to the best split value in the node [38].

Furthermore, the basic CART building algorithm is a greedy algorithm, which chooses the best discriminatory variable at each stage locally in the process. To

find the optimal splitting rules of each node, CART performs an exhaustive search of all the split values and the variables. Therefore, making it easy to recognize which variables are important to predict the outcomes. The splitting then stops at the pure node containing fewest samples.

11.4.1.5 C4.5

C4.5 is a statistical classifier algorithm developed by Ross Quinlan [41–43]. It is based on Quinlan's basic ID3 decision tree algorithm [44]. C4.5 is a more computationally efficient algorithm which is successful in avoiding overfitting issues and is also able to handle cases with continuous variables. The C4.5 algorithm generates rules by employing a divide and conquer method to data used for training into regions of disjoint variable space. The different variable space regions are created using the pre-assigned target labels of the training set [9]. Using the gain criterion the C4.5 algorithm performs a split on the best attribute. The gain criterion is based on the measure of randomness also known as entropy of the class distribution within the dataset. The greatest difference in entropy of the class probability distribution of the current subset S and the subsets generated by the split becomes the criterion.

$$Info(S) = -\sum_{i=1}^{n} p(k_i, S) \cdot \log_2 p(k_i -, S) \tag{11.5}$$

where $p(k_i, S)$ is the relative frequency of examples in S that belongs to class k_i. The split that reduces this value the most is considered as the best split. The output of the algorithm is a decision tree. Such decision tress can be further represented as a set of symbolic IF-THEN rules.

Note: The datasets contain variables of nominal categorical, such as gender and complication type. Gender is replaced by a binary variable where 0 is for male and 1 is for female. Every nominal value is coded from Yes/No to a binary value of 1/0 respectively. These values are also treated as individual attribute. Ten-fold cross-validation is used to measure the generalization quality and scalability of the rules. Each dataset is divided into 10 mutually exclusive subsets [45], and in each stage 9 of these subsets are used for training while the remaining subset is used for testing. Therefore, in this manner ten different trees are formed for each dataset.

11.4.2 Rule Performance Metrics

Upon having generated a variety of rules, the performance of each rule is measured as the probability of correct prediction. Assume that D is a dataset with each instance given by (x_i, y_i), where y_i is the actual survival outcome. Let D_r be the

training set, and a subset $D_t \in (D/D_r)$ be used for testing. The performance of the rule is calculated as:

$$acc_R = prob\left(y_i = y^R | (x_i, y_i) \in D_t\right) \tag{11.6}$$

where the outcome produced by induction is given by y^R which is the expected classification. The accuracy of the rule is measured by the number of positive matches in the testing set. The accuracy of the rules can also be estimated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11.7}$$

where the number of true positives is given by *TP* and the number of false positives is given by *FP*. *TN* is the number of true negatives and *FN* is the number of false negatives. Quality of these rules is then judged by the values computed by the sensitivity and specificity equations. Since these measures calculate the probability of false positives and false negatives separately, they can be very useful. When one of these measures is found to be considerably higher than the other, it may not be observed in a single average error measure. The equations of these measures are as follows.

$$Sensitivity = \frac{TP}{TP + FN} \tag{11.8}$$

$$Specificity = \frac{TN}{FP + TN} \tag{11.9}$$

In this application, high sensitivity is more important than high specificity. When patient lives are at stake—for example, while deciding the type of transportation to be provided for the patient—false positives are preferable to false negatives, despite the increased financial cost.

11.4.3 Improving Rule Quality

After having extracted the most valuable rules, to improve the quality of the rules direct maximum likelihood estimation with logistic regression is used. The expected probability of a dichotomy in the logistic function is calculated as:

$$\pi_i = pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}} \tag{11.10}$$

where X_i are variables with numeric values, Y is the outcome where 0 or 1 represents either alive/dead respectively, and the β's are the regression coefficients that quantify the contributions of the numeric variables to the overall probability [20].

Logistic regression provides knowledge of the relationships and strengths among the response variable and the multiple independent variables. It does not require any distribution on the independent variables; they do not have to be normally distributed. Furthermore, Logistic regression does not require linear relation or equal variance within each group. However, odds ratio is the most important interpretation from logistic regression, which measures the strength of the partial relationship between an individual predictor and the outcome event [46].

Logit function was introduced by Joseph Berkson in 1944. The logit function can be used as a special case of a link function in a generalized linear model of logistic regression: it is the canonical link function for the binomial distribution. Since the relationship between the logit and the predictors is linear, it is advantageous to use the logit scale for interpretation. Residual analysis and scatter plots are used to check the linearity assumption. It is seen in the results that there exists a linear relationship for all variables, although in comparison some relationships were found to be weaker than the others.

For sake of brevity, the results for only two variables: Head AIS and Age are presented. The scatter plot between the logit and its predictor is presented; also using regression analysis the residual plot between them is also given. A random variation without nay recognizable pattern can be expected from the residuals in the case where the linearity assumption is satisfied. In cases where a curve formation is observed in the residual plot, it can be assumed that there may be a nonlinear relationship in the variable. Statistical Analysis Software or SAS was the tool used to perform this analysis. Figures 11.2 and 11.3 present the scatter plots and residual plots using Age and Head AIS as the predictors for patient survival.

In the plots showing residuals against the predictors, if a curvature of some form is seen then a quadratic term should be used for testing the statistical significance thereby suggesting improved versions of the model. In fact, the quadratic term should also be included if the coefficient for the quadratic term is found to be significant.

Although the model presented here does not depict any string curvature, for the purpose of validating the results the Head AIS variable is tested using a quadratic term. The model is as follows:

$$logit = \alpha + \beta x + \gamma x^2 \tag{11.11}$$

where the intercept term is given by α, the parameter of the predictor is given by β, and the parameter of squared predictor is given by γ. For the Head AIS variable, the estimate of β is -0.1820 (p value $= 0.0015$), and the estimate of γ is -0.0124 (p value $= 0.2058$). The p values here show that the Head AIS does not require a term which is quadratic. Because of which, there exists a linear relationship between the logit and its predictor.

For the purpose of testing the significance of individual variables a comparison is performed between a full model using log likelihood test and a reduced model that drops one of the independent variables. The likelihood ratio test alone does not

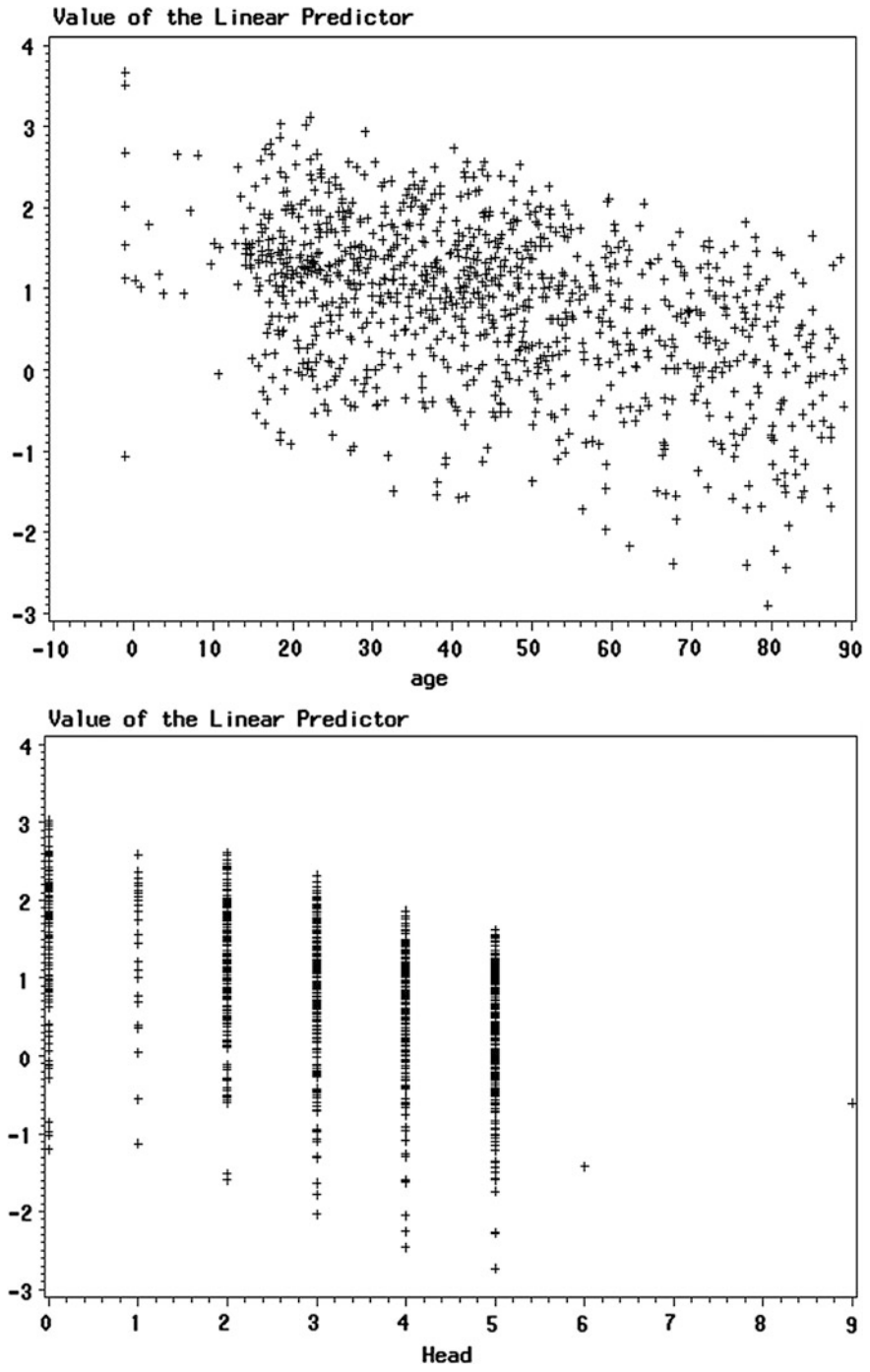


Fig. 11.2 Scatter plots of logits and predictors. This figure presents two scatter plots, used to demonstrate that the relationship between the logit and the predictors is linear. The first scatter plot is of logit versus age (a continuous variable), and the second is of logit versus head AIS (a discrete variable) (Figure source [23])

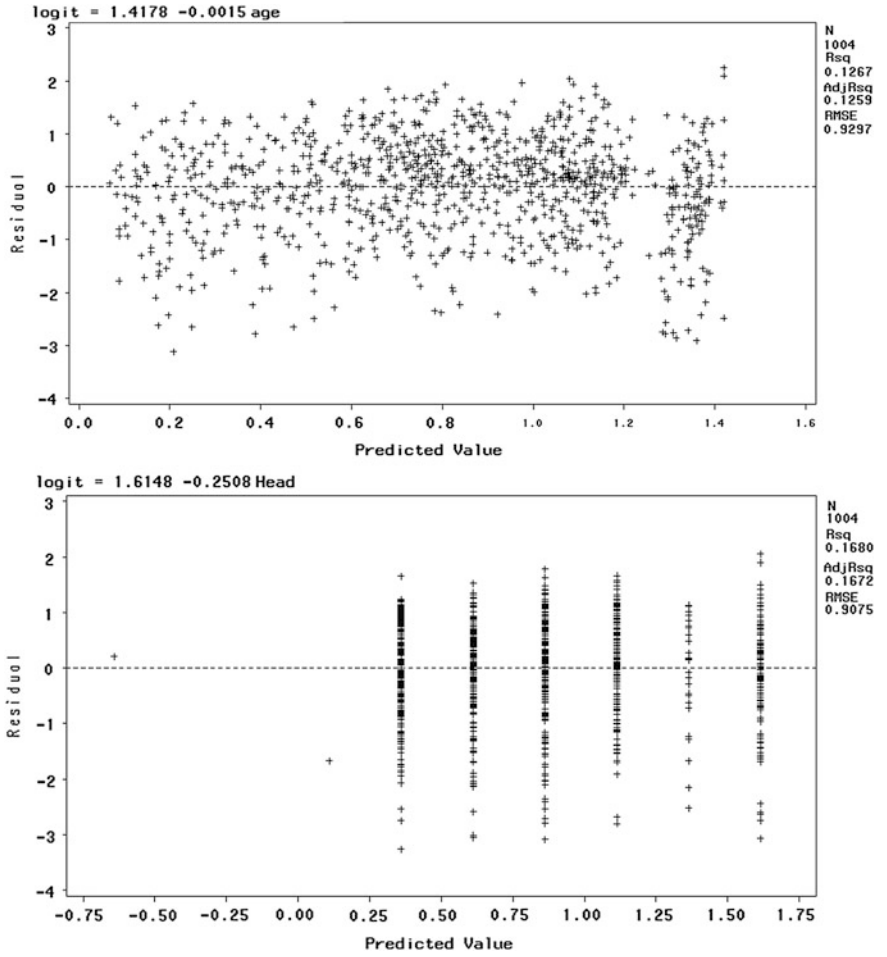


Fig. 11.3 Residual plots for logits and predictors. This figure presents two residual plots, used to demonstrate that the relationship between the logit and the predictors is linear. These plots were made using regression analysis. The first residual plot is between logit and age (a continuous variable), and the second is between logit and head AIS (a discrete variable) (Figure source [23])

reveal the importance of any particular independent variables over others. The difference between the results obtained for a nested reduced model which drops one of the independent variables and the results for the full model can be computed by estimating the maximum likelihood. An insignificant difference indicates that there was no effect on the performance of the model, thereby justifying the dropping of a particular variable. This is called directed MLE.

The test takes the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The resulting likelihood ratio is given by:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1) \tag{11.12}$$

If the Chi-square value for this test is significant, the variable is considered to be a significant predictor. Following these tests, only the significant variables (p value ≤ 0.05) are selected.

It is to be noted that other methods such as forward and stepwise model selections are also available for identifying the significance of individual attributes [17, 44]. The stepwise method considers all possible combinations of variables and is commonly used to find the best subset of variables for outcome prediction. However, the stepwise approach does not guarantee that the variables with the highest significance are always selected due to the repetition of insertion and deletion. For example, the variable 'age' may not be selected as an important variable; but physicians may consider that the patients' age is an important factor for deciding specific treatment options. Therefore, the use of directed MLE is preferred for such medical applications. Another reason for employing MLE is empirical; in previous study [10], it was found that the direct MLE method has a slightly higher accuracy in determining the significant variables as compared to stepwise or forward model selection. In this case as well, SAS was used as the statistical analysis tool to calculate the significance of each variable.

11.4.4 Constructing Reliable Rules

As mentioned previously, neural networks and SVM are not designed to produce any grammatical rules. Only C4.5 and CART methods designed for rule extraction. The variables which are recognized as significant are typically used as input variables to C4.5 and CART. Moreover, rules that are generated to adhere to only one or two examples may be considered to be unnecessarily specific for application to the entire population. Hence those rules with both a sufficiently large number of supporting examples and high accuracy are used to generate a rule base.

Note that AdaBoost, Neural Networks and SVM are still tested here for the reasons of performance comparison, despite the fact that they are not designed for generating rules. Due to the prevalence of these methods using them for comparison with C4.5 and CART algorithms helps in validating the stability and the accuracy of these rule based systems.

11.4.5 Analysis Results

The rule-based system employed in this study allows physicians and trauma experts to use these generated rules to predict the likelihood of patient's survival. Given the transparency of the reasoning behind these generated rules, the

physicians can also benefit by being able to allocate their resources in a more efficient manner. Following expert opinions from physicians, initially only rules with at least 85 % prediction accuracy on the testing set are included in the rule base, especially considering the total number of examples for training is rather small. However, rules with accuracy between 75 and 85 % are also incorporated.

There reasons for this are dual; firstly, the lack of accuracy of a rule may not be due to a flaw in the rule itself, rather it may be low due to the incompleteness of certain entities within the database. Secondly despite a rule having low accuracy, it might contain knowledge of unrecognized relationships amidst variables. As an example, almost all of the trauma experts consulted strongly suggested that a patient with an ISS score of over 25 has very little chance of survival. However a patient with a high ISS score but low thorax and head AIS score might have a higher probability of survival considering that prompt and appropriate medical treatment was provided to the patient. Rules with an accuracy lying between 75 and 85 % are usually considered as "supporting rules" for deciding and suggesting treatment options.

Average accuracy of survival prediction without the knowledge of any of the pre-existing conditions is around 73.9 %. However with the inclusion of the knowledge regarding pre-existing the accuracy increases to around 75.8 %. Hence for advanced prediction tests off-site data is also incorporated into the assessment since it has vital information about pre-existing conditions of the patients.

When using C4.5 and CART, the knowledge representing these conditions tends to appear at the highest levels of the tree. This indicates that they are important in the overall decision making process.

A good example of an important pre-existing condition would be coagulopathy or bleeding disorder. Patients with this disorder can have severe hemorrhage and thus could potentially be a life threatening condition. Therefore, the knowledge of the existence of this or similar disorders can be one of the most important factors to be considered for patients with TBI.

11.4.6 Significant Variable Selection

Identifying key variables in the dataset is essential in order to improve the rule quality and accuracy. Additionally, rules that are shorter and based on fewer and more significant variables are clinically useful for physicians. To accurately extract these key variables, direct MLE with logistic regression is used in the helicopter and off-site datasets. The results for the off-site dataset are presented in Table 11.4; here nine important variables are identified. To find the relationship between variables which can be expressed as a statistical model, Wald test is used. Wald Chi squares test is performed on each of the variable using standard deviations. The resulting odd ratios are identified as a strong relationship between the outcome and the independent variables. Table 11.5 presents the extracted significant and prominent variables from the helicopter dataset. As it turns out only five out of the eleven original variables are identified to be significant.

Table 11.4 Significant variables of off-site dataset

Variable	Coefficient	Walds χ^2	P value	Odd ratios	Mean ± SD
AIS Head	−0.58	23.61	<0.0001	0.56	3.25 ± 1.64
AIS Thorax	−0.13	4.37	0.003	0.88	2.33 ± 1.78
ID*	1.27	5.70	0.02	3.55	–
MI*	1.43	19.44	<0.0001	4.18	–
ARDS*	0.98	20.24	<0.0001	2.66	–
Cg*	0.63	24.96	<0.0001	1.88	–
Age	−0.03	29.22	<0.0001	1.03	44.15 ± 21.7
EDRTS	−0.27	4.94	0.03	0.77	12.1 ± 16.03
ISS	0.02	6.06	0.01	1.02	15.82 ± 19.03

Categorical variables are starred. *Cg* stands for Coagulapathy, *MI* for Myocardial Infarction, *ARDS* for Acute Respiratory Distress Syndrome, *ID* for Insulin Dependent, *EDRTS* for Emergency Department Revised Traume Score. (Table Source [23])

Table 11.5 Significant variable of helicopter dataset

Variable	Coefficient	Walds χ^2	P-value	Odd ratios	Mean ± SD
Age	−0.02	3.17	<0.0001	0.98	31.79 ± 17.5
Blood Pressure	0.01	2.85	0.01	0.01	129.45 ± 30.51
ISS-HN	0.01	0.003	0.25	1.11	3.00 ± 1.0
ISS	−0.14	36.47	0.02	0.87	19.56 ± 11.09

ISS stands for Injury Severity Score, *ISS-HN* for Head/Neck Injury Severity Score (Table Source [23])

In this study the scale of the data is small and several variables are unknown, so participating physicians assisted in identifying significant variables. Age, GCS, blood pressure, pulse rate, respiration rate, and airway were selected as the important factors.

11.4.7 Measuring Performance

The prediction results of five different machine learning methods are compared in Table 11.6. When only significant variables are used, the performance of each of the algorithm is very good. In addition, using only the most significant variables is shown to result in a more balanced testing-training performance. For physicians, being able to recognize and understand the reasoning behind decisions from such systems can be very useful. Especially when their decision matches the decision presented by the algorithm, their confidence in the system may be increased. Sometime, if the systems' reasoning is found clinically meaning or misguided, the physician can choose to disregard its recommendation. However, if the reasoning and decision of the system reveals some clinical merit, this may alert them to previously hidden factors affecting patient outcome.

Table 11.6 Performance comparison if five machine learning methods

	Logistic (%)	AdaBoost (%)	C4.5 (%)	CART (%)	SVM (%)	RBF NN (%)
All variables	69.4	70	68	75.6	73	67.2
Significant vars. only	72.9	73	75.2	77.6	79	79.04

The five chosen machine learning algorithms are AdaBoost, C4.5, CART, SVM, and RBF Neural Network (Table Source [23])

Table 11.7 Prediction results for outcomes and ICU days

	Logistic (%)	AdaBoost (%)	C4.5 (%)	CART (%)	SVM (%)	RBF NN (%)
Exact outcome	74.6	73	75.6	72	72.6	72.8
Days in ICU	80.6	78.7	77.1	77.4	80.1	77.4

This table compares the performance of logistic regression alone and the five chosen machine learning algorithms in predicting exact outcome (for off-site dataset) and ICU length of stay (for helicopter dataset) (Table Source [23])

Table 11.8 Performance comparison of AUC in ROC curve analysis

	Logistic (%)	AdaBoost (%)	C4.5 (%)	CART (%)	SVM (%)
All variables	63.7	63.1	58.1	60	64.5
Significant vars. only	66.9	67.5	63.2	64.6	67.6

Table Source [23]

Table 11.7 presents the performance accuracy in outcome prediction (rehabilitation or home) for the off-site dataset, and prediction of ICU days for the helicopter dataset. In both cases, only the significant variables are used. All available variables are not put to use, since the survival prediction test has already confirmed the improved performance when only significant variables are being utilized.

Receiver Operating Characteristic (ROC) curves is also generated in order to evaluate the model performance. ROC curves are plots of the true positive rate (sensitivity) versus the false positive rate (1-specificity). First, ROC analysis is performed on the patient survival prediction results. Table 11.8 compares the Area Under the Curve (AUC) for the ROC curves generated using all available variables and significant-only variables. Improvement is seen in the results when only significant variables are used in the model. Therefore, when dealing with the helicopter dataset, only ROC analysis is performed on the significant- variable-only model.

From Table 11.9, it can be seen that there does not exist any significant difference in ROC analysis results among the various machine learning methods. However, when the size of the given dataset is small—logistic regression outperforms the other methods as can be seen with the dataset used in this study for ICU days prediction. Figures 11.4 and 11.5 present sample ROC plots for logistic regression using only significant variables for survival and ICU days prediction respectively.

Table 11.9 ROC performance in exact outcome and ICU days' prediction

Variable	Logistic (%)	AdaBoost (%)	C4.5 (%)	CART (%)	SVM (%)
Exact outcome	76.8	76.4	71.9	71.5	68.7
Days in ICU	79.2	74.6	76.6	73	71.9

Table Source [23]

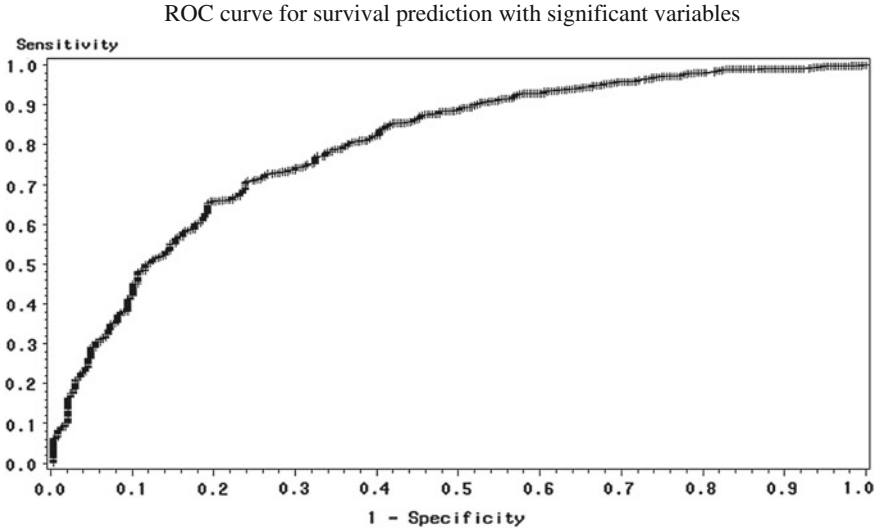


Fig. 11.4 ROC plot for Logistic regression on survival prediction. This figure presents the ROC plot obtained when applying logistic regression for survival prediction, using only significant variables. Tables 11.8 and 11.9 contain AUC (area under curve) results for the other machine learning methods and the other prediction scenarios (Figure source [23])

11.4.8 Constructed Database Using CART and C4.5

Using CART and C4.5 rule extraction algorithms, multiple rules were generated. Following discussion with trauma experts, it was identified that the robust rules are those with over 85 % accuracy. For survival prediction, the average rule accuracy using all available variables is 82, and 83.9 % when using only the most significant variables. Table 11.10 presents some examples of the most reliable generated rules for survival prediction (>85 % accuracy). Table 11.11 contains a few examples of survival rules with accuracy between 75 and 85 %. A more comprehensive list of the rules can be found in [23].

Table 11.12 presents some examples of the most reliable rules for generated outcome prediction (>85 % accuracy), and Table 11.13 contains examples of outcome rules with accuracy between 75 and 85 %. Finally, Table 11.14 presents some examples the most reliable generated rules for ICU days prediction (>85 % accuracy), and Table 11.15 contains ICU days rules with accuracy between 75 and

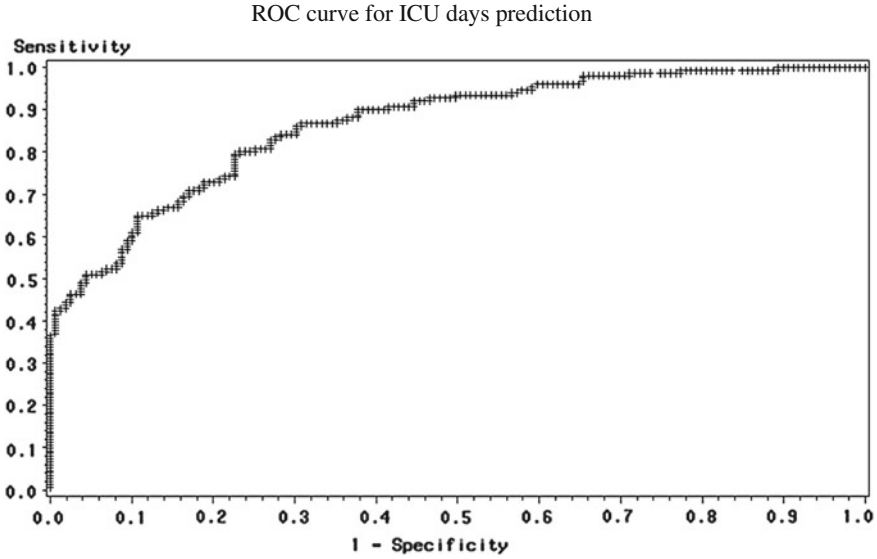


Fig. 11.5 ROC plot for logistic regression on ICU days prediction. This figure presents the ROC plot obtained when applying logistic regression for ICU day's prediction, using only significant variables (Figure source [23])

Table 11.10 Extracted reliable rules for survival prediction (>85 % accuracy)

Rules	Test accuracy	Method
(Cg = 'Yes') and HEAD < 2 and AGE < 76.65 Then Alive	29/34 (85.3 %)	CART
Cg = 'No') and (MI = 'No') and AGE < 61.70 and HEAD ≤ 4 and (ARDS = 'No')Then Alive	334/375 (89.1 %)	CART
(Cg = 'No') and (MI = 'No') and HEAD ≥ 5 and AGE < 22.35 Then Alive	55/64 (85.9 %)	CART
ISS ≥ 28 and (Cg = 'No') and THORAX ≤ 4 AND 62.25 ≤ AGE < 69.00 and EDRTS ≥ 2.88 Then Alive	10/11 (90.9 %)	CART
ISS ≥ 23 and (Cg = 'No') and THORAX ≤ 4 and 69 ≤ AGE < 72.35 Then Alive	13/15 (86.7 %)	CART
HEAD ≤ 2 and (MI = 'No') and (Cg = 'No') and AGE ≤ 62 Then Alive	182/206 (88.3 %)	C4.5
(MI = 'Yes') and AGE ≤ 62 and EDRTS > 5.39 and ISS ≤ 25 Then alive	19/209 (95 %)	C4.5
THORAX > 3 and HEAD ≤ 4 and (ARDS = 'No') and AGE ≤ 62 Then Alive	126/148 (85.1 %)	C4.5
THORAX ≥ 2 and EDRTS < 0.87 and ISS > 38 Then Dead	12/13 (92.3 %)	C4.5
(MI = 'Yes') and AGE > 82.6 Then Dead	16/18 (88.9 %)	C4.5
(MI = 'Yes') and ISS > 30 Then Dead	45/50 (90 %)	C4.5
HEAD > 4 and (MI = 'Yes') Then Dead	25/27 (92.6 %)	C4.5
(Cg = 'Yes') and HEAD ≤ 4 and AGE > 78 Then Dead	12/14 (85.7 %)	C4.5

Reliable rules are defined as those with accuracy greater than 85 %. *Cg* stands for coagulopathy, *MI* for myocardial infarction, *ARDS* for acute Respiratory Distress Syndropme, *EDRTS* for Emergency Department Revised Trauma Score, *ISS* for injury Severity Score, *ID* for Insulin-Dependent (Table Source [23])

Table 11.11 Extracted supporting rules for survival prediction (75–85 % accuracy)

Rules	Test Accuracy	Method
(Cg = 'Yes') and 2.5 ≤ HEAD < 3.5 and EDRTS < 6.07 and 35.65 ≤ AGE < 55.25 Then Alive	10/12 (83.3 %)	CART
(Cg = 'Yes') and HEAD ≥ 3 and EDRTS ≥ 6.07 and THORAX < 1 Then Alive	33/43 (76.7 %)	CART
(Cg = 'No') and (MI = 'No') and AGE < 61.70 and (ARDS = 'Yes') and HEAD < 3 Then Alive	50/59 (84.7 %)	CART
(Cg = 'No') and (MI = 'No') and ISS = 24 and 61.70 = AGE < 68.90 and HEAD ≤ 3 Then Dead	11/13 (84.6 %)	CART
AGE < 61.70 and HEAD ≤ 4 and (MI = 'No') Then Alive	625/793 (78.8 %)	CART
HEAD ≥ 5 and (Cg = 'No') and AGE < 22.85 Then Alive	60/73 (82.2 %)	CART
HEAD ≥ 5 and (Cg = 'No') and EDRTS < 5.02 and 22.85 ≤ AGE < 28 and ISS ≥ 33 Then Dead	11/13 (84.6 %)	CART
ISS ≥ 23 and (ID = 'Yes') and AGE ≥ 80.50 Then Dead	42/51 (82.4 %)	CART
(MI = 'No') and (ID = 'Yes') and (Cg = 'Yes') and AGE > 61.6 Then Dead	24/32 (75 %)	C4.5
(ID = 'No') and HEAD ≤ 3 and AGE ≤ 82.6 and ISS ≤ 22 Then Alive	236/305 (77.4 %)	C4.5
HEAD ≤ 4 and (MI = 'No') and AGE ≤ 60.8 and ISS ≤ 38 Then Alive	504/607 (83 %)	C4.5
HEAD ≤ 4 and AGE ≤ 78 and EDRTS > 7.55 and ISS ≤ 30 Then Alive	207/263 (78.7 %)	C4.5
(MI = 'Yes') and ISS > 27 Then Dead	50/60 (83.3 %)	C4.5
HEAD ≤ 3 and AGE ≤ 78 and 11 < ISS ≤ 27 Then Alive	290/368 (78.8 %)	C4.5
(Cg = 'No') and HEAD ≤ 3 and AGE ≤ 78 Then Alive	353/459 (76.9 %)	C4.5
(MI = 'Yes') and EDRTS ≤ 5. 39 Then Dead	41/51 (80.4 %)	C4.5
(Cg = 'No') and (MI = 'No') and 2 < HEAD ≤ 4 and EDRTS ≤ 1.47 and (ARDS = 'Yes') and ISS ≤ 41 Then Dead	13/17 (76.5 %)	C4.5

Though these rules are not reliable enough for practical use, they can contain pattern information which may be of interest to physicians. *Cg* stands for coagulopathy, *MI* for myocardial infarction, *ARDS* for Acute Respiratory Distress Syndrome, *EDRTS* stands for Emergency Department Revised Trauma Score, *ISS* for Injury Severity Score, *ID* for Insulin-Dependent (Table Source [23])

85 %. Note that the rules with accuracy between 75 and 85 % may not be sufficiently reliable, yet may contain useful pattern information, as described in the discussion section. A more comprehensive list of rules for each of these tables can be found in [23].

11.5 Discussion

In this study computer-aided rule-base system was developed using significant variables selected via logistic regression. It was seen that the rule quality was increased with approximations of the variables. The intent of this study is to

Table 11.12 Extracted rules for outcome prediction (>85 % accuracy)

Rules	Test accuracy	Method
HEAD $\leq$ 3 and AGE $<$ 43.45 and FSBP $<$ 143.50 and ISS $\leq$ 33 and EDRTS $<$ 0.87 and THORAX $\geq$ 2 Then Rehab	17/19 (89.5 %)	CART
EDRTS $<$ 5.36 and HEAD $\leq$ 3 and 33 $\leq$ FSBP $\leq$ 143 and ISS $\geq$ 33.50 Then Rehab	69/79 (87.3 %)	CART
HEAD $\geq$ 4 and FSBP $<$ 171 and EDRTS $<$ 2.25 Then Rehab	125/135 (92.6 %)	CART
2.25 $\leq$ EDRTS $<$ 5.36 and HEAD $\geq$ 4 and FSBP $<$ 171 and AGE $\geq$ 10.90 Then Rehab	45/52 (86.5 %)	CART
EDRTS $\geq$ 5.36 and AGE $<$ 48.15 and THORAX $\geq$ 1 and ISS $\leq$ 21 Then Home	23/27 (85.2 %)	CART
EDRTS $\geq$ 5.36 and AGE $\geq$ 48.15 and EDGCSTOTAL $\geq$ 9 and ISS $\leq$ 25 Then Rehab	61/65 (93.8 %)	CART
EDRTS $<$ 5.02 and HEAD $\leq$ 3 and 11.65 $\leq$ AGE $<$ 24.40 and EDGCSTOTAL $\leq$ 8 and FSBP $\geq$ 108 and THORAX $\leq$ 4 Then Rehab	24/28 (85.7 %)	CART
EDRTS $<$ 5.02 and HEAD $\leq$ 3 and 26.05 $\leq$ AGE $<$ 37.30 and EDGCSTOTAL $\leq$ 8 and FSBP $\geq$ 108 Then Rehab	22/24 (91.7 %)	CART
EDRTS $<$ 5.02 and HEAD $\geq$ 4 Then Rehab	179/201 (89.1 %)	CART
EDRTS $<$ 2.69 and HEAD $\leq$ 3 and AGE $<$ 38.30 and 108 $\leq$ FSBP $<$ 192 Then Rehab	38/43 (88.4 %)	C4.5
EDRTS $<$ 2.69 and HEAD $\geq$ 4 Then Rehab	132/146 (90.4 %)	C4.5
EDRTS $\geq$ 2.69 and AGE $<$ 48.15 and 84 $\leq$ FSBP $\leq$ 93 Then Rehab	18/21 (85.7 %)	C4.5
2.69 $\leq$ EDRTS $<$ 4.75 and 11.65 $\leq$ AGE $<$ 48.15 and FSBP $\geq$ 122 and (ARDS = 'No') Then Rehab	33/36 (91.7 %)	C4.5
EDRTS $\geq$ 2.69 and AGE $\geq$ 48.15 and ISS $\geq$ 26 Then Rehab	66/71 (93.0 %)	C4.5
EDGCSTOTAL $\leq$ 5 and ISS $\geq$ 15 and FSBP $\leq$ 177 and THORAX $\geq$ 4 Then Rehab	252/284 (88.7 %)	C4.5
EDGCSTOTAL $\geq$ 6 and AGE $\geq$ 48.15 and ISS $\geq$ 26 Then Rehab	66/72 (91.7 %)	C4.5
THORAX $\leq$ 2 and AGE $\leq$ 33.9 and EDRTS $\leq$ 5.03 Then Rehab	62/72 (86.1 %)	C4.5
(ID = 'Yes') and (Cg = 'No') Then Rehab	11/12 (91.7 %)	C4.5
HEAD $\leq$ 0 and THORAX $\leq$ 1 and AGE $\leq$ 59.7 and ISS $>$ 5 Then Rehab	28/32 (87.5 %)	C4.5

Reliable rules are defined as those with accuracy greater than 85 %. *FSBP* represents initial blood pressure, *ISS* stands for Injury Severity Score, *EDGCSTOTAL* is the total Glasgow Coma Score recorded in the emergency department, *EDRTS* is the Emergency Department Revised Trauma Score, *ARDS* stands for Acute Respiratory Distress Syndrome (Table Source [23])

develop a computer-assisted decision making system which extracts and formulates diagnostic knowledge into equivalent sets of transparent decision rules which presents a clear reasoning behind every decisions. Direct maximization likelihood estimation is employed along with logistic regression in this method to extract the most significant variables among all possible variables. The comparison of performances between AdaBoost, CART, SVM, C4.5 and RBF Neural Network, reveals that by using only significant variables for the computation, a considerable

Table 11.13 Extracted supporting rules for outcome prediction (75–85 % accuracy)

Rules	Test accuracy	Method
EDRTS $\geq$ 5.36 and EDGCSTOTAL $\geq$ 9 and ISS $\leq$ 24 and THORAX $\leq$ 3 and AGE $\geq$ 53.95 and FSBP $\geq$ 93 Then Rehab	49/62 (79.0 %)	CART
EDRTS $\geq$ 7.12 and AGE $<$ 47.55 and THORAX $\geq$ 1 and 28 $\leq$ ISS $<$ 35 and 94 $\leq$ FSBP $\leq$ 135 Then Rehab	16/20 (80.0 %)	CART
EDRTS $\geq$ 2.69 and AGE $<$ 22.80 and THORAX $\geq$ 1 and ISS $\geq$ 22 and 123 $\leq$ FSBP $\leq$ 139 Then Rehab	11/13 (84.6 %)	CART
EDRTS $\geq$ 7.70 and 22.80 $\leq$ AGE $<$ 45.90 and THORAX $\geq$ 1 and ISS $\geq$ 28 and FSBP $\geq$ 76 Then Rehab	31/39 (79.5 %)	CART
5.02 $\leq$ EDRTS $<$ 7.12 and AGE $<$ 45.90 and THORAX $\geq$ 1 and 22 $\leq$ ISS $\leq$ 39 Then Rehab	9/12 (75.0 %)	CART
EDRTS $\geq$ 7.12 and AGE $<$ 48.15 and ISS $\geq$ 25 and HEAD $\leq$ 4 and THORAX $\geq$ 1 and 69 $\leq$ FSBP $<$ 98 Then Rehab	15/19 (78.9 %)	CART
EDRTS $\geq$ 2.69 and AGE $<$ 47.80 and ISS $\leq$ 24 and HEAD $\leq$ 2 and Then Home	43/56 (76.8 %)	CART
2.69 $\leq$ EDRTS $<$ 5.02 and 26.75 $\leq$ AGE $<$ 47.80 and ISS $\geq$ 25 and HEAD $\geq$ 1 Then Rehab	28/34 (82.4 %)	CART
(ID = 'Yes') and AGE $>$ 44 and (ARDS = 'Yes') Then Rehab	30/39 (76.9 %)	C4.5
THORAX $\leq$ 3 and ISS $>$ 18 Then Rehab	342/431 (79.4 %)	C4.5
(Cg = 'No') and 18.4 $<$ AGE $\leq$ 59.7 and ISS $>$ 30 Then Rehab	162/199 (81.4 %)	C4.5

Though these rules are not reliable enough for practical use, they can contain pattern information which may be of interest to physicians. *FSBP* represents initial blood pressure, *ISS* stands for Injury Severity Score, *EDGCSTOTAL* is the total Glasgow Coma Score recorded in the emergency department, *EDRTS* is the Emergency Department Revised Trauma Score, *ARDS* stands for Acute Respiratory Distress Syndrome, *Cg* for coagulopathy (Table Source [23])

improvement can be seen in performance as opposed to the performance of these machine learning algorithms using all available variables. The proposed selection method seems robust and efficient sing all five methods show improvement across all-available and only-significant-variables.

By comparing the performance measure of each rule, it can be established that the rules with accuracy greater than 85 % can be considered as reliable rules. All rules that were selected were acknowledged as reliable only if the frequency of cases within the dataset matching the rule was greater than a specified threshold. After measuring rule sensitivity and specificity, for the given outcome pairs (home/rehab, severe/non-severe and alive/dead) the sensitivity was found to be 87.4 % while the specificity was found to be 88.4 %. This validates the good performance of the presented method. To improve rule quality some additional factors may be needed. In particular, large and well balanced datasets across all outcome classes could improve overall quality, as well as sensitivity and specificity. The sensitivity and specificity results for each of the datasets are presented in Table 11.16.

Table 11.14 Extracted reliable rules for ICU days prediction (>85 % accuracy)

Rules	Test accuracy	Method
(AIRWAY = 'Need') and 115 ≤ ED-BP < 156 and AGE ≥ 47.05 and Then ICU stay days ≥ 3	14/15 (93.3 %)	CART
(AIRWAY = 'Need') and 115 ≤ ED-BP < 156 and ED-RESP < 18 and 4.35 ≤ AGE < 14.5 Then ICU stay days ≥ 3	12/12 (100 %)	CART
(AIRWAY = 'No Need') and ED-RESP ≥ 21 and 45 ≤ AGE < 55.85 and Then ICU stay days ≤ 2	10/11 (90.1 %)	CART
(AIRWAY = 'Need') and ED-BP < 91 Then ICU stay days ≥ 3	14/14 (100 %)	CART
(AIRWAY = 'Need') and 93.5 ≤ ED-BP < 156.5 and ED-PULSE ≥ 60.5 and AGE ≥ 54.2 Then ICU stay days ≥ 3	10/10 (100 %)	CART
(AIRWAY = 'Need') and 94 ≤ ED-BP < 156 and ED-PULSE ≥ 61 and ED-RESP < 19 and 18.45 ≤ AGE < 44.5 Then ICU stay days ≥ 3	60/76 (86.6 %)	CART
(AIRWAY = 'No Need') and AGE < 52.9 and ED-BP ≥ 107 and ED-GCS ≥ 11 Then ICU stay days ≤ 2	175/192 (91.1 %)	CART
(AIRWAY = 'Need') and ED-BP < 150.5 and ED-RESP < 19 and AGE ≥ 4.9 and ED-PULSE ≥ 138 Then ICU stay days ≥ 3	18/20 (90 %)	CART
(AIRWAY = 'Need') and ED-RESP < 19 and ED-RESP < 19 and ED-PULSE < 138 and ED-bp < 115 AND 10.9 ≤ AGE < 47.3 Then ICU stay days ≥ 3	31/33 (93.9 %)	CART
(AIRWAY = 'No Need') and AGE < 37.1 and ED-GCS ≥ 11 and ED-BP ≥ 125 Then ICU stay days ≤ 2	89/90 (98.9 %)	CART
Age ≤ 42 and (Airway = 'No Need') and ED-PULSE ≤ 137 and ED-RESP > 19 Then ICU stay days ≤ 2	100/116 (86.2 %)	C4.5
Age > 37 and ED-BP ≤ 95 Then ICU stay days ≥ 3	14/14 (100 %)	C4.5

Reliable rules are defined as those with accuracy grater than 85 %. *ED-BP* is Emergency Department Blood Pressure, *ED-RESP* is Emergency Department Respiratory Rate, *ED-PULSE* is Emergency Department Pulse Rate, *ED-GCS* is Emergency Department Glasgow Come Score (Table Source [23])

The issue of dealing with rules with accuracy below 85 % is also an important factor for extracting certain knowledge. Since, when using only rules with accuracy over 85 %, some medical knowledge in the database might have been ignored. The accuracy of certain rules might have been low due to the lack of a comprehensive database, rather than a flaw within the rule itself. Therefore, rules with accuracies lesser than 85 % are not completely removed from the rule based system. Instead such rules are utilized as additional "supporting rules" in suggesting possible treatments and procedures. For example, according to trauma experts, patients with a high ISS score (>25) are least likely to survive. However, some rules with surprising implications were found. For instance, one of these "counterintuitive" rules pointed to the fact that there are 52 alive cases (3.3 %) with ISS high scores (38). Of these 52 patients, 33 (63.5 %) have high AIS head

Table 11.15 Extracted supporting for ICU days prediction (75–85 % accuracy)

Rules	Test Accuracy	Method
(AIRWAY = 'No Need') and ED-RESP < 21 ED-BP < 142 ED- PULSE < 79 and 37.05 ≤ AGE < 44.15 Then ICU stay days ≤ 2	13/16 (81.3 %)	CART
(AIRWAY = 'No Need') and AGE ≥ 52.9 and ED-BP ≥ 141 Then ICU stay days ≤ 2	12/15 (80 %)	CART
(AIRWAY = 'Need') and 117 ≤ ED-BP < 135 and ED-RESP < 19 and 68 ≤ ED-PULSE < 138 and 15.05 ≤ AGE < 46.4 Then ICU stay days ≥ 3	23/28 (82.1 %)	CART
(AIRWAY = 'Need') and 136 ≤ ED-BP < 150 and ED-RESP < 19 and ED-PULSE < 138 and 15.05 ≤ AGE < 23.25 Then ICU stay days ≥ 3	10/13 (77 %)	CART
(AIRWAY = 'No Need') and 96 ≤ ED-BP < 163 and 39.15 ≤ AGE < 69.05 Then ICU stay days ≤ 2	44/55 (80 %)	CART
(AIRWAY = 'Need') and ED-BP < 156 and AGE ≥ 24.35 Then ICU stay days ≥ 3	76/96 (79.2 %)	CART
(AIRWAY = 'Need') and ED-BP < 146 and AGE < 17.85 and 135 ≤ ED-PULSE < 181 Then ICU stay days ≥ 3	9/11(81.2 %)	CART
(AIRWAY = 'Need') and ED-BP < 146.5 and AGE < 17.85 and ED-PULSE < 131 and ED-RESP < 18 Then ICU stay days ≥ 3	15/20 (75 %)	CART
(AIRWAY = 'No Need') and ED-BP ≥ 141 Then ICU stay days ≤ 2	223/265 (84.2 %)	CART
(AIRWAY = 'Need') and Ed-BP < 114 Then ICU stay days ≥ 3	44/52 (84.6 %)	CART
(AIRWAY = 'Need') and 114 ≤ ED-BP < 135.5 and ED-PULSE < 97 and 17.2 ≤ AGE < 46.95 and ED-RESP < 7 Then ICU stay days ≥ 3	10/13 (77 %)	CART
(AIRWAY = 'No Need') and AGE ≥ 52.9 and ED-BP ≥ 141 Then ICU stay days ≤ 2	12/15 (80 %)	CART
(AIRWAY = 'Need') and ED-BP < 114 Then ICU stay days ≥ 3	44/52 (84.6 %)	CART
(AIRWAY = 'Need') and 110.5 ≤ ED-BP < 180.5 and ED-PULSE ≥ 62 and 4.35 ≤ AGE < 44.5 and ED-GCS < 10 Then ICU stay days ≥ 3	35/46 (76.1 %)	CART

Though these rules are not reliable enough for practical use, they can contain pattern information which may be of interest to physicians. *ED-BP* is Emergency Department Blood Pressure, *ED-RESP* is Emergency Department Respiratory Rate, *ED-PULSE* is Emergency Glasgow Coma Score (Table Source [23])

Table 11.16 Rule sensitivity and specificity

Predictive outcome	Off-site dataset Alive/Dead (%)	Off-site dataset Home/Rehab	Helicopter dataset ICU stay days
Sensitivity (>85 % rules)	91.9	88.7	90.6
Specificity (>85 % rules)	89.2	87.7	91
Sensitivity (75–85 % rules)	86.2	79	82.5
Specificity (75–85 % rules)	80.4	80.1	80.4

Table Source [23]

scores ($\geq$4), and 38 patients (73 %) are male. Considering the above conditions, surviving patients have lower thorax (average score = 2.61) and lower abdomen AIS scores (average score = 1.03) than those patients under fatal cases. These fatal cases typically have a higher head AIS score (average score = 5.08) than surviving patients (average head score = 3.90). In addition, it was found that none of the surviving patients have complications such as coagulopathy, and only a few had a pre-existing disease (in particular, Insulin Dependency and Myocardial Infarction).

Usually while considering an impact factor in predicting survival, only Acute Respiratory Distress Syndrome (ARDS) is considered as a significant factor. However, according to the rules created, pre-existing conditions, Acute Respiratory Distress Syndrome (ARDS), Insulin Dependency, Myocardial Infarction, and Coagulopathy all have a significant impact on survival prediction. Furthermore, airway status (needed/not needed) was identified as a primary factor in predicting the number of ICU days for patients transported via helicopter. Note that for predicting the ICU length of stay, 74.6 % of patients stayed at in ICU less than 2 days. Only 25.4 % of patients stayed more than 2 days, and only 2.9 % of those were in ICU for more than 20 days. This reinforces Eckstein's point [47] that many patients are transported via helicopter unnecessarily.

By using advanced image and signal processing systems, such as the ICP estimation and midline shift detection discussed earlier, more astute information can be added into the decision making process thereby increasing the precision and reliability of the rules generated. Therefore, the use of accurate prediction rules for potential number of ICU days may help improve the efficiency and reduce operational cost of helicopter transportation, as well as provide timely treatment for patients in critical condition.

11.6 Conclusion

The results from this study provide a framework to improve the physicians' diagnostic accuracy with the aid of machine learning algorithm. The resulting system is effective in predicting patient survival, and rehab/home outcome. A method has been introduced that creates a variety of reliable rules that make sense to physicians by combining CART and C4.5 and using only significant variables extracted via logistic regression.

A novel method for assessment of Traumatic Brain Injury (TBI) has also been presented. The ability of such a system to assess levels of Intracranial Pressure (ICP) as well as predict survival outcomes and days in ICU, together encompasses a wholesome diagnostic tool, which can help improve patient care as well as save time and reduce cost. The resulting computer-assisted decision-making system has significant benefits, both in providing rule-based recommendations and in enabling optimal resource utilization. This may ultimately assist physicians in providing the best possible care to their patients.

References

1. Faul M, Xu L, Wald MM, Coronado VG (2010) Traumatic brain injury in the United States: emergency department visits, hospitalizations, and deaths. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Atlanta
2. Centres for Disease Control and Prevention (2004) Facts about concussion and brain injury and where to get help. Atlanta, GA
3. Finkelstein E, Corso P, Miller T et al (2006) The incidence and economic burden of injuries in the United States. Oxford University Press, New York
4. Expert Working Group (2000) Traumatic brain injury in the United States: assessing outcomes in children. Atlanta, GA
5. Anderson RN, Minino AM, Fingerhut LA, Warner M, Heinen MA (2001) Deaths: injuries. Natl Vital Stat Rep 52(21):1–87
6. Fabian TC, Patton JH, Croce MA, Minard G, Kudsk KA, Pritchard FE (1996) Blunt carotid injury: importance of early diagnosis and anticoagulant therapy. Ann Surg 223(5):513–525
7. Jagielska I (1998) Linguistic rule extraction from neural networks for descriptive data mining. In: Proceedings of 2nd international conference on knowledge-based intelligent electronic systems, 21–23 Apr 1998, Adelaide, pp 89–92
8. Cunningham P, Rutledge R, Baker CC, Clancy TV (1997) A comparison of the association of helicopter and ground ambulance transport with the outcome of injury in trauma patients transported from the scene. J Trauma 43(26):940–946
9. Ruggieri S (2002) Efficient C4.5. IEEE Trans Knowl Data Eng 14(2):438–444
10. Ji SY, Huynh T, Najarian K (2007) An intelligent method for computer-aided trauma decision making system. In: ACM-SE 45—proceedings of the 45th annual southeast regional conference 2007, pp 198–202
11. Haug PJ, Gardner RM, Tate KE (1994) Decision support in medicine: examples from the HELP system. Comput Biomed Res 27(5):396–418
12. Fitzmaurice JM, Adams K, Eisenberg JM (2002) Three decades of research on computer applications in health care: medical informatics support at the agency for healthcare research and quality. J Am Med Inform Assoc 9(2):144–160
13. Quinlan J (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
14. Clarke JR, Hayward CZ, Santora TA, Wagner DK, Webber BL (2002) Computer-generated trauma management plans: comparison with actual care. World J Surg 26(5):536–538
15. Najarian K, Darvish A (2006) Neural Networks: Applications in Biomedical Engineering. Wiley Encyclopedia of Biomedical Engineering
16. Andrews PJ, Sleeman DH, Statham PF, McQuatt A, Corruble V, Jones PA, Howells TP, Macmillan CS (2003) Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. J Neurosurg 97(2):440–442
17. Kuhnert PM, Do K, McClure R (2000) Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Comput Stat Data Anal 34(3):371–386
18. Signorini DF, Andrews PJD, Jones PA, Wardlaw JM, Miller JD (1999) Predicting survival using simple clinical variables: a case study in traumatic brain injury. J Neurol Neurosurg Psychiatry 66:20–25
19. Hasford J, Ansari H, Lehmann K (1993) CART and logistic regression analyses of risk factors for first dose hypotension by an ACE-inhibitor. Therapie 48(5):479–482
20. Guo HM, Shyu YI, Chang HK (2006) Combining logistic regression with classification and regression tree to predict quality of care in a home health nursing data set. Stud Health Technol Inf 122:891
21. Kelemen A, Liang Y, Franklin S (2002) A comparative study of different machine learning approaches for decision making. In: Mastorakis E (ed) Advances in simulation, computational methods and soft computing. WSEAS Press, Piraeus

22. Snedecor GW, Cochran WG (1989) Statistical methods, 8th edn. Iowa State University Press, Ames
23. Ji SY, Smith R, Huynh T, Najarian K (2009) A comparative analysis of multi-level computer-assisted decision making system for traumatic injuries. BMC Med Inf Decis Making 9:2
24. Pfahringer B (1995) Compression-based discretization of continuous attributes. In: Proceedings of 12th international conference machine learning, Tahoe City, pp 456–463
25. Chen W (2010) Automated measurement of midline shift in brain CT images and its application in computer-aided medical decision making, VCU ETD Archive, Aug 2010
26. Besag J (1986) On the statistical analysis of dirty pictures. J R Statist Soc B 48(3):259–302
27. Leung CK, Lam FK (1997) Maximum a posteriori spatial probability segmentation. In: IEEE proceedings—vision image and signal processing, vol 144, pp 161–167
28. Mac Namee B, Cunningham P, Byrne S, Corrigan OI (2002) The problem of bias in training data in regression problems in medical decision support. AI Med 24:51–70
29. Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcome. J Clin Epidemiol 49(11):1225–1231
30. Jagielska I (1998) Linguistic rule extraction from neural networks for descriptive data mining. In: Proceedings of 2nd international conference knowledge-based intelligent electronic systems, 21–23 Apr 1998, Adelaide, pp 89–92
31. Lu Y, Sundararajan N, Saratchandran P (1998) Performance evaluation of a sequential minimal radial basis function (RBF) neural network learning algorithm. IEEE Trans Neural Networks 9(2):308–318
32. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc Ser B (Methodol) 36(2):111–147
33. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proceedings of ECML-98, 10th European conference on machine learning 1998, Chemnitz, DE pp 137–142
34. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914
35. Lee C, Chung P, Tsai J, Chang C (1999) Robust radial basis function neural networks. IEEE Trans Syst Man Cybern B Cybern 29(6):674–685
36. Freund Y, Schapire R (1999) A short introduction to boosting. J Japan Soc for Artif Intel 14(5):771–780
37. Fan W, Stolfo SJ, Zhang J (1999) The application of AdaBoost for distributed, scalable and on-line learning. In: KDD '99 Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, pp 362–366
38. Breiman L (1993) Classification and regression trees. Chapman & Hall, Boca Raton
39. Loh WY, Vanichsetakul N (1988) Tree-structured classification via generalized discriminant analysis. J Am Stat Assoc 83(403):715–725
40. Fu CY (2004) Combining loglinear model with classification and regression tree (CART): an application to birth data. Comput Stat Data Anal 45(4):865–874
41. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
42. Quinlan JR (1987) Generating production rules from decision trees. In: Proceedings of 10th international joint conference artificial intelligence (IJCAI-87), Milan, pp 997–1003
43. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
44. Quinlan JR (1983) Learning efficient classification procedures and their application to chess end games. In: Michalski RS, Carbonell JG, Mitchell TM (eds) Machine learning—an artificial intelligence approach. Tioga Publishing Company, Palo Alto, pp 463–482
45. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc Ser B (Methodol) 36(2):111–147
46. Hosmer D, Lemeshow S (1989) Applied logistic regression (Wiley series in probability and statistics), Chap 1. Wiley, New York
47. Eckstein M, Jantos T, Kelly N, Cardillo A (2002) Helicopter transport of pediatric trauma patients in an urban emergency medical services system: a critical analysis. J Trauma 53(2):340–344

Chapter 12
Supervised Learning Methods for Fraud Detection in Healthcare Insurance

Prerna Dua and Sonali Bais

Abstract Fraud in the healthcare system is a major problem whose rampant growth has deeply affected the US government. In addition to financial losses incurred due to this fraud, patients who genuinely need medical care suffer because of unavailability of services which in turn incur due lack of funds. Healthcare fraud is committed in different ways at different levels, making the fraud detection process more challenging. The data used for detecting healthcare fraud, primarily provided by insurance companies, is massive, making it impossible to audit manually for fraudulent behavior. Data-mining and Machine-Learning techniques holds the promise to provide sophisticated tools for the analysis of fraudulent patterns in these vast health insurance databases. Among the data mining methodologies, supervised classification has emerged as a key step in understanding the activity of fraudulent and non-fraudulent transactions as they can be trained and adjusted to detect complex and growing fraud schemes. This chapter provides a comprehensive survey of those data-mining fraud detection models based on supervised machine-learning techniques for fraud detection in healthcare.

Keywords Healthcare fraud · Fraud detection · Supervised methods · Unsupervised methods

P. Dua (✉)
Department of Health Informatics and Information Management, Louisiana Tech University, Ruston, LA, USA
e-mail: prerna@latech.edu

P. Dua
School of Biological Sciences, Louisiana Tech University, Ruston, LA, USA

S. Bais
Department of Computer Science, Louisiana Tech University, Ruston, LA, USA

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_12,

12.1 Introduction

Healthcare fraud, a severe and challenging problem faced by medical providers, can be defined as an offense committed by an individual or group of individuals who place false medical claims for services that have never been used to gain unauthorized financial benefits. According to the data provided by Centers for Medicare and Medicaid Services (CMS), the United States spent overall $2.5 trillion after healthcare during the 2009 fiscal year. This expenditure signifies $8,086 per person or 17.6 % of the Gross Domestic Product (GDP), increased from 16.6 % in the 2008 fiscal year [1]. It is estimated that over five billion health insurance claims were paid that year [2], and some of them were fraudulent. Even though these fraudulent claims composed only a small portion of the claims, they carried a very high cost value. It is predicted by CMS that the healthcare expenditure may increase up to $4.14 trillion by 2016, signifying 19.6 % of the GDP [3]. Information provided by National Health Care Anti-Fraud Association (NHCAA) shows that approximately $60 billion, or 3 %, of healthcare spending has been expended on healthcare fraud [4]. This quantity of money is more than the GDP of 120 countries including Kenya, Ecuador, and Iceland [5]. If steps against healthcare fraud are not taken, such expenses can affect quality of life and national economies. The Federal Bureau of Investigation (FBI) approximates that between $70 and $234 billion are stolen from US citizens in healthcare annually due to fraud in the healthcare system [6]. Even if the financial loss is disregarded, healthcare fraud can hamper the healthcare system of the US from offering good services and care to patients. Hence, the efficient detection of fraud is vital, as it allows for quality enhancements and the lowering the expenditures to healthcare services.

To detect fraud within the healthcare system, the process of auditing is followed by investigation. If accounts are carefully audited, it is possible to identify suspicious policy holders and providers. Ideally, all claims should be audited carefully and individually. However, it is impossible to audit all claims by any practical means as these form huge piles of data involving sorting operations and complex computation [23]. Moreover, it is difficult to audit service providers without clues as to what auditors should be looking for. A sensible approach is to make short lists for scrutiny and audit patients and providers based on these lists. A variety of analytic techniques can be used to compile audit short lists. Fraudulent claims frequently build into patterns that can be perceived using predictive models.

12.2 Types of Fraud in Health Care

Healthcare fraud can be divided into four types: (2.1) medical service providers, (2.2) medical resource providers, (2.3) insurance policy holders, and (2.4) insurance policy providers. Figure 12.1 demonstrates the overview of fraudulent activities found in healthcare.

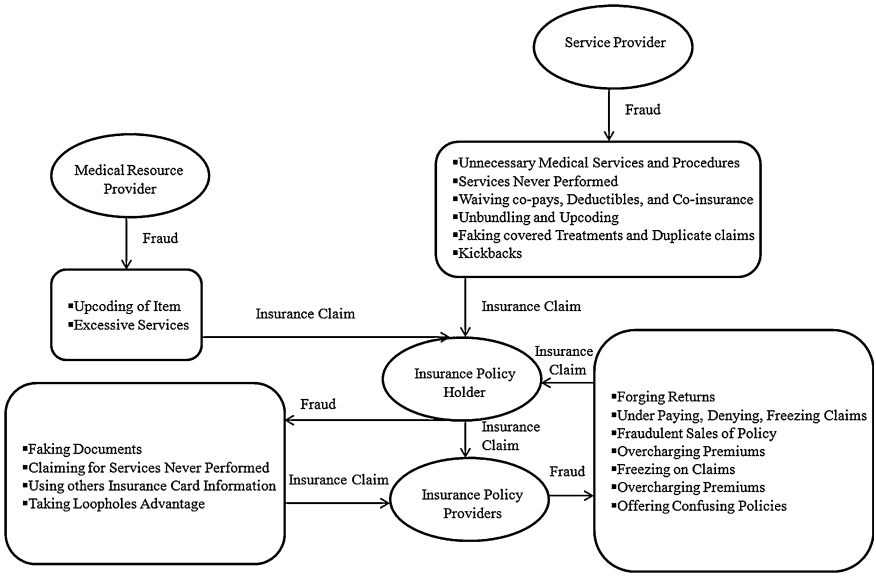


Fig. 12.1 Types of fraud in healthcare system

12.2.1 Medical Service Providers

Medical service providers can be hospitals, doctors, nurses, radiologists and other laboratory service providers, and ambulance companies. Activities involving Medical Service Providers may include:

- Forging a patient diagnoses to rationalize services and procedures that are not medically required [7],
- Billing for services which were never performed using authentic patient information, aiding in identity theft, or modifying claims with extra charges for the services or procedures that were never performed [7],
- Billing the insurance companies more by waiving patient co-pays, deductibles, or co-insurance [7],
- Billing for unnecessary services or procedures, such as daily checkups rather than monthly ones, only to create insurance payments [3, 7],
- Billing each phase of a process as if it were a different process, known as unbundling [7], for example billing tests within test "sessions" as if they were independent sessions [3],
- Billing for expensive services instead of billing for low cost services or procedures which were performed, known as upcoding or coding a patient diagnosis to a more critical and expensive charge and applying charges with false CPT codes, such as charging a 30 min group therapy as a 50 min personal therapy [3, 7],

- Billing for duplicate claims for a particular service or an item, in which the provider changes part of the claim for example the date on which the service was provided to fool the health insurance company into paying for a single service or item twice [3], and
- Accepting illegal kickback schemes in which healthcare providers trade money or something of value for patient referrals, for those healthcare services that can be paid by Medicaid or Medicare [3].

An example of fraud by healthcare providers is the Federal indictment of the Benitez brothers, owners of a chain of medical clinics in the Miami area that treated HIV-infected Medicare beneficiaries. The Benitez brothers were accused of submitting fraudulent insurance claims that cost Medicare approximately $110 million. For these claims, they performed unnecessary procedures on the patients. They also offered kickbacks to the patients in exchange of their Medicare information, which was used to submit false claims to Medicare for compensation [9].

12.2.2 Medical Resource Providers

Medical resource providers can be pharmaceutical companies, medical equipment companies that supply products like wheelchairs, walkers, specialized hospital beds and medical kits. Activities involving Medical Resource Providers may include:

- Medical equipments manufacturer put forward complimentary products before the patients. Then they charge patients insurance company for the free product which is either not required or was never delivered [26],
- Sometimes needless or false tests are provided to the individuals at shopping malls, old-age homes, and health clubs and billed to insurance companies or Medicare [26],
- Medical resource providers also bills individuals insurance companies for resources never been delivered by altering the bills or submitting forge ones [26],
- The common target of medical resource providers are the senior citizens who are offered free products by these providers in exchange of their Medicare insurance number. In order to get paid by the Medicare, the doctor is required to sign a form certifying that the equipment is required by the patient as a part of medical treatment. To achieve this, medical resource provider either bribes the corrupt doctors or provide fake signature for billing Medicare for the equipments which are either not delivered or not required [26],
- Up-coding items, for instance sending a patient an essential, manually pushed wheelchair but billing the patient's health insurance plan for a power-driven wheelchair [3] and
- Providing excessive services that the patient does not need, for example, delivering and billing for 40 injury care kits per week for a patient in hospital who only needs a change of dressings once per day [3].

In August 2004, Pfizer employees were convicted of being involved in illegal kickback schemes and off-label endorsements. In this case, a Pfizer northeast regional manager directed around a hundred sales employees to market a painkiller called Bextra, which was prohibited by FDA. The FDA had approved Bextra for distribution in U.S., but citing augmented occurrences of strokes, heart attacks, and severe skin reaction to drug, petitioned Pfizer to withdraw this drug from U.S. market [3].

12.2.3 Insurance Policy Holders

Insurance policy holders consist of individuals and groups who carry insurance policies, including both patients and employers of patients. Activities involving Insurance Policy Holders may include:

- Faking documents related to employment or eligibility to obtain low premium rates and good benefits [20],
- Filing claims for medical services or procedures which never took place [20],
- Claiming insurance benefits illegally using someone else's coverage or insurance card information [20], and
- Taking advantage of insurance benefits by finding loopholes in the policy.
- In 2007, a fraud case was committed by falsely filing a life insurance claim. The fraudulent owner faked his own death in a canoeing accident and lived a secret life in his house for five years. His family claimed the money from the insurance company, so that he and his family could start a hotel business for canoeing holidays in Panama [27].

12.2.4 Insurance Policy Providers

Insurance policy providers are the entities that pay medical expenses to a policy holder in return for their monthly premiums. Insurance policy providers can be private insurance companies or government administrated healthcare departments including agents and brokers. Very little research has been conducted regarding fraud committed by insurance policy providers as most insurance fraud data are provided by the providers. It is estimated that around $85 billion are lost yearly due fraud committed by insurance companies [28]. Activities involving Insurance Policy Providers may include:

- Forging returns and benefit/service statements by under paying claims [20],
- Insurance company wrongfully denies valid claims to try to discourage the policy holder and hoping that the patient will eventually give up [28],
- Freezing claims without investigating the merits of claims [28],

- Overcharging the premiums from policy holders by misinterpreting the information to the client and making them pay for those coverage which they don't actually have [28],
- Creating confusing policies that mislead policyholders on coverage issues [28], and
- Making fraudulent sales of fake policies which are of no use to the policy holders and are mainly intended to get high premium from them.

During September 2009, an individual lost her health coverage by Blue Cross Insurance Company because the company discovered her pre-existing condition. This company terminated her coverage because she never mentioned her pre-existing condition, which she herself was unware of initially. Hence, the company abruptly cancelled her coverage after she was diagnosed of thyroid disorder and fluid in the heart leaving her in a debt of $25,000 in medical bills [40].

Among these four types of fraud discussed above, the service providers alone commit the majority of the fraud. Although most service providers are trustworthy, those few dishonest service providers commit fraud and cause the loss of millions of dollars to the healthcare system. In some cases, more than one of the above types is involved in committing healthcare fraud. Detecting fraud in such a hybrid cases can be complex and challenging [20]. Hence, it is urgent that researchers find effective ways to discover patterns and relationships in data that may be used to make a valid prediction about fraudulent claims. Due to this pressing need, high end data mining and machine learning techniques holds a promise to provide sophisticated tools to identify possible predictors that characterize the fraudulent behaviors based on the historical data [20].

12.3 Data Mining for the Fraud Detection in Healthcare

Data mining is a popular means for detecting fraud and abuse in the healthcare system. The vast quantities of data produced by healthcare insurance companies are difficult to process and evaluate using conventional methods. Data mining provides the techniques and expertise to convert these heaps of data into the useful collection of facts for decision making [8]. This kind of analysis is becoming increasingly important, as financial pressure has increased the requirement for healthcare industries to construct judgments based on the study of fiscal and clinical data. Information and analyses obtained through data mining can improve operating efficiency, decrease costs, and increase profits while preserving a high-level of care [30].

Additional reasons behind the increasing popularity of data mining include the use of fee and categorization systems. For example as an outcome of the Balanced Budget Act of 1997, CMS have to employ a potential fee system supported categorizing patients into case-mix clusters, with the help of empirical proof that supplies utilized within every case-mix cluster are comparatively constant. CMS

has utilized data-mining techniques to build a potential reimbursement system for inpatient treatment [31].

The data-mining applications generally establish norms for detecting fraud and abuse. Then, these applications identify unusual patters of claims by clinics, laboratories, and physicians. Along with other details, these data-mining applications can provide information about out of place referrals, prescriptions, medical claims and fraudulent insurance claims. For example, the Texas Medicaid Fraud and Abuse Detection System collected a great deal of data produced by millions of treatment courses, operations, and prescriptions to recognize abnormal behaviors and discover fraud. It recovered $2.2 million and successfully recognized 1,400 suspects for inquiry in 1998. This result is impressive considering that it was obtained after only a few months of use [32]. Due to this achievement, the Texas system was awarded a national prize for this innovative utilization of the expertise.

Data-mining techniques can be categorized into supervised methods and unsupervised methods.

Supervised Data-Mining Methods

Supervised machine-learning techniques consists of algorithms that reason from outwardly given instances to construct universal theorems which then predict upcoming instances. Supervised machine learning is used to construct a brief model of the allotment of class labels which refer to predictor features. Then, the testing instances are assigned class labels based on the resulting classifier, in which the predictor feature significance is known, but the value of the class label is unknown [22].

In this context, the weightage is given to those healthcare fraud detection models which implement supervised machine-learning techniques. Supervised methods like Neural Networks [13, 16], Association Rules [14, 15, 17], Genetic Algorithms [10], Fuzzy Bayesian classifier [11], Logical Regression methods [16, 21], Bayesian Networks [12, 19], KNN classifiers [10], and Classification Trees [16] have been used by researchers to detect fraud in the healthcare system.

Unsupervised Data-Mining Methods

Unlike supervised data-mining methods, unsupervised methods do not get any objective output or benefits from their surroundings. Although it is difficult to visualize how a machine can be trained without any response from its surroundings, these methods work well. It is very likely to build a proper model for unsupervised learning methods supported on the idea that the mechanism's aim is to use input characterization to foresee prospective input, effectively communicating the input to another mechanism, decision making, and so on. It can be said that unsupervised learning can find patterns in a data which can also be unstructured noise. Clustering and dimensionality reduction are the classic examples of unsupervised learning [22].

The benefit of using supervised techniques over unsupervised is that once the classifier has been trained, it can be easily utilized on any same kind of datasets [41]

which makes it a best choice for a fraud detection program which involves screening and monitoring. In this chapter, we only consider fsupervised machine-learning techniques and provide an in-depth survey of their application in detecting fraud in the healthcare system.

12.3.1 Data Sources

The raw data for detecting healthcare fraud is obtained from insurance companies which explain the reason for decreased research performed in detecting fraud committed by insurance companies. The insurance company designation can include government administered healthcare departments like Medicare or private insurance companies [20]. The following data sources that have been developed and published in various sources through supervised machine-learning techniques have helped shape the healthcare detection fraud model:

- Health Insurance Commission (HIC) of Australia [10, 14],
- Medicare Australia's Enterprise Data Warehouse, Prism [17],
- National Health Insurance (NHI) Program in Taiwan [15],
- Taiwan's National Health Insurance (NHI) System [16],
- ISAPRE (A private health insurance company in Chile) [13, 21] and Banmedica S.A. [13], and
- Taipei Health Insurance Bureau [11].

Generally, the raw data provided from the above sources consists of insurance claims. The content of insurance claims is related to the service provider and insurance subscriber. These databases contain rich features that are helpful to the fraud detection model in identifying fraudulent patterns of behavior by insurance holders and healthcare service providers. It is possible to gain an overall perspective of both insurance holder and healthcare service provider behavior over time using this information. This overall perspective helps detect fraud committed by these entities [20].

12.3.2 Algorithms

This section provides a discussion on the algorithms based on supervised models for detecting fraud in healthcare.

12.3.2.1 Neural Network

A neural network imitates just like the brain of human beings in order to forecast and classify data. A neural network is composed of a set of connected, simulated neurons, few of which get scalar information from other neurons and convert that information into a singular output signal. All the inter associations are weighted among them and customization is done to these weights while the neural network runs on training data. A neural network comprises of a *layered*, *feedforward*, *completely connected* network of artificial neuron, or nodes. The term feedforward suggests that the data flows in one direction from input to the output layer. A classic neural network for classifying data can contain two or more layers, although most contain three layers such as, input, hidden and an output layer [42]. Multilayer Perceptron (MLP) is a classic example of a multilayer feed-forward network with one or more hidden layers between the input and output layer. Neurons in the hidden layer obtains a weighted summation of the input variable, and convert that total to a signal in the form of output with the help of a threshold function such as a sigmoid or step function. The weighted summation, which is obtained from the hidden layer, is given to the single node of an output layer and is transformed into a classification signal. Neural networks set up an association amid the input and yielded data, and are efficient with noisy data [16].

Ortega et al. [13] proposed a fraud detection system which utilizes committee of MLP networks for each entity (such as, medical claims, affiliates, medical professionals and employers) involved in the fraud, in a Chilean private health insurance company. Figure 12.2 demonstrates the four sub-models which are a working group of neural networks for all entities. The inputs are pre-calculated attribute vectors that represent the specific exploitation and fraud sub-problem. Whenever a medical claim is obtained by the ISAPRE (a private pre-paid health insurance plan) system, the output of each committee conveys a predictive value. These values provide supplementary inputs to the sub-models, offering a response technique for merging the diverse outcomes. Models are assessed at fixed time intervals as per the predetermined agenda. The model representing medical claim is implemented every day to process inward entries while the other models are implemented once in a month. Every sub-model is trained again on monthly basis. A data renewal process has been defined for keeping the training models descriptive for historical and new fraud patterns. With the help of experts new training samples are selected and thoroughly classified. Then a subclass of equivalent amounts of normal and fraud conditions are chosen and included to the training dataset. Hence, the model retains knowledge of new kinds of fraud, and is able to prevent these emerging categories [13].

The drawback of using a neural network is that it cannot identify the importance of individual variables. To overcome this problem, Liou et al. [16] used neural nets in detecting fraud and claim abuse based on diabetic outpatient services, which helped perform sensitivity analysis among variables. The authors then presented those variables which were most significant for classification. The end result specified the order of every variable's relational significance in categorizing the

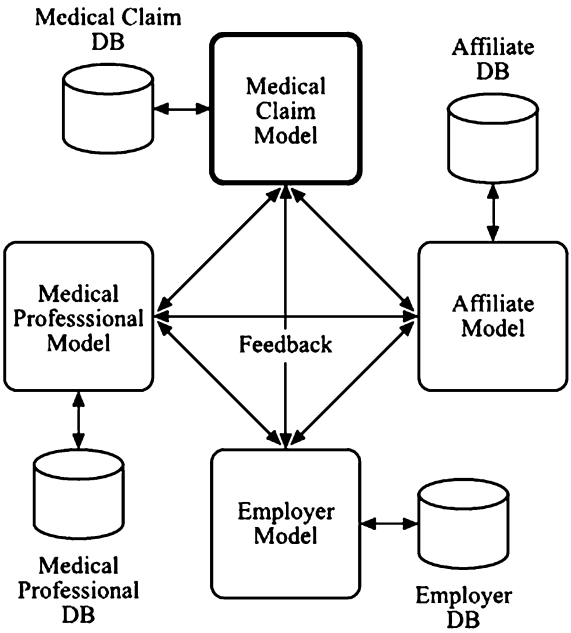


Fig. 12.2 Fraud detection scheme by using sub-models with feedback connections

data like averages of different types of claims like dispensary visiting service fees, diagnosis and prognosis fess, daily medical expenses, cost of medicine per patient, cost of medicine per patient per day, session and disease curing fees, healthcare expenditure, and insurance claimed. The results of applying neural network algorithm to the whole test sample and normal samples were 95.73 and 91.47 % respectively. This results further showed that the neural network model had an error rate of 9 % for classifying normal providers [16].

12.3.2.2 Bayesian Belief Network

Bayesian Belief Networks are dense networks of probabilities which capture information of the probabilistic relationships between variables as well as the past information about the relationships. This technique is very beneficial in those situations where some information is previously known and the incoming data is uncertain. These networks also give reliable semantics for characterizing effects and causes through an intuitive graphical representation. Due to these reasons it is used widely in those domains where automated reasoning is required [43].

In the research conducted by Ormerod et al. [12], a Mass Detection Tool (MDT) based on Bayesian Belief Network has been developed to detect healthcare fraud. This tool offers a real-time response as to the likelihood of several methods of fraud. It also recommends those unfamiliar indicators that can affect fraud possibilities and helps the claims handler improve online decision making.

Figure 12.3 shows the MDT framework, which helps link indicators to explicit types of fraud through a node-link topology. To forecast possible types of fraud for permutations of pointers, the network uses stored probability data. The most important advantage of the Bayesian Belief Networks is automated information updating. This updating is accomplished by the MDT in following three ways [12].

1. With the help of case result information from the other tool called Suspicion Building Tool (SBT), the provisional and previous possibilities of pointers and their associated fraud types are processed.
2. The claims handler classifies unexpected anomalies which uses the taxonomy educed in the ethnography. Then, each abnormality takes over the base rate possibility of identified fraud pointers in the similar subcategory. The information of prospective anomaly helps in reweighting the Bayesian Belief Networks, permitting the set of connections to acquire an explanation of unanticipated information as it arises.
3. A trace of anomalies and their results is developed over each instance. If an anomaly achieves better predictive power than a known pointer, then one is substituted by the other. Bayesian Belief Networks topology is updated with the help of SBT argumentation engine's output.

Hence, by this means the entire classification method maintains pace with the evolving types.

IBM teamed up with fraud investigation specialists and healthcare industry experts to develop a system to help detect fraud in insurance companies, health management organizations (HMOs) and for risk-bearing health care professionals. The system is known as Fraud and Abuse Management System (FAMS) and uses Fuzzy modeling along with decision support techniques for detection, investigation, prevention, and settlement [24]. Used with the Fuzzy modeling system, FAMS assigns a score to those providers who deviate from the normal behavior of their peer group. It has over 650 standard, individual behavior patterns such as percentage of specialty diagnoses and the average number of procedures per visit. For creating an analysis model, users choose and link behavior patterns suitable to the peer group they want to inspect from a library of functional objects. This model consists of around 25–30 behaviors. FAMS, used with analyses of claims data, calculates values for each provider in the model. Then, each value is allocated a score between 0 and 1,000 depending on the degree of deviation from the respective peer group norms. The score is higher if the deviation is greater. For scoring the behavior of each provider, FAMS uses Fuzzy membership functions. The values for each behavior pattern for all providers in the peer group are calculated, and the allocation of these values is analyzed by the system. Only providers are assigned scores with values greater than the peer group median. The investigation priority list is compiled of providers having the highest scores. By using the FAMS analysis tools, the suspected providers can be checked for fraudulent behavior [24]. The overall working of FAM is shown in Fig. 12.4.

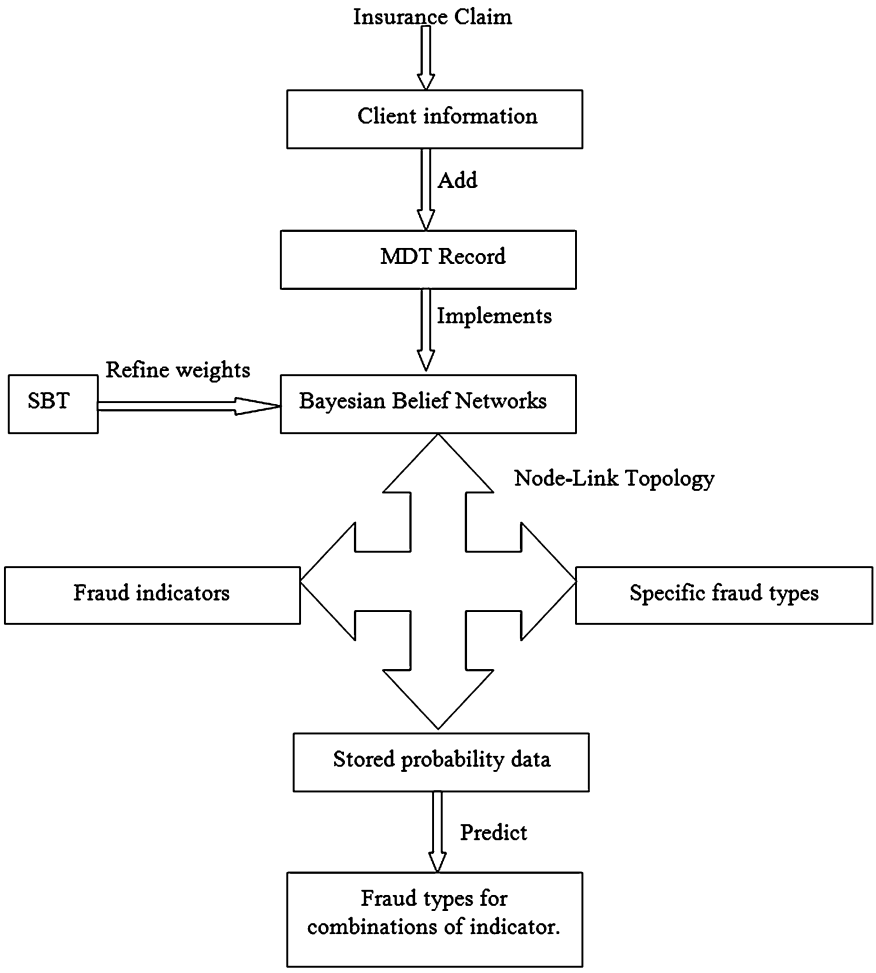


Fig. 12.3 Bayesian belief networks fuzzy modeling

12.3.2.3 Fuzzy Bayesian Classifier

Bayesian classifier is an important data mining technique, which can efficiently achieve optimal results when probability distribution is provided. Bayes rules can be used to computed the posterior from the likelihood and the prior, as the latter two is usually easy to be computed from a probability model [29].

Chan et al. [11] developed a new Fuzzy Bayesian classifier to audit health insurance fee data. Based on Bayesian inference, the Bayesian classifier acquires every attribute influencing the classification outcome. The Bayesian classifier classifies the case set more visibly by having an excellent control and understanding in the interpretation of the results. However, Bayesian inference requires

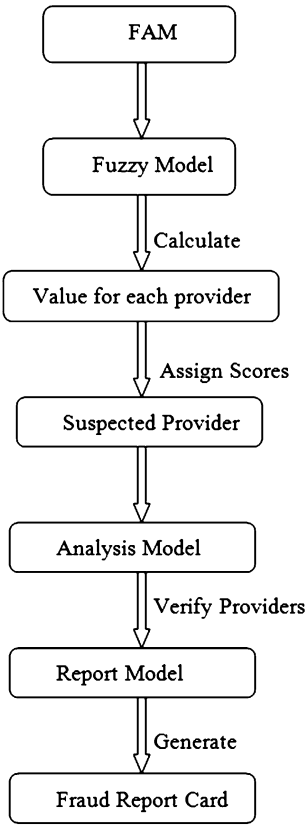


Fig. 12.4 FAMS fraud detection system

putting together diverse probability distributions while handling continuous attributes. Computation in this scenario will be very complex. This complexity can be overridden when this method is used along with Fuzzy set theory. With this hybrid method, it is possible to convert continuous attributes into discrete attributes. For their experiments, Chan et al. used 800 records containing health insurance fee data. Of these records, 166 were fraudulent and 634 were normal. They used three methods (80/20, 70/30, 60/40) to divide the training and testing datasets. Then, the training data was fed to Bayesian classifier for training the classification rules, and the remaining, testing, dataset was classified. The values of sensitivity, specificity, and accuracy were computed for all three methods. It was found that the quality of the proportion 80/20 was best, as it had the highest sensitivity (0.639), the highest specificity (0.968), and the highest accuracy (0.894), respectively. It is observed that the overall accuracy of the classifier is good, but sensitivity is slightly less than required. The reason for this low sensitivity is that the attributes chosen for detecting the fraud from the health insurance fee data are not adequately represented. Figure 12.5 demonstrates the overview of Fuzzy Bayesian classifier.

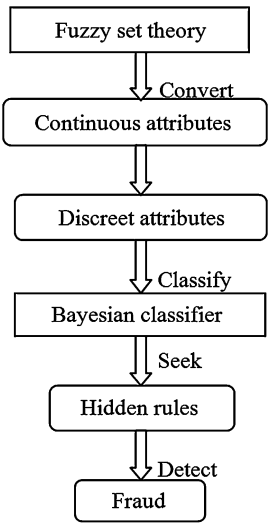


Fig. 12.5 Fuzzy bayesian classifier data mining model

12.3.2.4 Logistic Regression

Logistic Regression is a technique that is nonlinear and is used to represent the variables that are binary dependent. There can be only two values for classification variables, a value that signifies success or a value that signifies failure. The advantage of the logistic regression function is that it can be easily interpreted. The logistic regression technique was used by Liou et al. [16] to detect fraudulent and normal hospitals. In their experiments, they defined that claim was related with the value of zero if it was a regular claim and with the value of one if it was an irregular claim. As shown in Fig. 12.6 nine expense related variables were selected for the detection model. Logical regression was executed on all of these models independently to recognize the most efficient factors. Eight out of nine variables were observed to contain predictive power. The variable that was left out was the average medical expenditure. These eight detectors were then employed to form a complete logistic regression model. The model's successive rate was 100 % for detecting those hospitals involved in fraudulent activities. Along with this rate, the model also had an 84.6 % identification rate for normal hospitals. This rate shows that the logistic regression model has error rate of 15 % in classifying normal providers. The entire sample had a 92.2 % correct identification rate.

In another case, the logistic binomial regression technique was utilized by Table et al. [21] in detecting the health care fraud. Dependent, dichotomous variables were used to represent either fraudulent or non-fraudulent values. Four variables were selected from the database. They were (1) number of days of sick leave granted by the attending physician, (2) amount to be paid for granted days of sick leave, (3) classification of disease by whether it was diagnosable (true) or not (false), and (4) history of health insurance reimbursement claims. With these

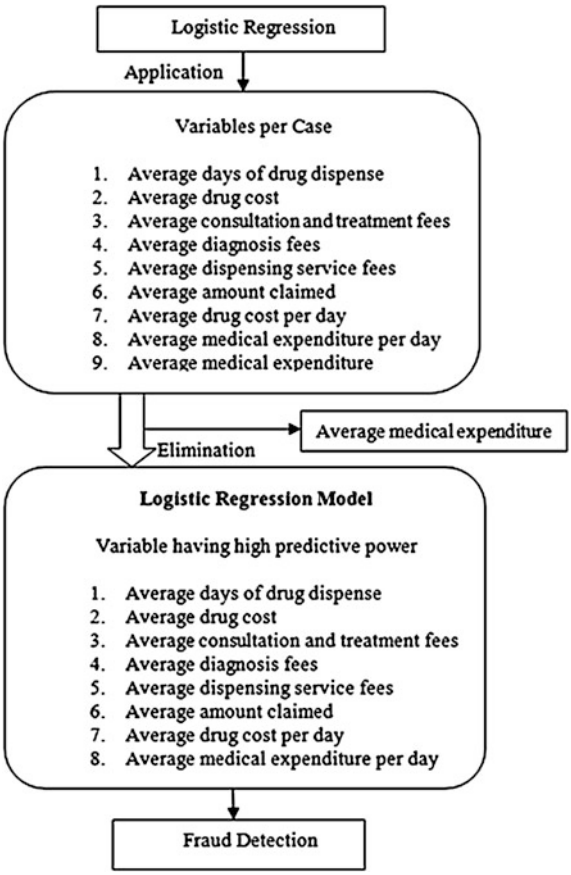


Fig. 12.6 Logistic regression. model for detection of fraudulent claims

variables, a binomial logistic regression model was developed with the help of following equation.

$$PFLMC = \beta_0 + (\beta_1 D) + (\beta_2 M) + (\beta_3 C) + (\beta_4 N),$$

where $PFLMC$ = probability of fraud, β_0 = constant regression, β_{1-4} = specific beta coefficients for each of the four independent variables, D = days of medical leave granted by the treating physician, M = amount to be paid in cash for sick days, C = diagnostic classification, and N = number of sick leave accumulated by an individual.

The variables with diagnostic classification and existence of multiple requests for leave of absence of an individual had strong predictive power, while variables demonstrating leave and total pay showed a marginally significant predictive power. The most characteristic features of these fraudulent behaviors were the existence of multiple requests for leaves of absence for one person and difficult diagnostic testing. If the number of sick days increase, then it was more likely that

fraud is taking place. The model performed well with sensitivity (99.71 %) and specificity (99.86 %). The positive predictive value, or percentage of fault detected correctly by the model, was around 98.59 %, and the negative predictive value, or the percentage of non-fraudulent cases detected correctly by the model, was 99.97 %.

12.3.2.5 Classification Tree

Classification trees are built with the help of rules, training sets, samples in the dataset, and can directly be applied in the form of a simple detection algorithm. The sequence of classification rules is signified by the "every probable pathway from the root node to a leaf node" technique as shown in Fig. 12.7. Liou et al. [16] were able to detect 100 % of fraudulent hospitals using a classification rule. The correct identification rate for the fraudulent hospitals was 99.30 % for the entire dataset and 98.73 % for the normal hospitals. This results showed 1 % fault rate in categorizing normal providers. Sequence of some classification rules for fraud detection can be seen in Fig. 12.7 and more rules depending on the requirement of the system can be added further.

12.3.2.6 Genetic Algorithm

Genetic algorithm (GA) is a search heuristic which is based on the process of natural selection and genetics. GA's not only perform better than other traditional methods in the majority of the problem link but also offers different methods in the majority of the problem link. GA has the ability of finding optimal parameters for the real world problems which is quite hard in the case of traditional methods [44].

The GA is considered as an ideal technique to solve optimization problems. Its application gets better matches amid when used with examination by expert consultants and the classifications of a system. He et al. [10] used the GA for selection, crossover, mutation, and cost functions to discover the best possible weighting of attributes utilized to categorize the practice profile of general practitioners. They utilized a validation dataset to optimize the weight. In this research, the GA proved to be effective, as each run required only a 2,000 generation, getting the preferred agreement rate for the given validation dataset. Figure 12.8 demonstrates the working of GA.

12.3.2.7 K Nearest Neighbor

K-nearest neighbor (KNN) is an extensively employed profile-matching method that establishes categorizations of every case on its nearest neighbors by applying several decision rules. As shown in Fig. 12.9, He et al. [10] used the weights of the features determined by genetic algorithm in the KNN technique to recognize the

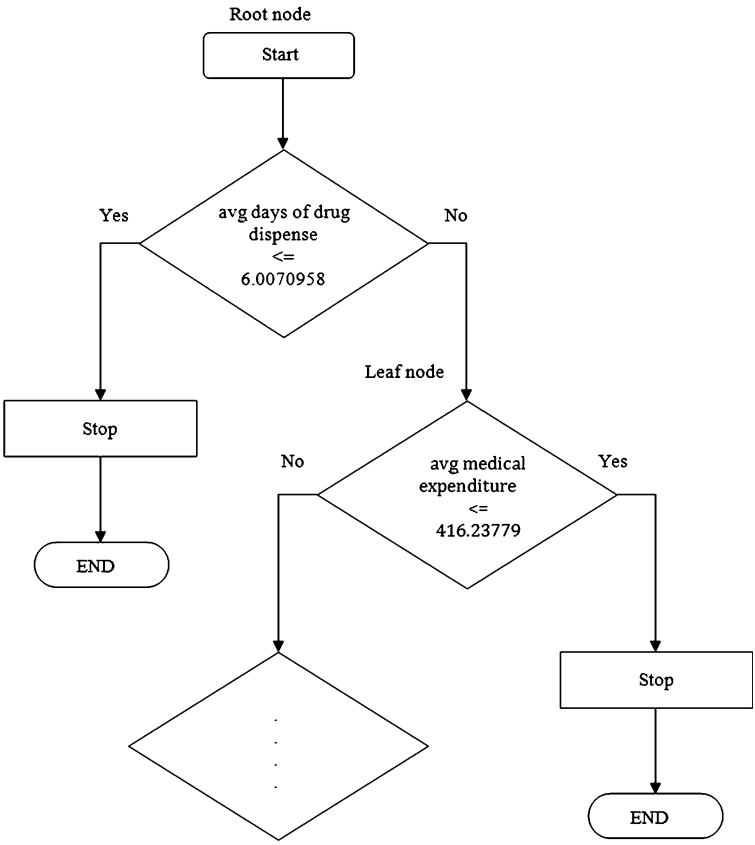


Fig. 12.7 A snapshot of rules involved in a classification tree

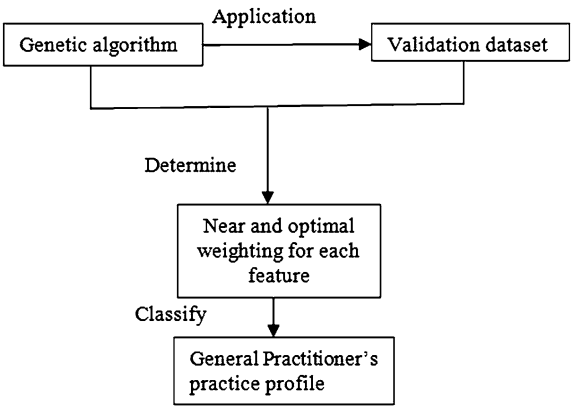


Fig. 12.8 Working of genetic algorithm

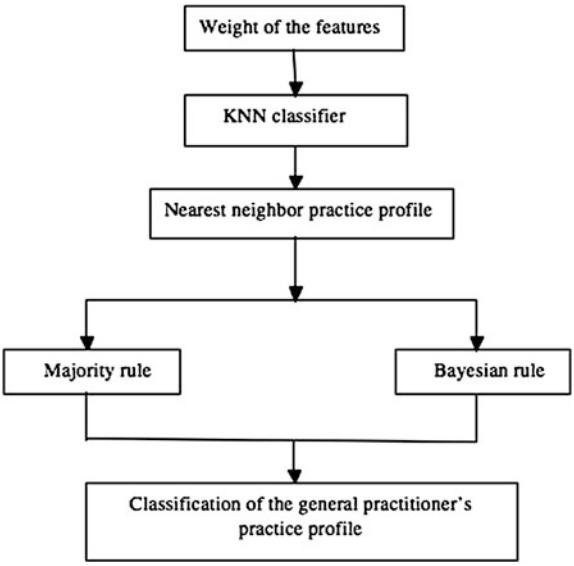


Fig. 12.9 Application of KNN classifier

nearest neighbor profile of the general practitioners. Then, the majority and Bayesian rules were used to classify the general practitioner profiles. The KNN classifier was trained with the help of the nearest neighbor examples provided by the profiles in the training dataset. The trained classifier was then tested on the test dataset. The KNN classifier, along with the genetic algorithm for optimizing the weights, uses Euclidean distance to improve the results of classification. The statistic employed to measure the efficiency of the KNN classifier and variants was the synchronization rate that is basically the percentage of synchronization amid the categorizations of both the KNN classifier and that of expert consultants partitioned by the number of cases in the dataset. Using the majority and the Bayesian rule along with the KNN classifier helped achieve the high agreement rate in this scenario.

12.3.2.8 Association Rules

The use of association rules is one of the many data-mining techniques applied to discover corresponding relationships and remarkable association among a huge array of data items. Association rules demonstrate feature value states which take place together in a known dataset repeatedly. This type of information is given in *if-then* statements by these rules. These rules are determined from the data set provided. Unlike, if-then rules that have logical characteristics, association rules have probabilistic attributes. Along with the precursor that is the 'if' part and the descendent 'then' part, an association rule has an additional two numbers which state the measure of improbability regarding the rule. In association analysis, the

precursor and descendent are itemsets which do not have mutual entries. The first number is identified as support for the rule which is basically the number of transactions that contain every item in the precursor and descendent section of the rule. Sometimes support is denoted as a fraction of the sum of entire records in the database. The second number is identified as the confidence of the rule which is the relative amount of the number of transaction that contain every items in the precursor as well as the descendent to the number of transactions that contain every items in the descendent [18].

Association rules have been widely used [14, 15, 17] for detecting fraud committed by medical service providers in the healthcare system. Until recently, positive association rules have been used to discover frequent patterns. However, the use of negative association rules [17] has proved effective in detecting fraud in the healthcare system. Shan et al. [17] identified around 215 association rules from the dataset, which consisted of 23 positive and 192 negative association rules for detecting inappropriate billing by specialists. The negative association rules outnumbered positive association rules because negative rules of both the presence and absence of the item were found, while only presence of item was considered for positive rules. In addition, negative rules were stronger than positive rules in terms of confidence; minimum confidence of negative association rules was 95.95 %, while for positive rules it was 80.25 %. The common patterns corresponding to negative rules were considered reliable, with the billing regulations enclosed under the Medicare Benefit Schedule. It was found that negative rules proved to be more intuitive and useful for locating violations than positive rules were. The violations of negative rules included billing items that were not generally billed by the majority of specialists. Those specialists who were found to be violating these rules repeatedly were marked unusual from their peers. Among 192 negative rules, 30 were found to have confidence value of 1.0 and were considered unimportant for fulfillment. Hence, these rules were removed. The remaining 162 rules were divided into three groups based on the probability of improper billing by a subject matter expert. A high rating suggested that a rule was crucial, and if this rule was broken then there was a high probability of inappropriate billing to Medicare Australia. On the other hand, a low rating suggested that if a rule was broken, then this behavior may indicate inappropriate billing or another valid explanation for billing may also exist. Hence, it was determined that a low rating may not be powerful for finding inappropriate billing information, but may be helpful in obtaining useful information on identifying specialists that had related compliance activities. It was proven from the experiment that more than half of the rules, i.e. 56.18 %, were considered high and/or medium rated and was regarded as suitable for detecting inappropriate billing. With the help of a consultant, 162 negative rules were rated from high to low. The specialists who were discovered breaking these rules more, were often classified as high risk providers. These rule violations helped indicate how much one specialist deviated from their peers. Using a database called Program of Research Informing Stroke Management (PRISM), maintained by Medicare Australia and containing information from those specialists who were approached for previous compliance activities, were

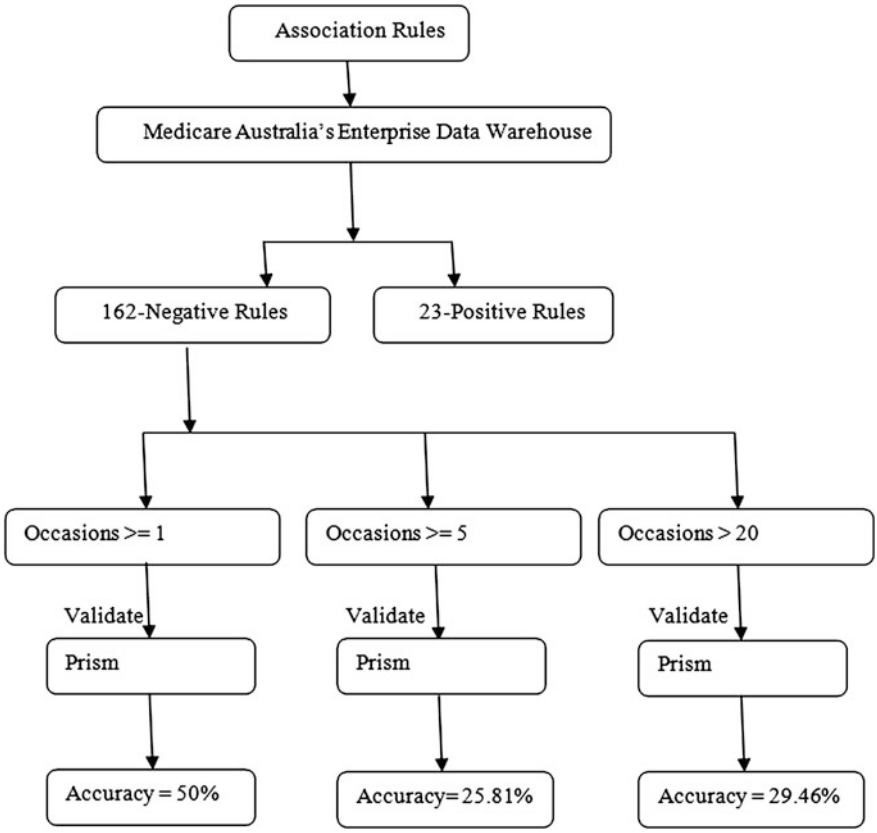


Fig. 12.10 Application of association rules

matched with suspected specialists who broke the rules. The specialists who broke rules were further divided into three classes:

1. Occasions $> = 1$,
2. Occasions $> = 5$, and
3. Occasions > 20.

It was found that ten specialists broke the rules on more than twenty occasions. Of these, five had a record in PRISM. This number showed the accuracy of these association rules to be 50 %. For those who broke the rules on more than five occasions, the accuracy was found to be 25.81 %, whereas the accuracy was 29.46 % for the ones who broke rules on one or more occasion. Hence, these results suggest that even breaking one negative rule can be a good indication of noncompliant practice. Figure 12.10 demonstrated the application of Association Rules on PRISM database.

In another research conducted by Viveros et al. [14] association rules were applied to the episode database used for pathology services, in which the visit of each patient was linked to a record in the database. Hence, with a unique identifier, a database tuple was obtained. This tuple could consist of one or more medical tests performed in any instance of time with a maximum number of 20 tests per episode. Association rules were obtained with the setting of 50 % minimum confidence and values of 1, 0.5, and 0.25 % for minimum support. With a minimum confidence of 50 % and a minimum support of 1 %, 24 association rules were obtained. For a minimum confidence of 50 % and a minimum support of 0.5 %, 64 association rules were obtained. Further using a minimum confidence of 50 % and a minimum support of 0.25 %, 135 association rules were obtained. It was found that more knowledge of the behavior patterns was obtained by setting the minimum support 0.5 % rather than to 1 %.

12.4 Summary

In summary, the Supervised data-mining methods can be effective in discovering fraudulent transactions in healthcare system as they require accurate identification of fraudulent transactions and are trained to discriminate non-fraudulent and fraudulent transactions. Among all the supervised data-mining methods mentioned above, the neural network and association rules perform the best and, thus, are more often used by researchers to detect fraudulent patterns in healthcare data. Table 12.1 provides a summary of the supervised methods discussed in this chapter.

12.4.1 Advantages of Fraud Detection

The following are the advantages [25] that can be gained by efficiently detecting fraud in the healthcare system using information technology related tools like data mining and machine learning techniques:

- Yearly recovery of expenditures by the government and private sector for fraudulent claims,
- Reduction of the probability of fraudulent up-coding,
- Detection of new leads from increased accessibility of supplementary digital fingerprints,
- Authentication and confirmation of genuine services by call-back or web-based services,
- Digital authentication of services submitted by patients and providers,
- Decrease in record gathering time by use of a common identifier and growing digital media,

Table 12.1 A summary of supervised methods for fraud detection in healthcare

Algorithms based supervised method	Advantages	Disadvantages
Neural network	It requires less formal training. It has ability to sense complex non-linear relationships between independent and dependent variables. It also has an ability to detect all potential interactions between predictor variables, and the availability of multiple training algorithms [33]	It has black box type of nature and need high computation power. It can be prone to overfitting and its model development is of empirical nature [33]
Association rules	It can handle both table form data and transaction form data. It doesn't require the whole database to be fetch into the main memory [35]	Only database entries which exactly match the candidate patterns may contribute to the support of the candidate pattern. This creates a problem for databases containing small variations between otherwise similar patterns and for databases containing missing values [35]
Logistic regression	It is robust in nature and also independent variables does not required to be normally distributed, unbounded or interval. It can handle nonlinear effect [33]	It needs lots of data for achieving stable and meaningful results [33]
Classification tree	It has built in feature selection method. Models are easily interpretable, and are robust to the effect of outliers by easily dealing with missing values [34]	It can be unstable with small variation in the data. Some tree models can be very large and complex. It is expensive to be trained [34]
KNN classifier	As the entire process in KNN classifier is transparent, it is easy to implement and debug. It has special noise reduction technique which helps in getting accurate results [36]	For large training set it can have poor runtime performance as all the work in KNN classifier is done at run time. It is very sensitive to redundant features [36]
Genetic algorithm	It can rapidly scan a huge solution set. Bad proposals do not influence the end solution negatively as they are simply discarded. It works by its own rules without the need of knowing the specific rules of the problem. It is very useful for complex problems. Also its one of the main strength is the parallel nature of its stochastic search [38]	It is usually slower than traditional techniques [38]
Fuzzy logic	It allows the use of vague linguistic terms in the rules [37]	It is difficult to estimate the membership function [37]
Bayesian network	It offers a natural and principled means of merging prior information with data. It also offers suitable settings for wide range of models [39]	It does specify about how it is selecting a prior and can generate posterior distributions that are greatly controlled by the priors. It is having high computational cost [39]

- Computerized digital verification for claims billing and payment,
- Real-time billing and confirmation of eligibility reimbursement,
- Decrease in customer time used in dealing with fraudulent claims,
- Prevention of duplication of imaging and laboratory tests,
- Prevention of creation of superfluous information which is already accessible digitally,
- Decrease in labor time to validate qualification,
- Decrease in material and effort to examine paper documentation,
- Decrease in time to store and retrieve paper records,
- Decrease in time used by a consumer in phone trees and taping redundant information,
- Decrease in communal medical fee and loss of life due to medication faults,
- Decrease in communal medical fee and loss of life due to clinical faults,
- Decrease in communal medical fee of duplicate diagnostic tests,
- Decrease in communal medical fee and loss of unnecessary medical surgeries,
- Decrease in referral visits to monitor prospective care provider through screening from a pay for performance initiative, and
- Decrease in provider time, bundling, accumulating, and forwarding of documentations to health plans, providers, and patients.

Hence, many benefits can be gained from new fraud detection techniques, which can help provide better medical services to authentic patients, help save money, and improve the healthcare experience for patients with genuine needs.

12.5 Conclusion

One of the most crucial problems facing the US government is fraud in healthcare system. Due to a large amount of data, it is impossible to manually audit for fraud. Hence, many statistical approaches have been proposed to overcome this problem. As fraud can be committed in complex and numerous ways, fraud detection is challenging, and there is a greater need for working models for fraud detection, including types of fraud that are not yet in use, as these models will not be outdated quickly.

To establish a well-functioning healthcare system, it is important to have a good fraud detection system that can fight fraud that already exists and fraud that may emerge in future. In this chapter, an attempt has been made to classify fraud in the healthcare system, identify data sources, characterize data, and explain the supervised machine-learning fraud detection models. Even though a large amount of research has been done in this area, more challenges need to be worked out. Fraud detection is not limited to finding fraudulent patterns, but to also providing faster approaches with less computational cost when applied to huge-sized datasets.

References

1. CMS (2011) Research, statistics, data and systems: national health expenditure data. NHE fact sheet
2. CMS (2011) Medicare: HCPCS–general information
3. FBI (2009) Reports and publications: 2009 financial crimes report
4. NHCAA (2007) The NHCAA fraud fighter's handbook: a guide to health care fraud investigations and SIU operations
5. IMF (2008) World economic and financial surveys: world economic outlook
6. Database NHCAA (2010) Combating health care fraud in a post-reform world: seven guiding principles for policymakers
7. NHCAA The problem of health care fraud, consumer alert: the impact of health care fraud on you, report of national health care anti-fraud association (NHCAA)
8. Koh H, Tan G (2005) Data mining applications in healthcare. j healthc inf mgmt 19(2):64–72
9. OIG (2011) Medical fraud cases: OIG most wanted fugitive
10. He H, Hawkins S, Graco W, Yao X (2000) Application of genetic algorithms and k-nearest neighbor method in real world medical fraud detection problem. J Adv Comput Intell Intell Inf 4(2):130–137
11. Chan CL, Lan CH (2001) A data mining technique combining fuzzy sets theory and bayesian classifier—an application of auditing the health insurance fee. In: Proceedings of the International conference on artificial intelligence, pp 402–408
12. Ormerod T, Morley N, Ball L, Langley C, Spenser C (2003) Using ethnography to design a mass detection tool (MDT) for the early discovery of insurance fraud. In: Proceedings of the ACM CHI conference, 650–651
13. Ortega PA, Figueroa CJ, Ruz GA (2006) A medical claim fraud/abuse detection system based on data mining: a case study in chile. In: Proceedings of international conference on data mining, 224–231
14. Viveros MS, Nearhos JP, Rothman MJ (1996) Applying data mining techniques to a health insurance information system. In: Proceedings of the 22nd VLDB conference, Mumbai, India, pp 286–294
15. Yang WS, Hwang SY (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31:56–68
16. Liou F, Tang Y, Chen J (2008) Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Manage Sci, 353–358
17. Shan Y, Jeacocke D, Murray D, Sutinen (2008) A mining medical specialist billing patterns for health service management. In: Roddick J, Li J, Christen P, Kennedy P, (eds) Proceeding 7th Australasian data mining conference (AusDM 2008), Glenelg, South Australia. CRPIT, 87. ACS 105–110
18. Sokol L, Garcia B, West M, Rodriguez J, Johnson K (2001) Precursory steps to mining HCFA health care claims. In: Proceedings of the 34th Hawaii International conference on system sciences
19. Yang WS (2002) Process analyzer and its application on medical care. In: Proceedings of 23rd International conference on information systems (ICIS02), Spain
20. Li J, Huang K, Jin J, Shi J (2008) A survey on statistical methods for health care fraud detection. Health Care Manage Sci, 275–287
21. Table F, Raineri A, Maturana S, Kaempffer A (2008) Fraud in the health systems of chile: a detection model. Am J Public Health, pp 56–61
22. Ghahramani Z (2004) Unsupervised learning
23. Rosella (2011) Predictive knowledge and data mining: healthcare fraud detection
24. Hall C (1996) Intelligent data mining at IBM: new products and applications. Intell Softw Strateg 7(5):1–11
25. Report on the use of health information technology to enhance and expand health care anti-fraud activities. Foundation of research and education of AHIMA

26. FBI (2011) Scams and Safety: common fraud schemes
27. London: The Guardian (2007) The mystery of John Darwin
28. Herb Denenberg (2005) The denenberg report: the insurance commissioners, other government agencies, and the insurance companies focus on insurance fraud committed by policyholders, but nothing is done about the multi-billion dollar racket of insurance fraud committed by insurance companies
29. Bhuvaneswari R, Kalaiselvi K (2012) naive bayesian classification approach in healthcare applications. Int j comput sci telecommun, 3(1):106–112
30. Silver M, Sakata T, Su HC, Herman C, Dolins SB, O'Shea MJ (2001) Case study: how to apply data mining techniques in a healthcare dataware house. J Healthcare Inf Manage 15(2):155–164
31. Relles D, Ridgeway G, Carter G (2002) Data mining and the implementation of a prospective payment system for inpatient rehabilitation. Health Serv Outcomes Res Method 3(3–4):247–266
32. Anonymous (1999) Texas medicaid fraud and abuse detection system recovers $2.2 million, wins national award. Health Manag Technol 20(10):8
33. Tu J (1995) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol pp 1225–1231
34. Lewis R (2000) An introduction to classification and regression tree (CART) analysis. Presented at annual meeting of the society for academic emergency medicine
35. Nayak J, Cook D (2001) Approximate association rule mining. In: Proceedings of the 14th International florida artificial intelligence research society conference
36. Cunningham P, Delany S (2007) k-Nearest neighbour classifiers. Technical report, UCD-CSI-2007-4
37. Russel S, Norvig P (2003) Artificial intelligence: a modern approach. Prentice-Hall, 2nd edition
38. Vose D (1995) The simple genetic algorithm: foundations and theory
39. Berger J (2006) The case for objective bayesian analysis. Bayesian Anal 1(3):385–402
40. Vick K (2009) As rescissions spawn outrage, health insurers cite fraud control. The Washington post, http://www.washingtonpost.com/wp-dyn/content/article/2009/09/07/AR2009090702455.html, Information Accessed on May 2012
41. Jeffries D, Zaidi I, Jong B, Holland M, Miles D (2008) Analysis of flow cytometry data using an automatic processing tool. Cytometry Part A 73A:857–867
42. Larose D (2005) Discovering knowledge in data, An introduction to data mining. Wiley InterScience
43. Niedermaye D (2008) An introduction to bayesian networks and their contemporary applications, innovations in bayesian networks. Springer, pp 117–130
44. De Jong KS, Spears WM, Gordon DF (1993) Using genetic algorithms for concept learning. Mach learn 13:161–188

Chapter 13
Feature Extraction by Quick Reduction Algorithm: Assessing the Neurovascular Pattern of Migraine Sufferers from NIRS Signals

Samanta Rosati, Gabriella Balestra and Filippo Molinari

Abstract A migraine is a neurological disorder that can be caused by many factors, including genetic mutations, life-style, cardiac defects, endocrine pathologies, and neurovascular impairments. In addition to these health problems, an association between some types of migraines and increased cardiovascular risk has emerged in the past 10 years. Moreover, researchers have demonstrated an association between migraines and impaired cerebrovascular reactivity. It is possible to observe carbon dioxide dysregulation in some migraineurs, while others show a markedly decreased vasomotor reactivity to external stimuli. Therefore, the assessment of the cerebrovascular pattern of migraineurs is important both for the onset of a personalized therapy and for follow-up care. Near-infrared spectroscopy is a widely used tool for the non-invasive monitoring of brain oxygenation. It can be used to track hemodynamic changes during external stimulation (i.e. vaso-active maneuvers such as hypercapnia or hyperventilation). Unfortunately, near-infrared spectroscopy (NIRS) signals acquired during vaso-active maneuvers are non-stationary and require a time–frequency processing approach. To fully describe the cerebrovascular patterns of migraineurs, we extracted several parameters from the NIRS signals. Using these parameters, we compiled a dataset in which complexity was very high and the clinical/physiological information was impossible to track.

13.1 Introduction

In this chapter, we present our latest feature selection approach. This approach is based on the quick reduction algorithm (QRA). We compared the results of this automated technique with the traditional ANOVA (ANalysis Of VAriance

S. Rosati · G. Balestra · F. Molinari (✉)
Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24 10129 Torino, Italy
e-mail: filippo.molinari@polito.it

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_13,

between groups) analysis. An artificial neural network (ANN) was employed in order to classify the subjects, using the extracted features as input parameters for the net. Our QRA-based system correctly classified 97.5 % of subjects using only nine of twenty-six features from the dataset. The ANOVA analysis extracted three features, but classified only 75 % of subjects correctly. Our QRA-based procedure is fully automated and has proven effective when tested on real clinical data. The extracted features will be used in a real clinical applications for the cerebrovascular assessment of migraineurs with and without aura disturbances.

13.1.1 Background

It has been shown that a migraine is a neurological disorder that correlates with an increased risk of subclinical cerebral vascular lesions [1]. Epidemiological studies showed that migraineurs are prone to an increased risk of vascular accidents [2], and this has led many researchers to consider a migraine as a systemic vasculopathy [3]. The association between a migraine and impaired cerebral autoregulation or vasomotor tone has been investigated and assessed widely [4–6].

However, there is a difference in the cardiovascular and cerebrovascular risks associated to the two types of migraines. Subjects suffering from migraine with aura (MwA) showed greater impairments than subjects suffering from migraines without aura (MwoA) [1–6]. Given the correlation between migraines and vascular disorders, migraine sufferers usually undergo an assessment of the cerebrovascular status. The accurate assessment of the cerebrovascular reactivity can be of paramount importance for the onset of a personalized and proper therapy.

Near-infrared spectroscopy (NIRS) is a non-invasive, real-time, and cost-effective monitoring technique for the assessment of the cerebral autoregulation of subjects [4]. In NIRS, infrared light is injected into the skull, and the changes in the concentration of oxygenated (O_2Hb) and deoxygenated (or reduced) (HHb) hemoglobin are instantaneously measured. The assessment of cerebral vasomotor reactivity (i.e., the arteries' capacity of compensating systemic blood pressure alterations) is of primary importance to assess the overall status of the artery bed. Usually, the active maneuvers like breath-holding (BH), hyperventilation (HYP), or Valsalva, are performed during the monitoring phase in order to assess cerebral autoregulation and vasomotor reactivity [4, 7, 8]. Such maneuvers are easy to perform and are safe for pathological subjects. Specifically, BH is a stimulus that determines vasodilation, because it increases the concentration of the carbon dioxide in the blood. Conversely, HYP triggers cerebral vasoconstriction, because of the increase of oxygen in the blood. Therefore, overall, the NIRS is a suitable system for long-term, bedside, or home monitoring and assessment. The use of NIRS for the assessment of migraineurs is gaining clinical importance. Recently, Watanabe et al. used NIRS to monitor the hemodynamical changes occurring during a migraine attack after the administration of sumatripan [9]. Viola et al.

studied the pathophysiology of prolonged migraine attacks by monitoring the cerebral oxygenation [10].

However, the accurate assessment of the cerebral autoregulation of migraineurs is complicated by many factors: it has been shown that cerebral oxygenation changes in the presence of smoke habits [2], of patent foramen ovale and other atrial septal defects [11–13], and of mutations of the 677-MTHFR gene [14]. In a recent paper, Giustetto et al. showed that there is a relationship between the vascular pattern of migraineurs, as assessed by NIRS, and some hematological parameters [14]. These relationships were different between subjects suffering from MwA and MwoA.

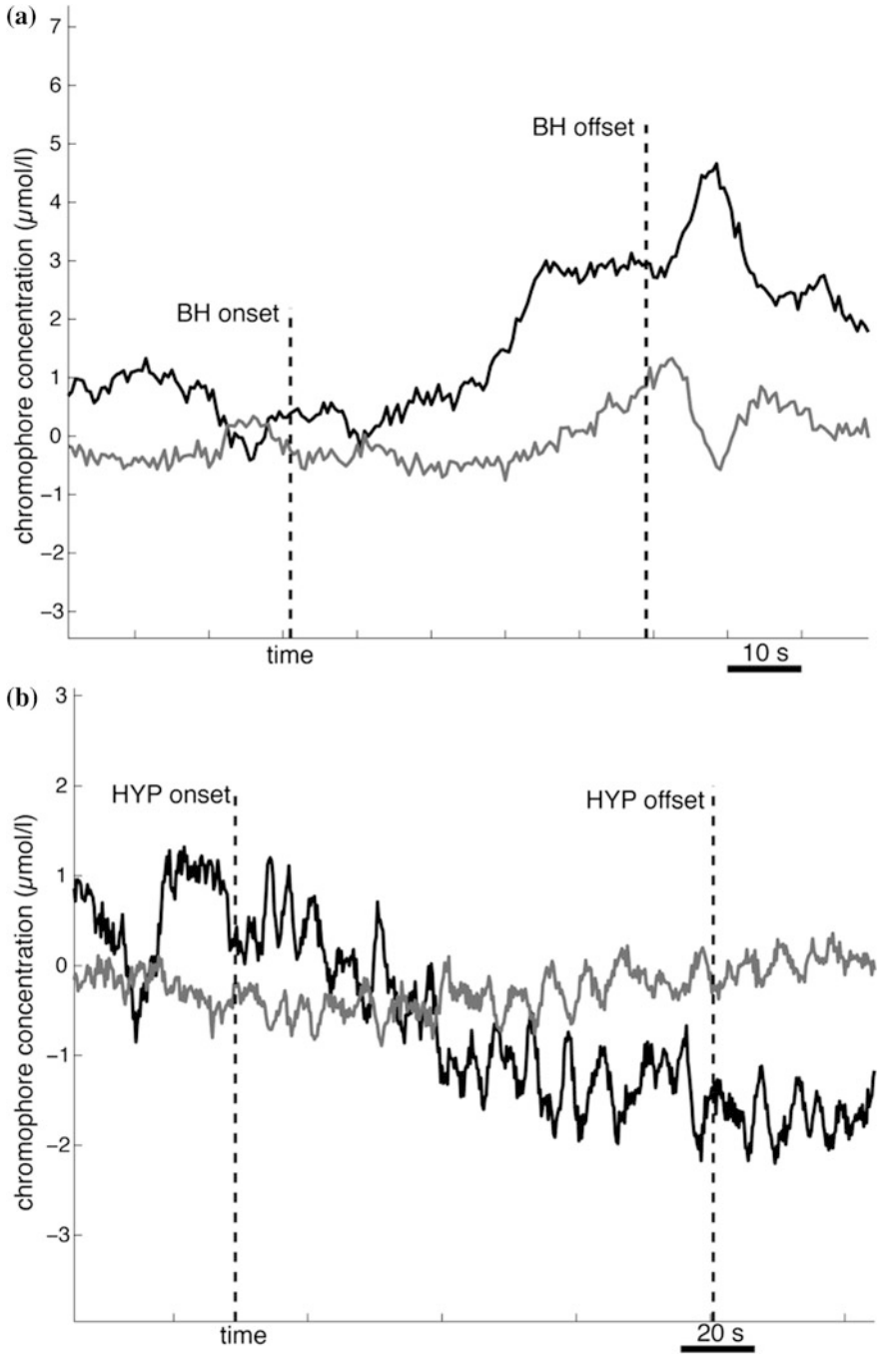
Another complication is given by the fact that NIRS concentration signals are usually considered nonstationary when recorded during vaso-active maneuvers (i.e., breath-holding and hyperventilation). Researchers studied the spontaneous cerebral low-frequency oscillations of NIRS signals recorded during active maneuvers [15]. They studied a group of healthy volunteers, which were used to derive the basis of frequency-derived parameters used to assess cerebral autoregulation. In fact, several studies showed that different cerebral hemodynamic signals (including NIRS signals, transcranial Doppler signals, and the fMRI BOLD signals) show a power spectrum essentially consisting of two bands [16].

1. A very low-frequency band (VLF—also called B-waves) that reflects the long-term autoregulation. Even though the actual origin of this band is still debated, VLFs are thought to be generated by brain stem nuclei, which modulate the lumen of the small intra-cerebral vessels. In humans, the frequency range of the VLF band is typically 20–40 mHz.
2. A low-frequency band (LF—also called M-waves) that is common to most mammalians. This frequency band reflects the systemic oscillations of the arterial blood pressure and, therefore, mainly reflects the sympathetic system activity. The frequency range of the LF band is typically 40–140 mHz.

Figure 13.1a shows an example of NIRS signals recorded during the BH, whereas Fig. 13.1b shows an example of HYP (the subject was a healthy volunteer).

In this study, we present a robust feature extraction technique for the assessment of the vascular pattern of migraineurs. The idea was based on the well-known fact that increasing the number of features used to build a classifier does not mean increasing its accuracy: several attributes may be irrelevant or, even worse, may introduce noise that decreases the classifier performance. To improve the classification accuracy, it is then important to select the useful features that reduce the number of attributes. This reduction can be obtained through selection or construction [17].

In construction, new attributes are created on the basis of some original features. Construction has the disadvantage that the results are difficult to interpret because they do not correspond to the original features.


(a)
chromophore concentration (μmol/l)
BH onset
BH offset
time
10 s
(b)
chromophore concentration (μmol/l)
HYP onset
HYP offset
time
20 s

◀ **Fig. 13.1** NIRS signals recorded on a healthy volunteer during breath-holding (**a**) and hyperventilating (**b**). The *red line* represents the O2Hb concentration signal, the *blue line* the HHb signal. The black vertical *dashed lines* mark the onset and offset of the breath—holding (**a**) and hyperventilation (**b**). In a recent paper, Molinari et al. measured the VLF and LF power changes in the NIRS signals of migraineurs during active stimulations [25] and documented the cerebral hemodynamics observed between MwA and MwoA sufferers. They applied principal component analysis (PCA) to the power spectral parameters of the NIRS signals recorded during BH and HYP. Results showed a clear correlation between vascular parameters and pathology. The methodology, however, was based on a strong feature reduction procedure, and it was shown that by changing the features used for PCA, the association between the vascular pattern and pathology changed

Feature selection is based on the idea of reducing the number of attributes by collecting the smallest number of important features from the original set without significantly deteriorating the classification accuracy. When feature selection is needed, there must be an appropriate and well-defined criterion to measure the relevance of the chosen features. However, as the number of initial features is usually large, it is computationally impossible to test all possible subsets of them, even if the criterion is simple to evaluate. A heuristic procedure is then applied to find a good set of features in a reasonable amount of time. Another important difference among the possible feature selection methods is the hypothesis that some systems are linear, and this requirement is needed in order to provide good results. Moreover, because most real situations are non-linear, two features may be useless when taken individually but will become highly predictive when used together. In addition, the required training set should be larger for a larger number of features [18, 19].

We chose two of several approaches available for dimensionality reduction. The first method, based on the Rough-Set Theory (RST), consists of the Quick-Reduct Algorithm (QRA); the second method, based on a linear model of data, is the ANalysis Of VAriance (ANOVA). RST offers a formal methodology for feature selection. Computationally, the approach is efficient. Unlike statistical correlation-reducing approaches, it requires no human input, and it preserves the semantics of the data, which makes the results more understandable.

In this study, we present a comparison of the two method's performances when applied to a dataset of features that describe the time and frequency changes of the hemoglobin (both in its oxygenated and reduced form) concentrations as measured by NIRS in a population of women suffering from MwA and MwoA.

13.2 NIRS System and Measurement Protocol

NIRS is a spectroscopic technique used for the monitoring of the concentration of oxygenated (O_2Hb) and reduced (HHb) hemoglobin in the human brain. NIRS allows non-invasive and real-time examinations. The possibility of quantifying the concentration of O_2Hb and HHb is given by the different optical properties of the

two hemoglobin types. In fact, the two hemoglobins have different absorption spectra. Therefore, it is possible to differentiate their concentration by irradiating the brain with two light wavelengths. A substance that absorbs light at a particular wavelength is called *chromophore*. The most important cerebral NIRS chromophores are O_2Hb, HHb, and the cytochrome-c-oxidase (which is a neuronal metabolic marker). These three chromophores have absorption peaks at wavelengths shorter than water, lipids, plasma, muscles, and bones. Hence, most of the tissues of the head and brain complex can be neglected since their absorption peaks are far from the infrared region [20, 21]. Being the cytochrome-c-oxidase mainly a metabolic marker (hence linked to a functional aspect of brain activation rather than to a hemodinamic aspect), we did not consider it in this present study.

In NIRS systems, an electromagnetic field in the infrared band (wavelengths ranging from 650 to 870 nm are usually used) is used to irradiate the skull. The source is usually a photoemitter (LED or laser diodes can be used) placed on the scalp. The receiver is usually placed few centimeters aside the source. The measurement of the chromophores concentrations is made by comparing the emitted light intensity to the received intensity. The light photons travelling in the tissues can be absorbed or scattered. Since the NIRS systems have the receiver placed close to the source, they work on the scattering basis. Only in newborn infants, who have a very soft and not fully calcified skull, it is possible to perform an absorption-based NIRS analysis.

Scattering introduces errors and forces the modification of the traditional absorption equation, which cannot be used in its raw version to measure the chromophore concentration changes. The traditional absorption Beer–Lambert law is redefined as (modified Beer–Lambert law):

$$\Delta A(\lambda) = L(\lambda) ln(10) \sum_i \in_i (\lambda) \Delta c_i, \tag{13.1}$$

where

- $\Delta A(\lambda)$ is the attenuation change at the wavelength λ;
- $L(\lambda)$ is the total pathlength (mm) traveled by the photons at wavelength λ;
- Δc_i is the concentration change ($\mu mol \cdot l^{-1}$) of the i-th chromophore at the wavelength λ;
- $\in_i (\lambda)$ is the decadic extinction coefficient ($\mu mol^{-1} \cdot l \cdot mm^{-1}$) of the i-th chromophore at the wavelength λ.

The solution of Eq. (13.1) gives the chromophore concentration change Δc_i. In fact, in Eq. (13.1) the attenuation and the chromophore concentration changes are linearly dependent. The total distance $L(\lambda)$ models the actual path the photons travel into brain. This path depends on the source—detector distance and it is usually modified by considering a specific coefficient (given by scattering). This numeric multiplier is called *differential pathlength factor*. In Ref. [22] modeled the propagation of infrared photons in adult human skull and proposed the value of 5.97.

13.2.1 Studied Population

The study involved 80 subjects, divided in 3 groups based on pathology: 15 healthy subjects as healthy controls (age: 29.2 ± 8.5), 14 women who suffered from MwoA (age: 44.4 ± 9.7), and 51 women who suffered from MwA (age: 38.0 ± 12.1). All diagnoses were performed by expert and in accordance to the criteria of the International Headache Society [23]. Migraine subjects were tested in their interictal periods (i.e., when they were free of pain).

The IRB of the Gradenigo Hospital of Turin (Italy) (where we conducted all the tests) approved the experimental protocol described below. All the patients were instructed about the examinations and about the aims of this study and they were asked to sign an informed consent.

13.2.2 Experimental Protocol

The recordings were performed in a quiet room, with dimmed lighting and with a constant room temperature of 24–25 °C. The subjects were asked to lay in a supine position, with their eyes closed and breathing room air. Before and after active maneuvers, they performed a resting period of 120 s.

A commercial NIRS device (NIRO 300, Hammamatsu Photonics, Australia) was used to perform the experiments. The light source was placed on the left side of the forehead, about 2 cm alongside the midline and 3 cm above the supraorbital ridge. The distance between the source and the receiver was equal to 5 cm, and the differential path length factor was set to 5.97. The sampling frequency of the signals was equal to 2 Hz.

13.3 Feature Extraction

As discussed in the introduction, the assessment of the cerebral autoregulation of migraineurs is complicated by many factors, ranging from life-style habits (e.g., current smoke) to genetic mutations. Therefore, to obtain a complete description of the overall "system", many instrumental, biochemical, and genetic data are required. As happens in the analysis of complex systems, often researchers come up with a wide database consisting of inhomogeneous data. In this section, we will describe our processing procedure for the analysis of the NIRS signals and feature extraction paradigms.

13.3.1 Time-Frequency and Coherency Analysis

As indicated in Fig. 13.1, the vasodilatation due to BH produces an increase in the O_2Hb and a decrease in the HHb concentration (Fig. 13.1a), whereas the vasoconstriction induced by HYP causes a decrease in the O_2Hb and an increase in the HHb concentration (Fig. 13.1b). Moreover, in Fig. 13.1b a periodic trend is visible on both the signals reflecting the respiratory rate.

Because of the rapid changes in NIRS signals, especially during active stimuli, the signals must be considered nonstationary. Hence, traditional Fourier-based spectral analysis cannot be applied for NIRS signal processing, and time-frequency distributions should be preferred. Specifically, a time-frequency distribution belonging to the Cohen's class $D_{xx}(t, f)$ was chosen in this study, for which the generic definition can be given as:

$$D_{xx}(t,f) = \iint_{-\infty}^{+\infty} x\left(t' - \frac{\tau}{2}\right) x^*\left(t' + \frac{\tau}{2}\right) g(\tau,\theta) e^{-j2\pi\theta(t'-t)} e^{-j2\pi f\tau} dt' d\theta d\tau, \quad (13.2)$$

when $x(t)$ is the signal, θ and τ are the frequency and time lags respectively, and $g(\tau, \theta)$ is the kernel of the time-frequency distribution.

The Choi-Williams transform used in this study [24] has a kernel that is expressed as:

$$g(\tau,\theta) = e^{-(\tau^2\theta^2/\sigma)}, \quad (13.3)$$

where σ is a parameter influencing the kernel selectivity: a lower σ value will create a higher attenuation of interference terms. A Choi-Williams representation of the HHb NIRS signal during BH for a healthy subject is reported in Fig. 13.2.

Moreover, the time-frequency squared coherence function (SCF), between the concentration signals of O_2Hb and HHb, was computed on the basis of the Choi-Williams representation of the two signals as:

$$SCF_{xy}(t, f) = \frac{|D_{xy}(t, f)|^2}{D_{xx}(t, f) \cdot D_{yy}(t, f)} \quad (13.4)$$

where $D_{xy}(t, f)$ is the cross time–frequency transform of the O_2Hb and HHb concentration signals, and $D_{xx}(t, f)$ and $D_{yy}(t, f)$ are the time-frequency representations of the O_2Hb and HHb signals, respectively. As the SCF is a quadratic function, it assumes only values between 0, if the two signals are uncorrelated, and 1, if they are totally correlated. Figure 13.3 represents an example of time-frequency SCF between the two NIRS signals during BH relative to a healthy subject.

We analyzed signal epochs of 256 s having the active maneuvers (either BH or HYP) in the middle of the window (see Fig. 13.1). We obtained a theoretical spectral resolution that was slightly better than 4 mHz. This value was shown to be suitable to clearly separate the two frequency bands of interest [4, 15].

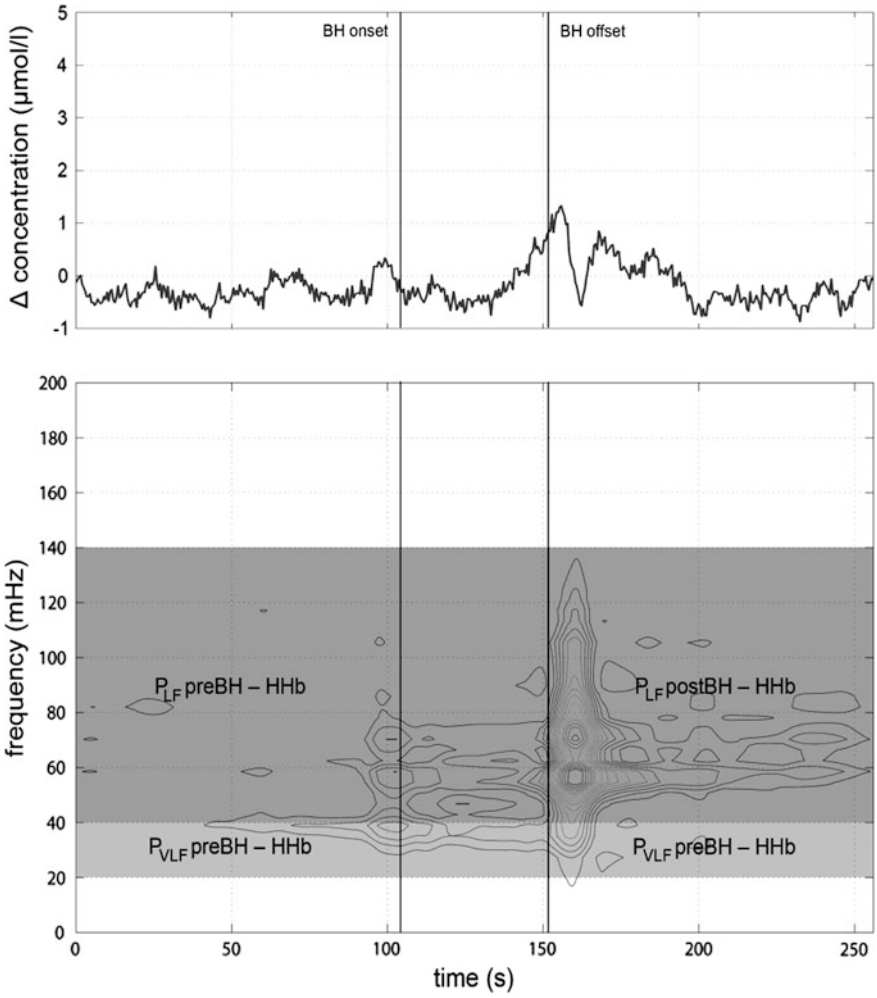


Fig. 13.2 HHb concentration signal (*upper panel*) recorded on a healthy subject and lasting 256 s with the BH in the middle of the analysis window. The onset and the offset of the event are marked by *vertical lines*. The *lower panel* shows the Choi-Williams distribution of the signal ($\sigma = 0.05$) by 15-level curves. The *yellow zone* represents the VLF band (20–40 mHz), while the *pink zone* indicates the LF band (40–140 mHz). The *graphs* show that the NIRS signals become nonstationary as a consequence of the active stimuli

Before calculating their time-frequency distributions, pre-elaboration was performed on all the signals, in order to remove the mean value and the trend. The trend removal was obtained by means of a high-pass Chebychev filter with a ripple in the stopband and a cutoff frequency equal to 15 mHz.

The time-frequency distributions, concerning both the Choi-Williams transforms of the two signals and the SCF, were analyzed in two specific bands, VLF

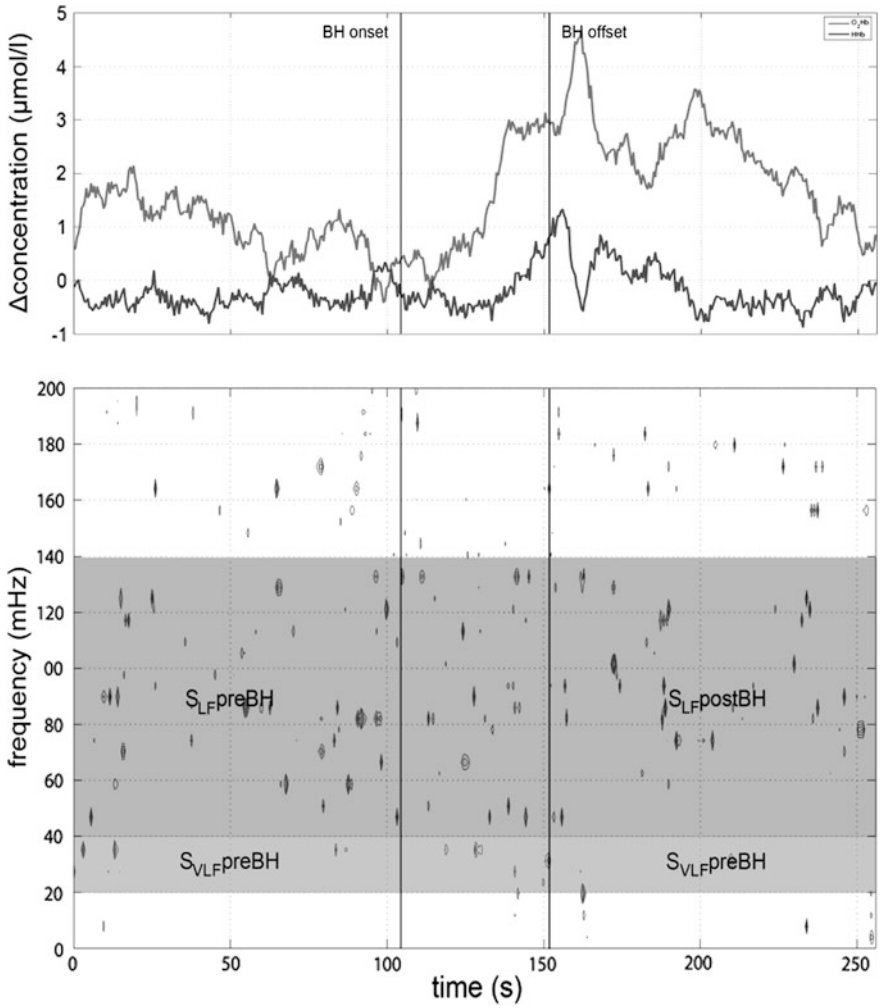


Fig. 13.3 The HHb (*blue line*) and O_2Hb (*red line*) signals during BH (*upper panel*) relative to a healthy subject. The onset and the offset of the event are marked by *vertical lines*. The *lower panel* shows the 15-level contour plot of the time-frequency SCF between the two signals. The *yellow zone* represents the VLF band (20–40 mHz), and the *pink zone* indicates the LF band (40–140 mHz)

and LF, before and after BH and HYP. The percentage of signal power in the two bands (referred to the total power of the signal) was calculated before and after each event. In this way, the following 24 variables deriving from the time-frequency representations have been measured:

- the HHb and O_2Hb power in the VLF and LF bands, before and after BH (for 8 variables),

- the HHb and O_2Hb power in the VLF and LF bands, before and after HYP (for 8 variables), and
- the O_2Hb and HHb SCF value in the two bands, before and after BH and HYP (for 8 variables).

Moreover, two additional variables, derived from the analysis in the time domain, were measured: the BHI indexes for HHb and O_2Hb signals calculated as:

$$BHI_{O_2Hb} = \frac{[O_2Hb]_{BH} - [O_2Hb]_{BASE}}{D_{BH}}, \tag{13.5}$$

where the numerator represents the variation of the O_2Hb concentration as an effect of BH, while the denominator represents the BH duration. Similarly, it is possible to calculate the BHI_{HHb}. A deeper description of the methods for signal analysis used in this chapter can be found in [25]. The 26 final variables, calculated as described above, are reported in the first column of Table 13.1.

13.3.2 Feature Selection

In this study, the *QuickReduct Algorithm* (QRA), a feature selection procedure based on the *Rough-Set Theory* (RST), was performed. This method was compared with the *ANalysis Of Variance* (ANOVA), assuming a linear model of data.

Both procedures have been implemented in a MATLAB environment. Data were organized in a matrix in which each row corresponds to the number of patients involved in the study and each column contains the measured features.

13.3.2.1 Rough Set Theory

Rough set theory (RST), as defined by Pawlak in Ref. [26], is a powerful tool that models imperfect and incomplete knowledge and does not require any a priori information or model assumptions about data.

In RST, a *decision system* (or *decision table—DT*) is defined as pair $DT = (U, A)$, where U is a non-empty set of objects (*the universe* of discourse) and A is a non-empty set of *attributes*. The latter is made up of a certain number of *conditional attributes* C, which represent the input features, and a *decision attribute* D, which is the class the objects belong to. This can be formally depicted as $A = \{C \cup D\}$ and $\{C \cap D\} = \emptyset$.

The basic principle of RST says that if two objects have the same values for a certain set of conditional attributes, then these objects should be classified into the same class. Hence, it is necessary to introduce the *indiscernibility* relation with respect to a non-empty subset $P \subseteq A$ as:

Table 13.1 Feature selection results

Features	QRA results	ANOVA results (p-value) (%)
P_{VLF} preBH O_2Hb		74.09
P_{VLF} postBH O_2Hb		65.42
P_{LF} preBH O_2Hb		84.20
P_{LF} postBH O_2Hb		22.18
P_{VLF} preBH HHb		51.45
P_{VLF} postBH HHb	1	71.68
P_{LF} preBH HHb		21.96
P_{LF} postBH HHb		50.65
P_{VLF} preHYP O_2Hb		17.16
P_{VLF} postHYP O_2Hb		70.92
P_{LF} preHYP O_2Hb	1	69.94
P_{LF} postHYP O_2Hb		2.94*
P_{VLF} preHYP HHb		73.27
P_{VLF} postHYP HHb		55.18
P_{LF} preHYP HHb		94.65
P_{LF} postHYP HHb	1	46.71
S_{VLF} preBH	1	79.05
S_{VLF} postBH		75.17
S_{LF} preBH		38.53
S_{LF} postBH		84.55
S_{VLF} preHYP	1	13.66
S_{VLF} postHYP	1	90.97
S_{LF} preHYP	1	62.81
S_{LF} postHYP	1	27.25
BHI_{O2}	1	4.93*
BHI_{CO2}		0.00*

The results of two feature-selection procedures. The *first column* contains the 26 variables used as input for the feature selection strategies. In the *second column*, results of QRA are reported (1: feature selected). Results of one-way ANOVA analysis, in terms of p-value, are collected in the *third column*. The significant parameters (p-value < 5 %), obtained considering as independent variable the subject pathology (no migraine, MwA and MwoA), are indicated with *asterisk*

$$IND(P) = \{(x, y) \in U^2 : \forall_{a \in P},\ a(x) = a(y)\} \tag{13.6}$$

where $a \in A$ denotes the value of attribute a for a specific object and (x, y) is a pair of objects indiscernible with respect to the P attributes. $IND(P)$ determines a partition of the universe U denoted as $U/IND(P)$ or simply U/P.

According to RST, each $X \subseteq U$ can be divided into two disjoint subsets named *lower* and *upper approximations*, using only the information contained in P and defined, respectively, as:

$$\underline{P}X = \{x \in U | [x]_P \subseteq X\}, \tag{13.7}$$

$$\overline{P}X = \left\{x \in U | [x]_P \cap X = \emptyset\right\}, \tag{13.8}$$

where $[x]_P$ denotes the equivalence classes of the indiscernibility relation with respect to P. The *P-lower approximation* of X is the complete set of objects certainly belonging to the target set X, according to the information carried on P, while the *P-upper approximation* of X includes the objects of U which may belong to X. The couple $(\underline{P}X, \overline{P}X)$ defines a rough set.

Based on the two approximation concepts depicted above, three regions can be defined:

- the *positive region* $(POS_P(D))$, including all the objects of universe that can be certainly classified into U/D classes by using only the attributes P:

$$POS_P(D) = \cup_{X \in U/D} \underline{P}X, \tag{13.9}$$

- the *negative region* $(NEG_P(D))$, being the complete set of objects that certainly does not belong to the U/D classes, according to the attributes P:

$$NEG_P(D) = U - \cup_{X \in U/D} \overline{P}X \tag{13.10}$$

and,

- the *boundary region* $(BND_P(D))$, containing the objects that can possibly, but not certainly, be classified into U/D classes:

$$BND_P(D) = \cup_{X \in U/D} \overline{P}X - \cup_{X \in U/D} \underline{P}X \tag{13.11}$$

Let $P \subset A$ be a subset of conditional features and $D \subset A$ be the decision feature. It is possible to measure the importance of P in classifying the objects of U into D by means of the *dependency degree*:

$$\gamma_p(D) = \frac{|pos_p(D)|}{|U|} \tag{13.12}$$

where $||$ denotes the cardinality of set. If $\gamma_P(D) = 1$, Q depends on the attributes in P, whereas values of $\gamma_P(D)$ between 0 and 1 mean that D only partially depends on P.

The minimal subset of conditional attributes $R \subset A$ is called a *reduct* of the whole set of conditional features C if $\gamma_R(D) = \gamma_C(D)$ [27]. A reduct indicates that no attribute can be removed from the subset without affecting the dependency degree, formally:

$$\gamma_{R-\{a\}}(D) \neq \gamma_R(D) \text{ for all } a \in R. \tag{13.13}$$

As for a given dataset, many reduct subsets may exist. The intersection of all reducts is called the *core* and is made up of those attributes that cannot be eliminated without information loss.

In the past, RST has found various areas of application, such as machine learning [28], knowledge acquisition [29, 30], decision analysis [31, 32], pattern

recognition [33], knowledge discovery from databases, and expert systems [34]. Recently, feature selection has been one of the most important fields in which RST has been employed, with satisfactory results.

Feature selection is a procedure that permits the dimensional reduction of multivariate data, in order to extract the most significant information from a high-dimensional dataset. By using RST, it is possible to find the most informative subset (*reduct*) of the original attributes, starting from a dataset with discretized attribute values; all other attributes can be removed from the dataset with minimal information loss [27]. In such a way, it is possible to highlight the relevant features while reducing the computational time and maintaining the quality of object classification.

The most obvious way to find the reduct with the minimal cardinality is to generate all possible reducts and then choose the smallest one. As this method is not effective and is often inapplicable for large datasets, several techniques for attribute reduction have been developed, as depicted in Ref. [35]. Section 13.3.2.2 focuses on the simplest and most used method for feature selection based on the RST: the *QuickReduct Algorithm* (QRA).

13.3.2.2 QuickReduct Algorithm

QRA, introduced in Ref. [36], is a basic tool for resolving reduct search problems without generating all possible subsets. QRA is based on the dependency degree measured between a decision attribute D and the subset of conditional features C analyzed in order to be a reduct.

The algorithm starts from an empty subset of features and adds to it the best attributes, until a stopping criterion is satisfied. As the goal of QRA is to find a reduct with the same dependency degree as the entire set of attributes, this parameter is chosen as a stopping criterion. The maximum dependence value results in 1 if the dataset is consistent. Consequently, attributes added to the reduct subset are those producing a larger increase in the dependency degree. The pseudo-code of QRA [27] is depicted in Fig. 13.4.

This algorithm, however, is not guaranteed to find a minimal reduct, because the obtained feature subset might still contain irrelevant attributes. It was shown that the classification performance could be degraded by feature subsets with irrelevant features [37].

13.3.2.3 ANOVA Analysis

ANOVA allows a comparison of two or more datasets, by analyzing the variance within and between data groups. In this study the one-way ANOVA analysis was performed using the pathology as an independent variable and the 26 parameters for each subject as dependent variables, one at a time. Then, the observations with

Fig. 13.4 QuickReduct Algorithm pseudo-code

```
QUICKREDUCT(C,D)
Input: C -> set of all conditional features
       D -> set of decision features
Output: R -> feature subset
(1) R <- {}
(2) while γR(D) ≠ γC(D)
(3)     T <- R
(4)     foreach x ∈ (C - R)
(5)         if γR∪{x}(D) > γT(D)
(6)             T <- R ∪ {x}
(7)     R <- T
(8) return R
```

a *P*-value greater than 5 % were neglected for the study. This technique allows feature reduction based on a linear model extracted from data.

13.3.3 Artificial Neural Networks

In this chapter, two feature selection strategies were used, and their performances were compared using an artificial neural network (ANN). The purpose of this method was to use a good feature selection procedure to remove redundant features. With this method, the reduct provides the same quality of classification of the original set [38] or even improves it. Specifically, we built three networks, using all 26 attributes as input data in one network, those selected by means of QRA in a second network, and those resulting from the ANOVA analysis in the third network. For the ANN structure, we chose only one hidden layer with a number of neurons approximately equal to half the input neurons. As for the neuron activation functions, we used a logarithmic sigmoid function for the hidden layer and a linear function for the output layer. Back-propagation was chosen as the learning algorithm, and the mean squared error was used as the performance function. The initial values of the interconnection weights were set randomly. As we only wanted to have a tool to compare the performances of the two feature selection methods, we did not optimize the parameters of the ANNs, but used three comparable ANNs. A schematic description of the three ANNs is reported in Fig. 13.5.

The ANNs were implemented using the Neural Network Matlab toolbox. Sixty percent of input data randomly collected was used as a training set, and the

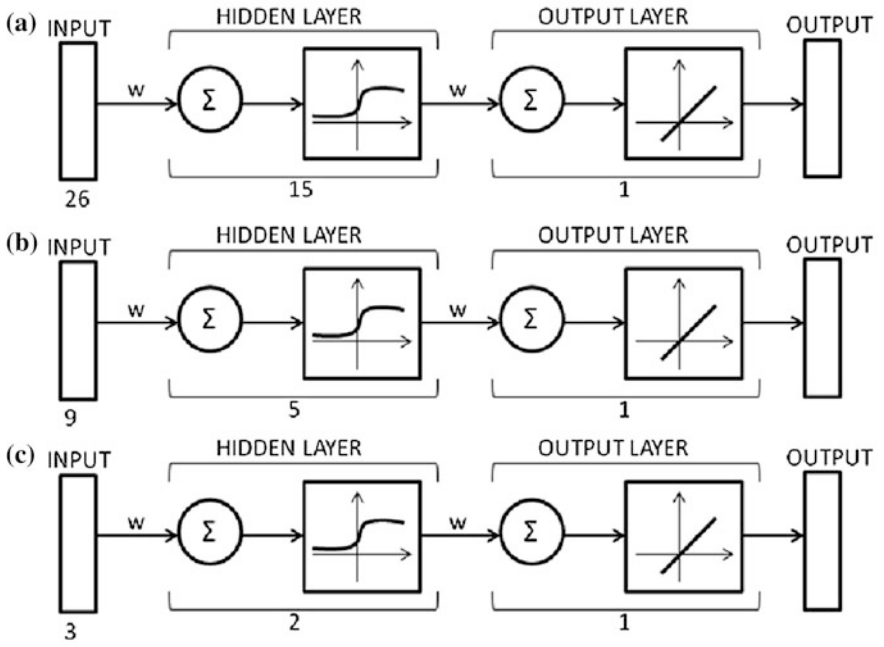


Fig. 13.5 Schematic description of ANNs employed in order to test the two feature selection strategies. **a** ANN using the whole set of 26 available features as input data. **b** ANN using the QRA subset of features (9 parameters) as input data. **c** ANN using the ANOVA subset of features (3 parameters) as input data. All three networks consist of a hidden layer composed of a number of neurons that are variable according to the number of input features and with logarithmic sigmoid activation function. The output layer is made up of only one neuron with a linear activation function

complete dataset was used as a test set. The sample numerosity for each class was non-homogeneous, which could influence the training step. As a matter of fact, it is possible that some classes were less represented than others in the sample used for the neural network training. This aspect, beyond our control, obviously influences the ANN final performances. To avoid this problem, we repeated the training ten times and gave as result the best performance for each ANN.

13.4 NIRS Features Reduction and Subjects Classification

The QRA returned nine variables: the CO_2 power in the VLF band after BH (P_{VLF} post BH–HHb), the O_2 power in the LF band before HYP (P_{LF} pre HYP–O_2Hb), the CO_2 power in the LF band after HYP (P_{LF} post HYP–HHb), the coherence value in the VLF band before BH (S_{VLF} pre BH), the coherence values in the VLF and LF band before and after HYP (S_{VLF} pre HYP, S_{VLF} post HYP, S_{LF} pre HYP,

S_{LF} post HYP), and the BHI_{O2}. A one in the second column of Table 13.1 indicates the selected features.

The third column of Table 13.1 reports the results of the ANOVA analysis considering the subject pathology as an independent variable and the 26 previously described parameters as dependent variables. In this study, only the features with P-values lower than 5 % were deemed as descriptive of the subjects' classification. This allowed for removing the features with a poor correlation with the independent variably (i.e. migraine type). Three of these features emerged from this analysis (marked with asterisk in third column of Table 13.1): the O_2 power in the LF band post HYP (P_{LF} pre HYP–O_2Hb), the BHI_{O2}, and the BHI_{CO2}. Only the last variable was recognized as relevant by the QRA procedure.

In order to test the performance of the two feature-selection procedures, ANNs were used, and the results for each net are reported in Fig. 13.6.

The whole set of features gives the correct classification for 100 % of subjects. Using the nine parameters highlighted with the QRA, the classification accuracy of the subjects is 97.5 %. The network performance decreases when it takes the three features selected with ANOVA analysis as input. In this case, the correct classification rate drops to about 75 %.

13.5 Data Interpretation and Discussion

As often happens in the analysis of complex physiological systems, many variables must be considered in order to gain a description of the system that is as complete as possible. When the physiological systems comprise pathology (or, as in this study, two pathologies) large feature datasets are required. We analyzed the vascular pattern of subjects suffering from MwA and MwoA by the time-frequency analysis of cerebral oxygenation NIRS signals. Overall, we derived a dataset of 26 variables (Table 13.1).

The purpose of this study was to compare the performances of two feature selection methods for identifying a minimal subset of variables able to keep the same amount of relevant information contained in the entire set of parameters derived from NIRS signals. Such a procedure allowed emphasizing the attributes that were relevant to obtain a reliable classification. The subject classification was assessed by means of ANNs, an unsupervised procedure in which knowledge is acquired by the network through a learning process.

The results described in the Sect. 13.4 lead to the observation that the features selected by QRA give the best results. From a physiological point of view, the results of QRA are in accordance with previously published results. It is widely accepted that migraine (and, particularly, MwA) is a pathology with a vascular component involving carbon dioxide dysregulation [6, 25, 39]. Among the nine features that were considered most important by QRA, five are relative to the coherence between oxygenated and reduced hemoglobin in vasoactive maneuvers. For example, it is interesting to notice that MwA sufferers differed from MwoA

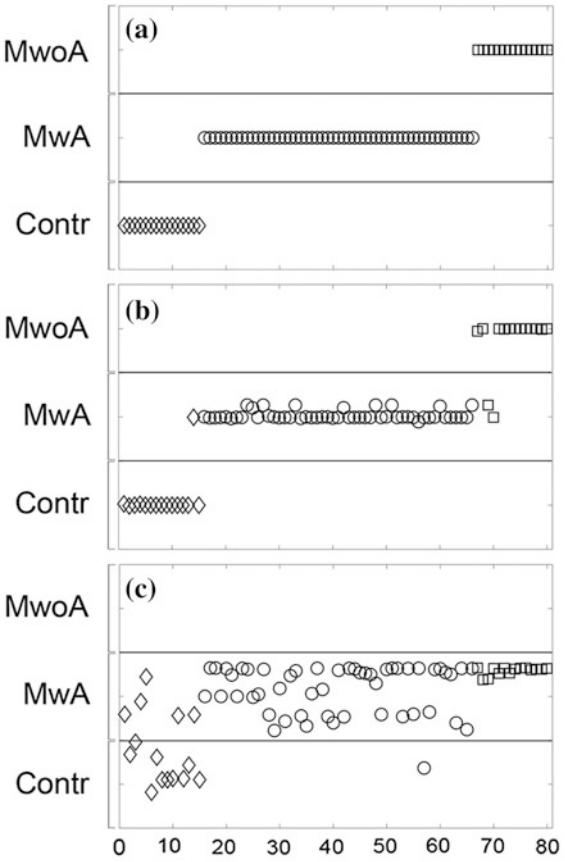


Fig. 13.6 ANNs result in terms of subject classification. Diamonds represent healthy subjects, circles are relative to MwA, and squares indicate MwoA. **a** ANN outputs using all features as input, **b** ANN outputs obtained with the nine features extracted by QRA, **c** ANN outputs relative to the three features highlighted with ANOVA analysis

sufferers in the coherence levels before and after hyperventilation, in both VLF and LF band. Hyperventilation is a stimulus causing a strong vasoconstriction and, therefore, a strong autoregulation is needed in order to maintain the correct proportion between the two hemoglobin types in the brain tissues. Since MwA is associated with impaired autoregulation [6, 25, 39], we believe that QRA effectively found the most important variables associated with the physiological system under analysis.

The principal innovation of this study is the development of a completely automated procedure for feature set reduction. Comparison of performance was made against a linear discriminator based on ANOVA and PCA, which we previously developed. The feature set extracted by this technique, which was coupled to an ANN, outperformed the previous method. A major drawback of the previous method was the need for the selection of a significance threshold, which we fixed equal to 5 %. Therefore, all the features that did not explain at least 5 % of the data variance with respect to the independent variable were discarded. This

selection was arbitrary and unsupported by any clinical observation. Moreover, ANOVA cannot correctly deal with non-linear correlation among the variables.

Using this approach, we overcame such limitations, because the feature selection procedure was able to cope with nonlinearities. In addition, this approach is less arbitrary than other methods because the only choices the user must face are what discretization ranges to use for each feature. It may become fully user independent using a computerized procedure for the discretization based on a validated training set.

Unfortunately, a direct comparison of our results with literature is impossible, because we could not find any other classification or feature extraction study based on NIRS signals in migraines.

A possible limitation of this study is the lack of comparison with other feature extraction strategies. Because classification is not the main purpose of this paper, we simply implemented an ANN, and we did not test further classifiers. We are now working to enlarge the database and build other classification schemes, to further validate the QRA methodology in this specific application. The application scenario that we foresee for this approach is the home monitoring by NIRS of subjects affected by chronic neurological or cerebrovascular impairments.

In conclusion, we applied an automated feature selection strategy to NIRS signals recorded from migraine patients during vasoactive maneuvers. The objective of this step was the dimensional reduction of a dataset of 26 variables, derived from the NIRS signals, which were correlated to the vascular pattern of the patients.

Our method extracted nine features, which lead to a classification accuracy of 97.5 %. Moreover, the extracted features were relevant for the pathology, since they were the most correlated to the carbon dioxide dysregulation typical of migraine with aura sufferers.

References

1. Kruit MC, van Buchem MA, Hofman PA, Bakkers JT, Terwindt GM, Ferrari MD, Launer LJ (2004) Migraine as a risk factor for subclinical brain lesions. JAMA 291(4):427–434
2. Scher AI, Terwindt GM, Picavet HS, Verschuren WM, Ferrari MD, Launer JL (2005) Cardiovascular risk factors and migraine: the GEM population-based study. Neurology 64(4):614–620
3. Tietjen GE (2009) Migraine as a systemic vasculopathy. Cephalalgia 29(9):987–996
4. Liboni W, Molinari F, Allais G, Mana O, Negri E, Grippi G, Benedetto C, D'Andrea G, Bussone G (2007) Why do we need NIRS in migraine? Neurol Sci 28:S222–S224
5. Nowak A, Kacinski M (2009) Transcranial Doppler evaluation in migraineurs. Neurol Neurochir Pol 43(2):162–172
6. Vernieri F, Tibuzzi F, Pasqualetti P, Altamura C, Palazzo P, Rossini PM, Silvestrini M (2008) Increased cerebral vasomotor reactivity in migraine with aura: an autoregulation disorder? a transcranial Doppler and near-infrared spectroscopy study. Cephalalgia 28(7):689–695

7. Molinari F, Liboni W, Grippi G, Negri E (2006) Relationship between oxygen supply and cerebral blood flow assessed by transcranial Doppler and near-infrared spectroscopy in healthy subjects during breath-holding. J Neuroeng Rehabil 3:16
8. Silvestrini M, Baruffaldi R, Bartolini M, Vernieri F, Lanciotti C, Matteis M, Troisi E, Provinciali L (2004) Basilar and middle cerebral artery reactivity in patients with migraine. Headache 44(1):29–34
9. Watanabe Y, Tanaka H, Dan I, Sakurai K, Kimoto K, Takashima R, Hirata K (2011) Monitoring cortical hemodynamic changes after sumatriptan injection during migraine attack by near-infrared spectroscopy. Neurosci Res 69(1):60–66
10. Viola S, Viola P, Litterio P, Buongarzone MP, Fiorelli L (2010) Pathophysiology of migraine attack with prolonged aura revealed by transcranial Doppler and near infrared spectroscopy. Neurol Sci 31(1):S165–S166
11. Diener HC, Kurth T, Dodick D (2007) Patent foramen ovale and migraine. Curr Pain Headache Rep 11(3):236–240
12. Liboni W, Molinari F, Allais GB, Mana O, Negri E, D'Andrea G, Bussone G, Benedetto C (2008) Patent foramen ovale detected by near-infrared spectroscopy in patients suffering from migraine with aura. Neurol Sci 29(1):S182–S185
13. Rothrock JF (2008) Patent foramen ovale (PFO) and migraine. Headache 48(7):1153
14. Giustetto P, Liboni W, Mana O, Allais G, Benedetto C, Molinari F (2010) Joint metabonomic and instrumental analysis for the classification of migraine patients with 677-MTHFR mutations. Open Med Inform J 4:23–30
15. Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhaupl K, Villringer A (2000) Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults. Neuroimage 12(6):623–639
16. Sliwka U, Harscher S, Diehl RR, van Schayck R, Niesen WD, Weiller C (2001) Spontaneous oscillations in cerebral blood flow velocity give evidence of different autonomic dysfunctions in various types of headache. Headache 41(2):157–163
17. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17(4):491–502
18. Su CT, Yang CH (2008) Feature selection for the SVM: an application to hypertension diagnosis. Expert Syst Appl 34:754–763
19. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37
20. Duncan A, Meek JH, Clemence M, Elwell CE, Tyszczuk L, Cope M, Delpy D (1995) Optical pathlength measurements on adult head, calf and forearm and the head of the new-born infant using phase resolved optical spectroscopy. Phys Med Biol 40(2):295–304
21. Leung TS, Tachtsidis I, Smith M, Delpy DT, Elwell CE (2006) Measurement of the absolute optical properties and cerebral blood volume of the adult human head with hybrid differential and spatially resolved spectroscopy. Phys Med Biol 51(3):703–717
22. Okada E, Firbank M, Schweiger M, Arridge SR, Cope M, Delpy DT (1997) Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head. Appl Opt 36(1):21–31
23. Headache Classification Committee (1988) Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. Headache Classification Committee of the International Headache Society. Cephalalgia 8:1–96
24. Cohen L (1989) Time-frequency distributions—a review. Proc IEEE 77(7):941–981
25. Molinari F, Rosati S, Liboni W, Negri E, Mana O, Allais G, Benedetto C (2010) Time-Frequency Characterization of Cerebral Hemodynamics of Migraine Sufferers as As-sessed by NIRS Signals. EURASIP J Adv Sig Proc. doi:10.1155/2010/459213
26. Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11(5):341–356
27. Jensen R, Shen Q (2008) Computational intelligence and feature selection: rough and fuzzy approaches. Wiley, Hoboken
28. Moradi H, Grzymala-Busse JW, Roberts JA (1998) Entropy of english text: experi-ments with humans and a machine learning system based on rough sets. Inf Sci 104:31–47

29. Feng L, Wang GY, Li XX (2010) Knowledge acquisition in vague objective information systems based on rough sets. Expert Syst 27(2):129–142
30. Matsumoto Y, Watada J (2009) Knowledge acquisition from time series data through rough sets analysis. IJICIC 5:4885–4897
31. Greco S, Matarazzo B, Slowinski R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129:1–47
32. Pawlak Z, Sowinski R (1994) Rough set approach to multi-attribute decision analysis. Eur J Oper Res 72(3):443–459
33. Swiniarski RW, Skowron A (2003) Rough set methods in feature selection and recognition. Pattern Recogn Lett 24(6):833–849
34. Tsumoto S (1998) Automated extraction of medical expert system rules from clinical databases based on rough set theory. Info Sci 112:67–84
35. Thangavel K, Pethalakshmi A (2009) Dimensionality reduction based on rough set theory: a review. Appl Soft Comput J 9(1):1–12
36. Shen Q, Chouchoulas A (2000) Modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. Eng Appl Artif Intel 13(3):263–278
37. Chen Y, Miao D, Wang R (2010) A rough set approach to feature selection based on ant colony optimization. Pattern Recogn Lett 31:226–233
38. Chen Y, Miao D, Wang R, Wu K (2011) A rough set approach to feature selection based on power set tree. Knowl -Based Syst 24:275–281
39. Liboni W, Molinari F, Allais G, Mana O, Negri E, Bussone G, D'Andrea G, Benedetto C (2009) Spectral changes of near-infrared spectroscopy signals in migraineurs with aura reveal an impaired carbon dioxide-regulatory mechanism. Neurol Sci 30(1):S105–S107

Chapter 14
A Selection and Reduction Approach for the Optimization of Ultrasound Carotid Artery Images Segmentation

Samanta Rosati, Gabriella Balestra, Filippo Molinari, U. Rajendra Acharya and Jasjit S. Suri

Abstract The segmentation of the carotid artery wall is an important aid to sonographers when measuring intima-media thickness (IMT). Automated and completely user-independent segmentation techniques are gaining increasing importance, because they avoid the bias coming from human interactions. However, automated techniques still underperform semi-automated IMT measurement methods. Automated techniques cannot reproduce human expertise in selecting the optimal point where IMT should be measured. Hence, superior intelligence must be embedded into automated techniques in order to overcome the performance limitations. A possible solution is to extract more information from the image, which could be obtained by an accurate analysis of the image at pixel level. In this study, we applied a feature selection and reduction approach to ultrasound carotid images, and measured 141 features for each image pixel and supposed that a pixel could belong to one of three classes: artery lumen, intima or media layer, or the adventitia layer. Among several approaches that are available for dimensional reduction, we chose to test three based on the Rough-Set Theory (RST): the QuickReduct Algorithm (QRA), the Entropy-Based Algorithm (EBR) and the Improved QuickReduct Algorithm (IQRA). QRA achieved the best performance

S. Rosati · G. Balestra · F. Molinari (✉)
Biolab, Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24 10129 Torino, Italy
e-mail: filippo.molinari@polito.it

U. Rajendra Acharya
Department of ECE, Ngee Ann Polytechnic, Singapore, Singapore

U. Rajendra Acharya
Faculty of Engineering, Department of Biomedical Engineering, University of Malaya, Kuala Lumpur, Malaysia

J. S. Suri
Global Biomedical Technologies, Inc, Roseville, CA, USA

J. S. Suri
Biomedical Engineering Department, Idaho State University, Pocatello, ID, USA

S. Dua et al. (eds.), *Machine Learning in Healthcare Informatics*,
Intelligent Systems Reference Library 56, DOI: 10.1007/978-3-642-40017-9_14,

and correctly classified 97.5 % of the pixels on a reduced testing image dataset and about 91.5 % for a large validation dataset. On average, QRA reduced the complexity of the system from 141 to 8 or 9 features. This result could represent a pilot study for developing an intelligent pre-classifier to improve the image segmentation performance of automated techniques in carotid ultrasound imaging.

Keywords Ultrasound imaging · Intima-media thickness · Atherosclerosis · Segmentation · Feature extraction · Feature selection · Quickreduct algorithm · Entropy · Rough set · Artificial neural networks

14.1 Background

Atherosclerosis is a life threatening disease that may result in the loss of elasticity of the arterial wall and the deposition within the wall itself, which include lipids and other blood-borne molecules [2, 44]. This loss of elasticity results, in a range of about 5–10 years, in possible impairments to the blood circulation that could damage the principal organs (i.e. liver, kidneys, heart and brain).

The ultrasound scan of the arterial bed is the most widely diffused clinical examination in the field of atherosclerosis prevention and monitoring [7]. Large arteries, such as the carotid artery, femoral artery, brachial artery, and aorta are imaged through acoustic waves in order to visualize inner wall composition. This imaging is performed because the intima-media thickness (IMT) of the major arteries is an important indicator of atherosclerosis [3, 31, 35]. The most widely used atherosclerosis indicator is the IMT of the carotid artery (CA), which has been used in several different multi-center studies around the world [3, 10, 27, 33–36, 42].

The clinical measurement of the carotid IMT is not a trivial task. Usually, a trained sonographer acquires a longitudinal projection of the CA and manually measures the IMT by placing two markers in correspondence of the two most evident interfaces of the image:

- The lumen-intima (LI) interface (depicted by the white line in Fig. 14.1), and
- The media-adventitina (MA) interface (depicted by the black line in Fig. 14.1).

The IMT is then defined as the geometrical distance (expressed in mm) between the LI and the MA interfaces.

Manual measurements, besides being prone to errors and subjective interpretations, are time consuming and little adapt to the new quality standard required by modern clinical guidelines. For this reason, work following the pioneer research of [30] has used computer methods to aid clinicians in measuring the carotid IMT. Recently, Molinari et al. reviewed the most used IMT measurement techniques for the carotid wall segmentation and IMT measurement from ultrasound images [20–22]. The most diffused IMT measurement techniques are semi-automated, which means that the human operator interacts with the computer program to drive (and

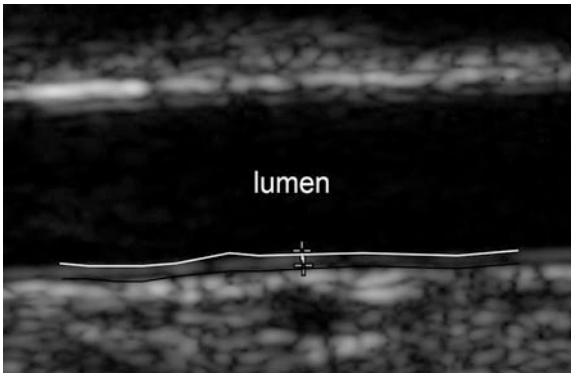


Fig. 14.1 Longitudinal projection of the CA. The *white line* corresponds to the lumen-intima (LI) interface while the *black line* marks out the media adventitina (MA) interface

optimize) the IMT measurement. The interaction between expert sonographers and computer methods increases the accuracy and repeatability of IMT measurements [8], but still lacks complete automation. Thus, the result depends on the human operator.

In last five years, fully automated measurement techniques have been widely developed. This growth in development is mainly because such techniques allow for a complete independence from the user; thus, they are indicated for the processing of large datasets, typical of multi-centric and epidemiological studies. Automated techniques still underperform user-driven algorithms, because it is very difficult for an automated algorithm to mimic the behavior of an expert sonographer.

In fact, when human operators drive the segmentation and an IMT measurement in ultrasound images, they select the optimal image region based on experience. In this scenario, skilled operators can select the correct morphological region (within 1 cm of the carotid bulb), with the lower influence given by noise. It is extremely hard to reproduce this same process using an automated strategy. As a consequence, usually user-independent methods segment image regions with suboptimal characteristics (e.g., with artifacts, excessive noise, and defocused wall layers). To give some normative data, considering the IMT thickness of 1 mm (in presence of atherosclerosis), the average IMT measurement error for user-driven techniques can be about 0.02 ± 0.01 mm [8], about 2 % of the nominal value. Despite intelligent and optimized segmentation strategies, most of the automated techniques reach performances that are about 0.03 ± 0.10 mm [20, 22–25], which means about 3 % of the IMT value. However, it must be noted that the measurement reproducibility (i.e., the standard deviation of the error) is roughly ten times higher for automated methods.

A possible solution to cope with this underperformance of automated methods is the extraction of further information from the ultrasound image. In a theoretical and optimal processing pipeline, such information should be used by the segmentation strategy to mimic the human operator decision process, and thus to optimize segmentation. The idea presented in this chapter is to start analyzing the information content of the ultrasound image at the pixel level.

In this study, we present a feature extraction technique for the classification of ultrasound carotid artery pixels. This idea is based on hypothesis that increasing the number of features used to build a classifier does not automatically increase classifier accuracy, as several attributes may be irrelevant or, even worse, may introduce noise which decreases classifier performance. To improve the classification accuracy, it is then important to select useful features. A procedure performed through selection or construction [16] usually reduces the number of features. During construction, new attributes are created on the basis of some original features. The disadvantage of this phase is that the results are difficult to interpret because they do not correspond to the original features.

Feature selection is based on the idea that the number of attributes can be reduced by collecting the smallest number of important features from the original set, without negatively affecting classification accuracy. When feature selection is needed, there must be an appropriate and well-defined criterion to measure the relevance of the chosen features. However, the number of initial features is usually large. It is computationally impossible to test all possible subsets, even if the criterion is simple to evaluate. A heuristic procedure is then applied to find a good set of features in a reasonable amount of time. Another important difference among the possible feature selection methods is the fundamental hypothesis of linearity, which is required by some of these. Most real situations are non-linear; for example, two features may be useless individually but will become highly predictive if used together. In addition, it is important to notice that the required training set should be larger for a larger number of features [12, 38].

From among the several approaches that are available for dimensional reduction, we chose to test three algorithms based on the Rough-Set Theory (RST): QuickReduct Algorithm (QRA), Entropy-Based Algorithm (EBR), and Improved QuickReduct Algorithm (IQRA) [14]. RST provides a formal methodology for feature selection. Computationally, the approach is efficient and, unlike statistical correlation-reducing approaches, requires no human input and preserves the semantics of data making the results more understandable.

The purpose of this study is to promote calculating the large and overabundant number of parameters extracted from ultrasound carotid images and then select a smaller subset to classify the pixels into three classes (lumen, intima-media complex, and adventitia). The selection was obtained through a feature selection method based on rough set theory. In particular, we describe the use of QRA, EBR, and IQRA and compare their performances.

14.2 Features Extraction and Selection

We tested a database consisting of 300 images from two institutions. One hundred images were acquired at the Cyprus Institute of Neurology (Nicosia, Cyprus) from 100 healthy patients (age: 54 ± 24 years; range: 25–95 years) using a Philips ATL HDI 3000 ultrasound scanner equipped by a linear 7–10 MHz probe. These images

were resampled at a density of 16.67 pixels/mm, obtaining a pixel size of 60 μm. The remaining 200 images were acquired at the Neurology Dept. of the Gradenigo Hospital of Torino, Italy, from 150 patients (age: 69 ± 16 years; range: 50–83 years) using a Philips ATL HDI 5000 scanner. Resampling was set to 16 pixels/mm, leading to a calibration factor of 62.5 μm/pixel. Both institutions obtained written, informed consent from the patients prior to enrolling them in the study. The respective local ethical committees approved the acquisition of the images, and all the subjects gave their informed consent. Three expert sonographers (a neurologist, vascular surgeon, and cardiologist) manually segmented the images. They traced the boundaries of the lumen-intima (LI) and media-adventitia (MA) interfaces. The average tracings were considered ground-truth (GT). The images were first auto-cropped to remove the surrounding black frame, which created the region of interest containing only the ultrasound data. The structure of the auto-cropping technique is beyond the scope of this chapter [19].

This database included both healthy and pathological vessels. In addition, all the possible carotid morphologies were represented: straight and horizontal vessels (Fig. 14.2a), inclined vessels (Fig. 14.2b), and curved arteries (Fig. 14.2c).

14.3 Feature Extraction

In order to build the dataset used for feature selection, we identified three classes of pixels according to their physiological meaning: lumen, intima-media complex, and adventitia from ultrasound carotid images.

Fifty of the images were randomly selected. From each image, ten pixels per class were chosen, for a total of 1,500 pixels. For each single pixel, we considered intensity and parameters based on the intensity of the pixels around each test pixel. That is if intensity and parameters belonged to statistical moments, estimates, and texture features.

Texture is naturally used by humans when analyzing an image [1] and, in this context, texture features are a set of digital parameters based on the spatial displacement of the intensity levels in an image. They are based on the *gray level co-occurrence matrix* (GLCM) [40], which can be calculated on an image I as:

$$GLCM_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{m}\sum_{q=1}^{n}\begin{cases} 1 \; if \; I(p,q) = i \; and \; I(p+\Delta x, q+\Delta y) = j \\ 0 \; otherwise \end{cases}, \quad (14.1)$$

where $m \times n$ is the image size and $\delta = (\Delta x, \Delta y)$ is the displacement. This matrix reports how often a pixel with a gray level i occurs at distance Δx and Δy from another pixel with a grey level j.

Image descriptors used in this work include:

Intensity of the single pixel,

First order statistics, including mean value, standard deviation, skewness and kurtosis, and

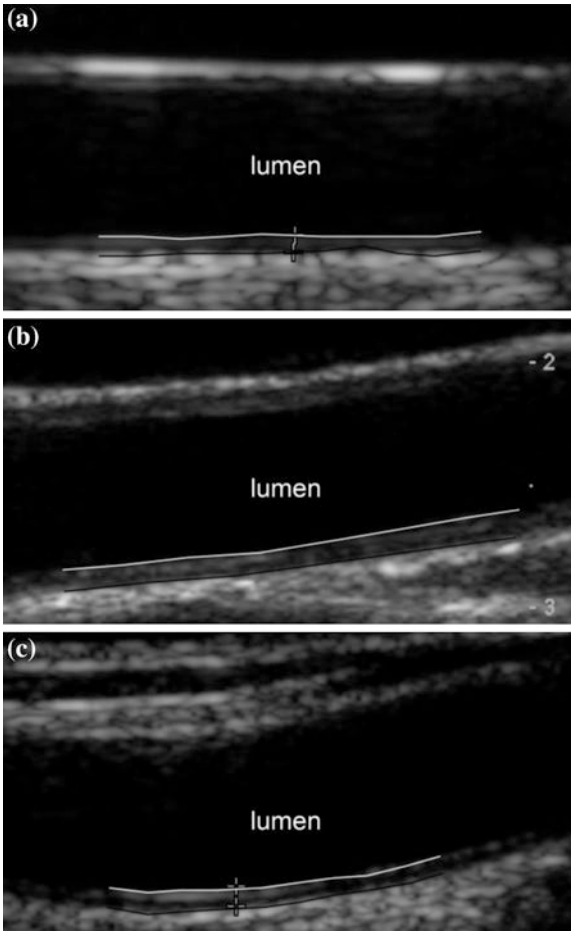


Fig. 14.2 Examples of the possible carotid morphologies: straight and horizontal vessel (**a**), inclined vessels (**b**), and curved arteries (**c**). The *white line* corresponds to the lumen-intima (LI) interface while the *black line* marks out the media-adventitina (MA) interface

Texture features based on the *spatial* G = *gray level dependence matrix* (SGLDM) [6]:

$$SGLDM_{\Delta x,\Delta y}(i,j) = \frac{GLCM_{\Delta x,\Delta y}(i,j)}{\sum_i \sum_j GLCM_{\Delta x,\Delta y}(i,j)} \tag{14.2}$$

We calculated the following parameters [40] with a displacement $\delta = (0,1)$:
Energy:

$$E = \sqrt{\sum_i \sum_j \left[SGDLM_{\Delta x,\Delta y}(i,j)\right]^2} \tag{14.3}$$

Contrast:

$$Co = \sum_i \sum_j (i-j)^2 \cdot SGDLM_{\Delta x,\Delta y}(i,j) \tag{14.4}$$

Homogeneity:

$$H = \sum_i \sum_j \frac{SGLDM_{\Delta x,\Delta y}(i,j)}{1+(i-j)^2} \tag{14.5}$$

Entropy:

$$En = -\sum_i \sum_j SGLDM_{\Delta x,\Delta y}(i,j) \cdot \log SGLDM_{\Delta x,\Delta y}(i,j) \tag{14.6}$$

Moments m_1, m_2 and m_4:

$$m_g = \sum_i \sum_j (i-j)^g \cdot SGLDM_{\Delta x,\Delta y}(i,j) \tag{14.7}$$

Each of these parameters was based on the *gray level difference method* (*GLDM*) [40]:

$$GLDM_\delta(k) = \sum_i \sum_j GLCM_{\Delta x,\Delta y}(i,j), \tag{14.8}$$

where $k = |i-j|$, $k = 0,\ 1, \ldots,\ n-1$, n is the number of gray levels and δ is the difference between two pixels, we calculated the following descriptors [6] with a displacement $\delta = (0,1)$:

Contrast:

$$Con = \sum_{k=1}^{n-1} GLDM_\delta(k) \tag{14.9}$$

Angular Second Moment:

$$ASM = \sum_{k=0}^{n-1} GLDM_\delta(k)^2 \tag{14.10}$$

Entropy:

$$Ent = -\sum_{k=0}^{n-1} GLDM_\delta(k) \cdot \log GLDM_\delta(k) \tag{14.11}$$

Mean:.

$$Mean = \sum_{k=0}^{n-1} k \cdot GLDM_\delta(k) \tag{14.12}$$

We calculated these measures based on the *gray level run length matrix* (*GLRLM*) [6], in which each cell $GLRLM_\delta(i,j)$ represents the number of occurrences of the j adjacent elements in direction δ with gray level i. We calculated the following features [6] for δ equal to 0, 45, 90 and 135°:

Short Run Emphasis:

$$SRE = \sum_i \sum_j \frac{GLRLM_\delta(i,j)}{j^2} \Big/ \sum_i \sum_j GLRLM_\delta(i,j) \tag{14.13}$$

Long Run Emphasis:

$$LRE = \sum_i \sum_j j^2 \cdot GLRLM_\delta(i,j) \Big/ \sum_i \sum_j GLRLM_\delta(i,j) \tag{14.14}$$

Gray Level Distribution:

$$GLD = \sum_i \Big[\sum_j GLRLM_\delta(i,j)\Big]^2 \Big/ \sum_i \sum_j GLRLM_\delta(i,j) \tag{14.15}$$

Run Length Distribution:

$$RLD = \sum_j \Big[\sum_i GLRLM_\delta(i,j)\Big]^2 \Big/ \sum_i \sum_j GLRLM_\delta(i,j) \tag{14.16}$$

Run Percentages:

$$RP = \sum_i \sum_j GLRLM_\delta(i,j) \Big/ A, \tag{14.17}$$

where A is the area of interest. Each of the above described features was calculated on four areas centered on the selected pixel, with sizes 7×15, 15×7, 7×3, and 3×7 pixels. In this way, we obtained a total of 141 features for all analyzed pixels.

14.4 Feature Selection

A literature survey provided us with applications for QRA, EBR, and IQRA, but did not help us choose an experiment, and all of these methods were used for experimentation.

All procedures were implemented in a MATLAB environment. Data was organized in a matrix with each row corresponding to the extracted pixel and the columns contain the 141 measured features.

14.5 Rough Set Theory

Rough set theory (RST) is a powerful tool to model imperfect and incomplete knowledge. As defined by Pawlak (2009), it does not require any *apriori* information or model assumptions about data. In RST, data are organized in a *decision system* (or *decision table-DT*), defined as a pair $DT = (U, A)$ where U is a non-empty set of objects (*the universe* of discourse) and A is a non-empty set of *attributes*. The attributes belong to two categories: *conditional attributes* C, which represent the input features, and *decision attribute* D, which is the class of the objects. These attributes can be formally depicted as $A = \{C \cup D\}$ and $C \cap D = \emptyset$.

The principle of RST is that if two objects are indiscernible with respect to a certain variable, then they should be classified in the same class. Hence, the *indiscernibility* relation needs to be introduced with respect to a non-empty subset $P \subseteq A$ as:

$$IND(P) = \left\{(x, y) \in U^2 : \forall_{a \in P,} a(x) = a(y)\right\} \tag{14.18}$$

where $a \in A$ denotes the value of attribute a for a specific object and (x, y) is a pair of objects indiscernible with respect to the P attributes. $IND(P)$ determines a partition of the universe U denoted as $U/IND(P)$ or simply U/P.

Let $X \subseteq U$, X can be divided into two disjoint subsets using only the information contained in P. The *lower* and *upper approximations* are defined respectively as:

$$\underline{P}X = \left\{x \in U | [x]_P \subseteq X\right\} \tag{14.19}$$

$$\bar{P}X = \left\{x \in U | [x]_P \cap X = \emptyset\right\}, \tag{14.20}$$

where $[x]_P$ denotes the equivalence classes of the indiscernibility relation with respect to P. The *P-lower approximation* of Xs is the complete set of objects certainly belonging to the target set X, according to the information carried on P, while the *P-upper approximation* of X include the objects of Us which may possibly belong to X. The couple $\left(\underline{P}X, \overline{P}X\right)$ defines a rough set.

Using the two approximation concepts depicted above, three regions can be defined: positive region, negative region, and boundary region.

- The *positive region* $(POS_P(D))$, includes all the objects of universe that can be certainly classified into U/D classes by using only the attributes P:

$$POS_P(D) = \bigcup_{X \in U/D} \underline{P}X. \tag{14.21}$$

- The *negative region* $(NEG_P(D))$ is the complete set of objects that certainly does not belong to the U/D classes, according to the attributes P:
-

$$NEG_P(D) = U - \cup_{X \in U/D} \overline{P}X. \tag{14.22}$$

- The *boundary region* ($BND_P(D)$) contains objects that can possibly, but not certainly, be classified into U/D classes:

$$BND_P(D) = \cup_{X \in U/D} \overline{P}X - \cup_{X \in U/D} \underline{P}X. \tag{14.23}$$

Let $P \subset A$ be a subset of conditional features and $D \subset A$ be the decision feature, then the *dependency degree* between *P*and D can be measured as:

$$\gamma_P(D) = \frac{|POS_P(D)|}{|U|}, \tag{14.24}$$

where $||$ denotes the cardinality of set. This parameter expresses the importance of P in classifying the objects of U into D. If $\gamma_P(D) = 1$ all values from D are uniquely determined by values of attributes P [14] and the dataset is defined as consistent. Real datasets are usually not consistent so the maximum value for $\gamma_P(D)$ is less than 1 because D partially depends on P.

The minimal subset of conditional attributes $R \subset A$ is called a *reduct* of the set of conditional features C if $\gamma_R(D) = \gamma_C(D)$ [14]. As a reduct, no attribute can be removed from the subset without affecting the dependency degree, formally:

$$\gamma_{R-\{a\}}(D) \neq \gamma_R(D) \, forall \, a \in R. \tag{14.25}$$

As for a given dataset many reduct subsets may exist; the intersection of all reducts is called the *core* and is composed of those attributes that cannot be eliminated without information loss.

In the past, RST has found different areas of application, such as machine learning [26], knowledge acquisition [9, 18], decision analysis [11, 29], pattern recognition [39], knowledge discovery from databases, and expert systems [43]. Recently, feature selection has been an important field in which RST has been employed, with satisfactory results.

Feature selection is a procedure that permits the dimensional reduction of multivariate data, in order to extract the most significant information from a high-dimensional dataset. The main concept is that, given a dataset with discretized attribute values, it is possible to find a subset (*reduct*) of the original attributes, using RST, which is the most informative; all other attributes can be removed from the dataset with minimal information loss [14]. Using this method, it is possible to highlight relevant features while reducing computational time and maintaining the quality of object classification. All feature selection strategies based on the rough-set method can be divided in two steps: (1) discretization of real numerical features, and (2) application of a feature selection technique.

As the classical rough-set approach uses only discrete data, we discretized the continuous features by looking to the data plots of each feature. For each variable, different intervals of values have been identified enabling the passage from continuous values to a number of discrete elements. The discretized dataset has then

Fig. 14.3 QuickReduct algorithm pseudo-code

```
QUICKREDUCT(C,D)
Input: C -> set of all conditional
features
       D -> set of decision features
Output:  R -> feature subset
(1) R <- {}
(2) while γR(D) ≠ γC(D)
(3)     T <- R
(4)     foreach x ∈ (C - R)
(5)         if γR∪{x}(D) > γT(D)
(6)             T <- R ∪ {x}
(7)     R <- T
(8) return R
```

been used for the feature selection and evaluated by means of the dependency degree measure.

The most obvious way to find a reduct with the minimal cardinality is to generate all possible reducts and then choose the smallest. As this method is not effective and often inapplicable for large datasets, several techniques for attribute reduction have been developed over the past few years, as depicted in [41]. The next section focuses on the simplest and most used method for feature selection based on the RST: the *QuickReduct Algorithm* (QRA).

14.6 QuickReduct Algorithm

QRA, introduced in [37], is a basic tool that allows users to resolve reduct search problems without generating all the possible subsets. The method is based on the dependency degree measured between a decision attribute D and the subset of conditional features C analyzed to be a reduct.

The algorithm starts from an empty subset of features and adds the best attributes to it until a stopping criterion is satisfied. As the goal of QRA is to find a reduct, with the same dependency degree of the entire set of attributes, this parameter is chosen as stopping criterion. The maximum dependence value results in 1 if the dataset is consistent. Consequently, attributes added to the reduct subset are those producing a larger increase in the dependency degree. The pseudo-code of QRA [14] is depicted in Fig 14.3.

This algorithm, however, is not guaranteed to find a minimal reduct, that is the feature subset discovered may contain irrelevant attributes. The classification accuracy may be degraded when designing a classifier using a feature subset with irrelevant features [5].

```
EBR(C,D)
Input: C -> set of all conditional features
       D -> set of decision features
Output: R -> feature subset
(1) R <- {}
(2) while H(D|R) ≠ H(D|C)
(3)     T <- R
(4)     foreach x ∈ (C - R)
(5)         if H(D|R ∪ {x}) < H(D|T)
(6)             T <- R ∪ {x}
(7)     R <- T
(8) return R
```

Fig. 14.4 Entropy-based reduct algorithm pseudo-code

14.7 Entropy-based Reduct

A further easy feature selection algorithm, with structure similar to the QRA, is the *Entropy-based Reduct* (EBR) algorithm, developed by [13]. This algorithm is based on the measure of *conditional information entropy* $(H(D|A))$ produced by an attribute A with respect to the decision feature D:

$$H(D|A) = -\sum_{j=1}^{m}\left(p(a_j)\sum_{i=1}^{n}p(c_i|a_j)\cdot\log_2 p(c_i|a_j)\right), \tag{14.26}$$

where $a_1 \ldots a_m$ and $c_1 \ldots c_n$ are the values of attributes A and D, while $p(a_j)$ is the probability that the value a_j occurs and $p(c_i|a_j)$ is the conditional probability of a_j given c_i.

This measure allows the evaluation of the information content generated by an information source [17]. The equation written above can be extended from one conditional attribute to the whole set of attributes, so that it can be used as a stopping criterion for the EBR algorithm. For a consistent dataset, the maximum value of entropy is 0. In this way, an algorithm similar to QRA can be implemented adding to the current subset, in each iteration, those features resulting in a higher decrease of entropy. The reduct search stops when the resulting subset reaches the same entropy of all the available attributes. Figure 14.4 shows the EBR pseudo-code [14].

As this algorithm has the same structure of QRA, it suffers from the same limits of the QRA and does not guarantee to find a minimal reduct.

14.8 Variable Precision Rough-Set

Although widely used for classification tasks and feature selection problems, RST has some limitations, and may manage only objects that have entirely correct or certain classifications. This requirement means that no degree of classification uncertainty is admitted, even if the information about data are only partial. In addition, RST assumes that the entire universe U is composed only of the data under consideration. Then, the conclusions derived from this model are applicable exclusively to this set of elements [45].

Reference [45] introduced the variable precision rough set (VPRS) theory, a generalization of the standard RST, which surpasses the above depicted limits and admits a degree of misclassification. Thus, RST becomes a special case of VPRS.

Let X and Y be non-empty subsets of a universe U of objects. VPRS introduces the measure of *relative degree of misclassification* $c(X, Y)$ of set X with respect to set Y:

$$c(X,Y) = 1 - \frac{|X \cap Y|}{|X|}, \tag{14.27}$$

where $||$ denotes the set cardinality. If $0 \leq \beta < 0.5$ is the *admissible classification error*, it is possible to redefine all rough-set concepts starting from the *majority inclusion relation* that may be generalized as:

$$Y \overset{\beta}{\supseteq} X \text{ if and only if } c(X,Y) \leq \beta \tag{14.28}$$

By replacing the standard inclusion relationship with this generalized relation, the β-approximations and the β-regions can be defined as:

$$\begin{aligned}
\underline{P}_\beta X &= \left\{ x \in U | [x]_P \overset{\beta}{\subseteq} X \right\} = \left\{ x \in U | c\left([x]_P, X\right) \leq \beta \right\} \\
\overline{P}_\beta X &= \left\{ x \in U | c\left([x]_P, X\right) \leq 1 - \beta \right\} \\
POS_{P,\beta}(D) &= \bigcup_{X \in U/D} \underline{P}_\beta X = \bigcup_{X \in U/D} \left\{ x \in U | c\left([x]_P, X\right) \leq \beta \right\} \\
NEG_{P,\beta}(D) &= U - \bigcup_{X \in U/D} \overline{P}_\beta X = \bigcup_{X \in U/D} \left\{ x \in U | c\left([x]_P, X\right) \geq 1 - \beta \right\} \\
BND_{P,\beta}(D) &= \bigcup_{X \in U/D} \overline{P}_\beta X \\
&\quad - \bigcup_{X \in U/D} \underline{P}_\beta X = \bigcup_{X \in U/D} \left\{ x \in U | \beta < c\left([x]_P, X\right) < 1 - \beta \right\}
\end{aligned} \tag{14.29}$$

Let $P \subset A$ be a subset of conditional features and $D \subset A$ the decision feature, it is possible to measure the *β-dependency degree* as:

$$\gamma_{P,\beta}(D) = \frac{|POS_{P,\beta}(D)|}{|U|} \tag{14.30}$$

The standard RST definitions are obtained from the above equations setting $\beta = 0$.

```
IMPROVED QUICKREDUCT(C,D)
Input: C -> set of all conditional features
       D -> set of decision features
       U -> object dataset
Output: R -> feature subset
(1) R <- {}
(2) Count=0
(3) DT=U
(4) while γ_{R,0}(D) ≠ γ_{C,0}(D)
(5)     AvailableSet=C-R
(6)     β=0
(7)     ε=0.1
(8)     T <- R
(9)     foreach x ∈ AvailableSet
(10)        if γ_{R∪{x},β}(D) > γ_{T,β}(D)
(11)            T <- R ∪ {x}
(12)    If T=R
(13)        β = β + ε
(14)        If β < 0.5
(15)            goto (8)
(16)        else
(17)            R=R ∪ {first attribute in AvailableSet}
(18)    else
(19)        R <- T
(20)    POSPARTIAL = POS_{R,0}(D)
(21)    DT = DT - POSPARTIAL
(22)    Count = Count + |POSPARTIAL|
(23)    γ_{R,0}(D) = Count/|U|
(24)    return R
```

Fig. 14.5 Improved QuickReduct Algorithm pseudo-code

14.8.1 Improved QuickReduct Algorithm

As with RST, QRA has some limits due to the assumption about the monotonicity of dependency degree. This assumption means that γ increases at each iteration and differs from zero at the first iteration. If these conditions are not satisfied, a random choice of features is performed, leading to a reduct with more attributes. Moreover, QRA ignores the redundancy in the dataset objects and the objects included in the positive region in an intermediate iteration will not contribute/add any more knowledge to the rest of iterations [32]. Deleting redundant elements from dataset reduces computational time for QRA.

These limitations led practitioners to modify the standard algorithm using different strategies, in order to improve the feature selection phase and consequently the classification task. One such approach, based on VPRS, is proposed in [32]. In this study the authors present the improved quick reduct algorithm (IQRA), which deletes redundant elements from the analyzed dataset. The IQRA pseudocode is depicted in Fig. 14.5.

IQRA is similar to QRA. The algorithm starts with an empty features subset and then adds, at every iteration, those attributes which induce the greatest increment

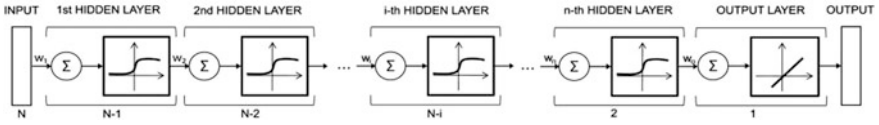


Fig. 14.6 Schematic description of ANNs employed in order to test different feature selection strategies. The network consists of an input layer made up of a number of neurons equal to the number of input features, a certain number of hidden layers with logarithmic sigmoid activation function and with a number of neurons progressively reduced of one element so that the output layer results made up of only one neuron with a linear activation function

in the dependency degree $\left(\gamma_{R,1}(\mathrm{D})\right)$. If no increase during iteration occurs, the VPRS is taken into account. In such a situation the dependency degree with tolerance β $\left(\gamma_{R,\beta}(D)\right)$ is calculated, until a new feature is found or β cannot be reduced any further. In the latter case, the first available feature is included in R. After including a new attribute, the positive region with tolerance 1 $\left(POS_{R,1}(D)\right)$ is calculated and the objects belonging to this set are deleted from the decision table (DT). These elements would be redundant for the next iterations. At the end of the current cycle, the algorithm restarts until the dependency degree with tolerance 1 reaches the maximum value.

14.9 Classification

In this chapter, three different feature selection strategies were used and their performances were compared using an artificial neural network (ANN). The idea behind this method is that a good feature selection procedure allows for removing redundant features so that the reduct provides the same quality of classification of the original set [5] or even improves it.

For the ANN structure, we started the network with one neuron for each feature in the input layer and terminated with one output neuron, progressively reducing the number of neurons of one element through every hidden layer. As for the neuron activation functions, we used a logarithmic sigmoid function for the hidden layers and a linear function for the output layer. Back-propagation was chosen as the learning algorithm and the mean squared error was used as a performance function. The initial values of interconnection weights were set randomly. The ANNs were implemented by means of the Neural Network Matlab toolbox, using the entire input dataset as training set. A schematic description of the three ANNs is reported in Fig. 14.6.

14.10 Feature Selection Process

The process performed to optimize feature selection for pixel classification consisted of several steps that are depicted in Fig. 14.7. Each pixel has 141 associated features and it belongs to one of the three classes.

We began with the application of QRA, EBR, and IQRA to an initial dataset (DS1) of 500 elements for each class. As shown in Table 14.1, each tested reduction method, applied on the same dataset, returned a different subset of features (respectively, FS_{QRA}, FS_{EBR}, and FS_{IQRA}) containing 10 or 11 attributes with a dependency degree slightly lower than 1.

The performances of the subsets were then compared using ANNs. For this portion of the experiment, we built three networks, each with a structure similar to that depicted in the previous section, using the attributes selected by means of QRA, the attributes chosen from EBR, and the attributes extracted by IQRA as input data. All networks were tested with the same dataset used as training set.

Although the accuracy for correct classification was more than 91 % for all methods, FS_{QRA} achieved a classification accuracy of 97.47 %.

Figure 14.8 shows the percentage of correct classification for pixels belonging to each class for each feature selection procedure. Table 14.2 shows that while more than 95 % of lumen pixels are classified in the right class, the percentage of correct classification slightly decreases for the other two classes, while remaining above 85 % (Fig. 14.8).

Two more datasets (DS2 and DS3) similar to the first one but containing the characteristics of different pixels were used first to validate the results and then to classify all the pixels of 50 test images. The classification results on the three datasets were equivalent. The classification errors related to each dataset corresponded to different pixels, indicating the need of further increasing the classification accuracy. To improve the correct classification of the image pixels, we built a classifier based on the combination of three ANNs each trained with a different dataset and a voting system [15]. Classifier results depicted the class with at least two votes. If each ANN output was different, the pixel was not classified. The image pixel classification obtained with this classifier was better than the classification obtained separately by each ANN. Figure 14.9a shows the results achieved by applying the classifier to a portion of the carotid image. Similar results were obtained on all the other images.

To reduce the classification error further, we decided to increment the initial datasets adding 300 pixels manually selected, on the tested images, among those resulting in a wrong. Such a procedure increased the number of pixel in each dataset up to 1,800. An example of a pixels zone added to the 1,500 pixels datasets is showed in Fig. 14.9a.

Each incremented dataset (we called DS1a the incremented dataset obtained by DS1; the same was for DS2a, and DS3a, which were obtained by DS2 and DS3, respectively) was used to perform new feature extractions using only QRA. This procedure returned three feature subsets (FSDS1a, FSDS3a, and FSDS3a), with a

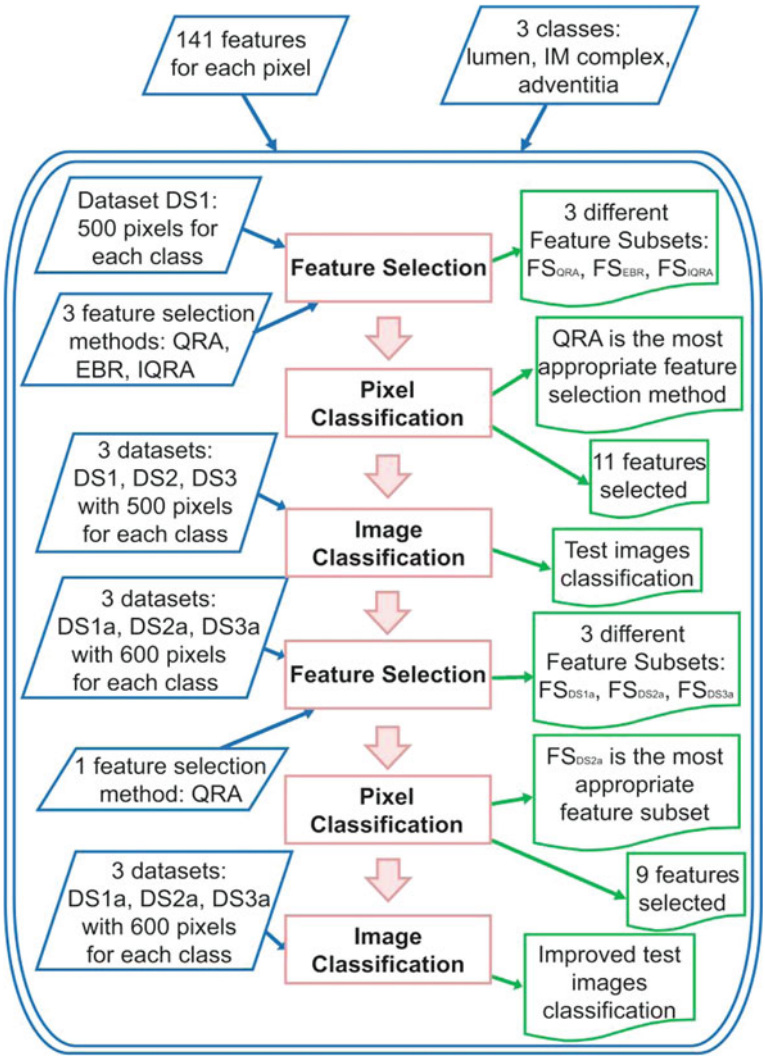


Fig. 14.7 Schematic representation of the process performed to optimize feature selection using the classical symbols of flowcharts. In this image, the process symbols are identified the procedural steps, the data symbol represents the input of the single step, and the document symbol indicates the output of the steps. The two data symbols on the diagram top represent the input of the entire system

between eight and nine selected features and a dependency degree equal to 1 in all three cases (Table 14.3).

The three subsets were then compared using the ANNs in which each dataset was used both as testing set and as training set. The percentage of correct

Table 14.1 Features selected from QRA

	FS_{QRA}	
Rectangle dimension	7 × 15	*Standard deviation*
		SGLDM—m_4
		GLDM-SRE 0°
	15 × 7	*SGLDM—*m_1
		GLDM-Con
		GLDM-GLD 0°
		GLDM-SRE 90°
	7 × 3	*Standard deviation*
		SGLDM—m_1
	3 × 7	*Mean value*
		GLRLM-RLD 0°

Features extracted applying QRA on a dataset made up of 1,500 pixels (FSQRA)

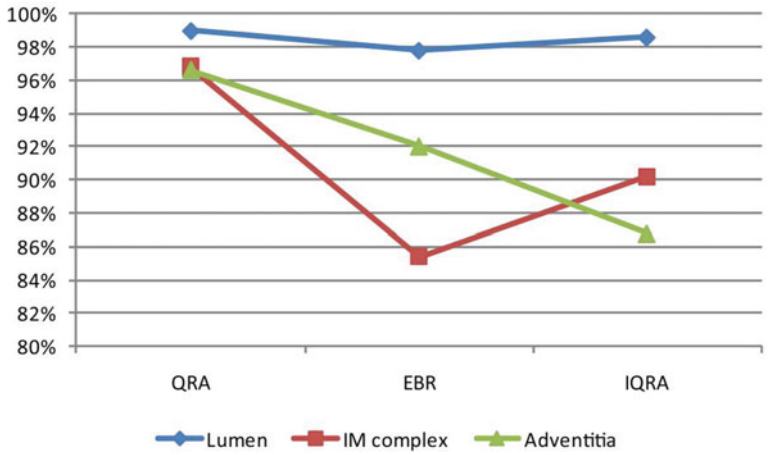


Fig. 14.8 The percentage of correct classification for pixels belonging to lumen, IM complex and adventitia for three tested feature selection procedures

Table 14.2 Feature selection method performances

FS method	Number of features	γ	Percentage of correct classification (%)
QRA	11	0.996	97.47
EBR	10	0.996	91.73
IQRA	10	0.995	91.87

The table below shows the results of three tested reduction methods, listed in the first column, applied on the same dataset made up of 1,500 pixels. The second and third columns contain the number of selected features and the reduct dependency degree, respectively. The last column reports the method performances in terms of percentage of correct classification

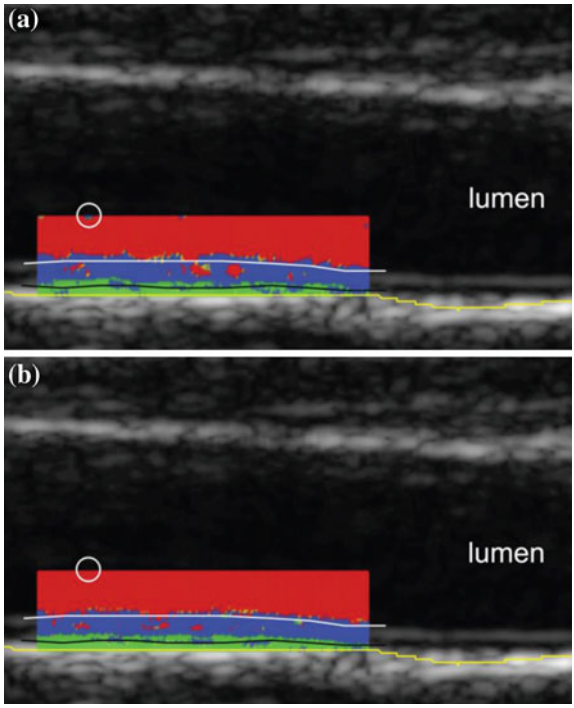


Fig. 14.9 Comparison between the first version of voting classifier, **a** trained with three datasets of 1,500 pixels each and the final voting classifier, **b** trained with three datasets of 1,800 pixels each. *Red*, *blue*, and *green points* mark the pixels classified as belonging to lumen, IM, and adventitia classes, respectively. Pixels not classified by the voting procedure are indicated with *yellow points*. The *white circle* highlights an example of pixels zone added to the initial datasets. These pixels are chosen among those pixels resulting in a wrong class from the first voting classification (**a**) and in the last classifier the same pixels result in correct class (**b**). The *white line* corresponds to the lumen-intima (LI) interface, the *black line* marks out the media-adventitina (MA) interface, and the *yellow line* delimits the far adventitia layer

Table 14.3 QRA results

Dataset used for FS	Number of features	γ	Percentage of correct classification (%)
DS1a	9	1	82.94
DS2a	8	1	91.56
DS3a	9	1	86.06

Results of feature reduction performed by means of QRA on three datasets made of 1,800 pixels each. The second and third columns contain the number of selected features and the reduct dependency degree, respectively. The last column reports the methods performances in terms of percentage of correct classification

classification resulted higher than 82 % for all nets (Table 14.3, third column), while the best performances were obtained with FSDS2a that allowed identifying the right class for the 91.56 % of pixels.

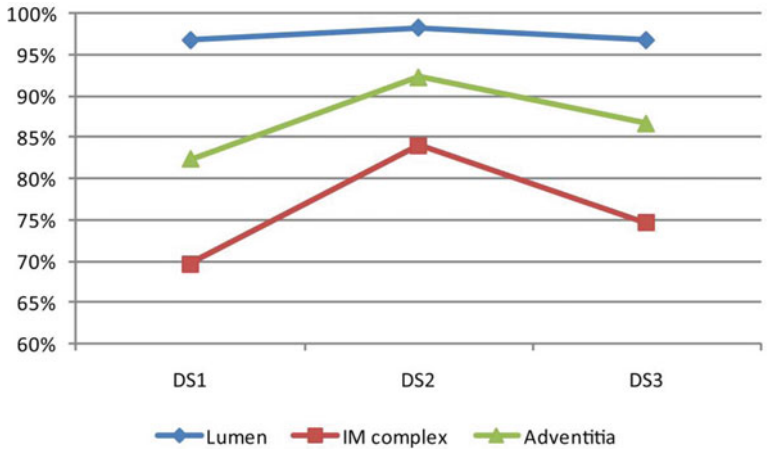


Fig. 14.10 Percentage of correct classification for pixels belonging to lumen, IM complex, and adventitia using three feature subsets extracted by means of QRA applied on three datasets

Table 14.4 Features selected: features extracted applying QRA on the second dataset made up of 1,800 pixels (FSDS2a)

	FS_{DS2a}	
Rectangle dimension	7 × 15	*Mean value*
		GLDM-Con
		GLRLM-LRE 0°
		GLRLM-GLD 0°
	15 × 7	*Standard deviation*
		GLDM-ASM
	7 × 3	*GLRLM-GLD* 45°
		GLRLM-RLD 90°
	3 × 7	*GLRLM-SRE* 0°

Figure 14.10 shows the percentage of correct classification of pixels belonging to the different classes for each extracted subset. It can be noted that while more than 95 % of lumen pixels are classified in the right class, the percentage of correct classification slightly decreases for the other two classes.

The features extracted from DS2a, listed in Table 14.4, were selected to perform a new validation step and then to classify all the pixels of the test images. Again we built a classifier based on three ANNs and a voting system. By comparing the results of the two classifiers, it can be seen that the total error was reduced. Fig. 14.9b shows the pixel classification performed on a portion of the US carotid artery image with the new classifier. Even if there were still misclassified pixels, the performances of the new classifier were acceptable and similar on all the test images. As an example, we report two images of a carotid without (panel A) and with plaque (panel B) in Fig. 14.11.

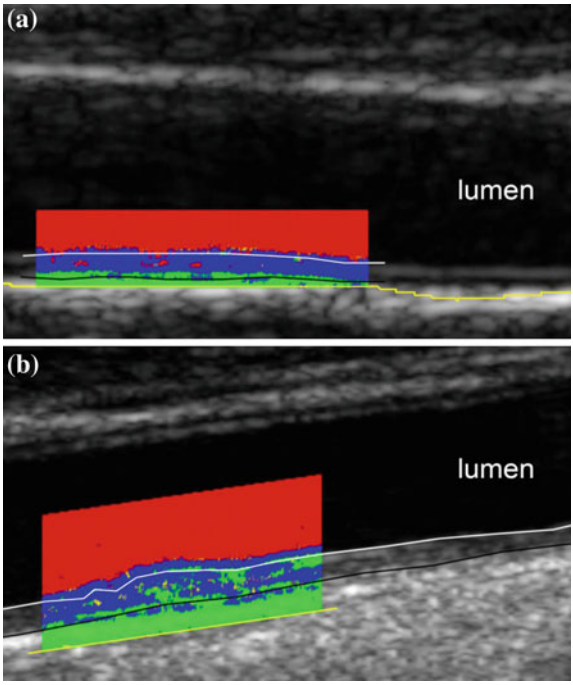


Fig. 14.11 The final voting classifier performance in classifying a portion of two carotid images without plaque (**a**) and with plaque (**b**)

From this analysis, the subset with the best performance was selected and a new voting classifier was built based on the outcomes of three ANNs. As for the previous classifier, the three networks were obtained using three datasets as a training set.

14.11 Data Interpretation and Discussion

Because of the complexity and the variety of the typology of analyzed images, many variables must be considered in order to gain a complete description of the system. For this study, we derived 141 features entailing the construction of large datasets.

First, we compared the performances of three feature selection methods in identifying a minimal subset of variables to keep the same amount of relevant information contained in the set of parameters derived from the US carotid images. Such a procedure allowed us to emphasize the relevant attributes to obtain reliable classification. The pixel classification was assessed by means of ANNs, an unsupervised procedure in which knowledge is acquired by the network through a learning process.

Results presented in the previous section lead to the observation that the features selected by QRA give the best results. Although the single pixel classification using the two feature subsets derived from QRA and listed in Tables 14.2 and 14.4 is encouraging, the selected subset of features changes using different data sets;

because, there is no correlation among the variables included in the subsets and the classification applied to the test images was acceptable.

A possible limitation of the procedure implemented in this study is its low robustness in the feature extraction stage: if the dataset is too small, the procedure could not extract all the features required for a proper classification. This limitation may be removed by increasing the number of elements that compose the dataset, so that all possible typologies of pixels can be taken into account for the feature selection. This step is important when dealing with multi-institutional and multi-ethnic databases, where a wide variety of pixel characteristics could be expected, for example in ultrasound imaging. Ultrasounds are user-dependent imaging modalities, since different sonographers could acquire different images of the same patient. Variability in ultrasound images is also increased by noise (and, particularly, speckle noise, which is typical of the multi-scattering of the ultrasound pulse), by the settings of the ultrasound device (i.e., intensity compensation, overall gain, time compensation, grayscale settings, dynamic range), and by the type and frequency of ultrasound probe. In summary, many causes determine the variability of the pixel intensities, distribution, and classes. Therefore, extensive and large validation studies are required to characterize the performance of feature selection approaches fully.

Despite the challenges that still need to be faced, this method is a novel approach toward automated feature selection in ultrasound carotid imaging that was specifically designed to improve the overall far wall segmentation performance.

In this optic, the results we showed, even though still preliminary, are encouraging. Figures 14.9 and 14.11 show the first classification samples. It can be observed that the overall classification of the pixels is good, despite the presence of some misclassified points in the three classes. However, Figs. 14.9 and 14.11 demonstrate that the class boundaries are correctly traced in correspondence of the manually traced LI and MA ground-truth profiles. The classification procedure was not influenced by the artery morphology and could correctly process normal (Fig. 14.11a) as well as plaqued (Fig. 14.11b) carotids.

In conclusion, selection and reduction could be an important pre-processing strategy for increasing the segmentation performance of automated ultrasound techniques. This step, which is to be performed prior of the actual IMT measurement, can foster the image knowledge discovery at pixel level, thus providing a reduced and organized set of parameters to segmentation techniques.

References

1. Amadasun M, King R (1989) Textural features corresponding to textural properties. IEEE Trans Syst Man Cyb 19(5):1264–1273
2. Badimon JJ, Ibanez B, Cimmino G (2009) Genesis and dynamics of atherosclerotic lesions: implications for early detection. Cerebrovasc Dis 27(1):38–47
3. Barnett H et al (1991) Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. North American symptomatic carotid endarterectomy trial collaborators. N Engl J Med 325:445–453

4. Chen Y, Miao D, Wang R (2010) A rough set approach to feature selection based on ant colony optimization. Pattern Recogn Lett 31:226–233
5. Chen Y, Miao D, Wang R, Wu K (2011) A rough set approach to feature selection based on power set tree. Knowl-Based Syst 24:275–281
6. Conners RW, Harlow CA (1980) A Theoretical comparison of texture algorithms. IEEE Trans Pattern Anal Mach Intell 2:204–222
7. de Groot E, van Leuven SI, Duivenvoorden R, Meuwese MC, Akdim F, Bots ML, Kastelein JJ (2008) Measurement of carotid intima-media thickness to assess progression and regression of atherosclerosis. Nat Clin Pract Cardiovasc Med 5:280–288
8. Faita F, Gemignani V, Bianchini E, Giannarelli C, Ghiadoni L, Demi M (2008) Real-time measurement system for evaluation of the carotid intima-media thickness with a robust edge operator. J Ultrasound Med 27:1353–1361
9. Feng L, Wang GY, Li XX (2010) Knowledge acquisition in vague objective information systems based on rough sets. Expert Syst 27:129–142
10. Fisher M, Martin A, Cosgrove M, Norris JW (1993) The NASCET-ACAS plaque project. North American symptomatic carotid endarterectomy trial. Asymptomatic carotid atherosclerosis study. Stroke 24:I24–I25
11. Greco S, Matarazzo B, Slowinski R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129:1–47
12. Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22:4–37
13. Jensen R, Shen Q (2001) A rough set-aided system for sorting WWW bookmarks. Web Intell: Res Dev 2198:95–105
14. Jensen R, Shen Q (2008) Computational intelligence and feature selection: rough and fuzzy approaches. Wiley, Hoboken
15. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley, Hoboken
16. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17:491–502
17. Małyszko D, Stepaniuk J (2010) Adaptive multilevel rough entropy evolutionary thresholding. Inform Sci 180:1138–1158
18. Matsumoto Y, Watada J (2009) Knowledge acquisition from time series data through rough sets analysis. IJICIC 5:4885–4897
19. Molinari F, Liboni W, Giustetto P, Badalamenti S, Suri JS (2009) Automatic computer-based tracings (ACT) in longitudinal 2-D ultrasound images using different scanners. J Mech Med Biol 9:481–505
20. Molinari F, Zeng G, Suri JS (2010) A state of the art review on intima-media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound. Comput Methods Programs Biomed 100:201–221
21. Molinari F, Zeng G, Suri J (2010) Inter-greedy technique for fusion of different segmentation strategies leading to high-performance carotid IMT measurement in ultrasound images. J Med Syst 35(5):905–919
22. Molinari F, Zeng G, Suri JS (2010) Carotid wall segmentation and IMT measurement in longitudinal ultrasound images using morphological approach. Paper presented at the 2010 IEEE international symposium on biomedical imaging: from Nano to Macro, Rotterdam, The Netherlands, April 2010, pp 14–17
23. Molinari F, Acharya UR, Zeng G, Meiburger KM, Suri JS (2011) Completely automated robust edge snapper for carotid ultrasound IMT measurement on a multi-institutional database of 300 images. Med Biol Eng Comput (in press)
24. Molinari F, Meiburger KM, Zeng G, Nicolaides A, Suri JS (2011) CAUDLES-EF: carotid automated ultrasound double line extraction system using edge flow. J Digit Imaging 24(6):1059–1077
25. Molinari F, Liboni W, Pantziaris M, Suri JS (2011) CALSFOAM-completed automated local statistics based first order absolute moment" for carotid wall recognition, segmentation and

IMT measurement: validation and bench-marking on a 300 patient database. Int Angiol 30:227–241
26. Moradi H, Grzymala-Busse JW, Roberts JA (1998) Entropy of english text: experiments with humans and a machine learning system based on rough sets. Inf Sci 104:31–47
27. Naqvi TZ (2006) Ultrasound vascular screening for cardiovascular risk assessment. Why, when and how? Minerva Cardioangiol 54:53–67
28. Pawlak Z (1982) Rough Sets. Int J Comput Inform Sci 11:341–356
29. Pawlak Z, Sowinski R (1994) Rough set approach to multi-attribute decision analysis. Eur J Oper Res 72:443–459
30. Pignoli P, Longo T (1988) Evaluation of atherosclerosis with B-mode ultrasound imaging. J Nucl Med Allied Sci 32:166–173
31. Poredos P (2004) Intima-media thickness: indicator of cardiovascular risk and measure of the extent of atherosclerosis. Vasc Med 9:46–54
32. Prasad PS, Rao CR (2009) IQuickReduct: an improvement to quick reduct algorithm. Lect Notes Comput Sci 5908:152–159
33. Roquer J, Segura T, Serena J, Castillo J (2009) Endothelial dysfunction, vascular disease and stroke: the ARTICO study. Cerebrovasc Dis 27:25–37
34. Rothwell PM, Warlow CP (1999) Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European carotid surgery trialists' collaborative group. Lancet 353:2105–2110
35. Rothwell PM, Gibson RJ, Slattery J, Warlow CP (1994) Prognostic value and reproducibility of measurements of carotid stenosis. A comparison of three methods on 1001 angiograms. European carotid surgery trialists' collaborative group. Stroke 25:2440–2444
36. Schargrodsky H, Hernandez–Hernandez R, Champagne BM, Silva H, Vinueza R, Silva Aycaguer LC, Touboul PJ, Boissonnet CP, Escobedo J, Pellegrini F, Macchia A, Wilson E (2008) CARMELA: assessment of cardiovascular risk in seven Latin American cities. Am J Med 121:58–65
37. Shen Q, Chouchoulas A (2000) A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. Eng Appl Artif Intel 13:263–278
38. Su CT, Yang CH (2008) Feature selection for the SVM: an application to hypertension diagnosis. Expert Syst Appl 34:754–763
39. Swiniarski RW, Skowron A (2003) Rough set methods in feature selection and recognition. Pattern Recogn Lett 24:833–849
40. Tan JH, Ng EYK, Acharya UR, Chee C (2010) Study of normal ocular thermogram using textural parameters. Infrared Phys Tech 53:120–126
41. Thangavel K, Pethalakshmi A (2009) Dimensionality reduction based on rough set theory: a review. Appl Soft Comput J 9:1–12
42. Touboul PJ, Hernandez–Hernandez R, Kucukoglu S, Woo KS, Vicaut E, Labreuche J, Migom C, Silva H, Vinueza R (2007) Carotid artery intima media thickness, plaque and Framingham cardiovascular score in Asia, Africa/Middle East and Latin America: the PARC-AALA study. Int J Cardiovasc Imaging 23:557–567
43. Tsumoto S (1998) Automated extraction of medical expert system rules from clinical databases based on rough set theory. Inform Sciences 112:67–84
44. World Health Organization (2011) Cardiovascular disease. W. H. Organization. http://www.who.int/cardiovascular_diseases/en/. Accessed June 2011
45. Ziarko W (1993) Variable precision rough set model. J Comput Syst Sci 49:39–59