Advanced Technologies and Societal Change

Bipin Kumar Rai
Gautam Kumar
Vipin Balyan *Editors*

# AI and Blockchain in Healthcare

Springer

# Advanced Technologies and Societal Change

This series covers monographs, both authored and edited, conference proceedings and novel engineering literature related to technology enabled solutions in the area of Humanitarian and Philanthropic empowerment. The series includes sustainable humanitarian research outcomes, engineering innovations, material related to sustainable and lasting impact on health related challenges, technology enabled solutions to fight disasters, improve quality of life and underserved community solutions broadly. Impactful solutions fit to be scaled, research socially fit to be adopted and focused communities with rehabilitation related technological outcomes get a place in this series. The series also publishes proceedings from reputed engineering and technology conferences related to solar, water, electricity, green energy, social technological implications and agricultural solutions apart from humanitarian technology and human centric community based solutions.

*Major areas of submission/contribution into this series include, but not limited to:* Humanitarian solutions enabled by green technologies, medical technology, photonics technology, artificial intelligence and machine learning approaches, IOT based solutions, smart manufacturing solutions, smart industrial electronics, smart hospitals, robotics enabled engineering solutions, spectroscopy based solutions and sensor technology, smart villages, smart agriculture, any other technology fulfilling Humanitarian cause and low cost solutions to improve quality of life.

Bipin Kumar Rai · Gautam Kumar · Vipin Balyan
Editors

# AI and Blockchain in Healthcare

Springer

*Editors*
Bipin Kumar Rai
ABES Institute of Technology
Ghaziabad, India

Gautam Kumar
CMR Engineering College
Hyderabad, India

Vipin Balyan
Cape Peninsula University of Technology
Cape Town, South Africa

ISSN 2191-6853 ISSN 2191-6861 (electronic)
Advanced Technologies and Societal Change
ISBN 978-981-99-0376-4 ISBN 978-981-99-0377-1 (eBook)
https://doi.org/10.1007/978-981-99-0377-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

In today's scenario, humankind has entered the domain of the Industrial revolution requirements with advanced tools and techniques. Artificial intelligence (AI) plays a pivotal role in the simulation of human intelligence to process, especially in computing systems. The performance of AI has various applications to apply in expert systems, natural language processing, speech recognition, machine vision, and more. In addition, some specific applications need special attention to deploy because AI helps data to process intelligently and outer protection coverage with some technology, like Blockchain. We have considered one of the Blockchain technologies to apply its principles to AI. This book provides basic concepts and applications of state-of-the-art Blockchain and AI research in healthcare. Its primary focus is on challenges and solutions to apply for the most-intensive application protections with AI techniques as a human safety integration with the healthcare area.

In this digital age, for all of our daily life, security and privacy are the most crucial parts. Dealing with tools and technologies is one of the significant challenges because so many attacks are reported on all kinds of computer systems and networks. It is becoming increasingly important to develop more robust, adaptive, scalable, reliable, private, and secured mechanisms for applications in their related areas. In relation to the same, it is imperative to understand the fundamentals of how AI is applicable with security principles, vulnerabilities protection, handful solutions & optimized solutions applicability use as defense mechanisms.

The objective of this book is to collect and address a variety of problems related to Healthcare Mechanisms, because, in the fast-growing environment, the research trends in this area are always having great demand in the form of prospective mechanisms. The contributors address theoretical and practical aspects of the challenges and opportunities of the application to strengthen the development of platforms. This book aids readers in gaining insight and knowledge about providing security and solutions to different challenges in healthcare using AI and Blockchain technology.

Therefore, we have an attempt to provide a significant effort in the form of a present book entitled "AI and Blockchain in Healthcare". The book contains fourteen chapters (14).

Chapter 1 covers drug discovery and manufacturing with machine learning (ML), which is used at each phase of drug development to speed up research on reducing risk and minimizing the cost in clinical trials. The ML techniques discussed herein are expected to increase the ML roles in drug discovery and manufacturing processes to a new level with the aid of advanced computer intelligence.

Chapter 2 discusses finding one of the best strategies with the Analytical Network Process (ANP) approach, whose results will help epidemiologists and healthcare center workers to take relevant measures. In addition, a case study is presented to confirm the measures extracted in the healthcare centers of Qeshm Island.

Chapter 3 discusses the Era of Blockchain in Healthcare, which is one of the most important application areas where Blockchain is expected to have a significant impact on other applications. It has allowed for more effective and efficient patient care administration. Blockchain helped this sector due to its useful characteristics: peer-to-peer, protected, and transparent technologies.

Chapter 4 presents Securing Healthcare records, Applications, and Challenges using Blockchain with the benefits of adopting blockchain technology for securing healthcare data and emphasizes its important characteristics. The work also lists the challenges impeding blockchain implementation in the healthcare sector.

Chapter 5 discusses the authentication schemes for healthcare data using emerging computing technologies, where results show the main interests while presenting data security and verification techniques, as well as validating the focus of past studies towards authentication schemes using the applications of emerging computing techniques.

Chapter 6 discusses biomedical data classification using fuzzy clustering designed for problems where the very nature of data is unclear and uncertain. It also focuses on the impact of related technologies on human life that has made a tremendous impact in enforcing acceptance of such smart intelligence technologies in various aspects of our lives.

Chapter 7 presents a case study of lung cancer diagnosis through deep learning. The literature for lung cancer detection employs features using deep residual networks, and a comparison between existing techniques is presented and discussed.

Chapter 8 endeavors the machine and deep learning models experimented on cardiotocograph (CTG) data. It has played a huge role in understanding the data and corresponding requirements of data preprocessing. It is also classified into various models that are used in experimenting the data augmentation with clear benefits in terms of performance analysis.

Chapter 9 focuses on the Blockchain framework, applications of Blockchain and machine learning in clinical research for health and well-being, impact, and future prospects of the healthcare ecosystem. The security issues of e-healthcare systems and various Security Protocols have been discussed.

Chapter 10 discusses a breast cancer-based recommendation system. It will provide insights into recommendation scenarios and recommendation approaches. The examples are from the prediction and treatment of breast cancer and recommendation for drug and rehabilitation. Finally, the challenges concerning the development of recommender systems in the future are discussed.

Chapter 11 concentrates on Real-Time Data Mining-Based Cancer Disease Classification Using the KEGG (Kyoto Encyclopedia of Genes and Genomes) Gene Dataset. This study's goal is to develop an effective computational strategy for identifying the sort of cancer tumors that will develop to find a wide variety of cancer-related disorders, their examination, and training of bioinformatics, as well as the memorization of knowledge about genomics, met genomics, and metabolomics.

Chapter 12 introduces the solution architecting on remote medical monitoring with AWS cloud and IoT. The findings use amazon web services while patients-doctors are online in the healthcare industry by cutting response time. The data is collected via an IoT that takes care of most of the concerns.

Chapter 13 this chapter presents a domain-oriented framework for the prediction of diabetes disease and the classification of diet. The proposed approach is available for Diabetic Prediction and Diet using Machine Learning Techniques with 10-fold cross-validation. The collected data is from the PIMA database for its classification.

Chapter 14 proposes an implementation to identify the most widely spread illness globally, known as swine flu, using a database of treatment patients. Swine flu is a respiratory sickness requiring a large number of tests from the patient to identify an illness. In order to justify this, it is proposed that the situation may be remedied with the use of advanced information mining techniques.

We hope that the works published in this book will be able to serve the concerned communities working in the fields of Healthcare, AI, Blockchain, Security, and IoT.

Bipin Kumar Rai
ABES Institute of Technology
Ghaziabad, India
bipinkrai@gmail.com

Gautam Kumar
CMR Engineering College
Hyderabad, India
gautam21ujrb@gmail.com

Vipin Balyan
Cape Peninsula University
of Technology
Cape Town, South Africa
balyanv@cput.ac.za

**Acknowledgements** The editors are thankful to the authors and reviewers of the concerned chapters who contributed to this book with their scientific work and valuable comments, respectively.

# Contents

## Part I Role of AI and Blockchain in Healthcare

## Part II Application of AI and Blockchain in Healthcare

# Part I
# Role of AI and Blockchain in Healthcare

# Chapter 1
# Machine Learning for Drug Discovery and Manufacturing

**Bogala Mallikharjuna Reddy**

## 1.1 Introduction

Drug discovery and manufacturing is a very difficult multi-objective optimization issue in which a candidate molecule must meet numerous criteria, including efficacy against a biological target, appropriate biophysical and pharmacokinetic characteristics, and safety [1]. Drug discovery and manufacturing are time-consuming and costly, with a high clinical failure rate (>90% from phase I to drug launch) [2]. According to Bernard H Munos' research published in Nature Reviews Drug Discovery, US Food and Drug Administration (USFDA) authorized just 1222 new drugs during 1950–2008, and the rate of new drug manufacturing by corporations has remained consistent over the previous 60 years [3]. For example, in 1996, 2003, and 2009, US and EU authorities reviewed 131, 72, and 48 applications for novel active chemical and biological substances, yet only 56, 27, and 25, respectively, were authorized by the USFDA [3]. At the same time, the Tufts Center for the Study of Medicine Development (TCSDD) estimates that the cost of successfully creating a new prescription drug that receives marketing clearance is $2.8 billion (USD) [2]. This entire process takes over a decade (11–16 years), which must be weighed against the therapeutic molecule's 25 year patent duration [2]. There is economic pressure to move quickly due to impending patent expiration and often fierce competition, as well as social pressure to make new biologics available to patients worldwide, and new technologies are thus required to generate high quantities of in-speciation products to meet market demand, at affordable prices, and in shorter timescales. With the commencement of the COVID-19 pandemic in early 2020, these demands have

B. M. Reddy (✉)
Technology Business Incubator, Abinnovus Consulting Pvt. Ltd., University of Madras, Guindy Campus, Chennai, India
e-mail: mrbogala@crimson.ua.edu

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_1

become much more apparent [1]. As a result, certain new technologies are desperately needed to address this costly and time-consuming drug development issue while also lowering the rate of clinical failure.

Physiological systems are complicated sources of information during the stages of illness development and treatment with appropriate drugs. This data is now being systematically monitored and mined at unprecedented levels, thanks to a slew of 'omics' and smart technologies. The introduction of these high-throughput methods to medicine and illness confronts the pharmaceutical industry with both problems and possibilities, to identify credible therapeutic hypotheses from which to produce the proper chemical molecule for healing the condition. In pharmaceutical companies, data digitization has increased dramatically in recent years. However, the difficulty of gathering, evaluating, and utilizing knowledge to solve complicated healthcare problems arise with digitalization [4]. Recent breakthroughs in computer science and technology (particularly artificial intelligence or AI) have heightened interest in the application of machine learning (ML) technologies in the pharmaceutical industry. With improved automation, ML can manage massive amounts of pharmaceutical data [4]. The significant expansion in the types and volumes of data sets, which may be the foundation for ML along with endlessly scalable storage, has enabled pharmaceutical companies to access and organize far more data. Patients' disease images (CT scans, X-rays, digital data, etc.), textual information (medical prescriptions, body fluid test results, etc.), biometrics, and other information from wearable (band-aids, implants, prosthetics, etc.), assay and clinical trial information (patients tested, completely cured, etc.), and high-dimensional 'omics' data are typical examples of pharmaceutical data types [5].

This book chapter contains literature on different ML techniques and approaches that are being used during the drug research and manufacturing processes to minimize time, risk, failure, and cost in clinical trials. To recover correct results, ML approaches increase the decision -making of pharmaceutical information with numerous applications, such as hit discoveries, QSAR studies, and de novo drug designs. This study considers prognostic biomarkers, target confirmation, and digital pathology under problem statements. ML problems must be relevant for the primary reason of insufficiency in interpretability outputs, which may limit drug discovery applications. Complete and practical data need to be collected in clinical trials to solve numerous mysteries in authenticating ML methods, enhancing decision-making, raising awareness of ML approaches, and determining possible breakdowns in drug discovery and manufacturing.

## 1.2 Machine learning (ML)

### *1.2.1 Artificial Intelligence (AI)*

Artificial intelligence (AI) is sometimes known as machine intelligence in computer science because computers may be educated or programmed to do tasks similar to those performed by the human brain [6]. AI is a technology-based system that uses a variety of advanced tools and networks to simulate human intellect. Still, it does not threaten to entirely replace human physical presence with computer-controlled machinery [1]. AI employs systems and software that can read and learn from input data to make independent judgments for achieving certain goals. AI is a discipline that relates to the broad variety of applications and layouts of various algorithms for understanding and gaining information from data. The AI idea is inextricably linked to several subjects like probability theory, pattern recognition, statistics, machine learning, and various processes such as fuzzy models and neural networks, all of which are referred to together as "Computational Intelligence" [6]. AI methodologies like classification, regression, predictions, and optimization approaches are used in a variety of complex applications.

### *1.2.2 Machine Learning Techniques (MLTs)*

AI encompasses multiple approach fields, including reasoning, knowledge representation, solution search, and a core machine learning (ML) paradigm. To effectively use any type of information, machine learning must first be updated, which means that a specific model must be defined along with input parameters. As a result, from the training data, machines can achieve expertise in the model with accessible parameters. Furthermore, the model can forecast data in the future to retrieve information from data [6]. ML employs algorithms that recognize patterns in a previously categorized collection of data. Deep learning (DL) is an area of ML that employs artificial neural networks (ANNs). These are a collection of interconnected sophisticated computer units that include 'perceptons', which are equivalent to human physiological neurons and simulate the transmission of electrical impulses in the human brain [4]. ANNs are a collection of nodes that receive different inputs before converting them to output, either singly or multi-linked, employing methods to solve issues [4]. ANNs come in a variety of forms, including multilayer perceptron (MLP) networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), and can be trained supervised or unsupervised [4].

### *1.2.3 Classification of ML*

The MLP network has applications like optimization assistance, pattern recognition, process identification, and controls. It is often taught using supervised training processes that operate in just one direction and may be utilized as universal pattern classifiers [4]. RNNs, like Boltzmann parameters and Hopfield networks, are closed-loop networks that can memorize and retain information [4]. CNNs are a class of dynamic systems with local connections that are differentiated by their topology. They are used in image and video processing, biological system modeling, complicated brain function processing, pattern recognition, and advanced signal processing [4]. Networks (Kohonen, RBF, LVQ, counter-propagation, and ADALINE) are some of the more complicated variants [4]. Figure 1.1 shows the classification of ML approaches [4].

The capacity of ML to detect patterns in complicated and large-volume data sets is key to its success [7]. Furthermore, ML techniques (MLTs) may be built using common programming languages like Python and R, which are accessible to the majority of academics. Furthermore, third-party software, such as Apple's Create ML, provides access to ML approaches for researchers who are uncomfortable with coding. Despite their ease of use, third-party software is restricted in its ability to conduct ML methods and other components of the ML pipeline. Traditional MLTs have been intensively investigated in drug development [7]. Examples of supervised and unsupervised MLTs are k-Nearest Neighbor (kNN), decision tree, random forest, support vector machines (SVM), artificial neural networks (ANN), principal component analysis (PCA), and k-means. Their attractiveness arises from their simplicity, which is computationally undemanding yet provides better prediction accuracy than typical predictive algorithms [7]. Noncomputer scientist researchers can also understand the underlying processes of traditional methodologies. For example, with kNN, the user controls only one parameter, the k value, which sets the categorization search

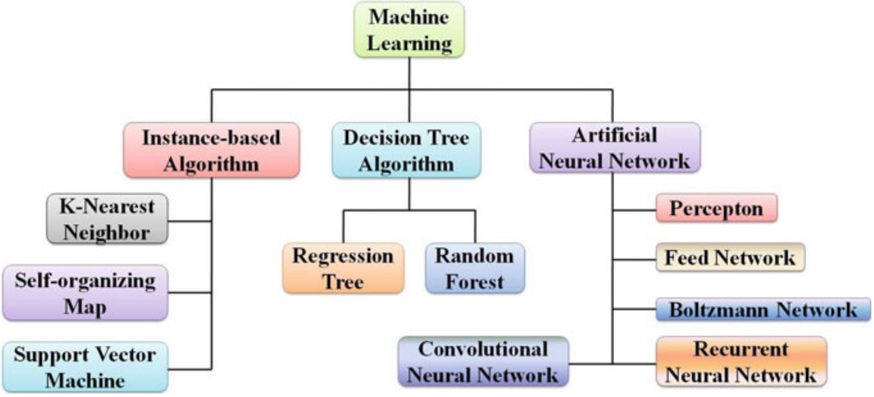


**Fig. 1.1** Classification of ML techniques [4]

space based on a plurality vote. Another example is SVM, which delineates categories by combining a hyperplane and support vectors to minimize the distance between them. SVM benefits from the kernel technique, which allows for non-linear data mapping and has been frequently utilized for non-linear data sets [7]. The method is also applicable to PCA (kernel PCA; kPCA) [7]. A recent study discovered that kPCA may be utilized to enhance the classification of linear models at a substantially quicker pace than non-linear models.

Traditional MLTs, despite their simplicity, have limitations. For example, kNN suffers from the curse of dimensionality, which occurs when the prediction performance begins to deteriorate in a high-dimensional space [7]. Similarly, once the number of dimensions exceeds the sample size, SVM performance begins to deteriorate [7]. Increasing the tree number in random forest enhances predicted accuracy, but a large number of trees results in an inefficient algorithm for real-time monitoring. However, there are two major objections to MLT: their need for huge data and their lack of transparency. Addressing these constraints is necessary since data collecting may be difficult, expensive, and time-consuming. Furthermore, transparency may aid the user's knowledge of the discovery process and reduce their dependence on ML to do so. Another disadvantage of traditional MLTs is their lack of autonomy. Supervised learning, for example, necessitates labeling of the target variable (i.e., the variable to be predicted). Furthermore, once implemented, such as web-based software, it will need post-production maintenance, particularly as the data collection expands. To solve these restrictions, research groups have embraced new methodologies, with promising results. These sophisticated approaches are expected to broaden the scope of ML applications. Deep learning has evolved in recent years as a technology capable of generating high accuracies from massive data while managing both organized and unstructured data. As previously stated, the use of ML in drug development is expanding. Several methods for drug discovery and manufacturing have been created based on the networks that comprise the basic architecture of ML systems. The Watson supercomputer at International Business Machine (IBM) is one such tool created with ML technology [4]. It was created to aid in the study of a patient's clinical data and its association with a large database, culminating in the recommendation of cancer treatment techniques. This technique may potentially be used to diagnose illnesses quickly [4]. Thus, MLs applications in the pharmaceutical area are constantly being expanded, with the ultimate goal of achieving AI in the drug development pipeline [7].

## 1.3 ML for Drug Discovery and Manufacturing

### *1.3.1 ML and Drug Discovery*

Recently, ML techniques have been used in a variety of industries, including the pharmaceutical industry. Figure 1.2 depicts the numerous ML disciplines of drug discovery and manufacturing [6]. To express therapeutic notions, each phase was carried out in the manner of a pipeline. The various phases indicate distinct repetitions (in time) and expense investment. Every stage is done to reveal the efficiency of the restorative treatment. Using ‘omics’ and ‘smart automation techniques, clinical data were extracted and precisely assessed. Extending these ML approaches into the biological sector expands both potential and problems in the pharmaceutical industry. Because the goal of many pharmaceutical companies is to differentiate the convincing clinical hypothesis, practitioners or physicians can create drugs based on the findings. ML algorithms are used to evaluate the performance of any type of medicine in pharmaceutical companies. With infinite storage and improvements in datasets like size and kind, ML may be given premises. As a result, ML may have access to massive amounts of data from the pharmaceutical industry. The huge chemical database, which contains useful information of >1060 molecules, supports the synthesis of new drugs [4]. The need for novel technology hinders drug discovery and manufacturing, thus making it slow and costly, which may be solved by employing ML [4]. ML can identify lead and hit molecules, allowing for faster confirmation of the therapeutic drug and optimization of its structure design [4].

Because the human genome has enabled cloning methodologies and improved protein refinement in large numbers, high-throughput screening with numerous libraries has often grown. Reverse pharmacology is the procedure of screening activity for big molecules via biological targets to effect change in disease. Numerous hits are created via screening activities to supply cells, and tests for adequacy have also been undertaken in animals. Nowadays, drug discovery is concerned with the identification of screening hits, as well as optimization strategies that can improve drug efficacy, affinity, and metabolic stability. If the chemical meets all requirements, a specific drug will be created in medical tests once the drug is effective. Target identification and validation, lead optimization, hit finding, and clinical trials are all required steps in the drug development and discovery process [6]. The developing cost of a new medicine may be over 2.56 billion USD, and it is a slow method since it takes roughly 10–15 years to sell it in the market [6].

Many investors are pouring a lot of money into pharmaceutical sector for generating precise progress in clinical trials to achieve a limited number of compounds in drug development. Even still, the 13% accuracy rate is disappointing. Clinicians have used the Computer-assisted Drug Design (CADD) method to overcome this problem [6]. By employing this method in drug development, artificial procedures not only give theoretical molecular qualities (i.e., absorption, distribution, selectivity, bioactivity, side effects, metabolism, and excretion), but also lead to products with optimal features in silico. Furthermore, the attrition cost in preclinical stage can be reduced by using multi-objective optimization approaches. In drug discovery, computational intelligence provides many approaches for studying, learning, and elucidating drug discovery with ML for identifying multiple pharmaceuticals in a planned and integrated structure [6]. As a result, numerous pharmaceutical companies have shown increased interest in contributing to technology and resources for obtaining precise drug discovery findings. Finally, this book chapter suggests ML strategies for drug discovery that use deep learning techniques to target many uses in drug discovery and

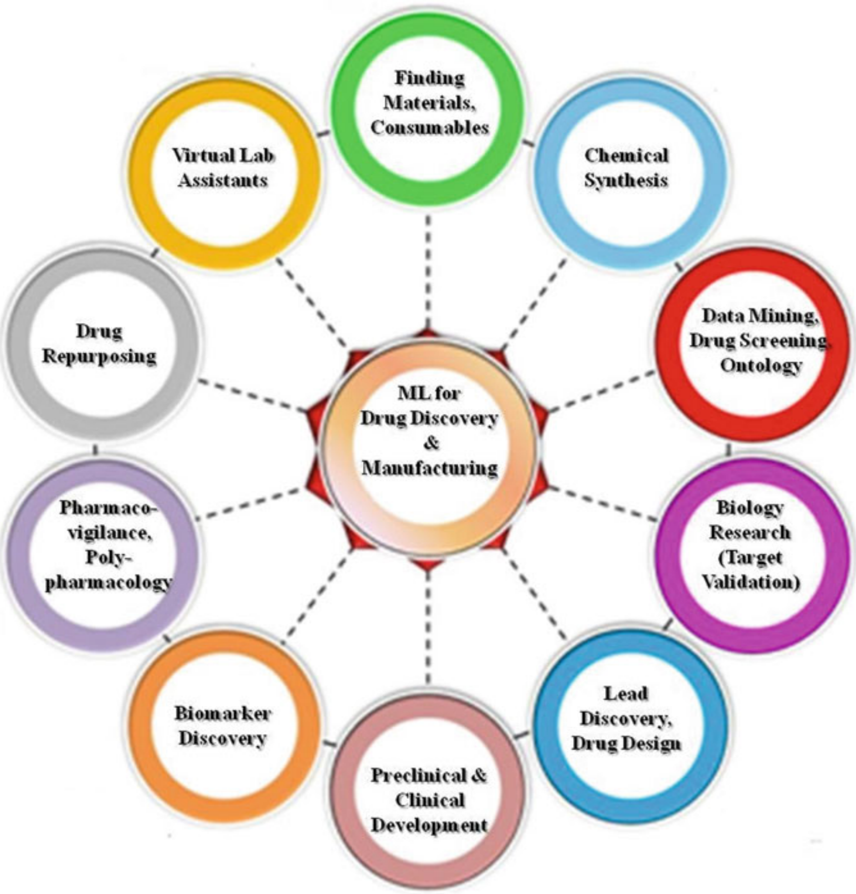

Finding Materials, Consumables
Chemical Synthesis
Data Mining, Drug Screening, Ontology
Biology Research (Target Validation)
Lead Discovery, Drug Design
Preclinical & Clinical Development
Biomarker Discovery
Pharmaco-vigilance, Poly-pharmacology
Drug Repurposing
Virtual Lab Assistants
ML for Drug Discovery & Manufacturing

**Fig. 1.2** ML disciplines of drug discovery and manufacturing [6]

manufacturing. In addition, ML predicts results in terms of computer intelligence in drug discovery and manufacturing.

### *1.3.2 MLTs Role in Drug Discovery*

Table 1.1 lists different types of MLTs and their role in drug discovery [6]. A variety of classifier and regression strategies, i.e., supervised learning approaches used to respond to required expectations in continuous or categorical data variables, as well as unsupervised methods used in developing a model that empowers clustering data, are available.

**Table 1.1** MLTs types and their role in drug discovery [6]

| Machine learning | Types | Models/Methods | Drug discovery role |
|---|---|---|---|
| Unsupervised learning | Clustering | Hidden Markov model | De novo molecular design |
| | | Hierarchal clustering | Deep feature selection for biomarkers |
| | | K-means | Feature reduction in single-cell data to identify cell types |
| | | Neural Networks (DAENS and autoencoders) | Cell types and biomarkers from single-cell RNA data |
| Supervised learning | Classification | NLP Kernel methods | Target-disease-drug associations |
| | | NLP Bayesian classifier | Target drug ability based on pharmaco kinetics (PK) properties |
| | | SVMs | Target drugability-based protein structure or sequence |
| | | Nearest neighbor | Tissue specific biomarkers from gene expression signatures |
| | Regression analysis | Random forests | Molecular features that predict cancer drug response |
| | | Linear regression | Targets for Huntington's disease |
| | | Decision trees | Disease and target drug ability from multi-dimensional data |
| | | Sparse linear regression | Gene expression data that predict clinical trial success |

### *1.3.3 Examples of MLTs in Drug Discovery*

Many ML computations are used in drug discovery and manufacturing to analyze and predict data. The principles of a few common models, such as Support Vector Machine (SVM), random forest (RF), multilayer perception model (MLP), and Deep learning (DL) techniques, and their effective usage in drug discovery and manufacturing was highlighted in Fig. 1.3 [6]. MLTs innovation is a major priority in drug creation, as it collects pharmacological data. ML does not depend on hypothetical advancements, but it is increasingly important in translating clinical data into research such as reusable approaches. Within the framework of ML, various methods like naïve Bayesian Classification (NBC), Random Forest (RF), Logistic Regression

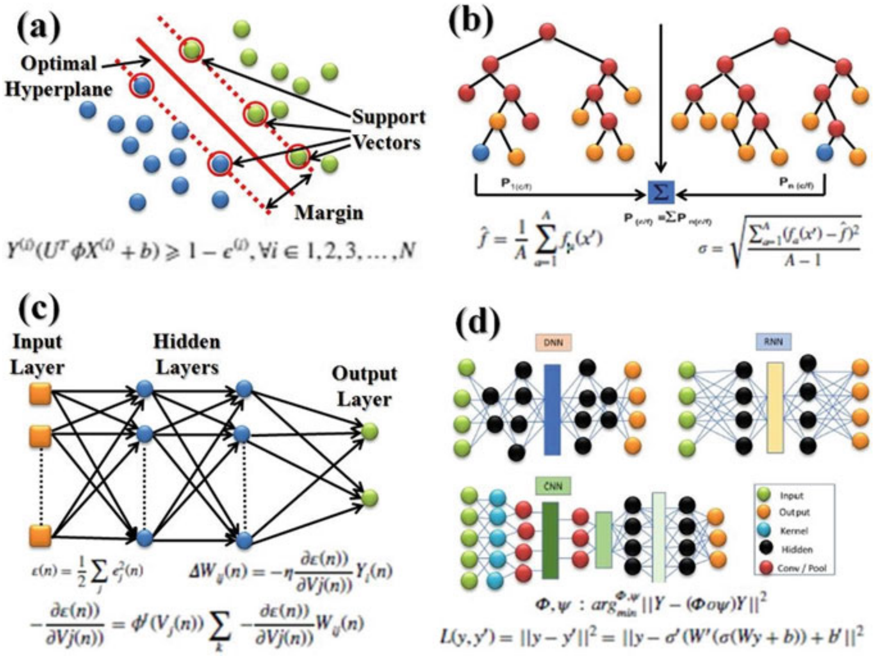

(a)
Optimal Hyperplane
Support Vectors
Margin
$Y^{(i)}(U^T\phi X^{(i)}+b)\geq 1-\epsilon^{(i)}, \forall i\in 1,2,3,\ldots,N$
(b)
$P_{1(c/r)}$
$P_{n(c/r)}$
$P_{(c/r)}=\Sigma P_{n(c/r)}$
$\hat{f}=\frac{1}{A}\sum_{a=1}^{A}f_a(x')$
$\sigma=\sqrt{\frac{\sum_{a=1}^{A}(f_a(x')-\hat{f})^2}{A-1}}$
(c)
Input Layer
Hidden Layers
Output Layer
$\varepsilon(n)=\frac{1}{2}\sum_j e_j^2(n)$
$\Delta W_{ij}(n)=-\eta\frac{\partial\varepsilon(n)}{\partial Vj(n)}Y_i(n)$
$-\frac{\partial\varepsilon(n)}{\partial Vj(n)}=\phi'(V_j(n))\sum_k-\frac{\partial\varepsilon(n)}{\partial Vj(n)}W_{ij}(n)$
(d)
DNN
RNN
CNN
Input
Output
Kernel
Hidden
Conv / Pool
$\Phi,\psi : \arg\min_{\Phi,\psi}\|Y-(\Phi\circ\psi)Y\|^2$
$L(y,y')=\|y-y'\|^2=\|y-\sigma'(W'(\sigma(Wy+b))+b'\|^2$

**Fig. 1.3** MLT models for drug discovery and manufacturing, **a** Support Vector Machine, **b** Random Forest, **c** Multilayer Perceptron, and **d** Deep Learning [6]

(LR), Multiple Linear Regression (MLR), Probabilistic Neural Networks (PNN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), etc., are considered [6]. ML improvements are notably applied as a deep learning approach in drug design to enhance capacity in feature extraction and feature generalization. Because multi-layer feature extraction techniques are utilized to turn basic features into complicated features, deep learning systems will automatically accelerate various features using available initialization data. One advantage of adopting deep learning algorithms was the presence of the low amount of generalization errors, which resulted in more precise findings. CNN, RNN, Auto Encoder, DNN, and RBN (radial basis function) are examples of deep learning approaches.

### *1.3.4 Support Vector Machine*

The Support Vector Machine (SVM) model is a supervised learning approach that is used to predict class-labeled data or binary data. In SVM, $x$ is regarded as a feature vector or input to the SVM model. Then, $x$ R, where $n$ is a dimension feature vector. $Y$ serves as a class, or output for SVM $Y \in \{-1, 1\}$. In this case, binary values are used to perform a classification task. SVM $u$ and $b$ parameters were examined for learning data in the training set. The $i^{\text{th}}$ sample in the dataset is (*X(i)*, *Y(i)*). The optimization of the quadratic equation (Fig. 1.3a) yields *Y*. The SVM and its different kernels are commonly employed in drug discovery [6]. Several problems are addressed, including (i) gene interaction and screen radiation protection using SVM-RBF, (ii) assessing target-ligand interactions using regression-SVM, (iii) identifying drug-target interaction via biased SVM, (iv) prediction of the drug sensitivity using ensemble SVM, and (v) the linear SVM used in identifying novel drug targets, anti/non-anticancer molecule classification, and kinase mutation activation [6, 8]. Based on cancer cell parameters, the SVM technique was utilized to measure anticancer drugs [6]. Twenty-four medicines were evaluated on cancer cell lines to better understand the association between cancer cell characteristics and treatment resistance [6]. The SVM-RBF technique has been utilized in the cure of oral tumors to locate medicinal chemicals from a wide collection of public datasets; the RBF is a prominent kernel function employed in numerous learning algorithms [6]. SVM is used in therapeutic activities to assist locate the active component at various phases of drug discovery and manufacturing. SVM, unlike other ANNs, exhibited the capacity to evaluate drug similarity predictions for a wide range of chemicals. The SVM excelled in the job due to this collection of descriptors, and it was also revealed that the predicted SVM model improved the quality of enzyme inhibitors for conventional QSAR [6].

### *1.3.5 Random Forest*

A supervised algorithm is the random forest (RF) algorithm. The random forest technique has the important benefit of applying it to categorization and regression problems. Overfitting can occur naturally throughout the organization and regression tasks, resulting in a worse output. Random forests can make use of trained algorithmic approaches such as bagging. The training set consists of X = *X1, X2,…,Xn* and *Y* = *Y1, Y2,…, Yn* (Fig. 1.3b). Then, for fitting a random forest tree, samples can be chosen at random from the training set data. In addition, the random forest may be used in medicine for determining the proper segments of grouping in therapy, and studying patient data can help in identifying infections [6]. Using random forests to increase scoring function performance in binding affinity of ligand–protein interaction [6]. The key challenges in the QSAR model are the depiction of exact drug models and compound structures [6]. Once descriptors are selected, the optimum mathematical model for accurate fitting in the structure–activity connection must be established. RF method was employed to increase fitting criteria in a mathematical model (Fig. 1.3b). The molecular descriptors selection is regarded as a key stage in the virtual screening method for identifying bioactive chemicals throughout the drug discovery and manufacturing process. Because the descriptors set produces less accurate predictions, the random forest approach was employed in the pharmaceutical industry to enhance prediction and then pick molecular descriptors (trained) for enzymes, hormone receptors, kinase ligands, etc. [6].

### *1.3.6 Multilayer Perceptron*

A feed-forward neural network is another name for the multilayer perception model (MLP), which generates a result depending on a set of input sources. The backpropagation method is used to train any type of information. Because nodes (input and/or output) are coupled with weights, this model is analogous to a directed graph [6]. Following data processing, the perceptron can vary each linked weight in the network. The existence of inaccuracy in the actual output may, therefore, be compared to the intended result. Figure 1.3c illustrates how to use specialized algorithms to separate a few characteristics in input data. It learns ideal weights as a result, and then input characteristics are augmented with available weights to determine if a given neuron was terminated or not. Multilayer perceptron employs a backpropagation approach with the activation function in this manner [6]. A multilayer perceptron was used in this context to anticipate the action of the drugs. This model has one benefit, in that, it does not need knowledge about the structure of drugs because it predicts accuracy using experimental data [9]. In addition, MLP was used to create a novel drug design. This model can independently generate various compounds with sophisticated features [6]. In general, MLP may be utilized quickly and simply, but performing its obligations in training is tough, and MLP does not provide any assurance of global minimum performance [6]. Protein secondary structure provides a significant edge in identifying protein function and drug design. The MLP technique was more interested in classifying success in that research. However, identifying the protein secondary structure in the experimental research is more difficult and costly. Thus, the MLP training data findings were presented as a success when compared to the experimental categorization [6].

### *1.3.7 Deep Learning*

Deep learning (DL) is a sub-discipline of ML, which can extract a higher degree of characteristics from input data by utilizing numerous layers [6]. DL is a major field that is now producing enormous premium solutions. DL algorithms are recently been employed in numerous academic domains and have resulted in increased profitability in business market. In general, deep learning is the same neural network design that has numerous layers and may alter data between them. It remains a popular term, but the creativity behind it is genuine and developed systematically. Deep learning models may therefore be constructed using an approach known as greedy layer-by-layer [6]. Figure 1.3d compares sophisticated DL algorithms with pooling layers to identify significant concerns and develop the best solution even when the challenge is complicated. DL algorithms provide several models in drug discovery fields, such as DNN, CNN, RNN, and Autoencoder. Another component that impedes neural networks is the pooling layer. The pooling layer can lower the representation's spatial size to decrease boundary dimension and minimize system calculations and work independently on every attribute map (channel). The reason why max-pooling layers perform so well in diverse networks is that they enable the system to detect features extremely efficiently after down-testing an input structure and reduce overfitting.

The DNN design emerged from an extension of the ANN and includes many layers between input and output nodes [6]. The DNN architecture tracks the results in a mathematical model, which can be either non-linear or linear. Each mathematical model is assumed to be a layer in this case, and numerous layers are accessible in sophisticated DNN, hence the network is referred to be 'deep'. DL models are used in QSAR modeling to automatically obtain feature extractions and capabilities in chemical properties. ML approaches such as artificial neural networks have been used to predict peptide intestinal permeability. Positive controlled data acquired via the peroral phage approach comprise the intestinal permeability of peptides, and random sequence data may be prepared using negative controlled data. To provide acceptable predictions, several statistical measures such as specificity, sensitivity, ROC score, enrichment curves, and so on are validated. And the statistical findings revealed that the models are of high quality can segregate between random sequences and are permeable with high certainty. Finally, the ANN models outperformed the unexpected ones in terms of prediction. As a result, this approach may be useful for selecting intestine permeable peptides to produce peptidomimetics [6].

Multitask neural networks are incorporated into the 'DeepChem' platform which aids multi-task neural networks in drug production [6]. In addition, networks have a fine role in multi-task deep networks to be resilient. Finally, deep learning algorithm performance in QSAR models improved prediction performance. DNN also plays an important function in the hit-to-hit lead optimization study. CNN (a subclass of DNN) is commonly used to analyze visual pictures [10]. CNN is also known as shift-invariant ANN since it typically relies on the weights. It is a multi-layer perceptron that has been regularized. Multilayer perceptrons are fully linked networks in which each neuron (in the first layer) is correlated with the next layer. A network can overcome the overfitting problem by employing a fully connected method. The CNN algorithm evaluates the clinical field in such a way that each neuron in a human cell resembles the visual cortex [6]. Many researchers used the CNN model to predict protein–ligand affinity in ligand–protein interactions [6]. The best correlation in the dataset was predicted by affinity prediction [6]. The CNN algorithm predicted binding affinities in protein–ligand interactions, which can improve scoring function, but predictive capacities must be upgraded concurrently. RNN algorithm is a type of ANN in which connections can form among nodes (input and output). In this manner, a directed graph and a timed sequence may be generated in the network. Similarly, the RNN network uses internal memory to conduct input variable grouping [6].

The DeepDTnet is one among 15 chemical, phenotypic, and genetic variants and cellular profiles included in deep learning algorithms used to expedite drug repurposing and target discovery. DeepDTnet has been certified by the USFDA for identifying new targets for known drugs because of its excellent accuracy. Topotecan was licensed as an inhibitor for retinoic acid receptors in humans based on experimental data, reducing the transitory vacuum in drug development [6]. RNN was employed in screening candidates virtually to generate novel molecular libraries, which aided in the discovery of anti-cancer drugs via molecular fingerprints [6]. To create a de novo drug design, biological performance must be predicted. In this manner, the RNN algorithm was used to generate molecules [6]. Molecules might be found in the ChEMBL dataset. For sampling, produced molecules must be conditionally trained using the RNN algorithm. 'Deep Interact' was a domain-based integrative technique used to predict PPIs using a Deep Neural Network. Multiple PPIs are distributed from (KUPS) Kansas University Proteomics Service and Database of Interacting Proteins (DIP). Discovering and analyzing biological components in the specificity of connections and explicit molecular protein complexes is extremely important. The major objective is to perform large-scale high-throughput trials using in silico technique to enhance PPI uncovering levels. From the *Saccharomyces cerevisiae* dataset, 34,100 PPIs were validated, yielding good findings with a sensitivity of 86.85%, accuracy of 98.31%, specificity of 98.51%, and correctness of 92.67%. As a result, the Deep Interact technique was found to outperform previous ML algorithms in PPI prediction [6].

Autoencoders are a subclass of ANN that recovers data through unsupervised learning [6]. The autoencoder's goal is to represent the encoding data format in dimensionality reduction to keep a strategic distance from the network's 'noise' signal. In addition, the autoencoder must investigate incoming data before copying it to the output layer. Autoencoder is divided into two sections: Encoder and Decoder, as well as one hidden layer. The hidden layer is regarded as code in this context. The encoder sends input data to the buried layer. The decoder can extract data to reproduce the signal output. Autoencoders were the best choice for dimensionality reduction and learning data from generative models [6]. Autoencoders were primarily taught to reduce reconstruction mistakes (loss). Autoencoders are used in drug research as a novel design to distribute chemicals by conducting tests directly into vermin [11]. Deep learning models, such as autoencoders, were used to generate compounds in the design of de novo drugs. As a result, the autoencoder technique was used in conjunction with other classifiers such as multilayer perceptron to generate automatically novel compounds with relevant attributes [6].

## 1.4 ML Applications in Drug Production

MLTs can be involved in the discovery and manufacturing of drugs from the laboratory research to the industrial scale because they can aid in rational drug design, decision-making, determining the appropriate treatment for a patient, which include personalized medicines, and managing medical information produced and its use in future drug discovery and manufacturing [4]. Eularis' E-VAI is a decision-making and analytical AI-based platform that uses ML algorithms and a user-friendly interface for building systematic roadmaps depending on key stakeholders, current market share, and competitors to anticipate key drivers in pharmaceutical sales, assisting marketing executives in allocating resources for maximum market share gain, reversing poor sales, and anticipating where to invest [4]. Figure 1.4 summarizes various applications of ML in drug discovery and manufacturing [4]. The ML-based drug discovery is further subdivided into task performance of ML and their applications such as target identification, hit discovery, hit-to-lead, and lead optimization strategies. The drug design strategies rely on databases, which are created using various ML algorithms. In the drug development stage, accurate training, validation, and deployment of ML algorithms deliver an enthusiastic output by lowering the intricate error-prone methods. Most drug design procedures use machine learning approaches to minimize time and manual intervention. The finest example is QSAR, in which massive data gathering and dataset training are seen as rate-limiting phases in creating ligand-dependent virtual screening techniques and have been substituted by de novo design approaches.

The interaction of side-chain amino acid residues in the homology modeling/prediction of protein folding of secondary structures like $\beta$-sheets and α-helices is particularly important in maintaining the smooth functioning of three-dimensional proteins. NMR spectroscopy, X-ray crystallography, and electron microscopy (cryogenic) are employed to acquire an accurate protein folding as well as its primary active ligand site (Cryo-EM). The UniProt server stores information regarding the main amino acid sequences of proteins/enzymes/receptors, both soluble and insoluble, as well as their targets and biological activities. The major role of proteins is recognized based on medicinal chemistry, pharmacological and biochemical investigations, and this knowledge is also the core unit for building protein folding prediction studies using software or experimental studies. In contrast,

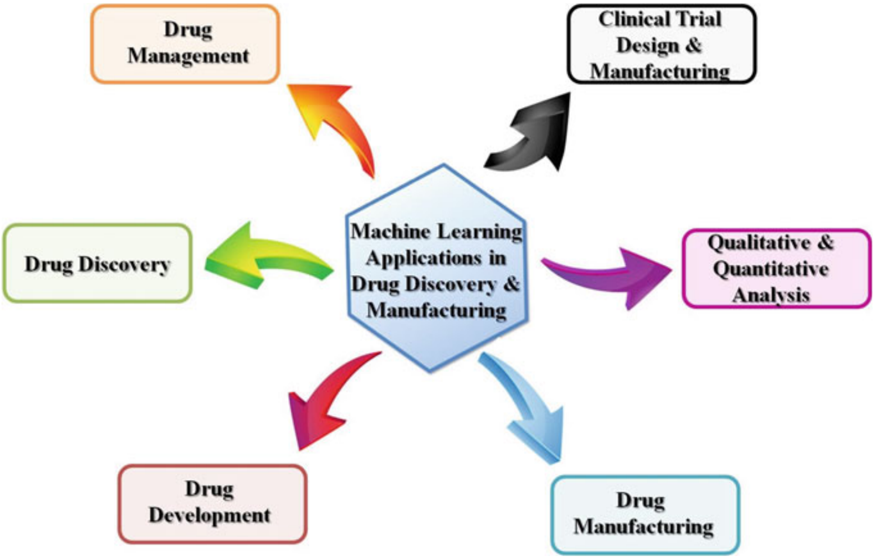

Drug Management
Clinical Trial Design & Manufacturing
Drug Discovery
Machine Learning Applications in Drug Discovery & Manufacturing
Qualitative & Quantitative Analysis
Drug Development
Drug Manufacturing

**Fig. 1.4** ML applications from drug discovery to drug management [4]

the findings of protein folding based on amino acid sequence (UniProt) were evaluated from its empirically determined PDB homologs, which became a promising way to computationally filter the novel protein models and this process is known as “homology modeling”. The homology modeling or comparative modeling is assessed by numerous methods that must be executed via software modules (PRIME) or web-based servers such as SWISS-MODEL and EXPASY to predict the folding of secondary protein structure with good precision among specified templates. However, Ramachandran analysis is used to fine-tune the derived homology models or template-based models, which may be handled by commercial modules (PRIME) or online servers (QMEAN, PROCHEK). Furthermore, picking the best homologous model derived from the preceding procedure is a significant effort that may be accomplished using SVQMA (Support-vector-machine Protein single-model Quality Assessment) servers, ProQ3 or ERRAT, which are powered by Deep learning algorithms. Following the completion of the preceding procedures, the best 3D protein template may be employed in any basic drug chemistry investigation to find hits as part of a structure-based virtual screening strategy.

Because of a lack of understanding about their off-targets, such as enzymes, ion channels, proteins, or receptors, identifying targets for novel chemical entities (NCEs) is a difficult undertaking. Another important challenge for computational/bioinformatics investigations where more than one active site exists in the protein is NCE binding site identification. In the aforementioned instances, the predefined web-based servers (FTMap), as well as specific modules such as “Sitemap” designed with the assistance of algorithms, can determine the preferential binding site to accelerate the drug discovery process. POCASA, GHECOM, SURFNET, Pocketome, LIGSITE, ConCavity, PASS, Fpocket, and Q-SiteFinder are a few more online tools that estimate the possible binding sites within the specified protein templates. The metaPocket 2.0 application makes use of the aforementioned platforms to provide the most dependable ligand binding locations existing on templates. Furthermore, ML models like FD/DCA can anticipate molecular docking locations of drugs in biological macromolecules. DeepDTnet has recently been explored as novel target identification in drug repurposing. DeepDTnet is created by combining multi-disease cellular targets, pathogenic genes (genomics), and drugs (chemical spaces) used in their therapy.

The hit finding method has evolved in terms of success, with the progress in drug discovery. Small compounds are evaluated as hits for target binding in this process to discover the best-altered functionalities. In the existing drug discovery and manufacturing paradigm, hit detection using various algorithms is now dominating as a robust approach. In one such technique, the use of multivariate parameters (k-nearest neighbors (kNN) and support vector machine (SVM) on high content screening (HCS) analysis provided a wide range of positives against neurological problems. DeepDTnet’s training parameters exceed other existing target identification approaches and require few FDA-approved drugs to generate the positive therapeutic effects of existing enzyme inhibitors. The deepDTnet method also transfers multiple FDA drugs against GPCR with novel focused pharmacological effects. DeepDTnet is thought to be far superior to KBMF2K and NetLapRLS techniques, as well as SVM, Random Forest Nave Bayes, and kNN algorithms. “Repurpose” means to “reprocess/reuse/recycle”. Drug repurposing is defined as “finding new indications for drugs that are currently in the development stage”. It decreases the amount of time and risk involved in drug discovery [6]. An important rationale for using the drug repurposing idea in drug development is that it is extremely beneficial to have many targets in one molecule that correlate to different effects [6]. As a result, it presents a wide range of drug-disease connections. Other components used in drug repurposing include drug targets and disease related genes.

ML methods are used in drug discovery and manufacturing to locate tiny compounds with distinct bind structures for a therapeutic target. In drug research, virtual screening uses software and algorithms to distinguish hits from proprietary chemical libraries and to get unique hits efficiently. Following the identification of new hits, a subsequent step is required to purify molecules with unfavorable scaffolds (framework) [6]. Furthermore, it comprises just a few tactics such as docking-based, similarity searching, pharmacore-based, and ML approaches [6]. Based on the methodologies described above, classification has adopted two approaches: structure-based and ligand-based virtual screening. When 3D protein structures were available, the molecular docking procedure could be widely used [6]. Many applications linked to docking-dependent virtual drug screening have been successfully developed with no negative consequences [6]. This technique may have certain hurdles like the scoring function, which cannot accurately assess binding affinities (bond/relationship) because inadequate configurations and entropy effects have reduced protein flexibility, making it more difficult [6]. Finally, several docking models took binding affinities into account, although refusals such as docking score and distance-time remained [6]. In comparison to docking-based virtual screening, ligand-based virtual screening is unable to conform to the 3D protein structure. Its purpose is to create bioactivity domains based on molecular characteristics [6]. The goal of this technique is to consistently enhance yields while decreasing false hit rates [6]. The SVM approach was extensively used in virtual drug screening to achieve this goal [6]. DL techniques were used to obtain high classification capacity, low generalization error, and strong feature extraction ability [6]. Following target selection, routine procedures include virtual screening (HTVS) and molecular docking techniques incorporated in free energy perturbations, sampling, and scoring algorithms. Knowledge of the enzyme’s active site for the protein/receptor where the ligand would bind to mimic/antagonize the physiological role is required to start the HTVS process. Similarly, another fundamental approach, ligand-based virtual screening (LBVS), is based on the physicochemical features of chemical databases.

Mapping ability characteristics in machine learning may offer considerable success in extracting physical, geometric, and chemical aspects to obtain scores [6]. Based on scores, data-driven black box models are suggested to predict interactions in binding affinities while disregarding a few aspects of docking such as physical function [6]. Random Forest and SVM ideas were found using ML for improved scoring function performance. For example, instead of a linear additive technique connected to the energy terms notion, an SVM model can be used. Since an SVM can define the link among experimental binding affinities of drug candidates and their predicted energies, the eHiTS docking program may be derived. As a result, data provides improved execution in terms of scoring and screening power [6]. Many researchers began to employ the CNN model in image processing because it displayed higher performance and protein–ligand interactions by giving multiple characteristics to CNN for predicting protein–ligand affinities [6]. Jimenez et al. focused on the 3D visual representation of the CNN model and binding affinities in the estimate of protein–ligand affinities, which suggested better correlation behavior in data sets [12]. And, fundamentally, deep learning indicates its actual intensity to grow abstract characteristics from basic features, since key features for a compound-protein structure, such as molecule kinds, particle separation, and so on, must be represented. A framework Deep VS (based on CNN), deals with learned abstract features from basic characteristics to create docking programs like ICM and GLIDE SP [6].

In the early stages of drug research, hit-to-lead is known as lead generation. It finds interesting lead compounds by deficiently optimizing tiny molecules known as hits from the High Throughput Screen (HTS). The practical interface of a hit-to-lead optimization strategy combined with chemical synthesis, as well as the mapping algorithm “design layer” Random Forest regression, were used to develop new physiologically active chemical spaces by utilizing an existing kinase inhibitor library [6]. QSAR analysis was utilized in the hit-to-lead optimization process to uncover prospective lead compounds from hit analogs with bioactivity analog prediction [6]. And it is generally used in mathematical ideas to investigate quantitative mapping with physicochemical or structural objects, as well as biological activities. QSAR analysis is divided into four parts: the basis of mathematical models, the selection and advancement of molecular descriptions, evaluation and interpretation methods, and usage procedures [6]. In this case, mathematical models and chemical structural representations are regarded as problems in QSAR demonstration. After descriptors are selected, mathematical models must be located to fit connections in the structure–activity technique. As another component, it use characteristics such as physicochemical descriptors and linear regression models to elucidate the 2D structure–activity relationship [6]. ML algorithms like SVM and Random Forest are utilized in mathematical models [6]. Similarly, QSAR modeling used deep learning approaches to recover capabilities in chemical strings (ASCII) and extract the characteristics automatically. In 2012, Merck Molecular Activity Challenge was held, and a team led by George Dahl won using ensemble approaches such as multi-task DNN, gradient boosting machine, and Gaussian progress regression [6].

Because of the multi-task technique, neural networks acquire characteristics from diverse parameters, even if the tasks are comparable [6]. In drug development, multi-task neural structures were utilized to analyze performance, and the RF algorithm produced outstanding results [6]. Since multi-task neural architectures converged on a platform known as DeepChem, researchers used canvas descriptors to implement DNN. To obtain results in human-secretase-1 inhibitors, binding affinity prediction must be combined with a classification and regression model [6]. The use of a DNN model yields excellent results in the validation set, with a classification accuracy of

0.82, regression ability R2 of 0.74, and MAE (Mean Absolute Error) of 0.52. DNN models use 2D descriptors and produce superior outcomes than force-field-based techniques due to the usage of partial capability models in deep learning. Finally, QSAR models rely on deep learning approaches, which yield superior outcomes in the future prediction role of hit-to-lead optimization research. By changing or balancing the goal interest, De novo Drug Architecture advanced discovery of new chemical structures [6]. To create a novel molecule from scratch using the fragment-based technique, a common De novo model is used. If there are impracticalities and complexity in the molecular structure at this time, issue emerges in building the structure and the assessment of bioactivity becomes difficult. Deep learning models used their extensive knowledge and generative skills to create a new structure with relevant attributes [6]. Deep learning models serve as autoencoders in the de novo drug design process, generating an acceptable format for novel chemical entities (NCEs). As a result, an autoencoder embedding with a multilayer perceptron classifier is a value-added approach to the production of NCEs with preset physicochemical attributes. The syntax of the drug/chemical structure is generated in SMILES format, which can be difficult to interpret in many situations, and grammar variational autoencoder (VAE) solves this issue to accelerate the process. By changing the RNN model, the deep reinforcement learning approach was extended for predicting biological activities to build novel compounds [6]. To acquire SMILES syntax, an RNN model must be trained; molecules can be obtained via chemBL. To obtain a high reward for activity scoring, the SVM algorithm is used to improve a few ways based on ligands' idea in the training set. De novo was a sequence-based negative examination framework that learned the several viruses in PPI to forecast the novel one, where the shared host proteins may exploit.

De novo has attempted to test the PPIs with diverse domains to determine generalization. The de novo technique achieved 81% accuracy in minimizing noisy negative associations and 86% accuracy in viral protein prediction during the training phase. In intra-species and single virus–host prediction scenarios, the de novo method performed better. As a result, predicting the PPI for a contaminated individual becomes challenging, and the best accuracy is reached when testing for human–bacteria interactions [6]. The use of a multi-objective optimization process and ML to generate biological and chemical possibilities has yielded promising results by entrusting an automated de novo compound structure as a human-creative process. The RNN algorithm is employed to computerize exclusive molecules using a de novo structure built on common features identified among consistent physical and chemical properties for principal trade-offs in the work, which depends on multi-objective technology. Several chemical libraries are connected to de novo structure targeting neuraminidase and acetylcholinesterase, according to this viewpoint. Numerous quality measures were used to examine chemical feasibility, validity, drug similarity, and diversity content. Molecular docking has occurred in the de novo generating molecules for the evaluation of posing and scoring using X-ray cognate ligands with comparable molecular counterparts. Also, multi-objective optimization and ML are made available for use in readily adaptable design methodologies that are particularly useful for lead progression and generation [13].

The lead optimization process is an important phase in drug discovery and manufacturing in which the best medicinally active fragment hits are deemed leads to further medicinal chemistry projects. The primary goal of lead optimization is to minimize the side effects/unwanted effects of current active analogs with minimum structural alteration, resulting in a better and safer scaffold. One such example is the optimization of Autotoxin inhibitors, such as the therapeutic drug GLPG1690, which is now in human clinical trials to treat pulmonary fibrosis. Another example is improving active analog potency using tailor-made procedures. The different aspects of ADME/T are covered here, including chemical and physical properties, absorption, distribution, metabolism and excretion, toxicity, and the ADME/T multitask neural networks. Physical and chemical features have been used to prevent substantial failures in drug discovery and manufacturing pipelines. DL models are then used to lead optimization strategies to develop distinct procedures [6]. When a projected value was included as a parameter in the fundamental drug discovery paradigm, the ML algorithms emerged as the major scoring function in computational assessment. The site of metabolism may be predicted by numerous modules, including the "ADMET Predictor" of SimulationsPlus tools, which is based only on models built by artificial intelligence algorithms. The FAME3 (online server) predicts the area for a specific drug/compound that travels through metabolism-verified databases accumulating phase-1/phase-2 metabolic characteristics related with multiple databases authorized via evaluating by Matthews correlation coefficient (MCC) [14].

Numerous DL techniques are necessary for image analysis to accomplish specified tasks; hence, the integration of image analysis and DL algorithms may be accommodated for problem-solving. Although the employment of DL approaches can surpass the results in a variety of situations, it was not an image analysis tool due to a lack of flexibility. Another obstacle to digital pathology is the question of openness. In deep learning methods, black box is a well-known methodology. Decision-making is hazy in categorization tasks. Interpretable outcomes can be useful in discovering possible biomarkers and therapeutic targets for predicted responses in therapy for understanding various pathways in drug development. Furthermore, trust in the generation of combined features with interpretability should be increased. A major problem in clinical trials is the huge sample size necessary to use DL approaches legitimately for predicted response in treatment. In clinical studies, the DL needs a large number of sample instances. Integrating medical test data is sometimes possible, but the presence of bias might make the results difficult to understand. Several DL and image analysis models that can predict responses in therapy were applied, and the CNN model was used to detect characteristics in the succeeding chart and lymphocytes located in E&H-stained cells. DL will eventually include additional abilities for replacing nuclear identification and typical segmentation methods for delivering spatial context characteristics.

In addition, it is critical to understand the chemical modifications of drugs/NCEs that have undergone metabolism and may therefore be utilized to calculate dose regimen, dosage frequency, toxicity, and other favorable side effects. GLORY/GLORYx online services give exact information about the possibilities of novel metabolites and their associated production data concerning mitochondrial cytochromeP450 enzyme and conjugations [6]. Prediction of skin sensitivity is an important factor for evaluating the safety characteristics of novel drugs/compounds, and it varies from patient to patient. In this regard, ML models such as Random Forest-based MACCS (RF MACCS) and support vector machine (SVM)-based PaDEL (SVM PaDEL) algorithms were trained using roughly 1400 ligands connected to local lymph node assay (LLNA) data [14]. The ML trained on libraries of 265,000 product isolates (natural) and synthetic compounds, confirmed by MCC, is being utilized as a fundamental prediction model in the NP Scout web server, which will identify the uniqueness of the novel analogs. The use of NP Scout to anticipate resources for query drugs may offer data about their product resources (natural) and may become a part of natural product-based drug development [6].

Many specialists and clinicians faced problems in drug research and discovery, such as computational pathology data, target validation, and the identification of predictive biomarkers in clinical trials. Drugs for modifying the infection state can be created by controlling the target activity of drug molecules using the most advanced drug discovery and manufacturing approaches. Target identification necessitates the formulation of a therapeutic hypothesis for the regulation of drug targets in the aftermath of an infectious state. Target identification occurs when available evidence for that target is recognized. In vivo and ex vivo models are employed to authorize the target illness based on basic decisions. Although clinical preliminary results can be obtained for target validation, it is vital to focus on target validation efforts for successful projects. The disorders include metabolomic, transcriptomic, and proteomic profiles from in-patient clinical data. The clinical database's capacity to repurpose data through public sources enables basic target identification and target validation. Appropriate methodologies for producing genuine statistical models are required for predicting target identification. Because of the increase in data-driven target identification trials, ML methods are being applied to target identification. Recognizing the causal connection between disease and target is the first stage in target identification, i.e., by naturally or artificially altering the target disease (experimental). ML techniques may be used to make predictions based on known attributes of targets, causalities, and driven targets. In the realm of target identification, ML approaches may be applied from a variety of angles. SVM exposed feature selection and linear kernels are often better than supervised machine learning models for detecting biomarkers. ML may be used to construct predicted targets, i.e., blind drugs are utilized for therapeutic assumptions.

ML is the best extension for understanding biological factors for discovering therapeutic therapy through novel targets. The splicing signal model is one method used to treat Alzheimer's disease. To anticipate alternative signals, the DL splicing signal model is used [6]. Binding integrative splicing signals, such as RNA sequencing data and CLIP-seq splicing data, revealed knock-down findings [6]. To anticipate Alzheimer's disease variations, coding models such as complicated variants and de novo designs must be included [6]. ML can forecast the effects of cancer-related drugs [6]. As a result, ML looked at how DNA-methylation, somatic mutation data, and genome-wide data affect drug feedback. To predict drug response, scientists use logical models, ANOVA, and machine learning models such as random forests to find molecular properties. SVM's DL model was used to predict drug targets using various physicochemical attributes from protein sequences [6]. Proteins occupy specified sites in the PPI network to form unusual associations [6]. To reduce search space, ML systems used newly constructed targets to forecast blind drugs, however, drug targets require additional endorsements. Predicting the outcome of clinical trials in therapeutic targets is a difficult task for target validation and identification. Using ML techniques, omics data was used to choose 332 therapeutic targets, allowing it to fail or succeed in the third phase of clinical trials via multi-variate compound selection [6]. In clinical studies, gene expression data has been recognized as a good prediction across tissue layers with large variation and lower RNA mean expression. This validated the therapeutic target unique disease expression that can alter tissue area [6]. ML classifiers should be trained on open platforms to predict de novo therapeutic drug targets [6]. Important indicators include crucial data categories such as genetic data and gene expression for determining therapeutic drug targets. In such circumstances, ML techniques hampered by a lack of complete data and availability of insufficient data are primary reasons for the failure of drug discovery and manufacturing. Practically, it takes duration into account because of greater technological progress, novel models such as biologics (with added antibodies) can be accessible, and the design of tiny drug molecules cannot be the same as today.

Biomarker discovery is utilized to advance the success of clinical trials by distinguishing drugs and identifying pharmacological mechanisms for sensitive patients using the ML technique [6]. The last rounds of clinical studies occupy a considerable quantity of time and money. To overcome this problem, predictive models must be used, built, and validated in the early stages of clinical trials. The use of ML algorithms allows for the prediction of translational biomarkers in preclinical data collection. Following data validation, associated biomarkers and models must analyze patient indications and, finally, recommend medicine. Several studies in the literature gave information about prediction models and biomarkers, and just a handful was used in clinical trials. In a clinical environment, several elements such as model rebuilding, designing, data accessing, data quality, and model selection are required. The main issue was that ML techniques evaluate community efforts for constructing regression and classification models. Many years ago, at the last step of clinical trials, the FDA (Food and Drug Administration) led MicroArray Quality Control (MAQC II) have examined the ML algorithms for predicting gene expression data [6]. Predictive ML models have been presented in several studies in which biomarkers play a significant function in drug discovery and manufacturing. Alternative other oncology data types, predictive biomarkers showed improvement in ML. Multi-omics data should be used by ML algorithms to improve treatment responses in patients [6]. In addition, the gradient regression tree is used to enhance polygenic risk scores in clinical trial prediction [6]. Using ML techniques, feature selection made significant advances in biomarker identification. Table 1.2 summarizes the association of major pharmaceutical industries with ML organizations and their mutual projects for drug discovery and manufacturing [4].

Pathology is a realistic field, and each pathologist defined what can be seen via a glass slide by visual examination. Glass slides generate a lot of information, such

**Table 1.2** List of pharmaceutical industries and their collaboration with ML organizations [4]

| Pharmaceutical industries | ML organizations | Mutual projects |
|---|---|---|
| AstraZeneca | BenevolentAI | ML systems based on neural networks for the discovery and manufacturing of novel therapeutics for chronic renal disease and idiopathic pulmonary fibrosis |
| BAYER | Sensyne health, exscientia | Clinical development of novel cardiovascular disease medicines employing a proprietary clinical ML technology platform. Centaur Chemist™ ML drug discovery technology is being used to optimize new lead structures for prospective therapeutic candidates for cardiovascular and oncological illnesses |
| Boehringer Ingelheim | Healx | Healnet, an ML-based platform, is used to discover treatments for uncommon neurological illnesses |
| Eli Lilly and company | Atomwise | Discovery and manufacturing of drugs with novel targets |
| Janssen (Johnson & Johnson) | BenevolentAI | To discover, manufacture, and market innovative clinical-stage medication candidates. To obtain clinical data from Bavisant Phase IIb studies in Parkinson's disease patients with excessive daytime drowsiness |
| MATEON Therapeutics | PointR™ DATA | Later-stage melanoma, gliomas, and pancreatic cancer treatment |
| NOVARTIS | IBM Watson, Microsoft | To improve the health outcomes of breast cancer patients. Cell and gene treatments, generative chemistry, image segmentation and analysis for smart and tailored therapy delivery, and large-scale cell and gene therapy optimization |

(continued)

as which cell type is grouped in which tissue layer and in what geographic context. In this regard, it is often necessary to investigate the interactions between immune cells and immune-oncology tumors. Before selecting a patient to test with hundreds of chemicals in clinical trials, pharmaceutical companies must understand how the specific medicine may treat patient cells and tissues in the body. Because of the rapid improvements in clinical trials, discovering biomarkers has become increasingly important for victims, i.e., those who are ready to respond to therapy. Rapid advancements in digital pathology can lead to the finding of new biomarkers with

**Table 1.2** (continued)

| Pharmaceutical industries | ML organizations | Mutual projects |
|---|---|---|
| Pfizer | XtalPi, IBM Watson | The use of quantum mechanics and machine learning techniques in conjunction with cloud computing architecture to predict the 3D structure of molecules, including mechanical and chemical characteristics as well as binding to protein receptors. In immune-oncology, ML, natural language processing, and reasoning technologies are being used to discover new therapeutic targets, combination therapy, and patient selection strategies |
| Roche | OWKIN | Clinical trials based on ML networks for drug discovery and manufacturing |
| SANOFI | Exscientia | To find and develop bispecific small molecules for the treatment of diabetes and its complications |
| SERVER | Numerate | Using a unique algorithm-driven drug discovery platform, researchers created small-molecule modulators of the ryanodine receptor 2 (RyR2), a target discovered in cardiovascular illness |
| Takeda | Numerate | To find drugs for cancer, gastrointestinal, and central nervous system illnesses |

more reasonable, precise, and high-throughput behavior, saving drug development time and allowing patients to get treatment more quickly. Before using DL models, several image analysis methods have been used by pathologists.

Numerous computer scientists are required to manufacture graphical characteristics in computers to categorize tissue layers. The goal of digital pathology research is to identify etymological descriptors often used in eosin and hematoxylin (E&H) structures. To predict the treatment response, ML models and feature selection must be used. Because only one biopsy can be sufficient to train feasible pixels, the CNN model is highly suited for digital pathology work. As a result, DL models are trained using controlled features from diverse categorization tasks automatically [6]. M-CNN (Multi-scale CNN) was used in this model as a supervised learning strategy for phenotyping pictures with high content cells, where it limits a few models with their unique steps [6]. When picture pixel values were converted to phenotypic images, the M-CNN technique displayed greater accuracy at categorization levels. Several DL approaches are used in tubules, lymphocytes, mitotic activity, and cancer tumors seen in lung and breast malignancies to create objectives in image analysis [6]. DL models in digital pathology give information linked to other approaches. DL models are used to increase data capture in MRI (Magnetic Resonance Imaging) or to reduce radiation exposure in CT (Computed Tomography) image processing [6]. Because picture quality has grown significantly in terms of noise signal ratio and spatial resolution, applications such as victim stratification, illness prediction, and image certification have improved as well. Another study determines the use of altered genes named lung tumors from eosin and hematoxylin (E&H) stained pictures using the DL framework [15].

## 1.5 Challenges and Risks

There are several problems in drug development, but the majority of them can be overcomed by employing MLTs. Some of the issues are presented below, along with proposed solutions. Several ML techniques gave accurate results, although few parameters and structures caused problems throughout the training phase. The specific method cannot meet the accuracy and local optimum, especially when data is scarce during the training phase. To address this issue, a deep belief architecture, which is an unsupervised pre-trained model, must be applied to improve parameters and produce more effective outcomes [16]. Another obstacle in drug discovery is the issue of openness. Because decision-making is ambiguous in several categorization schemes, many mechanisms must be understood in drug development to interpret the results. As a result, it is more helpful in the discovery of new drug targets, and numerous assembled characteristics must boost faith in interpretability [5]. Numerous processes, such as SVM, MLR, RF, and DL approaches, can be used in drug development to understand and interpret the results. Hence, it is more helpful in discovering novel drug targets, and it has various integrated properties for increasing faith in interpretability.

Many references, particularly in the ‘omics’ domain, can provide access to integrated data. It’s becoming more difficult by the day, because not only is the data increasing, but this data type contains a great deal of variability among pharmaceutical industries [6]. Although public databases such as ZINC, BindingDB, PubChem, Drugbank, and REAL chemical databases are accessible, developers must construct pipeline architecture to combine these disparate data sources. Adaptable Clinical Trial Database, Integrated Genomic Database, SWISS-PROT, DataFoundry, dbEST, and SCoP are data warehousing tools that use ETL (Extract, Transform, and Load). Genome Information Management System, BIOMOLQUEST, PDB, SWISS-PORT, ENZIME, and CATH data [6]. Furthermore, homogeneous data might cause integration concerns, beginning with testing and logical issues, cross-platform normalization, and statistical issues, which can result in massive heterogeneity of information [6]. As a result, ML combined with big data analytics may be used to integrate heterogeneous data sources. Ontology-based integration technologies such as Ontology Web Language, Extensive Markup Language (XML), RDF Schema or Resource

Description Language (RDF), Unified Medical Language System, and others are available [6]. There are other weblink-based integration solutions available, such as Sequence Retrieval System [6]. Sequence search analysis using ChEMBL, NCBI Entrez, PubChem, Integr8, DisaseCard, and EMBL-EBI [6, 17]. Microsoft Power BI, IBM Cognos, Tableau, Zoho Analytics, Sisense, SAS Business Intelligence, and more visualization tools are also available. Because integration and visualization technologies aid in discovering bottlenecks and possible issues before they disrupt critical operations [6].

In pharmaceutical companies, research was extended from large molecules to individuals and typically relied on the integration of heterogeneous data, which presents its issues in different settings and sizes [6]. A high degree of artificial intelligence is required for handling several sources, and it must be upgraded with a greater grasp of the data acquired. As a result, current data connections are recommended to consolidate disparate data, and these data connectors ultimately aid in the allocation of original data. The availability of big data is critical to the success of ML since this data is utilized for the following training supplied to the system. Access to data from many database providers might entail additional expenditures for a corporation, and the data must be dependable and of good quality to enable accurate result prediction. Other barriers to full-fledged ML adoption in drug manufacturing include a lack of skilled personnel to operate ML-based platforms, a limited budget for small organizations, fear of replacing humans, which could result in job loss, skepticism about the data generated by ML, and at last the mysterious black box phenomenon (i.e., how the conclusions are reached by the ML platform in drug discovery and manufacturing?) [18].

Certain activities in drug discovery, manufacturing, supply chains, clinical trials, and sales will be automated over time, but they all fall under the category of ‘narrow ML’, in which ML must be trained using a huge volume of data and therefore, becomes fit for a specific activity. As a result, human interaction is required for the ML platform’s proper deployment, development, and operation. However, the apprehension of joblessness may be fiction, as ML is now taking away repetitious professions, allowing room for the human intellect to create more complex insights and creativity. Nonetheless, ML has been adopted by several pharmaceutical companies, and it is expected that revenue of more than US $2.2 billion will be generated by 2023 through ML-based solutions in the pharmaceutical sector, with the pharmaceutical industry investing more than US $7.2 billion across >300 agreements in 2013–2018 [4, 19]. Pharmaceutical companies must comprehend the potential of ML technology in solving issues after it has been applied, as well as the acceptable goals that may be met. To realize the full potential of the ML platform, skilled data scientists and software engineers with a solid understanding of ML technology, as well as a comprehensive awareness of the company’s aim and R&D goal, may be produced.

## 1.6 Conclusions and Future Perspectives

In everyday life, ML technology is used in pharmaceutical sectors, including ML algorithms and deep learning approaches. In drug development areas and health care centers, ML approaches have experienced several conflicts, particularly in omics data and image analysis. In clinical research, ML models forecast training data in a recognized framework, i.e., the multi-faceted structure involves the execution of alternative tools such as PPT inhibitors and macrocycles using standard methods. Furthermore, deep learning models may be used to examine compound structures and QSAR outcomes using pharmaceutical data, which was relevant for potential drug candidates with adequate attributes due to the forward success rate in clinical trials. ML technology has advanced by entering computer-aided drug research to obtain powerful data mining skills. Some concerns remained, for example, the effectiveness of DL methods can have a direct impact on data mining innovation because numerous deep neural networks are successfully trained on a vast amount of data. The primary goal is to address the automatic transfer learning problem. In deep learning principles, the "Black Box" model got perplexed. A counterfactual probe is the Local Interpretable Model-Explanations (LIME). The black-box model was unlocked using LIME. In this case, restricted data was required to explain deep learning models. DL algorithms expose data only in the early phases. Many parameters of neural networks are altered throughout the training stage, however, some theoretical and practical frameworks are out of reach for optimizing these models.

The advancement of ML, together with its outstanding tools, promises to continually minimize obstacles faced by pharmaceutical firms, altering the drug discovery and manufacturing as well as the total lifespan of the product, which might explain the surge in start-ups in this sector. The contemporary healthcare industry is confronting various complicated issues, such as rising medicine and therapy costs, and society needs considerable improvements in this area. With the incorporation of ML into pharmaceutical product manufacturing, individualized pharmaceuticals with the appropriate dosage, release characteristics, and other needed elements may be made based on each patient's need. Using the latest ML-based technologies will not only reduce the time required for products to reach the market but will also improve product quality and overall safety of the manufacturing process, as well as provide better utilization of available resources while remaining cost-effective, thereby increasing the importance of automation. The major concern about the adoption of these technologies is the job losses that would result, as well as the tight laws required for ML applications. However, these systems are intended only to make work easier and not to completely replace humans. ML can not only help with rapid and painless hit compound discovery, but it can also help with synthesis pathway suggestions, prediction of the required chemical structure, and comprehension of drug–target interactions and SAR.

Web innovation was combined with clinical research to increase decision-making capability and DL algorithms concerning biomarkers, adverse effects in medicines, and therapeutic benefits. Clinical trial success is accomplished by the use of certain applications. As a result, the incentive for future investment in pharmaceutical firms is performed. In the future, ML technology will encompass all elements of drug discovery and development. For development, automated ML must coordinate theoretical findings like chemistry information, omics data, and clinical data. Furthermore, we anticipate that additional confirmations will be generated for the medicine disclosure effort. ML may significantly contribute to the subsequent inclusion of the discovered medicine in its right dosage form, as well as its optimization, in addition to facilitating swift decision-making, resulting in faster manufacture of higher-quality goods with batch-to-batch consistency. ML may also help to establish the product's safety and efficiency in medical studies, as well as ensure correct market positioning and pricing through detailed market research and forecast. Although there are presently no pharmaceuticals on the market that have been produced using ML-based techniques, and particular hurdles to the adoption of this technology exist, it is projected that ML will become an indispensable tool for drug discovery and manufacturing in the near future.

**Acknowledgements** Author expresses thanks to Abinnovus Consulting Pvt. Ltd., Technology Business Incubator, University of Madras, Guindy Campus, Chennai, India.

## References

1. Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M., Arosio, P.: Machine learning for biologics: opportunities for protein engineering, developability, and formulation. Trends Pharmacol. Sci. **42**(3), 151–165 (2021)
2. Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L., Pei, J.: Transfer learning for drug discovery. J. Med. Chem. **63**(16), 8683–8694 (2020)
3. Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., Cao, R.: Survey of machine learning techniques in drug discovery. Curr. Drug Metab. **20**(3), 185–193 (2019)
4. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., Tekade, R.K.: Artificial intelligence in drug discovery and development. Drug Discov. Today **26**(1), 80 (2021)
5. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S.: Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. **18**(6), 463–477 (2019)
6. Dara, S., Dhamercherla, S., Jadav, S.S., Babu, C.H., Ahsan, M.J.: Machine learning in drug discovery: a review. Artif. Intell. Rev. **55**, 1947–1999 (2022)
7. Elbadawi, M., Gaisford, S., Basit, A.W.: Advanced machine-learning techniques in drug discovery. Drug Discov. Today **26**(3), 769–777 (2021)
8. Syed, K., Sleeman IV, W.C., Nalluri, J.J., Kapoor, R., Hagan, M., Palta, J., Ghosh, P.: Artificial intelligence methods in computer-aided diagnostic tools and decision support analytics for clinical informatics. In: Artificial Intelligence in Precision Health, pp. 31–59. Academic (2020)
9. Stokes, A., Hum, W., Zaslavsky, J.: A minimal-input multilayer perceptron for predicting drug-drug interactions without knowledge of drug structure. STEM Fellowsh. J. **6**(1), 19–23 (2021)
10. Valueva, M.V., Nagornov, N.N., Lyakhov, P.A., Valuev, G.V., Chervyakov, N.I.: Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Math. Comput. Simul. **177**, 232–243 (2020)

11. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., Volkov, Y.: Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. **37**(9), 1038–1040 (2019)
12. Jiménez, J., Skalic, M., Martinez-Rosell, G., De Fabritiis, G.: K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J. Chem. Inf. Model. **58**(2), 287–296 (2018)
13. Domenico, A., Nicola, G., Daniela, T., Fulvio, C., Nicola, A., Orazio, N.: De novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization. J. Chem. Inf. Model. **60**(10), 4582–4593 (2020)
14. Stork, C., Embruch, G., Šícho, M., de Bruyn Kops, C., Chen, Y., Svozil, D., Kirchmair, J.: NERDD: a web portal providing access to in silico tools for drug discovery. Bioinf. **36**(4), 1291–1292 (2020)
15. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. **24**(10), 1559–1567 (2018)
16. Ghasemi, F., Mehridehnavi, A., Fassihi, A., Pérez-Sánchez, H.: Deep neural network in QSAR studies using deep belief network. Appl. Soft Comput. **62**, 251–258 (2018)
17. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D., Lopez, R.: The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. **47**(W1), W636–W641 (2019)
18. Lamberti, M.J., Wilkinson, M., Donzanti, B.A., Wohlhieter, G.E., Parikh, S., Wilkins, R.G., Getz, K.: A study on the application and use of artificial intelligence to support drug development. Clin. Ther. **41**(8), 1414–1426 (2019)
19. Davenport, T.H., Ronanki, R.: Artificial intelligence for the real world. Harvard Bus. Rev. **96**(1):108–116 (2018)
20. Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., Fang, J., Huang, Y., Guo, H., Li, L., Trapp, B.D.: Target identification among known drugs by deep learning from heterogeneous networks. Chem. Sci. **11**(7), 1775–1797 (2020)

# Chapter 2
# Knowledge Strategies Influencing on the Epidemiologists Performance of the Qeshm Island's Health Centers

**Kamran Yeganegi, Maryam Ebrahimi, and Ahmed J. Obaid**

## 2.1 Introduction

By utilizing experience, abilities, and interaction outside of an organization, collective knowledge resulting from human expertise can be turned into the vital resources. Meanwhile, as two concepts focused on people, human resource management and knowledge management emphasize the knowledge use, sharing, and creation. It is often impossible to manage knowledge in an isolated environment without the presence of people. Besides, knowledge, skills, and behaviors as the human capitals, as well as routines, systems, and tacit knowledge as organizational capitals, are acquired over time and competitors are hardly able to imitate them. Knowledge and knowledge management, as an important asset and a significant process, respectively, is one of the issues in strategic management. In this regard, it is required to invest in developing human capital in order to enhance companies' performance [1].

The measures taken in the field of the performance of epidemiologists are still few and in the early steps of development [2]. On the contrary, in health centers, human resource management is of particular importance, and epidemiologists are no exception. According to the results of the studies, the most significant factor in the success of healthcare centers is their human resources.

K. Yeganegi (✉)
Department of Industrial Engineering, Islamic Azad University, Zanjan Branch, Zanjan, Iran
e-mail: yeganegi@iauz.ac.ir

M. Ebrahimi
Department of Information Technology Management, Islamic Azad University, Electronic Branch, Tehran, Iran
e-mail: ebrahimim@modares.ac.ir

A. J. Obaid
Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq
e-mail: ahmedj.aljanaby@uokufa.edu.iq

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_2

Recently, by accepting the vital role of human factors in comparison with technical issues in success, it can be said that human resources play a strategic role in the successful management of healthcare centers [3]. However, few experimental studies have been done in this field [4]. The three main factors of time, cost, and performance have been studied in previous studies as success factors of healthcare centers [5]. As a matter of course, ignoring other aspects of epidemiologists' performance in healthcare centers is one of the basic problems of the previous approaches [6, 7]. In consequence, the current research has been conducted with the aim of identifying the most significant aspects of the success of healthcare centers, including the performance of epidemiologists.

The process of managing the performance of epidemiologists, which includes the need for design, preparation, consultation, allocation of staff time, and discharge in healthcare centers, requires the epidemiologists' coordination in healthcare centers. Of course, these processes depend on the relevant activities, which can be manual or intellectual. Since the intellectual work entails great complexity due to the need for specific knowledge, more resources are spent on this work by epidemiologists. On the other hand, the output quality of these centers depends on their management. Recently, a large amount of empirical research has been conducted to study the effect of various factors of knowledge workers on the success of epidemiologists' performance. But few, if any, have examined and understood the strategies affecting the performance of epidemiologists.

Nowadays, with the transition from an industrial economy to an economy based on information and knowledge, improving the knowledge workers' productivity is one of the main challenges of healthcare centers [8]. The three components of Can, Will, and May constitute the main factors of knowledge workers in healthcare centers, and a scientific method is needed to classify the knowledge factors of employees in healthcare centers.

The objectives of this study are: (1) defining knowledge strategies for epidemiologists' performance success, (2) using Analytical Network Process (ANP) approach to prioritize knowledge strategies, and (3) validating the model in health centers in Qeshm (Gheshm) Island. This paper after introduction is categorized into the following sections: (1) reviewing the existing literature, (2) introducing Qeshm Island, (3) presenting an ANP-based approach for prioritizing knowledge strategies, (4) presenting the results, and finally (5) conclusion. Unlike Analytic Hierarchy Process (AHP), ANP offers a more generalized model of decision-making without assumptions about element independence [9]. Therefore, instead of using the conventional AHP approach to solve interdependence problems, it is recommended to use an ANP-based model to prioritize knowledge strategies.

## 2.2 Literature Review

As referred by [10], performance is the ability of a measuring object to produce an output that is consistent with its objectives. It is also stated that performance can be understood through the actual results or outputs of activities, as a form of how something is done, and as the ability of an individual to achieve results. Performance metrics reflect the effectiveness of measures.

As non-manual workers who are usually employed by healthcare managers to perform innovative activities, epidemiologists or knowledge workers use knowledge to increase worker productivity [11]. In other words, a knowledge worker refers to a person who works in the field of developing or using knowledge for the purpose of living [12].

Knowledge work is complex and those who do it need special skills and abilities in addition to having real and theoretical knowledge. These people need to have the ability to find, access, and apply information, communicate well with others, and have the motivation to gain and promote these skills. Although the prominence of one or more of these features might be different from job to job, all knowledge workers should have the following features [13]:

(1) Having real and theoretical knowledge,
(2) Being able to search and access information,
(3) Being able to utilize information,
(4) Having the ability to interact with others,
(5) Being motivated, and
(6) Having intellectual capabilities.

The performance of knowledge workers can be promoted through various ways such as providing availability to required information, education opportunities, and the balance between control and autonomy. Via information technology (IT), they can do different tasks relevant to information including capturing, processing, storing, retrieving, and sharing information. Information technology should be designed to save the time needed for accessing, managing, and manipulating information. With knowledge management, a company can direct information to knowledgeable employees [13]. The factors related to knowledge workers should be considered by managers of a healthcare center as part of the management process and as a strategic element in order to achieve the goals of healthcare centers' continuous improvement.

There are some studies about the linkages between knowledge management enablers (process, culture, and information technology) and performance [14, 15]. Besides, [16] focused on the complexities of using knowledge management enablers strategically or KM strategies for the effective knowledge management. In many studies, the relationship of KM strategies with performance criteria including organizational performance [17, 18] as well as market performance [16] is emphasized.

As a vital source, the knowledge importance encourages managers to pay more attention to the company's knowledge management strategies, which, if properly defined, align the organizational process, culture, and implementation of information technology (IT) related to knowledge management. In this regard, the aim of research No. [19] was to investigate the synergistic interaction between knowledge management strategies and their impact on organizational performance.

The scientific and systematic planning of human resources to meet the organization's operational and strategic requirement forms the core of knowledge management. In research No. [20], an attempt was made to determine how much knowledge sharing improves employee performance. It was also found in [21] that by managing the intellectual capital resulted from the implicit and explicit knowledge, we can assist the organization to learn from its environment and use the knowledge to interact with its business processes. The impact of knowledge process on organizational performance was investigated in [22], which was done with the aim of practically confirming the ability of knowledge management to improve the performance of organization. Based on the obtained results, the organization's competitive advantage that ultimately leads to organizational performance can be realized through the collection and sharing of new knowledge [23] concluded that knowledge management directly and indirectly affects organizational performance through strategic human resource actions.

## 2.3 Qeshm Island

Located a few kilometers from the southern coast of Iran (Persian Gulf), and close to the port cities of Bandar Abbas and Bandar Khamir, Qeshm Island, with a length of 135 km, hosts the jurisdiction of the Free Zone with an area of 300 km$^2$ (116 mi$^2$) and is strategically located in the Strait of Hormuz, with the distance of 60 km (37 mi) and about 180 km (112 mi) from Port Khasab in Oman and Port Rashid in the United Arab Emirates, respectively. The widest point of this island near its center, and the narrowest point, have 40 km (25 mi) and 9.4 km (5.8 mi) width, respectively. With an area of 1,491 km (576 mi$^2$), the island is almost twice Bahrain. In the distance of 22 km from Bandar Abbas and as the closest point of the island, which is only two km from the mainland, Qeshm city is located at the easternmost point of the island. Including 59 towns and villages, the island has a population of 117,774 people according to the 2011 census. The occupation of the local people of this island is fishing, shipbuilding, trade, and services, and about 30,000 people are in official or industrial jobs or students.

## 2.4 Research Methodology

Since little research has been done on investigating the knowledge strategies, a case study approach was used in the current descriptive research. In this research, a questionnaire was used as a research tool to collect the required data, including reference materials and questionnaire review, and the necessary sample for the study was selected in a simple random approach. Previous researches in this field have often applied approaches such as Analytical Network Process (ANP) method, measurement tools, and theoretical frameworks for case study analysis.

Interview requests and questionnaires were sent to a number of epidemiologists and employees of 3 healthcare centers on the Qeshm Island in April 2017. The researchers assured the respondents about the confidentiality of answers and raising questions in public areas and explained the purpose of the research before the interview and completing the questionnaire. Then, the interview and questionnaire were transcribed to ensure internal validity and sent to the employees of the healthcare centers in order to verify the authenticity and check the absence of sensitive business information in them.

Sampling: Since it was important to cover the opinions of all epidemiologists, diversity was observed in the selection of respondents, and as a result, different job groups in various departments were considered. Four government departments in healthcare centers were selected as employers, and since the healthcare centers in Iran's provinces range from 20 to 200 employees, the average number of epidemiologists and employees was 120 and 60, respectively. Epidemiologists interviewed include independent contractors, short-term employees hired by the organization, seasonal employees, and those who were hired directly by the organization and had temporary or fixed-term employment contracts in the positions like specialists, healthcare center managers, researchers, planners, and resource conservation officers. The contract worker interviewees are consisted of 135 men and 45 women ages ranged from the early 20 s to near retirement age. The respondents' education level was also as follows; 7 individuals with no postsecondary degree, 30 with an undergraduate degree, and 143 with a graduate degree.

Data analysis: Two researchers analyzed the answers independently to identify key problems and themes for the qualitative part of research. The ANP method is also used to select the strategies.

### *2.4.1 Analytical Network Process (ANP)*

Developed by "Saaty," ANP is proposed as a generalization of AHP, that like it provides a hierarchical structure framework with one-directional relationships that helps for considering complex internal dependencies among various decision-making and criterion levels [24]. The structural difference between hierarchy (a) and network (b) are shown in Fig. 2.1.

As a comprehensive and explanatory approach, ANP is used for the purpose of multi-objective decision-making and also for solving complex decision-making problems. ANP method is used in [24] to select internally dependent information systems for healthcare centers in which no requirement is considered to perform an ideal Zero–One programming. Furthermore, ANP is used in [25, 26] for quality activity development. A reflective state system can be explained by a network that Fig. 2.2 shows the structural model of such a network. The existent element in each cluster can affect all or some of the other cluster elements. Main clusters, middle clusters, and final clusters can be the components of a network. In this network, arrows and their direction indicate the relationships and dependence, respectively.

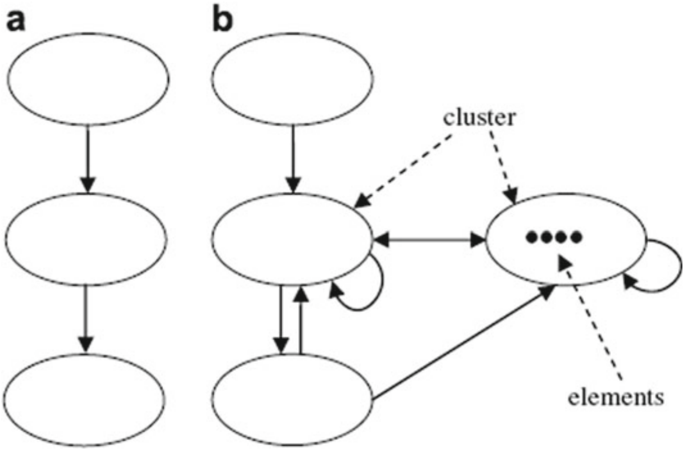


**Fig. 2.1** The hierarchy (**a**) and network (**b**) structural difference [24]

Also, the dependence among clusters and the internal dependence among elements of a cluster can be called external dependence and circle dependence, respectively [27, 28].

### *2.4.2 ANP-Proposed Algorithm*

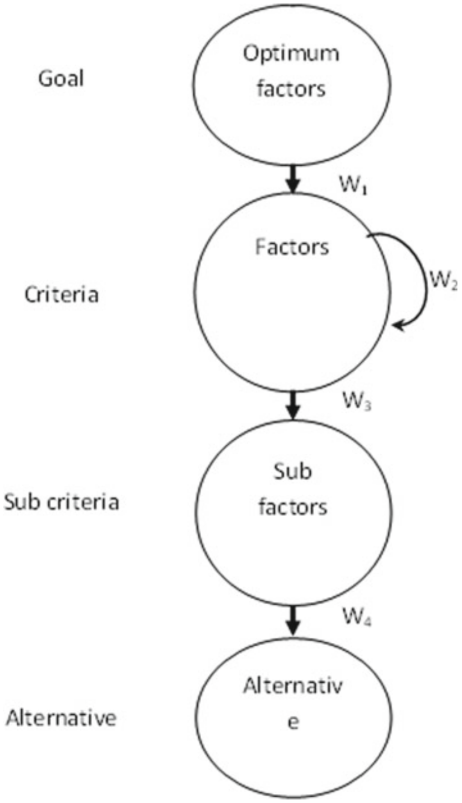
Figure 2.2 indicates the proposed four-level model hierarchy and network for knowledge strategies analysis. In this figure, the first level indicates the goal (most appropriate strategy), the second and third levels show the criteria, and the alternatives (strategy options) are indicated in the last level.

The matrix below shows the four levels of a knowledge strategies hierarchy supermatrix:

$$\mathrm{W} = \begin{array}{c} \text{Goal} \\ \text{Factors} \\ \text{Sub}-\text{Factors} \\ \text{Alternative} \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 \\ \mathbf{w}_1 & \mathbf{W}_2 & 0 & 0 \\ 0 & \mathbf{W}_3 & 0 & 0 \\ 0 & 0 & \mathbf{W}_4 & \mathbf{I} \end{bmatrix}$$

where the goal or aim effect vector, e.g., the most appropriate strategy selection according to element factors, the element factors' internal dependence matrix, the effect matrix of element factors on each of the element sub-factors, and the index of element sub-factors' effect on the strategic options are shown by W1, W2, W3, W4, respectively. The matrix functions detail the algorithm steps. ANP and the matrix functions are used to derive the proposed algorithm as follows.

Fig. 2.2 Model structure of network adopted from [24]



**Step 1**: Specifying the element sub-factors and the strategy options according to sub-factors,

**Step 2**: Assuming no dependencies among element factors, then using the numerical scale of 1–9 (W1 calculation) as the importance degree of element factors,

**Step 3**: Using the numerical scale of 1–9 to specify the element factors of the internally dependent matrix, considering other factors by the schematic view and internal dependencies among them (W2 calculation),

**Step 4**: Calculating $w_{factors} = W_2 \times W_1$ to determine the internal dependencies' priorities,

**Step 5**: Using the numerical scale of 1–9 (calculation of $\mathrm{W}_{\text{sub-factors(local)}}$) to determine the local importance degree of element sub-factors,

**Step 6**: Calculating $\mathrm{W}_{\text{sub-factors(global)}} = \mathrm{W}_{\text{factors}} \times \mathrm{W}_{\text{sub-factors(local)}}$ to determine the importance degree of sub-factors,

**Step 7**: Using the scale of 1–9 (W4 calculation) to determine the importance degree of the strategy options, considering each sub-factor,

**Step 8**: Calculating $W_{alternatives} = W_4 \times W_{sub-factors(global)}$ to determine the final priority of the strategy options derived from the internal relationships among element factors.

## 2.5 Results and Discussion

The steps of the proposed approach are provided below:

**Step 1**: First, a hierarchical structure, containing the strategy options and sub-factors for the next calculations using ANP, is used to depict the problem. At the first level of the ANP Model, the goal or aim is selected as the most appropriate strategy and at the second level, the element factors are specified. The three element sub-factors of CAN, WILL, and MAY are also considered in the third level. Finally, the level fourth presents 13 strategy options.

The strategy options are as follows:

**A–C** Motivating staff spiritually and financially by the output work level,

**A–D** Specifying the authority of knowledge workers and removing cumbersome rules,

**A–E** Creating a creative and trust-based environment for communication,

**A–F** Changing the view of knowledge workers as piece workers, not daily workers,

**B–D** Employees training and development,

**B–E** Work turnover in organization,

**B–F** Creating a reward and evaluation system for organizational employees,

**C–D** Establishing flexible structures,

**C–E** Transparency of activity and right ownership of intellectual property,

**C–F** Establishing proper informative and communicative structures,

**D–E** Providing cooperation opportunities,

**D–F** Improving organizational environment,

**E–F** Creating job security.

**Step 2**: Let's consider no dependency among the element factors: A-Intellectual capital, B-Expert management, C-Structure, D-Culture, E-Qualified workforce, and F-Motivation. The numerical scale of 1–9 is used to determine the factors' pair comparison matrix. All the pair comparisons are performed by an experts team. Expert Choice software is used to analyze the pair comparison matrix and to obtain the following special vector.

$$W_1 = \begin{bmatrix} A \\ B \\ C \\ D \\ E \\ F \end{bmatrix} = \begin{bmatrix} 0.366 \\ 0.231 \\ 0.170 \\ 0.114 \\ 0.078 \\ 0.041 \end{bmatrix}$$

**Step 3**: The comparison of the effect of factors on each other is used to specify the internal dependency among element factors. As mentioned, considering independence among the element factors is not always possible. Suitable and realistic results are obtained from the ANP technique and element analysis. An analysis of elements reveals the element factors' dependencies. A pair comparison matrix for factors is illustrated. The results obtained from the special vectors are depicted. The internal dependency of the element matrix, based on the calculated relative importance weights, is shown by $W_2$. While opportunities are only influenced by strengths, a pair comparison matrix cannot be formulated for the opportunities.

$$W_2 = \begin{bmatrix} 1 & 0 & 0.565 & 0.44 & 0.422 & 0.490 \\ 0 & 1 & 0 & 0.307 & 0.329 & 0.249 \\ 0.53 & 0 & 1 & 0.029 & 0.039 & 0.042 \\ 0.31 & 0.055 & 0.056 & 1 & 0.078 & 0.081 \\ 0.117 & 0.173 & 0.089 & 0.067 & 1 & 0.138 \\ 0.042 & 0.772 & 0.290 & 0.157 & 0.131 & 1 \end{bmatrix}$$

**Step 4**: The factor's internal dependencies priorities are calculated as shown below:

$$\begin{aligned} w_{\text{factors}w} &= W_2 * W_1 \\ &= \begin{bmatrix} 1 & 0 & 0.565 & 0.44 & 0.422 & 0.490 \\ 0 & 1 & 0 & 0.307 & 0.329 & 0.249 \\ 0.53 & 0 & 1 & 0.029 & 0.039 & 0.042 \\ 0.31 & 0.055 & 0.056 & 1 & 0.078 & 0.081 \\ 0.117 & 0.173 & 0.089 & 0.067 & 1 & 0.138 \\ 0.042 & 0.772 & 0.290 & 0.157 & 0.131 & 1 \end{bmatrix} * \begin{bmatrix} 0.366 \\ 0.231 \\ 0.170 \\ 0.114 \\ 0.078 \\ 0.041 \end{bmatrix} = \begin{bmatrix} 0.565 \\ 0.302 \\ 0.372 \\ 0.260 \\ 0.189 \\ 0.312 \end{bmatrix} \end{aligned}$$

Overall priority of the factors according to the $W_{\text{factors}}$ is as follows:

- A-Intellectual capital = 0.565,
- B-Expert management = 0.302,
- C-Structure = 0.372,
- D-Culture = 0.260,
- E-Qualified workforce = 0.189,
- F-Motivation = 0.312.

The priority vector of sub-factors is defined based on these priorities.

$$W_{\text{sub}-\text{fdactor}-A} = \begin{bmatrix} 0.308 \\ 0.192 \\ 0.151 \\ 0.133 \\ 0.108 \\ 0.108 \end{bmatrix}, \quad W_{\text{sub}-\text{fdactors}-C} = \begin{bmatrix} 0.35 \\ 0.29 \\ 0.15 \\ 0.13 \\ 0.08 \end{bmatrix},$$

$$W_{\text{sub}-\text{fdactor}-F}F = \begin{bmatrix} 0.342 \\ 0.211 \\ 0.178 \\ 0.105 \\ 0.077 \\ 0.055 \\ 0.032 \end{bmatrix} W_{\text{sub}-\text{fdactors}\,B} = \begin{bmatrix} 0.352 \\ 0.181 \\ 0.150 \\ 0.110 \\ 0.150 \\ 0.062 \\ 0.031 \\ 0.028 \\ 0.022 \\ 0.015 \\ 0.009 \end{bmatrix},$$

$$W_{\text{sub}-\text{fdactors}-D} = \begin{bmatrix} 0.255 \\ 0.202 \\ 0.132 \\ 0.123 \\ 0.102 \\ 0.095 \\ 0.085 \\ 0.072 \\ 0.033 \\ 0.028 \\ 0.018 \\ 0.012 \\ 0.008 \end{bmatrix}, \quad W_{\text{sub}\,f\,\text{fdactors}-\,E} = \begin{bmatrix} 0.208 \\ 0.119 \\ 0.113 \\ 0.122 \\ 0.106 \\ 0.095 \\ 0.084 \\ 0.052 \\ 0.034 \\ 0.025 \\ 0.018 \\ 0.012 \\ 0.008 \\ 0.003 \\ 0.001 \end{bmatrix},$$

**Step 5**: Multiplying the internal dependency priorities, obtained in Step 4, by the local priorities of element sub-factors obtained in Step 5, is used to calculate the general priorities of the element sub-factors. Vector $W_{sub-\,factors(global)}$, obtained from the general priority amounts, shows the results.

**Step 6**: The strategy options' importance degree is obtained from each element's sub-factor perspective. Special vectors are obtained by analyzing this matrix and matrix $W_4$.

**Step 7**: Finally, regarding the element factors' internal dependencies, the strategy options' general priorities are obtained as follows:

$$w_{alternatios} = \begin{bmatrix} \mathrm{A-C} \\ \mathrm{A-D} \\ \mathrm{A-E} \\ \mathrm{A}-F \\ \mathrm{B-D} \\ \mathrm{B-E} \\ \mathrm{B-F} \\ C-D \\ C-E \\ C-F \\ D-E \\ D-F \\ E-F \end{bmatrix} = W_4^* w_{sub-factor\,(global)} = \begin{bmatrix} 0.076 \\ 0.080 \\ 0.085 \\ 0.081 \\ 0.063 \\ 0.071 \\ 0.078 \\ 0.086 \\ 0.097 \\ 0.089 \\ 0.095 \\ 0.066 \\ 0.078 \end{bmatrix}$$

According to the Step 7, overall priority of the alternative scenarios are as follows:

- Spiritual and financial motivation based on the output work level = 0.076,
- Specifying the authority of knowledge workers and removing cumbersome rules = 0.080,
- Creating a creative and trust-based environment for communication = 0.085,
- Changing the view of knowledge workers as piece workers, not daily workers = 0.081,
- Staff training and development = 0.063,
- Work cycling in organization = 0.071,
- Creating a reward and evaluation system for organizational employees = 0.078,
- Establishing flexible structures = 0.086,
- Transparency of activity and right ownership of intellectual property = 0.097,
- Establishing proper informative and communicative structures = 0.089,
- Providing cooperation opportunities = 0.095,
- Improving organizational atmosphere = 0.066,
- Creating job security = 0.078.

The C–E strategy or transparency of activity and right ownership of intellectual property with a score of 0.097 is the most important strategy for the success of epidemiologists' performance based on the results of ANP analysis. Also, creating cooperation opportunities in organizations or D–E with a score of 0.095 is another important strategy. The important issue is that all the mentioned strategies should be used to improve the performance of epidemiologists for the purpose of the success of healthcare centers.

The inconsistency ratio of the pairwise comparison matrix, which is calculated using Expert Choice, should not be less than 0.1 in ANP, and in this study, all values of the inconsistency ratio are less than 0.1. In this regard, transparency of activity and the right ownership of intellectual property are the most important elements in the performance of epidemiologists, followed by compiling the organization's intellectual property document and implementing it. As far as we know, so far, no research has focused on the analysis of knowledge factors related to the success of epidemiologists' performance using the model proposed in this article, and this is one of the contributions of the present research.

Validation: Cronbach's alpha, which is obtained more than 98.03, approves the validity of the results and is confirmed by company directors as well as experts and managers, to 87% and 88, respectively.

## 2.6 Conclusion

The acceptance of the knowledge worker factors affecting the measurement and prediction of the performance of epidemiologists is emphasized in the present study as an important issue that allows any company to survive in a very competitive industry sector. In the meantime, based on the perspective consistent with the resource-oriented perspective of the enterprise and management of knowledge workers in knowledge-based industries, knowledge-based workers are a valuable resource, which also shows the necessity of using knowledge management as a factor of competitive advantage.

As a defining characteristic of the new era, this research seeks to understand the best strategy for the success of the epidemiologists' performance. Understanding more effective ways for managing the performance of epidemiologists as important and unique factors in the knowledge-based economy is one of the results of this study to help academics and organizations.

In this research, firstly, the identification and classification of the elements affecting the performance of knowledge workers were considered for the success of epidemiologists' performance, and then, Analytical Network Process (ANP) was used to analyze them. Finally, some strategies were presented for improving factors affecting the performance of knowledge workers, whose validity was confirmed during a case study in the healthcare centers of Iran's provinces.

## References

1. El-Farr, H., Hosseingholizadeh, R.: Aligning human resource management with knowledge management for better organizational performance: how human resource practices support

knowledge management strategies? In: Wickham, M., (ed.) Current Issues in Knowledge Management. Intechopen (2019)

2. Bredin, K.: People capability of project-based organizations: a conceptual framework. Int. J. Project Manag. **26**, 566–576 (2008)
3. Larson, E.W., Gobeli, D.H.: Significance of project management structure on development success. IEEE Trans. Eng. Manag. **36**(2), 119–125 (1989)
4. Raide´n, A.B., Dainty, A.R.J., Neale, R.H.: Balancing employee needs, project requirements and organizational priorities in team deployment. Constr. Manag. Econ. **24**, 883–895 (2006)
5. Belout, A., Gauvreau, C.: Factors influencing project success: the impact of human resource management. Int. J. Project Manag. **22**, 1–11 (2004)
6. Hackman, J.R.: The design of teams. In: Lorsch, J. (ed.) Handbook of Organizational Behaviour, pp. 315–342. Prenctice-Hall, Englewood Cliffs, NJ (1987)
7. Scott-Young, C., Samson, D.: Project success and project team management: evidence from capital projects in the process industries. J. Oper. Manag. **26**, 749–766 (2008)
8. Drucker, P.F.: The New Realities. Heinemann Professional Publishing, Oxford (1989)
9. Jharkhariaa, S., Shankar, R.: Selection of logistics service provider: an analytic network process (ANP) approach. Omega **35**, 274–289 (2007)
10. Vänni, K.: Health and Performance of Knowledge Worker. MBA thesis, Finland (2007)
11. Stuhlman, D.: Helping you turn data into knowledge: knowledge Management Terms (2006) Accessed form: www.home.earthlink.net/~ddstuhlman/defin1.htm
12. Adkoli, A.: An agenda for ICT-enabled education. Indian Manag. 44–50 (2006)
13. Mohanta, G.C., Kannan, V., Thooyamani, K.P.: Strategies for improving productivity of knowledge workers–an overview. Strength Based Strat. 77–84 (2006)
14. Ho, C.T.: The relationship between knowledge management enablers and performance. Ind. Manag. Data Syst. **109**(1), 98–117 (2009)
15. Theriou, N., Maditinos, D., Theriou, G.: Knowledge management enabler factors and firm performance: an empirical research of the Greek medium and large firms". Eur. Res. Stud. **14**(2), 97–134 (2011)
16. Choi, B., Jong, A.M.: Assessing the impact of knowledge management strategies announcement on the market value of firms. Inf. Manag. **47**(1), 42–52 (2010)
17. Liao, Y.S.: The effect of human resource management control systems on the relationship between knowledge management strategy and firm performance. Int. J. Manpow. **32**(5/6), 494–511 (2011)
18. Yang, J.: The knowledge management strategy and its effect on firm performance: a contingency analysis. Int. J. Prod. Econ. **125**(2), 215–223 (2010)
19. Choi, B., Poon, S.L., Davi, J.G.: Effects of knowledge management strategy on organizational performance: a complementarity theory-based approach. Omega **36**, 235–251 (2008)
20. Ranjan Meher, J., Kumar Mishra, R.: Examining the role of knowledge sharing on employee performance with a mediating effect of organizational learning. VINE J. Inf. Knowl. Manag. Syst. vol. Ahead of print, No. Ahead of print, 2059–5891 (2021)
21. Chennemaneni, A.: Determinants of knowledge sharing behavior: developing and testing a theoretical model. The University of Texas (2007)
22. Gold, A.H., Malhotra, A., Segars, A.H.: Knowledge management: an organizational capabilities perspective. J. Manag. Inf. Syst. **18**(1), 185–214 (2001)
23. Rajabi Ferjad, H., Najar, M.: The impact of knowledge management on organizational performance with regard to mediating role of strategic activities of human resource. J. Res. Hum. Resour. Manag **10**(3.33), 191–214 (2018)
24. Yüksel, T., Dagˇdeviren, M.: Using the analytic network process (ANP) in a SWOT analysis–a case study for a textile firm. Inf. Sci. **177**(16), 33 (2007)
25. Karsak, E.E.: Personnel selection using a fuzzy MCDM approach based on ideal and anti-ideal solutions. Multiple Criteria Decision Making in The New Millenium, Berlin (2001)
26. Partovi, F.Y., Corredoira, R.A.: Quality function deployment for the good of soccer. Eur. J. Oper. Res. **137**, 642–656 (2002)

27. Mikhailov, L., Singh, M.S.: Fuzzy analytic network process and its application to the development of decision support systems. IEEE Trans. Syst., Man, Cybern.-Part C: Appl. Rev. **33**, 33–41 (2003)
28. Expert Choice: Expert Choice, Analytical Hierarchy Process (AHP) Software, Version 9.5, Expert Choice, Pittsburg (2000)

# Chapter 3
# Healthcare: In the Era of Blockchain

**Gagandeep Kaur, Priyanka Choudhary, Lakshita Sahore, Sunil Gupta, and Veerpal Kaur**

## 3.1 Introduction

Blockchain is an open-source digital ledger that keeps track of transactions on numerous computers in a way that prevents any record from being changed retrospectively without also changing any subsequent blocks [1]. Immutability, decentralization, transparency, and distributed ledgers are a few characteristics that drive developers to blockchain technology. Prominent areas of blockchain include Cryptocurrency exchange [2, 3], secure sharing of medical data, better model building for prediction purposes using machine learning [4–6] and data mining [7–10], personal identity security, etc.

The role of blockchain is indispensable as well as vital in healthcare industry and it has the power to transform this area [11–14]. During medical treatment, privacy and data protection are important considerations. A platform for secure communication is necessary for a healthcare system to operate properly and efficiently hence blockchain first appeared as an emerging technology [15]. Figure 3.1 shows the healthcare systems enabled by blockchain [16].

The chapter is organized into the following sections. Section 3.2 gives an overview of related work while Sect. 3.3 focuses on application areas. Section 3.4 highlights

G. Kaur (✉) · P. Choudhary · L. Sahore · S. Gupta
Institute of Engineering and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
e-mail: gaganmalhotra1791@gmail.com

P. Choudhary
e-mail: priyanka0552.be20@chitkara.edu.in

L. Sahore
e-mail: lakshita0416.be20@chitkara.edu.in

V. Kaur
Department of Computer Science and Engineering, Lovely Professional University, Punjab, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_3

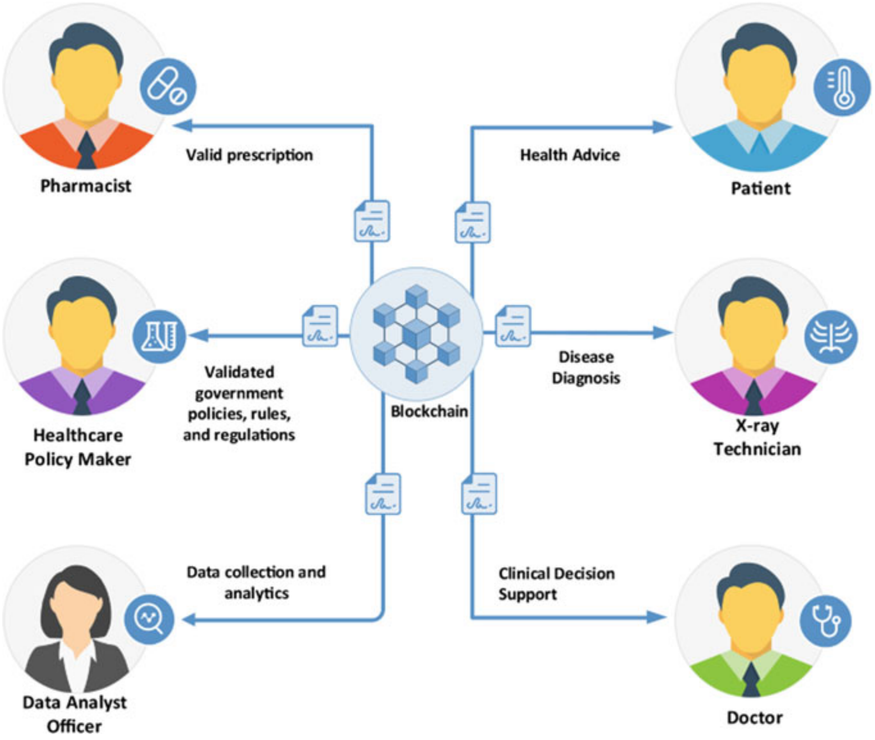


**Fig. 3.1** Blockchain-enabled healthcare system

the challenges and limitations of blockchain in healthcare. Conclusion of the chapter is listed in Sect. 3.5.

## 3.2 Related Work

The related work shown in Table 3.1 states how blockchain technology not only accelerated the establishment of the ecosystem life cycle, but it also catalyzes interactions between various fields. A variety of blockchain-based healthcare prototypes, uses, needs, methods, advantages, drawbacks, applications. and features have been highlighted in it.

McGhin [17] discussed due to strenuous legal requirements, healthcare blockchain applications require more validation, interoperability, and record sharing. As part of the work, it is also stated that a number of aspects of security are unique to the healthcare industry, including security in Internet of Things applications, wireless security, interoperability, access control, authentication, and non-repudiation. In the segment, it was stated that data integrity, immutability, and security are key features of

**Table 3.1** Description of the related work

| S. no. | Year | Title | Outcome |
|---|---|---|---|
| 1 | 2019 | Blockchain in healthcare applications: research challenges and opportunities | Highlight the various applications of blockchain with their issues and advantages |
| 2 | 2020 | Blockchain in healthcare and health sciences—a scoping review | Examines how blockchain technology can be used in the area of healthcare |
| 3 | 2018 | A systematic review of the use of blockchain in healthcare | Highlights the growing importance of blockchain technology in healthcare |
| 4 | 2020 | Blockchain in healthcare innovation: literature review and case study business ecosystem perspective | Describes how blockchain and its applications in healthcare have evolved and the evolution of stakeholders in the field |
| 5 | 2019 | Blockchain technology in healthcare: a systematic review | Depicts the various healthcare use cases with its prototypes |
| 6 | 2019 | Lightweight blockchain for healthcare | It proposes blockchain architecture and its role in different data management systems |
| 7 | 2018 | Applications of blockchain within healthcare | Highlights the current implementation and issues within the modern healthcare industry |
| 8 | 2016 | A case study for blockchain in healthcare: “MedRec” prototype for electronic health records and medical research data | Addresses the MedRec prototypes and its principles in the EHR and medical systems |
| 9 | 2019 | Blockchain in healthcare | Several insights about blockchain technologies and their potential applications for healthcare systems are uncovered |
| 10 | 2021 | Blockchain for healthcare data management: opportunities, challenges, and future recommendations | Provides insights into blockchain technology, its key features, and its application in healthcare |
| 11 | 2019 | Blockchain technology in healthcare: a comprehensive review and directions for future research | Describes the open research matters in the fast-growing field with revolutionizing the healthcare industry |
| 12 | 2019 | Blockchain in healthcare: a patient-centered model | It states about security and accessibility and its use in different healthcare venues and remote applications |
| 13 | 2020 | Blockchain in healthcare: a systematic literature review, synthesizing framework, and future research agenda | The work summarized the thematic trends of academic research on databases in the healthcare field and aimed to highlight their applications and objectives |

(continued)

**Table 3.1** (continued)

| S. no. | Year | Title | Outcome |
|---|---|---|---|
| 14 | 2019 | Applications of blockchain technology in medicine and healthcare: challenges and future perspectives | States the current and future developments of blockchain in healthcare with its applications and challenges |
| 15 | 2019 | A systematic review for enabling of develop a blockchain technology in healthcare applications: taxonomy, substantially analysis, motivations, challenges, recommendations, and future direction | Identifies the various types, uses, solutions, challenges, and applications among different healthcare sectors and medical institutes |
| 16 | 2020 | The benefits and threats of blockchain technology in healthcare: a scoping review | Discusses how blockchain technology can improve the sharing and storing of healthcare data |
| 17 | 2017 | Applying software patterns to address interoperability in blockchain-based healthcare apps | Depicts the features, challenges, and software patterns used in developing blockchain-based healthcare apps |
| 18 | 2019 | Comparison of blockchain platforms: a systematic review and healthcare examples | Addresses the benefits, features, methods, and development of biomedical blockchain applications |
| 19 | 2020 | Blockchain-based electronic healthcare record system for healthcare 4.0 applications | It proposes an access control policy algorithm for improving and implementing EHR sharing system |
| 20 | 2020 | Blockchain platform for industrial healthcare: vision and future opportunities | It summarizes the consensus algorithm with blockchain components and its types in the healthcare field |

blockchain technology. The research gaps in this study focus on the key management of decentralized privacy leak and lack of standardization. It also includes blockchain specific and software vulnerabilities.

Hasselgren [18] highlights that electronic health records and personal health records are two sectors targeted by blockchain technology. The most popular platform in this domain appears to be Ethereum and Hyperledger fabric. Research shows that blockchain technology is rapidly gaining traction in the health sector. Decentralization, Anonymity, and Persistency are considered as a major features of blockchain.

Kompara [19] describes the various distributed systems and its growing consensus in the blockchain technology. The work also discusses the technology's potential applications and challenges on the healthcare front as well as future directions, as the majority of which focuses on providing unique structural designs as frameworks, structures, or models. It provides an introduction to Bitcoin and how it ensures decentralized transactions without the requirement for a trusted central authority.

Chen [20] focuses on the blockchain technology and its potential paradigm shift being used in a variety of healthcare industries and its leveraging progress in the business ecosystem. It also defined the changing paradigm on the healthcare ecosystem from segments like Health Information Exchange (HIE), Digital Identity Management, Healthcare Supply Chain Management, Medical Research Data Exploitation, and Automation of Financial Transactions and Insurance Procedures.

Agbo et al. [21] states that broad range of industries and use cases have seen the benefits of blockchain, including identity management, dispute resolution, etc., states the various use cases stated in the blockchain which include, remote patient monitoring, medical record, etc.

Ismail et al. [22] addresses on how health information management has gained lots of potential in the field of blockchain due to its cost-efficient patient care and states its general agreement in the prospect of scalability, decentralization, transparency, immutability, traceability, privacy, and system vulnerability. The author also presents a method for avoiding forking which is common in the Bitcoin network using a Head Blockchain Manager (HBCM) which will provide security and privacy to the proposed architecture. Transactions, Blocks, Merkle tree root hash, Nodes, and Mining are a few of the various features of blockchain network that the study describes.

Bell et al. [23] provides an insight into how blockchain technology can improve healthcare and wellness. Among the topics addressed in the study are pharmaceutical traceability, data sharing, clinical trials, etc. Additionally, it discusses the current challenges within the modern healthcare industry, such as the interoperability of healthcare data, the tracking of medical devices, and the tracking of drugs.

Ekblaw et al. [24] explains that digital health records (EHRs) need to be stabilized with blockchain technology used by MedRec and its decentralized management system. In addition, it shows the unique characteristics of blockchain technology that allow MedRec to manage the identification, secrecy, data sharing, etc., which are all critical concerns when handling sensitive information. It also highlights the security testing and a bug bounty program which are outlined in the "Ubiquitous Secure Network Infrastructure."

Prokofieva and Miah [25] describes the various IS solution like Nebula Genomics, Secure Health Chain, and Doc.AI which have been utilized as blockchain applications in healthcare. The work describes the benefits of blockchain in healthcare where professionals and patients can access records easily without added cost with high security, less chances of errors, and data loss. It also addressed the various issues like privacy, coordination, time, and human factor which are being shown in current healthcare blockchain solutions.

Yaqoob et al. [15] describes the key features and benefits of blockchain in healthcare industry along with its opportunities and research challenges. Transparency, Decentralization, Immutability, Data provenance, Distributed ledger and consensus, Anonymity, and programmability are some of the key blockchain features that have been described. It also addresses various opportunities like improved drug traceability, clinical trials and precision medicine, patient and management, protecting telehealth systems, optimizing health insurance coverage, medical billing system, and maintaining consistent permissions. It highlights the various case studies and ongoing project which includes the Estonian e-health system.

Khezr et al. [26] presented an overview of blockchain technology's potential application in the Internet of Medical Things (IoMT), including data administration, storage device connectivity, and security. In addition to data sharing, data management, data storage, and EHRs, blockchain-based healthcare management applications are covered. A focus is also placed on blockchain technology's potential as a responsible and transparent online data storage system and how the distribution system provides opportunities to address some of the most pressing issues facing healthcare in terms of data privacy, security, and integrity.

Chen et al. [27] describes the role of blockchain in e-health and electronic health records. It describes the various services provided in IT which includes Healthcare Service Innovation, Telehealth, and different development methodologies. Security and confidentiality of data have been the two major focuses. The work describes how Indicator-Centric Schema (ICS) is used to organize and secure Healthcare Data Gateway (HDG) data, ensuring privacy and security for healthcare data.

Tandon et al. [28] stated that PsycINFO, PubMed, Scopus, and Web of Sciences were identified as the source of information in health informatics. In this work the key topical issues discussed were to develop an intelligent healthcare system with building predictive capabilities and technical improvements to blockchain architecture. It also states how the viability of applications has been established and tested. The author also highlights that the medical diagnostics, legal compliance, avoiding fraud, and increasing patient care in times of remote monitoring or emergency will all benefit from blockchain applications.

Siyal et al. [29] highlights how the blockchain technology is gaining interest in a variety of fields including data management, financial services, cyber security, IoT, etc. In the article, it discusses how Safe, Secure, and Scalable (SSS) data sharing is necessary for diagnosis and combined clinical decision-making in the healthcare profession. Additionally, it identifies the Strengths, Weaknesses, Opportunities, and Threats (SWOT) associated with blockchain technology used in healthcare.

Hussien et al. [30] identifies, evaluates, and recommends a unified taxonomy for blockchain technology research in healthcare applications. It discusses about the various types of blockchain for healthcare system which includes un-permissioned or public blockchains, permissioned or private blockchains, and consortium or federated blockchain. Problems of Medical Data security, privacy, integrity, access control, interoperability, and managing the massive volumes of patient care are the types of problems in healthcare applications which have been addressed. The various recommendations mentioned in this script are to improve blockchain scalability, security, latency, throughput, blockchain size, and computing power restrictions.

Abu-Elezz et al. [31] states the benefits of blockchain technology which include Personalized healthcare, Exchange of health information, Pharmaceutics supply chain, Clinical trials, etc. It discusses the lack of technical skills due to high energy consumption and transaction costs with its security and scalability issues. By providing insight into the current positive and negative contributions of blockchain to healthcare, this work was able to help stakeholders, agencies, and organizations better understand blockchain's contribution in that area.

Zhang et al. [32] stated that the development of a blockchain-based healthcare mobile app, this work highlights the features and challenges associated with healthcare interoperability. In this work, challenges associated with healthcare interoperability are explored, such as minimizing integration complexity and data storage, etc. It shows the use of DApp for Smart Health (DASH) web-based portal and how it helps the patients to access, update, and submit their medical records.

Kuo et al. [33] highlights the four main applications of clinical blockchain which include improvement of medical record management, enhancement of insurance claim process, acceleration of biomedical research, and advancement of healthcare through data ledgers. It describes the main blockchain platforms Ethereum, Hyperledger, and Multichain that are adopted in healthcare applications. Bitcoin, Zcash, Dash, Ripple, Peercoin, etc., are some of the blockchain platforms that have been described in the study. The main key benefit addressed was that blockchain applications rely on its "off-the-shelf" capabilities, which address a variety of real-world health science requirements.

Tanwar et al. [34] shows that blockchain technology has the potential to improve interoperability in the healthcare industry by improving access to patient medical records, device tracking, prescription databases, and hospital assets via the blockchain infrastructure, etc. It also depicts the potential of blockchain technology through Distributed ledger, Provenance, Smart Contract, Finality, and Consensus Mechanism and describes the evolvement of different eras with time and how it has contributed to create the healthcare system.

Farouk et al. [16] describes the importance of Internet of Healthcare Things (IoHT) and its uses in clinical services and support operations. Moreover, the work recommends the combination of IoHT with Augmented-Reality (AR) technology to create digital twins and to offer technicians and clinicians the ability to perform realistic hands-on training to improve the transparency, security, and efficiency of clinical decision-making. It addresses the various components of blockchain technology which include ledger, cryptography, immutability, consensus algorithm, Merkle trees, assets, and peer-to-peer networks.

## 3.3 Application Areas of Blockchain in Healthcare

Healthcare offers a wide range of applications, from facilitating secure records transfers and managing supply chains to predicting disease outbreaks and helping scientists crack genetic codes. Some of them are listed below.

1. Patient Data Protection and Security
   Healthcare is a sector that is concerned about data security. Blockchain technology consists of a distributed and unalterable ledger of records that is decentralized and distributed. As a result, blockchain technology enables providers

and patients to send and receive important healthcare data quickly and securely, while simultaneously maintaining privacy and transparency through encryptions and complex security codes. Here are a few examples of how it occurs:

- Smart contracts in blockchain can be used by healthcare providers and organizations to store patient records. A public key, often known as a unique ID, is created to gain access to the data.
- To access the information, a doctor needs to have the patient's public key. The data is only visible to the healthcare provider when the key or unique ID matches [35–37].

2. Smoother transition of patients among Care providers
   Individual patients might utilize the same information on the blockchain to easily unlock and share their health data with other doctors or organizations by sharing a shareable private key. This could help with health information technology (HIT) interoperability and collaboration among different users.
3. Electronic medical records (EMRs)
   The blockchain may offer a single transaction layer where organizations may submit and share data through one safe channel by retaining a specified set of standardized data on the chain, along with private encrypted connections to separately stored information such as radiographic or other images. Smart contracts and uniform authorization standards can substantially assist in providing seamless communication.
4. Data Safety
   The blockchain's security features can significantly increase the security of health data. Each person has a public identifier or key as well as a private key that can only be used when and for the amount of time specified.

   In addition, the obligation to target each user individually in order to obtain important information would put a stop to hacking. As a result, blockchains can supply medical data with an immutable audit trail.
5. Review of medical personnel
   Blockchain technology can be used to trace the origin of a medical commodity, it can also be used to track the experience of medical practitioners. The following are some of the most significant benefits of the blockchain system:

   - Healthcare organizations will be able to get accredited more rapidly during the hiring process.
   - Transparency and assurance for partners, such as organizations who subcontract locum tenens or innovative virtual healthcare models that educate patients about the experience of medical professionals [38].

## 3.4 Limitations and Challenges in the Adoption of Blockchain in Healthcare

Although blockchain technology has many uses in the healthcare sector, there are still several hurdles for newcomers to the sector to clear before implementing it. When attempting to integrate blockchain technology into their system, the stakeholders in the healthcare sector may encounter the following significant difficulties:

- Many people may find it challenging to comprehend blockchain technology. They need more blockchain app examples to fully grasp the technology, hence they are not yet ready for production. The processing of records is not streamlined for healthcare providers or insurance companies. Getting them to adopt blockchain technology would be challenging without such a mechanism.
- Having a large storage capacity is necessary for the size of health and medical data. The use of blockchain technology in healthcare IT solutions can occasionally be impractical for small or medium-sized organizations.
- Blockchain does not have any standardized procedures for dealing with data ownership concerns. Companies are still unsure of how healthcare blockchain rules will work with privacy laws.
- Having a good infrastructure, being interconnected, and having specialists are necessary for implementing blockchain technology. These could represent technical obstacles to success in the healthcare sector. Blockchain in the healthcare industry sounds really intriguing, but it hasn't been used anyplace to great success yet. Since the healthcare sector is so delicate and fragmented, changes cannot be made quickly. Therefore, difficult to say whether the adoption of this technology will be practical until this technology is successfully pilot operated by stakeholders in the industry [6, 39].

## 3.5 Conclusion

Blockchain technology has brought about numerous ground-breaking advancements in the healthcare sector. In this industry, blockchain has many benefits, which are applicable to solving different issues such as record sharing and security. By fostering the creation of more patient-centric solutions, it promises to hasten a beneficial shift. With blockchain technology's capacity to stop ineffective healthcare practices, ongoing data breaches, and rising hospital expenses, healthcare providers, and patients should continue to benefit from its advantages.

The healthcare industry has not fully addressed issues such as mining incentives that are a fundamental part of blockchain technology, as well as specific blockchain attacks that can disrupt the entire process. Therefore, in the future such issues will be addressed.

## References

1. Tandon, R., Verma, A., Gupta, P.K.: Blockchain enabled vehicular networks: a review. In: 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), pp. 1–6. IEEE (2022, November)
2. Kaur, V., Gupta, K., Baggan, V., Kaur, G.: Role of Cryptographic Algorithms in Mobile Ad Hoc Network Security: An Elucidation (2019)
3. Kaur, V., Kaur, G., Dhiman, G., Bindal, R., Mishra, M.K.: Adaptability of machine learning in cryptography. Solid State Technol. **63**(4), 2874–2880 (2020)
4. Gupta, S., Saluja, K., Goyal, A., Vajpayee, A., Tiwari, V.: Comparing the performance of machine learning algorithms using estimated accuracy. Meas. Sens. **24**, 100432 (2022)
5. Yadav, K., Gupta, S., Gupta, N., Gupta, S.L., Khandelwal, G.: Hybridization of K-means clustering using different distance function to find the distance among dataset. In: International Conference on Information and Communication Technology for Intelligent Systems, pp. 305–314. Springer, Singapore (2020)
6. Saluja, K., Bansal, A., Vajpaye, A., Gupta, S., Anand, A.: Efficient bag of deep visual words based features to classify CRC images for colorectal tumor diagnosis. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1814–1818. IEEE (2022)
7. Kaur, G., Kaur, H.: Black spot and accidental attributes identification on state highways and ordinary district roads using data mining techniques. Int. J. Adv. Res. Comput. Sci. **8**(5) (2017)
8. Kaur, G., Kaur, H.: Prediction of the cause of accident and accident prone location on roads using data mining techniques. In: 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2017)
9. Bindal, R., Sarangi, P.K., Kaur, G., Dhiman, G.: An Approach for Automatic Recognition System for Indian Vehicles Numbers Using k-Nearest Neighbours and Decision Tree Classifier (2019)
10. Kaur, G., Goyal, S., Kaur, H.: Brief Review of Various Machine Learning Algorithms. In: Proceedings of the International Conference on Innovative Computing and Communication (ICICC) (2021)
11. Rai, B.K.: Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records. Int. J. Reliab. Qual. E-Healthc. (IJRQEH) **11**(1), 1–15 (2022)
12. Rai, B.K.: Ephemeral pseudonym based de-identification system to reduce impact of inference attacks in healthcare information system. Health Serv. Outcomes Res. Methodol. 1–19 (2022)
13. Rai, B.K.: PcBEHR: patient-controlled blockchain enabled electronic health records for healthcare 4.0. Health Serv. Outcomes Res. Methodol. 1–23 (2022)
14. Rai, B.K., Tyagi, A., Arora, B., Sharma, S.: Blockchain based Electronic Healthcare Record (EHR). In: ICCCE 2021 Lecture Notes in Electrical Engineering, vol. 828. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7985-8_19
15. Yaqoob, I., Salah, K., Jayaraman, R., Al-Hammadi, Y.: Blockchain for healthcare data management: opportunities, challenges, and future recommendations. Neural Comput. Appl. 1–16 (2021)
16. Farouk, A., Alahmadi, A., Ghose, S., Mashatan, A.: Blockchain platform for industrial healthcare: Vision and future opportunities. Comput. Commun. **154**, 223–235 (2020)
17. McGhin, T., Choo, K.K.R., Liu, C.Z., He, D.: Blockchain in healthcare applications: research challenges and opportunities. J. Netw. Comput. Appl. **135**, 62–75 (2019)
18. Hasselgren, A., Kralevska, K., Glioroski, D., Pedersen, S.A., Faxvaag, A.: Blockchain in healthcare and health sciences—a scoping review. Int. J. Med. Informatics **134**, 104040 (2020)
19. Hölbl, M., Kompara, M., Kamišalić & Nemec Zlatolas, L.: A systematic review of the use of blockchain in healthcare. Symmetry **10**(10), 470 (2018)
20. Chang, S.E., Chen, Y.: Blokchain in health care innovation: literature review and case study from a business ecosystem perspective. J. Med. Internet Res. **22**(8), e19480 (2020)

21. Agbo, C.C., Mahmoud, Q.H., Eklund, J.M.: Blockchain technology in healthcare: a systematic review. In: Healthcare, vol. 7, No. 2, p. 56. Multidisciplinary Digital Publishing Institute (20190
22. Ismail, L., Materwala, H., Zeadally, S.: Lightweight blockchain for healthcare. IEEE Access **7**, 149935–149951 (2019). https://doi.org/10.1109/ACCESS.2019.2947613
23. Bell, L., Buchanan, W.J., Cameron, J., Lo, O.: Applications of blockchain within healthcare. Blockchain Healthc. Today 1 (2018)
24. Ekblaw, A., Azaria, A., Halamka, J.D., Lippman, A.: A case study for blockchain in healthcare:"MedRec" prototype for electronic health records and medical research data. In: Proceedings of IEEE Open and Big Data Conference, vol. 13, p. 13 (2016)
25. Prokofieva, M., Miah, S.J.: Blockchain in healthcare. Australas. J. Inf. Syst. 23 (2019)
26. Khezr, S., Moniruzzaman, M., Yassine, A., Benlamri, R.: Blockchain technology in healthcare: a comprehensive review and directions for future research. Appl. Sci. **9**(9), 1736 (2019)
27. Chen, H.S., Jarrell, J.T., Carpenter, K.A., Cohen, D.S., Huang, X.: Blockchain in healthcare: a patient-centered model. Biomed. J. Sci. Tech. Res. **20**(3), 15017–15022 (2019)
28. Tandon, A., Dhir, A., Islam, A.N., Mäntymäki, M.: Blockchain in healthcare: a systematic literature review, synthesizing framework and future research agenda. Comput. Ind. **122**, 103290 (2020)
29. Siyal, A.A., Junejo, A.Z., Zawish, M., Ahmed, K., Khalil, A., Soursou, G.: Applications of blockchain technology in medicine and healthcare: challenges and future perspectives. Cryptography **3**(1), 3 (2019)
30. Hussien, H.M., Yasin, S.M., Udzir, S.N.I., Zaidan, A.A., Zaidan, B.B.: A systematic review for enabling of develop a blockchain technology in healthcare application: taxonomy, substantially analysis, motivations, challenges, recommendations and future direction. J. Med. Syst. **43**(10), 1–35 (2019)
31. Abu-Elezz, I., Hassan, A., Nazeemudeen, A., Househ, M., Abd-Alrazaq, A.: The benefits and threats of blockchain technology in healthcare: a scoping review. Int. J. Med. Informatics **142**, 104246 (2020)
32. Zhang, P., White, J., Schmidt, D. C., Lenz, G.: Applying Software Patterns to Address Interoperability in Blockchain-Based Healthcare Apps (2017). arXiv:1706.03700
33. Kuo, T.T., Zavaleta Rojas, H., Ohno-Machado, L.: Comparison of blockchain platforms: a systematic review and healthcare examples. J. Am. Med. Inform. Assoc. **26**(5), 462–478 (2019)
34. Tanwar, S., Parekh, K., Evans, R.: Blockchain-based electronic healthcare record system for healthcare 4.0 applications. J. Inf. Secur. Appl. **50**, 102407 (2020)
35. Rai, B.K., Srivastava, A.K.: Security and privacy issues in healthcare information system. Int. J. Emerg. Trends Technol. Comput. Sci. **3**(6), 248–252 (2014)
36. Wadhwa, S., Babbar, H., Rani, S.: A survey on emerging software-defined networking and blockchain in smart health care. In: IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012056. IOP Publishing (2021)
37. Wadhwa, S., Rani, S., Verma, S., Shafi, J., Wozniak, M.: Energy efficient consensus approach of blockchain for IoT networks with edge computing. Sensors **22**(10), 3733 (2022)
38. Ben Fekih, R., Lahami, M.: Application of blockchain technology in healthcare: a comprehensive study. In: International Conference on Smart Homes and Health Telematics, pp. 268–276. Springer, Cham (2020)
39. Kumar, T., Ramani, V., Ahmad, I., Braeken, A., Harjula, E., Ylianttila, M.: Blockchain utilization in healthcare: key requirements and challenges. In: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–7. IEEE (2018)

# Chapter 4
# Securing Healthcare Records Using Blockchain: Applications and Challenges

**Inderpreet Kaur, Vanshit Gupta, Vansh Verma, and Supreet Kaur**

## 4.1 Introduction

Since its inception in 2008, the blockchain has transformed the way we think about problems. Blockchain is a decentralised ledger, a record book which holds the detail of every transaction that takes place on the network, tracking digital assets on network. It has not only changed the financial industry, but it has also shown to be a godsend for secure, efficient, and transparent peer-to-peer (P2P) information transmission. It has solved the key issues like trust in a network which allows any organisation to focus on the issues at hand. With the huge demand of the Internet and related technologies, a variety of Industry 4.0-based applications have been implemented around the world, in which sensors and actuators sense, calculate, and transfer data for industry automation via an open channel named internet. So it is obvious that security and privacy threats have been increased. Due to this, it is essential to consider issues like data integrity, data redundancy, and data heterogeneity [1].

Please note that the LNCS Editorial assumes that all authors have used the Western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

I. Kaur (✉) · V. Gupta · V. Verma · S. Kaur
Institute of Engineering and Technology, Chitkata University, Rajpura, Punjab, India
e-mail: kaur.inderpreet@chitkara.edu.in
URL: https://www.chitkara.edu.in/

V. Gupta
e-mail: vanshit2485.be20@chitkara.edu.in

V. Verma
e-mail: vansh2482.be20@chitkara.edu.in

S. Kaur
e-mail: supreet2464.be20@chitkara.edu.in

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_4

This technology has come up with better transparency and has enhanced the security system of various sectors.

Today, in the ever-advancing technological era, the healthcare system has progressed significantly. Increased the usage of Electronic Health Records systems has resulted in previously inconceivable levels of healthcare data breaches [2]. Earlier patients were worried about the safety of their medical information. Due to these worries, some people have gone so far as to refuse to share vital personal health data with their healthcare providers [3, 4].

The lack of an efficient, smart, and secure solution to address the problem of securely sharing data among healthcare professionals, patients, and medical researchers is the main roadblock to the widespread adoption of these Electronic Health Records. So, blockchain is a state-of-the-art digital ledger that can be bespoke to record nearly anything of value in addition to financial transactions. The digital data that is stored on a blockchain exists as a shared database that is continually logged.

In late 2019, a coronavirus (COVID-19) outbreak produced a global health emergency [5]. It has a sizable and global impact on healthcare, which makes it easier to embrace digital technologies to meet a variety of needs in the healthcare sector. Despite the fact that the COVID-19 bubonic plague has amply demonstrated the necessity for secure, decentralised, multipurpose platforms for orchestrating the large-scale transfer of sensitive information, such as contact tracking, vaccination status monitoring, and the issuance of COVID-19 health certificates, numerous organisational, technological, and governance obstacles may still prevent ubiquitous implementation of blockchain technology in the healthcare industry. These requirements create a powerful push for a deliberate effort to increase blockchain usage and remove some of the hurdles to widespread deployment [6], Rai et al. [4]. Blockchain in healthcare is all about removing the middleman. It has the potential to improve clinical trial data management by lowering regulatory approval delays and streamlining communication between various supply chain players.

When it came to data breaches, the healthcare industry was the most victimised industry [7]. As reported by several practitioners, between 2005 and 2019, 249.09 million people were impacted by healthcare data breaches. The Health Insurance Portability and Accountability Act (HIPAA) states that at least one data breach of healthcare records occurred every day in 2018. Between 2009 and 2018, data breaches exposed the healthcare records of more than 59 per cent of the US population. In terms of healthcare data breaches, 2015 was the worst year ever with more than 113.27 million records exposed, stolen, or improperly shared [8] (Fig. 4.1).

According to an IBM report titled ‘Healthcare rallies for Blockchain’, 16% of healthcare executives plan to deploy blockchain technology in 2017, with 56% pledging to do so by 2020. The creation of a unified database of healthcare data is on the agenda of world leaders. Doctors and providers can use this information across numerous Electronic Health Records (EHRs), and the data is more secure [9].

Internet of things (IoT) breakthroughs have undoubtedly resulted in continuous advancements in the healthcare business. However the fact that EHR/EMR data is scattered across numerous medical facilities has made it extremely difficult to

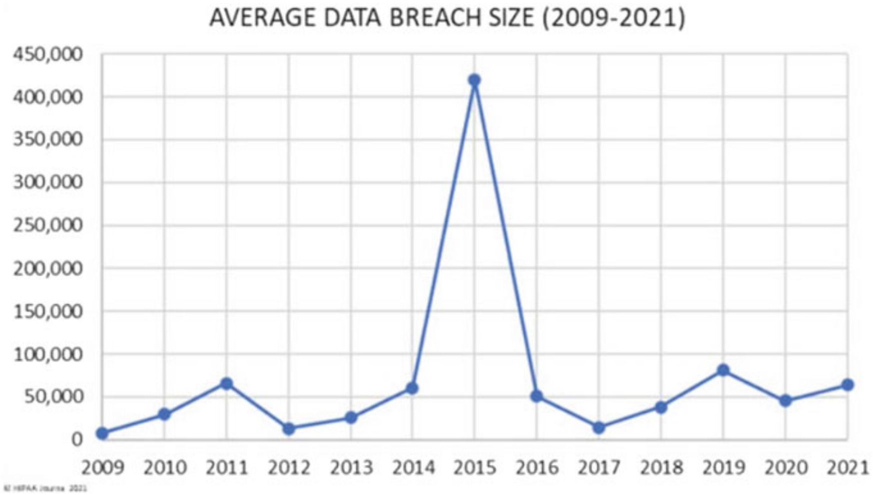


**Fig. 4.1** Data showing average data breaches from the year 2009–2021 [9]

process EHR/EMR in a secure manner. In light of these realities, blockchain has gained traction as a viable solution to such problems [10–12].

## 4.2 Why Blockchain?

Everyone is jumping the hype train. Many IT providers saw the potential lying in Blockchain technology and want their piece of the pie. This is so because of the below-mentioned features [13, 14].

1. Decentralisation

Decentralisation technology allows you to retain your assets in a network without relying on the oversight and control of a single person, organisation, or institution. As the system runs on algorithms, there is no chance for people to scam you out of anything. No one can utilise blockchain for their personal gains. The majority of contemporary healthcare organisations and facilities are largely built on centralised systems, which means that a single body wields far too much power.

2. Transparency

Transparency of data in healthcare industry can assist in the creation of a complete, accurate, and auditable ledger of transactions. The solutions for the management of healthcare data are now unable to simultaneously assure transparency, privacy, and security. Additionally to ensuring greater transparency, blockchain ensures the security of data and enables authorised authority over healthcare data and it can be achieve through encryption and control mechanisms.

3. Immutability

This striking feature of blockchain refers to the potential of a blockchain ledger to remain unaltered and untamperable. It has the power to remodel and alter the auditing process into one that is quick, efficient, and cost-effective. It makes almost impossible for any user on the network to modify, remove, or update the data. This feature can be achieved through cryptographic hashing.

4. Data Provenance

For the healthcare industry, data provenance is crucial to increase public trust in it by providing all pertinent information about how health data is created, accessed, and transferred. It is ensured by the ability to track data changes from its inception to its current state. Blockchain employs a timestamping procedure that entails evaluating the hash values of the provenance record, which is then forwarded to consensus nodes to verify a reliable ledger of all legal transactions is kept.

5. Minting

In essence, there are numerous ways to create a manipulation problem that blockchain can solve. Banks and international tech firms such as Google and Meta create a sense of dependability and accountability to people and corporations in the West. Although mining remains the most prevalent strategy, the prospects for blockchain are now larger in countries that have not yet reached a level.

6. Anonymity and Programmability

Anonymity and programmability are some distinctive attributes of blockchain. Anonymity assures that the individuality of senders and beneficiaries in transactions stay anonymous whereas smart contracts with programmability enable new regulations and transactions being automated. Self-executing programmes known as ‘smart contracts’ are built on agreements between purchasers and traders.

7. Distributed Ledger and Consensus

Through the integration of fundamental technology components like distributed ledgers and consensus techniques, blockchain can offer a variety of advantages. Consensus is a decision-making procedure for the network’s active nodes as when millions of nodes are validating the transactions simultaneously, that time this algorithm helps the system to run smoothly whereas Distributed Ledger System (DLT) makes the process visible and dependable by allowing anyone with the necessary access to observe the ledger (Fig. 4.2).

Blockchain networks have evolved into a formidable combination of security, anonymity, and openness since their inception. The potential for blockchain-based security systems is essentially endless. In a short period of time, technology has irreversibly changed the shape of the global economy, becoming indispensable for many banks, huge enterprises, and even governments.

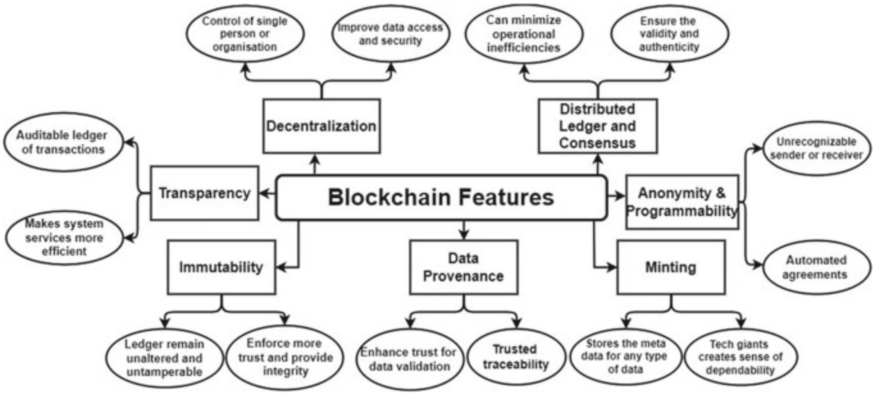

Blockchain Features
Decentralization
Control of single person or organisation
Improve data access and security
Distributed Ledger and Consensus
Can minimize operational inefficiencies
Ensure the validity and authenticity
Transparency
Auditable ledger of transactions
Makes system services more efficient
Anonymity & Programmability
Unrecognizable sender or receiver
Automated agreements
Immutability
Ledger remain unaltered and untamperable
Enforce more trust and provide integrity
Data Provenance
Enhance trust for data validation
Trusted traceability
Minting
Stores the meta data for any type of data
Tech giants creates sense of dependability

**Fig. 4.2** Features of blockchain

## 4.3 Applications

1. Electronic Health Record

An electronic health record is a computerised representation of a patient's medical file. An individual's health records, prescription details, therapy plans, examination details, laboratory results, test results, and other information are all stored in the Electronic Health Record. Data is dispersed among Electronic Health Record systems, and patients frequently contact with many healthcare providers, making it difficult to access previous records. Furthermore, multiple Electronic Health Record systems are used by different healthcare providers, and some of them are not entirely compatible. These factors make data sharing more challenging [14, 15].

With the use of blockchain technology, patients have full control over how their health records are shared. The blockchain's decentralised, self-trusted, and tamper-resistant structure makes sure that medical data is distributed and stored securely, while also drastically reducing the time it takes to share data and total costs. Patients will also focus more on their own healthcare because they can engage in their own health records.

2. Clinical trial

Recruiting patients for clinical trials is one of the most difficult aspects of any clinical study. There are various challenges that must be addressed from the sponsor's, patient's, and main investigator's viewpoints, resulting in the failure of the most of clinical studies to meet their recruiting targets on time. Conducting clinical trials with insufficient participants reduces the power of definitive conclusions or leads to early trial termination. Blockchain technology can play a crucial role in the future clinical trials as they have a complete set of medical history of the participant, allowing the clinical researcher to make an informed judgement about the trial [14, 15]. From clinical trial analysis results to patients' medical data and genetic information blueprints, the blockchain provides a safe and secure platform for storing and processing all forms of vital information [15–17].

3. Pharmaceutical supply chain

From acquiring active medicine ingredients to manufacturing the final product that is distributed and delivered to patients, the pharmaceutical supply chain follows a process from beginning to conclusion. The distribution of authentic and high-quality products is the fundamental duty of the members of supply chain as soon as possible because it has a direct impact on patients' health and safety. The pharmaceutical supply chain involves numerous parties and product delivery of goods frequently necessitates a complex process, making it more challenging to trace a medicine.

Blockchain gained enormous acceptance as they enable non-trusting stakeholders to keep a record of transactions in an immutable and transparent manner. The key advantage of blockchain technology is its capacity to monitor asset transactions recorded on a distributed ledger that is decentralised and encrypted with time stamps, enable transaction records to transfer and stored directly and digitally without the help of intermediate service firms [17, 18]. Several drug-tracking activities and processes are made possible by blockchain technology, ensuring accurate provenance, tracing, and tracking.

4. Remote patient monitoring

It is the process of gathering medical information from patients via IoT, body monitors, and mobile devices. Blockchain is useful for storing, distributing, and retrieving biomedical data collected remotely. The COVID-19 pandemic is driving acceptance of telemedicine and telecare technology because it allows for safe connection with doctors and other health professionals across digital channels, reducing infection spread. The data in today's telehealth and telemedicine systems is potentially exposed to several outer and inner data violations, threatening the services' dependability and accessibility [15–17].

Blockchain technology can assist in resolving such critical issues. The new blockchain technology uses a decentralised design to maintain a common database of patient records among multiple participants, with copies of each ledger being validated by and coordinated with each node of blockchain. The following are some of the key challenges that blockchain technology can address: monitoring pharmaceuticals and medical test kits through the supply chain, checking the qualifications of doctors, securing remote patient–doctor consultation records, tracking the locations visited by sick patients, tracking the provenance of defective Kits for medical tests, and more.

The integration of blockchain technology into existing telehealth and telemedicine systems can provide a number of benefits for secure healthcare digitisation, comprising the capability to verify the validity of users requesting patient data, the management of device IDs used for monitoring a patient digitally, the protection of patient confidentiality, and the automated settling of payments.

5. Cost saving

The cost of interacting with a doctor and the price of medicinal research and development can be used to calculate cost savings. For verification, a medical credentialing application must be sent to numerous organisations. As a result, it can be expensive and time-consuming. By using blockchain technology, hospitals, pharmaceutical businesses, and insurance organisations might save money. One of the key advantages of blockchain systems and electronic medical records is that doctors and patients have better access to medical records, which improves nursing efficiency and quality.

When it comes to patient expenditures and prescription administration, it can be helpful to give ledgers to various organisations, such as the insurance firms and dispensaries. Providing pharmacists with reliable, up-to-date prescription data will enhance logistics, particularly in the context of chronic disease care. Multi-site clinical trials at drug research facilities can save money on trials, and data management solutions based on blockchain intelligent contract technology can reduce the cost of operating multi-site trials in medicine.

6. Health insurance

Between payers, providers, and patients, there is a lack of confidence. Patients usually pay expensive rates while dealing with an absence of transparency and the inability to compare prices, as well as the potential of insurance fraud, which affects all parties involved. Due to its Practical Byzantine Fault Tolerance and multi-node co-maintenance characteristics, the blockchain can guarantee a high level of stability for consumers. The use of blockchain technology in medical insurance can minimise the complexity and expense of medical insurance, protect patients' rights, and reduce hospital uncollected funds and insurance company management costs.

## 4.4 Limitations

Despite all the advanced features blockchain provides, it still has a variety of limitations and issues that need to be addressed. Lack of standardisation, energy consumption, privacy leakage, scalability, cost, and supporting software vulnerabilities are among the specific issues to be examined in this study [18, 19].

1. Scalability

Communication gaps and information-sharing issues are major roadblocks to healthcare innovation and patient care quality. As users upload data, blockchain expands, recording all of the hashes connected with the newly added data in this scenario. The network may have fewer nodes with enough computer power to analyse and validate blockchain data as a result of the increasing storage and processing requirements. The trade-off between computer capacity and the number of medical transactions may limit the scalability of healthcare systems.

2. Insufficient standardisation

The lack of standards hinders widespread adoption and inhibits progress because this technology is still relatively young and in its infancy. [1]. It is possible to keep patient data, medical histories, and other organised data on the blockchain itself.

Any type of comprehensive medical documentation, including X-ray, scans, and unstructured doctor notes, may be stored outside of the blockchain to address scalability issues. Users may desire some kind of standardisation to manage and access such a wide variety of data dispersed around the firm. There must be a high degree of standardisation among the numerous parties concerned to enable all of these different infrastructures and applications. The issue of standardisation and criteria will become more important as more countries adopt blockchain as a solution.

3. Security and Privacy

Blockchain technology emphasises transparency, which may not be desirable in the health arena in some situations. Blockchains tackle the problem of transactions requiring trusted third parties, which exposes users to security and privacy risks. Although blockchains are being used to create smart contracts between healthcare providers to enable each other access to certain data or patient records, the question of who is accessing the data and whether they are authorised to do so remains.

Another significant issue that could imperil PHI and EMR is the nature of blockchain deployment, which does not guarantee the confidentiality of data stored or transferred off-chain.

4. Cost of operating blockchain

The development and operation of blockchain-based healthcare systems may be costly. For all engaged stakeholders, the government and healthcare industry still need to identify the many types of development, operations, and total deployment costs. As a result, finding the best strategies to lower the overall cost and resources required to develop such systems is critical.

5. Attacks/Vulnerabilities

Additionally, because of how the system is set up and built, blockchain technology has a few unique flaws. 51 per cent attacks, double spending attacks, selfish mining attacks, eclipse attacks, block discarding attacks, difficulty rising assaults, and anonymity issues in blockchain are all examples of blockchain vulnerabilities.

One of the simplest attacks that may be made on the blockchain is a 51% attack since it makes use of the consensus algorithm's legal purpose. An attacker has control of a PoW blockchain if they have 51% of the computational resources. They can launch double-spend attacks once they control the blockchain.

Blockchain is also susceptible to a number of common software flaws that enable for harmful attacks. Theft of data and identity are just two examples of other crimes that can be aided by these destructive acts. When a user's private key is stolen on the blockchain, identity theft happens since the criminal now has access to anything the victim has ever broadcast on the network.

6. Adoption and Incentive for Participation

In order to supply the required processing capacity for both the production of cryptocurrencies and transaction blocks, blockchain technology requires the utilisation of a network of networked computers.

Through incentive systems, participants ought to be compensated for contributing processing power. Additionally, health organisations may require encouragement to use blockchain. According to the quantity of participants, blockchain's influence will increase.

## References

1. Bodkhe, U., Tanwar, S., Parekh, K., Khanpara, P., Tyagi, S., Kumar, N., Alazab, M.: Blockchain for industry 4.0: a comprehensive review. IEEE Access **8**, 79764–79800 (2020)
2. Jain, U.: Blockchain: future of the anesthesia record? ASA Monitor **83**(3), 32–33 (2019)
3. Patel, V., Hughes, P., Barker, W., Moon, L.: Trends in individuals' perceptions regarding privacy and security of medical records and exchange of health information: 2012–2014. ONC Data Brief (33) (2016)
4. Rai, B.K., Verma, R., Tiwari, S.: Using open source intelligence as a tool for reliable web searching. SN Comput. Sci. **2**(5), 1–12 (2021)
5. Gunasekeran, D.V., Tseng, R.M.W.W., Tham, Y.C., Wong, T.Y.: Applications of digital health for public health responses to covid-19: a systematic scoping review of artificial intelligence, telehealth and related technologies. NPJ Digit. Med. **4**(1), 1–6 (2021)
6. Ahad, M.A.: Coronavirus-a global emergency. Med. Today **32**(2), 138–142 (2020)
7. Seh, A., Zarour, M., Alenezi, M., Sarkar, A., Agrawal, A., Kumar, R., Khan, R.: Healthcare data breaches: insights and implications. Healthcare (2020)
8. Zubaydi, H.D., Chong, Y.W., Ko, K., Hanshi, S.M., Karuppayah, S.: A review on the role of blockchain technology in the healthcare domain. Electronics **8**(6), 679 (2019)
9. Journal, H.: Hipaa Journal (2020). https://www.hipaajournal.com/healthcaredata-breach-statistics/
10. Kumar Rai, B., Sharma, S., Kumar, A., Goyal, A.: Medical prescription and report analyzer. In: 2021 Thirteenth International Conference on Contemporary Computing (IC3–2021), pp. 286–295 (2021)
11. Rai, B.K.: Ephemeral pseudonym based de-identification system to reduce impact of inference attacks in healthcare information system. Health Serv. Outcomes Res. Methodol. 1–19 (2020a)
12. Rai, B.K.: Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records. Int. J. Reliab. Qual. E-Healthc. (IJRQEH) **11**(1), 1–15 (2022b)
13. Kumar, T., Ramani, V., Ahmad, I., Braeken, A., Harjula, E., Ylianttila, M.: Blockchain Utilization in Healthcare: Key Requirements and Challenges (2018). https://doi.org/10.1109/HealthCom.2018.8531136
14. McGhin, T., Choo, K.K.R., Liu, C.Z., He, D.: Blockchain in healthcare applications: research challenges and opportunities. J. Netw. Comput. Appl. **135**, 62–75 (2019)
15. Bell, L., Buchanan, W.J., Cameron, J., Lo, O.: Applications of blockchain within healthcare. Blockchain Healthc. Today 1 (2018). https://doi.org/10.30953/bhty.v1.8, https://blockchainhealthcaretoday.com/index.php/journal/article/view/8
16. Gao, W., Hatcher, W.G., Yu, W.: A survey of blockchain: Techniques, applications, and challenges. In: 2018 27th International Conference on Computer Communication and Networks (ICCCN), pp 1–11 (2018). https://doi.org/10.1109/ICCCN.2018.8487348
17. Siyal, A.A., Junejo, A.Z., Zawish, M., Ahmed, K., Khalil, A., Soursou, G.: Applications of blockchain technology in medicine and healthcare: challenges and future perspectives. Cryptography **3**(1) (2019). https://doi.org/10.3390/cryptography3010003, https://www.mdpi.com/2410-387X/3/1/3

18. Farouk, A., Alahmadi, A., Ghose, S., Mashatan, A.: Blockchain platform for industrial healthcare: vision and future opportunities. Comput. Commun. **154**, 223–235 (2020). https://doi.org/10.1016/j.comcom.2020.02.058, https://www.sciencedirect.com/science/article/pii/S014036641931953X
19. G¨okalp, E., G¨okalp, M., G¨okalp, S., Eren, P.: Analysing Opportunities and Challenges of Integrated Blockchain Technologies in Healthcare, pp. 174–183 (2018)

# Chapter 5
# Authentication Schemes for Healthcare Data Using Emerging Computing Technologies

**Akanksha Upadhyaya and Manoj Kumar Mishra**

## 5.1 Introduction

Expert systems that can simulate the intelligence of the human brain without the requirement for human input are used in AI. It is used to learn from previous experiences in order to solve complex problems [1, 2]. To handle the huge amounts of data generated by IoT devices on a regular basis, most IoT applications have relied on cloud computing for data processing and storage [3]. Multi-access edge computing, which provides compute and storage resources at the network edge for low-latency and real-time applications, stands out as a feasible alternative for enabling IoT systems [4]. Furthermore, the combination of MEC and IoT resulted in EoT, or Edge of Things, for smart healthcare applications. Blockchain, as a new paradigm, has been found to be more robust and secure in reducing the danger of user or patient data being compromised [5]. Because patient data is constantly monitored and sent to healthcare systems, protecting it from malicious activity has become one of the most significant topics of research. Cloud computing provides patient data tracking, health monitoring, drug inventory and management, and other services using developing technologies such as BEoT, EoT, and IoT [6]. Another technology which is gaining popularity is AIoT, where data transmission to the cloud servers was sent to edge computing for effective decision-making by securely and efficiently utilizing the resources [7]. One of the most difficult problems in new technology is authentication and security. These technologies have been given the ability to make judgments, and a breach could result in considerable damage. Additionally, in the healthcare industry, privacy is a top priority [8].

A. Upadhyaya
Rukmini Devi Institute of Advanced Studies, New Delhi, India

M. K. Mishra (✉)
OP Jindal University, Raigarh, Chhattishgarh, India
e-mail: manoj.mishra@opju.ac.in

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_5

The paper aims to conduct an extensive review of the recent studies that have incorporated applications of emerging technologies for data authentication. The paper has been organized into three major sections; Sect. 5.2 investigates past studies concerning data authentication using applications of emerging technologies specifically in the healthcare sector. Section 5.3 is all about the validation of review using Word cloud analysis, where researchers aim to validate the relevance of the literature with the current study. At last, the Conclusion section summarizes the present study.

## 5.2 Related Work

Edge computing is an architecture that effectively deals with a variety of problems concerning processing and calculation [7]. Edge computing has been one of the technologies that provides effective data services with short delay by accelerating the transmission speed along with the computational power of IoT-enabled healthcare systems. The authors proposed a framework that uses a lightweight authentication scheme to authenticate IoT devices using the Edge servers in a Software Defined Networking-based Edge computing system for IoT-based health care. The authentication-enabled device collects and transmits the patient's data to Edge servers for further processing by first storing it. Using computer-based simulations, the proposed framework was evaluated for the effectiveness of IoT-driven healthcare systems. The simulation was performed on different parameters, and the authors first built an experimental setup and simulation scenario followed by performance evaluation metrics and at last analysis of simulation results. The performance evaluation metrics include the calculation of "Average response time", "Packet delivery ratio", "Average delay", "Throughput"', and "Control head" for the proposed scheme. "Average response time" concerning upload and download time, waiting time, and system load was calculated and compared for Edge and Cloud servers. The table shows a comparison of Edge and Cloud servers w.r.t. "Average response time".

| Average response time w.r.t | Edge servers | Cloud servers |
|---|---|---|
| Upload and download time | Half than cloud servers | Double than edge servers |
| Waiting time | Much less than cloud servers | Much higher than edge servers |
| System load | Much lower than coud servers | Much higher than edge servers |

Other parameters of performance evaluation metrics, i.e. "Packet delivery ratio", "Average delay", "Throughput", and "Control head" were performed for Traditional network, Edge computing network, and Software Defined Networking-based Edge computing networks. The comparison is shown in the table below.

| Performance evaluation metrics | Traditional network (low-powered IoT devices) | "Edge computing" network | "SDN-based Edge computing" |
|---|---|---|---|
| Packet delivery ratio | Goes down quickly | Better than traditional and shows improved ratio | Provides best result, takes intelligent decisions |
| Average delay | In this case, it is higher because of slow processing speed of "IoT devices" | Lower in comparison to traditional network (due to edge server collaboration) | Very low; because for network optimization, resource utilization, and load balancing it provides agility to edge servers |
| Throughput | Less effective in comparison to other two | Higher output than traditional | It is High in this case because with the use of SDN-based Edge computing it makes intelligent decisions about effective utilization of network resources, edge collaboration, and load balancing |
| Control overhead | Lower due to less number of control message exchange | Higher control overhead | Relatively higher control overhead in comparison to traditional and edge-based computing networks |

The authors suggested that based on simulation results for 3 different scenarios, it has been verified that the proposed scheme is efficient, and they further suggested that the proposed scheme can be enhanced to protect the data privacy of the patients.

Edge-based AI is one of the welcome additions to Cloud-based AI [8]. It eliminates the privacy problem of data streaming and data storage in the cloud. It allows real-time applications to be done where each millisecond is significant and offers AI capabilities to remote places with inadequate networking infrastructures by bringing intelligence to the Edge-enabled devices. The authors also mentioned that Edge-enabled AI applications like "self-driving cars" and "intelligent health care" have security as one of the major concerns. Attacking and hacking edge devices can cause substantial damage because they can interpret data and take action. In spite of that, the widespread use of computationally restricted devices in edge contexts, as well as the escalation of leakage attacks, provides significant security challenges. As a result, the authors have proposed an improved "Edge AI" security by creating and developing lightweight and "Leakage-Resistant Authenticated Key Exchange (LRAKE)" protocols. The protocol suggested in the proposed scheme can be easily integrated into several mainstreaming security and communication standards. For the proposed protocol, the research also offers prototypes and implementation details as well as a use case using Bluetooth 5.0. The authors recommended that in the future the proposed theoretical and implementation details will aid in the deployment of the LRAKE protocols in Edge AI applications.

The Edge of Things, which is enabled by the convergence of "Edge Computing" as well as the "Internet of Things" [9], is one of the technological areas where blockchain is gaining traction as a viable and revolutionary technological solution. The integration of blockchain and the IoT created a new paradigm known as Blockchain-enabled Edge of Things also known as BEOT, which is critical for allowing potential short delay and efficient security systems and services. The authors have proposed a revolutionary BEOT architecture for enabling industrial applications in a variety of IoT use cases, including smart homes, smart health care, smart grid, and smart transportation, monitored by the blockchain on the edge network. The authors also investigated the possibilities of BEOT in delivering security services, such as user authorization, confidentiality, vulnerability detection, and information security in the proposed research. The authors created aBEOT, a unique technology driven by MEC and blockchain that enables short delays in IoT-enabled applications along with the delivery of excellent security at the edge network.

In this research [10], a fall detection approach with improved accuracy was suggested, to integrate the data analysis as well as for estimating the fall detection using feature detection; three stages of data processing—convolution layer, pooling layer, and overall layer—have been implemented. The vision-based estimate was shown to be 98% accurate in the trial. The study uses a data collection of diverse photos to estimate the accuracy of fall detection using SVM and ANN. The CNN method improves classification and accuracy rate of detection by using IoT sensor data and data storage via fog-based retrieval [7]. The inclusion of Wi-Fi utilizing LoRA also includes the precision of detection in rural regions. The data analysis was conducted using an edge gateway to improve the sent data, and, as a result, the system's prediction performance has progressively increased. The raw data information received from the study of analyzing online data sequencing was used to report fall detection for clinical observation [7].

Smart systems and services have been widely used to help to develop urban populations with social-online behaviors [11], Flexi economies, industrial automation, and contemporary lifestyle, among several other things. The research also emphasized that smart system services must have a complex collection of attributes to meet the desired goals including security, simplicity of use, and user-friendliness along with management, adaptability, flexibility, decision-making, and customization. AI has recently been recognized as a technology-driven data and capable of providing effective information representation, cognitive and semantic modeling, and intelligent behavior support.

An edge computing environment paired with 5G speed and current computing methods is one of the solutions for excelling the performance and enhancing energy requirements for real-time capture and analysis of data concerning health care [12]. Past surveys on health care were more concentrated on emerging "fog architecture" and sensors, leaving out the issue of optimum computing techniques employed in on-device deployment in "Edge computing" architecture, such as classification, encryption, and authentication. The primary goal of the research study was to evaluate present and upcoming "Edge computing" architectures and approaches for healthcare applications, as well as to define the device needs and problems for distinct use cases. The health data was categorized to monitor physiological parameters and fall detection, and this was considered as the primary emphasis of the edge computing application in the proposed scheme. Other low-latency apps track particular symptoms for disorders like Parkinson's disease patients' gait problems. The current research also examines edge computing data activities such as transmission, security, verification, categorization, compression, and forecasting in depth.

"Edge AI"' and "IoMT" are sophisticated concepts that are now being employed in several aspects of health care in Smart Cities [13]. The applications of such technologies include tracking and managing networked health care, and they may also be highly useful because they reduce the number of human efforts and boost management efficiency. A review-based study was conducted by the researchers to explore the difficulties, and possibilities, of Edge AI for network-based health care in Smart Cities. The authors have methodically examined several kinds of research and separated them into two areas. Following an analysis of the studies, it was discovered that with the increase in the spread of disease and population, it is becoming crucial for healthcare professionals to perform data management, acknowledge the problems or issues, for emergency case management, patient history management, and diagnosis and treat diseases that react differently in patients. "Edge AI"', "IoMT", "5G", "Fog computing", and "Cloud computing", on the other hand, have been highlighted as ways to solve these growing issues. Nonetheless, a few areas have applied the most recent innovations, and the results have improved. The researchers also mentioned that these implementations had positive outcomes not just from the patient's point of view but also from the point of view of healthcare professionals. Furthermore, it is highlighted through an extensive review that the various models that have been suggested in several kinds of research must be verified further and applied in other domains to validate their efficacy and ensure that they can be implemented effectively in several areas.

Recent breakthroughs in ICT have offered intelligent approaches to several areas of life [14]. Smart apps and devices have become an indispensable part of each life, yet their use has resulted in a slew of long-term health difficulties in a modern context. Diabetes mellitus is one of the prominent health problems that affects people of all ages. Using RASGD, the proposed work attempts to improve an AI-based intelligent system for early disease prediction. The suggested work uses weight decay methods, such as "Absolute Shrinkage and Selection Operator" and "Ridge Regression"' methods, so as to improve the classification model. The RASGD uses a dissolute optimized model to reduce the classifier's cost function. The ASGD Classifier was also used with ridge regression to improve the classifier's convergence speed. Finally, the suggested scheme's findings were compared to the latest machine learning algorithms, namely Regression and SVM in order to validate the usefulness of the intelligent and agile system. The intelligent system showed a 92% accuracy rate, which is higher than the other classifiers tested.

These days, the "Internet of Things" is one of the most renowned ICT technologies [15]. Creating a secure and trustworthy authentication mechanism for IoT-based infrastructures, on the other hand, remains a difficulty Zargar et al. [16]. Another research demonstrated that [17, 18] schemes are vulnerable to password guessing attacks in off-line mode, user tracking attacks, and other types of attacks [19]. Based on the claimed vulnerability, a lightweight IoT authentication system was proposed. Additionally, the authentication scheme based on IoT withstands many types of threats and provides essential security features including mutual authentication, user audit, and session-based security. Moreover, when the scheme was evaluated, they discovered flaws. As a result, the researchers proposed an improved scheme based on their technique, which met the security requirements while also being resistant to well-known attacks. During the authentication phase, anonymity and authentication cannot be ensured. Then, to compensate for the Zhou et al. system [19], the researchers devised a new certification scheme. The suggested approach is resistant to typical attacks and includes features like "user anonymity" and "mutual authentication". In the initial step of the authentication phase, the researchers also added a new parameter that can identify whether the identity and password inputted by the user are correct at an early stage. Improved IoT-based cloud computing authentication was also proposed, with performance evaluation findings indicating that the approach has a significant level of computation and strong security. As a result, the proposed authentication system can be used with real-world IoT devices.

Greater quantity of healthcare information has been gradually started storing and communicating so as to reduce expenses and improve healthcare facilities [20]. The rise of e-Health clouds in this context opens up new possibilities for accessing medical data located in some remote places. However, this accomplishment brings with it a slew of additional dangers and concerns, such as ensuring the integrity, security, and confidentiality of extremely vulnerable healthcare data. Authenticating data is a crucial concern among these issues, ensuring that vulnerable medical data in clouds is not accessible to unauthorized users. Smart cards, passwords, and biometrics are three types of authentication that meet the need for high security. On e-Health clouds, a number of three-factor ECC-based authentication algorithms have been presented by the researchers. The majority of the protocols, on the other hand, contains substantial security weaknesses and incurs significant compute and communication overheads. As a result, the authors developed a novel scheme for the e-Health cloud that protects against a variety of threats, including guessing passwords, attacks on stolen smart cards, impersonation, and user anonymity. Furthermore, the researchers have used the Random Oracle Model to assess the protocol through formal security analysis (ROM). The researchers revealed that the proposed scheme is more effective with respect to computing and communication costs if compared with present protocols. As a result, the suggested protocol was said to be more effective, reliable, and secure.

With the fast use of Cloud Computing and IoT [21], it's becoming more important than ever to improve authentication procedures to decrease attacks and security vulnerabilities that impede application performance. In this study, the researchers have presented a new cloud computing-based-IoT-based authentication technique. The authors claim that Zhou et al.'s scheme protocol [19] is not safe and secure. They claimed that after investigating the protocol, as a result, they discovered some security flaws and shortcomings, rendering the system unsafe. To improve security, researchers suggest a new version that includes three phases. In the first phase, the user needs to login, thereafter mutual authentication will be performed, and thereafter key agreement will be done. They also incorporated a child phase called an "evidence of connection" attempt, which verifies the involvement of users and then the server's involvement. It is mentioned in the research that the new technique meets the parameters concerning security, and it is well resistant to most common but popular attacks, which was claimed to be an improvement on past efforts. Furthermore, the performance study shows that the novel system requires less communication cost during the registration and login stages than earlier authentication protocols.

Securely authenticating a remote user is a mechanism for verifying a user's identification via a reliable communication network by a remote server [22]. Various remote user authentication techniques have been presented since then, but each has its own set of benefits and drawbacks. Proposed authentication mechanisms have a big impact on real-time applications including "e-Health", "telemedicine", the "Internet of Things (IoT)", the "Cloud"', and "Multi-server" apps, in addition to their strengths and weaknesses. TMIS is one of the vulnerable systems that continues to leak privacy to unauthorized parties. One of the approaches, namely "Remote user authentications", meanwhile, has proven critical in speeding up IoT. Although IoT offers safe access to remote services, security is a big problem. "Cloud computing" services and a "Multi-server architecture" communicate data among several end-users over the Internet, which necessitates a high level of security. Even though significant work was put into developing remote user authentication schemes for health care, IoT, multi-server, and cloud applications, the bulk of these systems are vulnerable to security threats or lack important functionalities. The researchers proposed an analytical and detailed assessment of numerous remote user authentication systems and then categorized them according to their intended usage. Around 100 current remote user authentication techniques were examined and evaluated in terms of the benefits, essential features, computational cost, and storage cost, and transmission cost of state-of-the-art current remote user authentication approaches have been compared.

The growth of "Cloud computing" and the "Internet of Things (IoT)" concepts has enabled remote medical experts to monitor patients in real time, allowing patients to receive healthcare services at home [23]. Patient data is maintained at a centralized healthcare facility in these scenarios, where it may be accessed by a medical practitioner at regular intervals. However, the open Wi-Fi environment makes patient privacy a serious concern. Many authentication techniques for healthcare services have recently been presented in the literature, and most of them have been found to fall short of the security criteria. It was also claimed that they do not take into account how a medical practitioner might access data stored on a cloud server. Researchers presented a "Multi-factor authentication" formulated on the cryptographic elliptic curve technique in this work, which allows licensed medical practitioners having a licensed authorization to access healthcare data concerning patient information stored on a cloud server. The web-based "AVISPA" tool is also used for analysis and confirms that the proposed methodology is safe against various attacks such as "replay" and "man-in-the-middle attacks". A comparative analysis of performance and security features further shows that the scheme provides good protection against security threats while also achieving session key agreement.

Edge Computing [24], EC, will substantially encourage the growth of the IoT industry and enhance the diversification concerning the applications of the IoT ecosystem as one of the new computing paradigms that gives diverse answers to the difficulties that traditional cloud encounters. For high-quality IoT services, it is necessary to impose privacy protection and also those methods that can improve the security and reliability of EC. The survey of the use of AI and EC in IoT security is described in this study. The fundamental ideas and terminologies are first introduced. The IoT service architecture was then integrated with EC. Following that, the traditional and AI-driven privacy safeguarding of edge-based IoT is evaluated. The use of blockchain and AI to improve IoT security was also highlighted by the researchers. Lastly, the author talks about the open AI difficulties and issues for protecting IoT services in the EC.

Edge computing [25], which processes data at or near its source, has emerged as a result of the rapid expansion in the volume of data created by IoT sensors and devices. Keeping data confined to the edge node allows for decreased latency as well as data protection and privacy. However, many edge computing systems are vulnerable to a wide range of assaults due to resource-constrained hardware and software heterogeneities. Moreover, the growing trend of embedding intelligence into edge computing systems has resulted in its own set of security vulnerabilities, including data and model poisoning, as well as evasion attempts. The most critical risks to edge intelligence are discussed by the researchers in the proposed scheme. The researchers also mentioned the future scope of the study and mentioned that the latest technologies such as “Blockchain” and “Deep Reinforcement Learning” can be used to augment existing Edge-based AI systems.

Edge intelligence (EI) is one of the emerging technologies that integrates notions of AI with mobile edge computing practices through the applications of 5G and beyond 5G (B5G) systems [26]. Integrating EI with heterogeneous networks such as those equipped with wireless local area networks introduces additional security and privacy challenges. Authentication and reliability-based user equipment (UE) detection are two significant security components of EI-enabled, heterogeneous B5G networks examined in this research. The technical difficulties are discussed. A novel edge-computing-enabled, uniform authentication framework has been created that authenticates UEs consistently across heterogeneous networks while maintaining UE privacy.

Patients increasingly expect a complicated and advanced intelligent healthcare system that is tailored to their own health needs [27]. Edge computing, in conjunction with 5G and cutting-edge IoT sensors, enables real-time as well as smart healthcare solutions which are able to meet energy consumption and latency requirements. Previous surveys on smart healthcare systems included topics that deal with architectures based on technologies such as cloud computing, fog computing, data authentication, security, and various sensors and devices used in edge computing frameworks. Other than this, the past studies didn’t pay attention to the IoT applications for health care that was implemented in edge computing systems. To increase the significance of edge nodes, the GPU devices were embedded at the edge level. Additionally, the computational power of the model was dramatically improved by deploying GPU-powered nodes at the fog level. These GPU-powered nodes were also used at the fog level to boost the model's computational and data-processing capacities. When training-based DL models at the fog and edge levels are still not viable, workload distribution and deep network approaches have been recognized as one of the potential solutions in smart healthcare systems. The main purpose of the research was to look at existing and upcoming edge computing architectures and methodologies for smart health care, as well as to understand the expectations and constraints of various application scenarios. The authors looked into edge intelligence that uses state-of-the-art deep learning algorithms to target healthdata categorization with the tracking and recognition of vital signs. This research also includes a thorough examination of the usage of cutting-edge artificial intelligence-based categorization and prediction approaches in edge intelligence. Edge intelligence, besides its numerous benefits, has hurdles in terms of computing complexity and security. To provide an improved quality of patient life, the study identified prospective recommendations based on the current research for developing edge computing services for health care. Furthermore, the research also provides a quick summary of how IoT-driven solutions are being used in edge-based platforms for healthcare treatments.

The primary priority of the government all around the world is to effectively improve the healthcare system [28]. However, providing flexible services to people at the lowest possible cost is challenging. The upcoming wireless technologies and Edge computing, which can enable real-time and cost-effective patient remote monitoring, are among the most promising options for providing smart health care. The authors have shared research perspectives for using MEC for smart health applications. It was also mentioned that imagining a MEC-based architecture and explaining the advantages of implementing in-network and context-aware processing is very much feasible to meet smart health criteria. The study also disclosed how to use such architecture to provide multi-modal data compression and edge-based feature extraction for event detection in order to enable effective data transmission. The former enables efficient and low-cost compression, while the latter provides high dependability and quick reaction in the event of an emergency. Finally, we explore the primary difficulties and possibilities that edge computing may present, as well as future research areas.

This article explains the ways edge technology and artificial intelligence approaches are being used to improve smart healthcare systems [29]. Edge technology facilitates smart healthcare systems by decreasing latency, network stress, and power consumption by processing locally. AI also adds intelligence to the system. Certainly, these two technologies have improved the intelligence of smart healthcare system components and provided several benefits. The research presents a comprehensive smart healthcare framework that altogether employs edge technology and AI parallelly to improve healthcare systems. Processes are distributed from sensors to cloud servers in the proposed approach. When some activities are carried out on sensors utilizing a lightweight AI approach, the delay, complexity, and network load are reduced. Furthermore, using a flexible AI technique to do parallel computation on the edge layer reduces latency, allows for faster reaction and decision-making, and allows caregivers to get real-time warnings, all of which are critical in smart health care. The suggested model takes into account the security and privacy needs for complete smart health care, including sensitive medical data, network links, and user authentication and authorization. Although the suggested system architecture addresses some of the shortcomings of previous smart healthcare architectures, there are still certain issues to be addressed, such as data loss and autonomous network management.

Data digitization has become one of the recent trends in healthcare systems [30]. Blockchain programs such as Smart contracts have started playing a significant role in this scenario. It is essential to continuously process and monitor IoT data. Patient Health Data generated by the Healthcare IoT is diverse and reliable. As a result of processing such a large volume of data, communication between users and devices becomes unsafe. As a result, the topic of IoT-based healthcare data and device authentication was addressed in this work. Following that, the researchers describe a unique technique for securely processing Patient Health Data. The suggested technique enhances both the accuracy and the reliability of malicious node identification. Using Fog Computing and Blockchain, the researchers have offered a revolutionary approach. For authentication of healthcare Internet-of-Things devices, and Patient Health Data, a 3-layer framework, an applied mathematical model and framework, and one of the Encryption techniques, i.e. Advanced Signature-Based Encryption method based on Fog Computing were introduced in the research. The goal is to expand secure data transfer for IoT-based healthcare and real-time service users. The suggested architecture and algorithm will be capable of providing safe transaction and transmission services near the edge. The ASE algorithm easily outperforms the cloud-based and other existing framework and techniques. The novel ASE algorithm shows accuracy for parameters, namely reliability, false node detection, and throughput. 91% accuracy was reached for detecting false nodes using the ASE algorithm in the Fog computing environment, additionally, an 83 percent accuracy in the cloud. In FC, the ASE algorithm has a 95% reliability rate, whereas in the cloud, it has an 87 percent dependability rate. Two different simulators, namely iFogSim (Net-Beans) and SimBlock, are used to test the suggested technique.

Researchers are paying close attention to a robust certification ecosystem for health care in this modern era [31], because sensors, network-enabled devices (IoMT), and pervasive data acquisition, among other things, have nudged the healthcare industry to make diagnoses and remote monitoring easier for its patients. Two items to focus on in such an environment, namely information authenticity and verification, are difficult to achieve since no one can ensure secure communication without resolving these concerns and obstacles. It is impossible to assure data integrity, authorization, nonrepudiation, and user validity, as well as appropriately address information verification, without a strong authentication mechanism. As a result, we used WMSN to create enhanced, lightweight, and strong authentication methods for IoMT. The suggested approach addressed all of the shortcomings that had previously been identified in the literature. The protocol's robustness was tested using the ProVerif2.00 verification tools and the BAN logic of belief. The performance evaluation result, on the other hand, reveals that the suggested approach is both quick and secure. The part on comparative analysis reveals that the suggested protocol is light and secure, which is typically lacking in other techniques.

## 5.3 Text Analytics of Related Studies Using Word Cloud

### *5.3.1 Proposed Method*

In the present study, the researcher used Word cloud analysis as part of analyzing past literature conducted in a similar dimension. A total of 47 studies were found and after a thorough review of the literature, 21 were extracted for Word cloud analysis. Although there are no criteria specified for the number of literature that can be chosen for Word cloud formation, in general, performing a literature review for a good number of past studies provides more precise, relevant, and well-grounded research key terms. These key terms provide a clear picture of the studies that have been carried out and also build a notion of advanced research [32–34].

### *5.3.2 Data Collection*

The researcher collected data in the form of past studies that have been conducted concerning Data Authentication using emerging technologies such as Cloud computing, IoT, Fog computing, blockchain, and Edge AI. The search keywords used to extract the relevant literature were "Data Authentication using Edge AI", "Data Authentication in emerging technologies", and "Applications of Emerging technologies in Data Authentication". The target for searching and extracting the published past studies for extensive review was taken from the year 2019–2021. But after thorough analysis and reading of the past studies, only 21 were found to be most relevant. Therefore, these 21 literature reviews were taken up further for Word cloud formation.

### *5.3.3 Data Analysis*

An extensive review has been performed to identify the most frequent key terms which are extensively used by past researchers while working on the applications of Edge AI in health care. The current study aims to perform an extensive review for authentication of data recorded, managed, and controlled with the help of Edge AI and other emerging areas such as IoT, Cloud computing, Fog Computing, and

**Table 5.1** Word frequency of top 20 key terms

| Word | Length | Count | Weighted percentage (%) |
|---|---|---|---|
| Data | 4 | 1398 | 1.13 |
| Edge | 4 | 1132 | 0.92 |
| IoT | 3 | 1084 | 0.88 |
| Healthcare | 10 | 950 | 0.77 |
| Based | 5 | 902 | 0.73 |
| Authentication | 14 | 896 | 0.73 |
| Computing | 9 | 744 | 0.60 |
| Smart | 5 | 712 | 0.58 |
| Cloud | 5 | 673 | 0.55 |
| Scheme | 6 | 616 | 0.50 |
| Security | 8 | 603 | 0.49 |
| Proposed | 8 | 590 | 0.48 |
| Network | 7 | 503 | 0.41 |
| Devices | 7 | 493 | 0.40 |
| Key | 3 | 480 | 0.39 |
| System | 6 | 435 | 0.35 |
| User | 4 | 423 | 0.34 |
| Health | 6 | 413 | 0.34 |
| Fog | 3 | 405 | 0.33 |
| Blockchain | 10 | 345 | 0.28 |

*Source* Researcher own analysis

Blockchain. Using Word cloud, an exploratory tool in NVIVO, the most frequent terms have been identified, which have been discussed and highlighted by most of the past research [15]. Word clouds have evolved as a simple and aesthetically appealing way of text representation. They have been extensively used in a variety of scenarios to offer insight by reducing text to the most frequently occurring terms [35]. Word cloud analysis along with word count, as mentioned in Table 5.1 has been considered as one of the refined ways of the literature review analysis process so as to retrieve possible research synthesis on literature analyses [36].

### *5.3.4 Discussion*

In the present study, a Word cloud as one of the exploratory tools has been generated to explore the most frequent words and provide insights on the key terms frequently discussed [37] in the existing literature in relevance to the authentication of data using an application of Edge AI and other emerging technologies. Table 5.1 shows Edge, data, authentication, healthcare, IoT-based, and computing key terms frequently used and discussed in past studies. Therefore, it gives a clear understanding of the areas where research are going on concerning applications and implementation of emerging technologies for healthcare data authentication.

The quantitative results of text analytics are shown in Table 5.1. The results represent the length, count, and weighted percentage of the most commonly used key terms for the authentication of data in health care using emerging technologies. The length column represents the word length, count represents the total number of times a particular key term appeared or was discussed by the researchers, and weighted percentage represents the frequency of a word in relation to the total number of words counted. The higher the weighted percentage, higher is the weightage or frequency of occurrence of the word [38]. On performing text analytics using Word cloud in NVIVO, a total of 100 key terms of minimum length 3 has been generated. The minimum length of 3 sometimes generates conjunction or preposition words, but it was required as in many research the name of the technology was written in Acronym such as IoT and also some technology bears a three-character length name such as Fog computing. Table 5.1 represents the top 20 key terms retrieved by analyzing the 22 most relevant studies for authenticating healthcare data using Edge AI, IoT, Fog Computing, Cloud Computing, and other emerging technologies.

Authentication is one of the keywords that has been often mentioned by researchers in terms of suggesting robust authentication schemes [39], challenges, and issues of patient healthcare data authentication [27], accessing healthcare data, proposing mathematical model, and framework for authentication of healthcare data [30], an improvement on the existing scheme of data authentication [21, 40]. Similarly, Communication retrieved has been mentioned by the researchers in relevance to proposing a protocol for secure communication, representing communication overhead due to insecure channels [20], optimized communication cost during the process of data authentication [21], remorse user authentication using reliable communication channels [22]. Additionally, Edge, IoT, Fog, and Blockchain key terms have been mentioned as emerging technologies where prospects of data authentication have been taken further as the basis of proposing a robust framework for the security of healthcare data.

## 5.4 Conclusion

The evolution of interaction on the Internet may be divided into two categories: Mobility and Intelligence. Networked digital solutions moved to the cloud and intercommunicating devices as connectivity and speed of communication have been improved. Mobility has already been shown to be beneficial to individuals, businesses, and society, and with intelligence added, it will be even more valuable. The power of Push and Pull has made a significant positive impact on each sector including health care [41]. The healthcare sector has rigorously started using AI for decision-making, recommendations, diagnosis, and many more. To perform all these smart and automated activities, the AI-driven healthcare system requires data from an authenticated source and should be acquired by authenticated users. Thus, data authentication becomes one of the important dimensions of study. The present study focuses on an extensive review of past studies using Word cloud analysis to identify the focus of the researchers while proposing schemes to secure and verify the data. Word cloud has been applied to analyze the content of relevant papers [42] so as to see whether sufficient attention has been given to the authentication process. The analysis revealed that the studies are more focused on the authentication of data instead of user authentication, as shown in Table 5.1, where the occurrence of the "user" key term has less weightage in comparison to the key terms "data" and "authentication". Finally, a Word cloud analysis of the paper revealed that an extensive literature review appears to have mostly achieved its stated goal of prioritizing research papers [43] that are directly relevant to the authentication schemes proposed by the researchers using applications of emerging technologies. Therefore, the researchers have done the analysis in the right direction starting from exploring the literature review followed by extracting relevant review and thereby validating it with the help of Word cloud analysis in NVIVO.

## References

1. Han, J., Han, J.: Evaluation of artificial intelligence techniques applied in Watson and AlphaGo. Acad. J. Comput. Inf. Sci. **4**(8) (2021)
2. Gupta, I., Nagpal, G.: Artificial Intelligence and Expert Systems. Stylus Publishing, LLC (2020)
3. Yassine, A., Singh, S., Hossain, M.S., Muhammad, G.: IoT big data analytics for smart homes with fog and cloud computing. Futur. Gener. Comput. Syst. **91**, 563–573 (2019)
4. Porambage, P., Okwuibe, J., Liyanage, M., Ylianttila, M., Taleb, T.: Survey on multi-access edge computing for internet of things realization. IEEE Commun. Surv. Tutor. **20**(4), 2961–2991 (2018)
5. Pan, J., Wang, J., Hester, A., Alqerm, I., Liu, Y., Zhao, Y.: EdgeChain: An Edge-IoT framework and prototype based on blockchain and smart contracts. IEEE Internet Things J. **6**(3), 4719–4732 (2019)
6. Mutlag, A., Abd Ghani, M., Arunkumar, N., Mohammed, M., Mohd, O.: Enabling technologies for fog computing in healthcare IoT systems. Futur. Gener. Comput. Syst. **90**, 62–78 (2019)
7. Li, J., Cai, J., Khan, F., Rehman, A.U., Balasubramaniam, V., Sun, J., Venu, P.: A secured framework for sdn-based edge computing in IOT-enabled healthcare system. IEEE Access **8**, 135479–135490 (2020)
8. Zhang, J., Zhang, F., Huang, X., Liu, X.: Leakage-resilient authenticated key exchange for edge artificial intelligence. IEEE Trans. Dependable Secure Comput. **18**(6), 2835–2847 (2021)
9. Prabadevi, B., Deepa, N., Pham, Q.V., Nguyen, D.C., Reddy, T., Pathirana, P.N., Dobre, O.: Toward blockchain for edge-of-things: a new paradigm, opportunities, and future directions. IEEE Internet Things Mag. **4**(2), 102–108 (2021)
10. Vimal, S., Robinson, Y.H., Kadry, S., Long, H.V., Nam, Y.: IoT based smart health monitoring with CNN using edge computing. J. Internet Technol. **22**(1), 173–185 (2021)
11. Lee, D., Park, J.H.: Future trends of AI-based smart systems and services: challenges, opportunities, and solutions. J. Inf. Process. Syst. **15**(4), 717–723 (2019)
12. Hartmann, M., Hashmi, U., Imran, A.: Edge computing in smart health care systems: review, challenges, and research directions. Trans. Emerg. Telecommun. Technol. **33**(3), 1–25 (2019)

13. Kamruzzaman, M., Alrashdi, I., Alqazzaz, A.: New opportunities, challenges, and applications of edge-AI for connected healthcare in internet of medical things for smart cities. J. Healthc. Eng. **2022**, 1–14 (2022)
14. Deepa, N., Prabadevi, B., Maddikunta, P., Gadekallu, T., Baker, T., Khan, M., Tariq, U.: An AI-based intelligent system for healthcare analysis using rdge-adaline stochastic gradient descent classifier. J. Supercomput. **77**(2), 1998–2017 (2020)
15. Oesper, L., Merico, D., Isserlin, R., Bader, G.D.: WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. Source Code Biol. Med. **6**(1), 7 (2011)
16. Zargar, S., Shahidinejad, A., Ghobaei-Arani, M.: A lightweight authentication protocol for IoT-based cloud environment. Int. J. Commun. Syst. (2021)
17. Amin, R., Kumar, N., Biswas, G., Iqbal, R., Chang, V.: A light weight authentication protocol for IoT-enabled devices in distributed cloud computing environment. Futur. Gener. Comput. Syst. **78**, 1005–1019 (2018)
18. Maitra, T., Islam, S., Amin, R., Giri, D., Khan, M., Kumar, N.: An enhanced multi-server authentication protocol using password and smart-card: cryptanalysis and design. Secur. Commun. Netw. **9**(17), 4615–4638 (2016)
19. Zhou, L., Li, X., Yeh, K.-H., Su, C., Chiu, W.: Lightweight IoT-based authentication scheme in cloud computing circumstance. Future Gener. Comput. Syst. **91**, 244–251 (2019)
20. Minahil, A.M., Mahmood, K., Kumari, S., Sangaiah, A.: Lightweight authentication protocol for e-health clouds in IoT-based applications through 5G technology. Digit. Commun. Netw. **7**(2), 235–244
21. Martínez-Peláez, R., Toral-Cruz, H., Parra-Michel, J., García, V., Mena, L., Félix, V., Ochoa-Brust, A.: An enhanced lightweight IoT-based authentication scheme in cloud computing circumstances. Sensors **19**(9), 2098 (2019)
22. Rajasekar, V., Jayapaul, P., Krishnamoorthi, S., Saračević, M.: Secure remote user authentication scheme on health care, IoT and cloud applications: a multilayer systematic survey. Acta Polytechnica Hungarica **18**(3), 87–106 (2021)
23. Dhillon, P., Kalra, S.: A secure multi-factor ECC based authentication scheme for Cloud-IoT based healthcare services. J. Ambient. Intell. Smart Environ. **11**(2), 149–164 (2019)
24. Xu, Z., Liu, W., Huang, J., Yang, C., Lu, J., Tan, H.: Artificial intelligence for securing IoT services in edge computing: a survey. Secur. Commun. Netw. **2020**, 1–13 (2020)
25. Ansari, M.S., Alsamhi, S.H., Qiao, Y., Ye, Y., Lee, B.: Security of distributed intelligence in edge computing: threats and countermeasures. In: The Cloud-To-Thing Continuum, pp. 95–122. Palgrave Macmillan, Cham (2020)
26. Cui, Q., Zhu, Z., Ni, W., Tao, X., Zhang, P.: Edge-intelligence-empowered, unified authentication and trust evaluation for heterogeneous beyond 5G systems. IEEE Wirel. Commun. **28**(2), 78–85 (2021)
27. Amin, S., Hossain, M.: Edge intelligence and internet of things in healthcare: a survey. IEEE Access **9**, 45–59 (2021)
28. Abdellatif, A., Mohamed, A., Chiasserini, C., Tlili, M., Erbad, A.: Edge computing for smart health: context-aware approaches, opportunities, and challenges. IEEE Netw **33**(3), 196–203 (2019)
29. Hayyolalam, V., Aloqaily, M., Ozkasap, O., Guizani, M.: Edge intelligence for empowering IoT-based healthcare systems. IEEE Wirel. Commun. **28**(3), 6–14 (2021)
30. Shukla, S., Thakur, S., Hussain, S., Breslin, J., Jameel, S.: Identification and authentication in healthcare Internet-of-Things using integrated fog computing based blockchain model. Internet of Things **15**, 100422 (2021)
31. Jan, S., Ali, S., Abbasi, I., Mosleh, M., Alsanad, A., Khattak, H.: Secure patient authentication framework in the healthcare system using wireless medical sensor networks. J. Healthc. Eng. **2021**, 1–20 (2021)
32. Boyd, B.K., Solarino, A.M.: Ownership of corporations: a review, synthesis, and research agenda. J. Manag. **42**(5), 1282–1314 (2016)
33. Mazumdar, T., Raj, S.P., Sinha, I.: Reference price research: review and propositions. J. Mark. **69**(4), 84–102 (2005)

34. Rodell, J.B., Breitsohl, H., Schröder, M., Keating, D.J.: Employee volunteering: a review and framework for future research. J. Manag. **42**(1), 55–84 (2016)
35. Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: text analytics based on word clouds. In: 2014 47th Hawaii International Conference on System Sciences, pp. 1833–1842 (2014). https://doi.org/10.1109/HICSS.2014.231
36. Leech, N., Onwuegbuzie, A.: Qualitative data analysis: a compendium of techniques and a framework for selection for school psychology research and beyond. Sch. Psychol. Q. **23**(4), 587–604 (2008)
37. Kabir, A.I., Karim, R., Newaz, S., Hossain, M.I.: The power of social media analytics: text analytics based on sentiment analysis and word clouds on R. Informatica Economica **22**(1) (2018)
38. Dicle, M., Dicle, B.: Content analysis: frequency distribution of words. SSRN Electron. J. (2017)
39. Hou, J., Yeh, K.: Novel authentication schemes for IoT based healthcare systems. Int. J. Distrib. Sens. Netw. **11**(11), 183659 (2015)
40. Wu, H., Chang, C., Zheng, Y., Chen, L., Chen, C.: A secure IoT-based authentication system in cloud computing environment. Sensors **20**(19), 5604 (2020). https://doi.org/10.3390/s20195604
41. Shnurenko, I., Murovana, T., Kushchu, I.: Artificial Intelligence: Media and Information Literacy. Human Rights and Freedom of Expression. Published by TheNextMinds for the UNESCO Institute for Information Technologies in Education (2020)
42. Ramsden, A., Bate, A.: Using word clouds in teaching and learning. University of Bath. Accessed 18 Dec 2009
43. Atenstaedt, R.: Word cloud analysis of the*bjgp*. Br. J. Gen. Pract. **62**(596), 148–148 (2012). https://doi.org/10.3399/bjgp12x630142

# Chapter 6
# Biomedical Data Classification Using Fuzzy Clustering

**Shivani Sharma and Bipin Kumar Rai**

## 6.1 Introduction

When we analyze data, there is a task in which we find a group of data set objects which share the same features or characteristics. When we do so, the users get the knowledge of their data, they understand it, and can also decrease the nature of high dimensionality in the data. When we group these conceptual data into groups, they are called clusters.

The phrase “cluster analysis,” first introduced by Tryon in 1939, refers to a variety of distinct algorithms and techniques for classifying objects of a similar type. How to build taxonomies—that is, how to organize observed data into meaningful structures—is a general subject that scholars in a variety of fields of study grapple with.

In other terms, a cluster analysis is a study that seeks to group various items so that the degree of association between two objects is maximal if they are members of the same group and minimal otherwise. Given the aforementioned, cluster analysis can be used to find patterns in data without offering an explanation or interpretation. In other words, cluster analysis doesn’t actually explain why structures exist; it only finds them in the data [1].

In technical and scientific fields like health, computer vision, remote sensing, psychology, etc., clustering has attracted significant interest as a technique for knowledge discovery. Data patterns in one group are grouped in a way that makes them more comparable to one another than to data patterns in another group. To achieve clustering, which is determined by the purpose and data organization, it employs a dissimilarity metric.

Clustering affects practically every element of our daily lives. A cluster of individuals could be, for instance, a number of diners at the same table at a restaurant.

S. Sharma · B. K. Rai (✉)
Department of IT, ABES Institute of Technology, Ghaziabad, Uttar Pradesh 201009, India
e-mail: bipinkrai@gmail.com

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_6

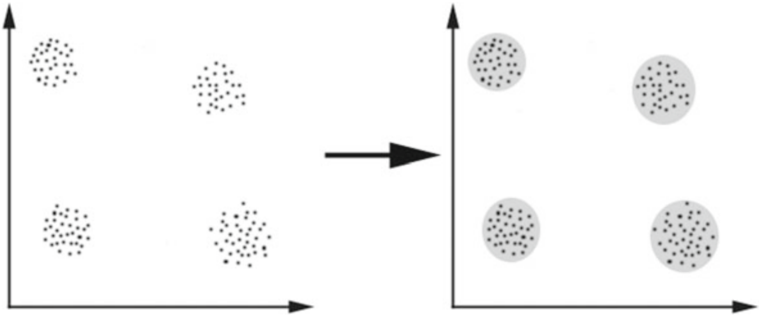


**Fig. 6.1** Cluster [3]

Similar products, like various types of meat or vegetables, are often exhibited side by side or in close proximity at grocery stores [1].

The importance of clustering can be seen in a virtually infinite number of situations. For instance, before a coherent explanation of the distinctions between animals is feasible, scientists must first classify the different kinds of animals.

According to the contemporary biological framework in use, man is a member of the primate, mammal, amniotes, vertebrate, and animal kingdoms [2]. Keep in mind that in this categorization, the less similar the members of the relevant class are to one another, the higher the level of aggregation. The closest relatives of mammals, such as dogs, are more distant from man than all other primates, such as apes. In other words, regardless of the type of your work, you will sooner or later have a clustering issue of some kind [2].

Clustering could be loosely defined as the practice of grouping objects into units that share characteristics. Therefore, a cluster is a group of things that are "dissimilar" from the objects in other clusters yet similar to one another. An example of a cluster has been shown in Fig. 6.1.

## 6.2 Fuzzy Logic and Biomedical Data

Sadegh-Zadeh [4] states that medical research is marked by "inescapable uncertainty" which can be attributed to a variety of factors, the most prominent of which is medicine's inherent vagueness [5].

The lack of knowledge, inaccurate knowledge, and conflicting nature of medical practice contribute to its complexity [6].

Beck and Melo [5] claim that the complexity of medical practices has grown as a result of the rise of significant advances in diagnostic methods and therapies in the twentieth century.

Nowadays, a physician assesses a patient's risk based on subjective variables and other information about the scenario [7].

In the diagnosis stage, essentially all clinicians employ an information system. This pool of information either enables them in recording data or assists them in taking important decisions. Making decisions is increasingly difficult and unpredictable due to the abundance and variety of data, which encompasses textual, quantitative, time series, and visual information [8], therefore it is required to handle this uncertainty by using special procedures [9].

## 6.3 Need for Fuzzy Logic in Biomedical Domain

Different sources of uncertainty of patient data can be classified as follows:

- The patient's medical history is very subjective and imprecise, as provided by the patient or his/her family members.
- During a physical examination, the physician normally obtains objective data; nevertheless, sometimes it is difficult to distinguish between a normal state and a pathological state.
- Diagnostic and Laboratory test results are susceptible to human error and inadequate medical conduct by the patient.
- Patient symptoms that are simulated, exaggerated, or underestimated, as well as failing to describe a few of them.
- Doctors underline the paradox between the rise in mental illnesses and the absence of a natural classification scheme.
- Classification is difficult in critical circumstances, especially when a categorical approach to diagnosis is used.

Fujita [10] mentioned additional reasons that patient's symptoms are fuzzy with many related alternatives, which can be listed as follows:

- Patients employ unclear phrases and words to convey their problem, and they are not exact in defining their issue in a practical mathematical approach.
- Doctors come from a variety of backgrounds and experiences; therefore, their interpretations may differ.
- Symptoms of the pathological process are vague and similar to those of other illnesses.

## 6.4 Various Types of Clustering on Biomedical Data

Utilizing technology in the healthcare sector can help firms run more effectively and fulfill the growing demand for better patient care. The Electronic Health Record offers all of these benefits (EHR). EHRs are a type of information system that stores all of a patient's medical records digitally [11]. In recent trends, we have seen a large amount of growth in biomedical data. This increasing amount of data is becoming a challenge for the patients and the medical workers as they are struggling to find the information they want. As discussed earlier, the solution for such challenges is the clustering technique which can help the patients and the medical workers in finding out the needed information from the large-scale bio-medical data.

In biomedical research, clustering tools have been utilized in manifold areas, among many others in expression analysis, disease subtyping or protein research. In biomedical research, clustering tools have been utilized in manifold areas, among many others in expression analysis, disease subtyping or protein research.

In biomedical research,

In the research of biomedical data, the clustering algorithms act as a tool to be utilized in many areas such as protein research or in disease sub-typing or in expression analysis. It is also helpful in figuring out the knowledge and the latent structure behind such a great volume of biomedical data [12, 13].

Clustering's goal is to maximize intra-cluster similarity while minimizing inter-cluster similarity which is commonly used in grouping gene expression data or automatic categorization of text.

Automatic clustering is a method for finding "natural" structures hiding in the data with an unsupervised technique. A set of unlabeled data samples are automatically grouped into clusters so that samples assigned to the same cluster are more comparable to one another than samples assigned to other clusters [14].

Researchers have developed numerous clustering techniques for biomedical data, which fall into two categories which are shown in Fig. 6.2.

(1) **Hard clustering**: Each data point in the hard clustering technique either fully belongs or does not belong to a cluster.
(2) **Soft clustering**: Instead of assigning each data point to a distinct cluster, soft clustering assigns a chance or likelihood of falling the data point in those clusters will be found.

In biological or medical data, the data is generally in table format where the objects are the rows of the tables while the attributes are the columns of the table. If

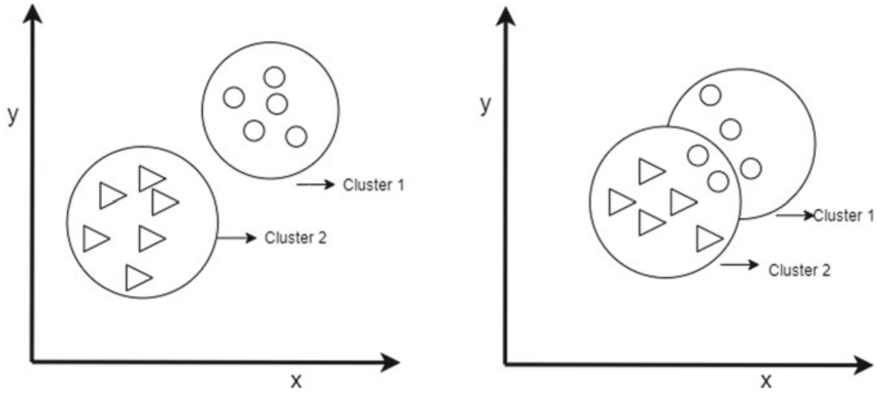


**Fig. 6.2** Hard clustering versus soft clustering

we talk about the hard clustering out of the complete data set, each object belongs to exactly one cluster but in soft clustering an object will belong to more than one cluster. Out of these two, soft clustering seems to be more reasonable as many times it is not justifiable to put an object into just one cluster. This can be understood with the help of an example as the atypical hyperplasia can be contemplated as normal endometrium or abnormal endometrium by various physicians [15].

### *6.4.1 Fuzzy c-Means Algorithm (FCM)*

Fuzzy c-Means Algorithm (FCM) is a type of soft clustering algorithm which one dimensional. Here, one-dimensional means when we group a disease-symptom table, the Fuzzy c-Means Algorithm presumes that there exists no relationship among the symptoms and it just categorizes the diseases based on their symptoms. It will not be justified if we overlook the correlations between the symptoms as an example; we can see that there exists a close relation between metabolic diseases and increased pulse pressure.

The fundamental goal of fuzzy c-means clustering is to divide data into a number of clusters. It operates by giving each and every cluster a membership function value [16].

The two processes involved in fuzzy c-means clustering are

i. The cluster centers are calculated.
ii. Making use of Euclidian distance, the assignment of points to these centers is done.

Until the cluster centers stabilize, the above two processes will be repeated.

In this algorithm, extensive research and implementation have been done by the researchers and the FCM clustering algorithm needs prior knowledge of the number of clusters. Hence, we can say that, whenever an object-feature contingency table is being analyzed for clustering, the object and feature dimensions both should be grouped [15].

The Fuzzy c Means (FCM) clustering has been used with the combination of other clustering algorithms which can be termed as the hybrid clustering techniques which give better results. The Hybrid forms are the automatic fuzzy clustering algorithm.

### *6.4.2 Hierarchical Clustering*

There are two main characteristics, i.e., small size sample data and high dimensionality which make the classification difficult in Biomedical Spectra such as the ones got from magnetic resonance (MR) spectrometers. Hierarchical clustering gives us powerful and strong results mainly when it works with small-size high-dimensional data sets [17].

The evolutionary method for identifying a data set's hierarchical structure. Previously, the cluster structure of the data set was represented linearly. A population of clustering hierarchies evolves, thanks to an evolutionary algorithm. As search operators, this method makes use of mutation and crossover. It takes into account a novel crossover operator and a binary tournament selection.

Since the data is present in tree-like structure, it is possible to optimize the number of clustering layers and the number of clusters on each level. The goal of cluster analysis is to divide a given data set into several groups so that the data within each group are more similar to one another than the items included in the various clusters. By organizing the data into a tree structure, hierarchical clustering creates a multi-level representation that can reveal links between clusters or inter-cluster relationships.

A linear sequence which was used earlier can now be used to present the cluster hierarchy. Using the sequence, a classification tree is created. A cluster is represented by each node in the tree. The solution then evolved represents trees of variable shapes and sizes. The classification tree's classes are identified by parsing each level from top to bottom and left to right. The entire data set is regarded as belonging to the class of the root node. The data set being present in each node must satisfy the chromosome length and the number of elements being present in the data set.

Each gene in the population can be altered in this method with a specific probability. A crossover operator is taken into account, and the mutation is merged with the crossover operator. Here, one-point crossover is taken into account. The intersection point is selected at random.

A mutation operator is utilized after crossover. Each gene in the point section is subjected to mutation, which disregards the condition. Mutation is applied with the restriction which ensures that no chromosome reparation, i.e., act for making up for something wrong that has been done is required.

The effectiveness of this algorithm will depend on its representation. The representation of the data set needs to be accurate. Thus, the solution accuracy will increase with the number of generations.

### *6.4.3 K-means Clustering Algorithm*

K-means clustering algorithm can be said as the unsupervised clustering method. It is commonly used because of the reason that it has comparatively low computational complexity and is simple to implement. As the number of clusters (K) is typically known for images of certain parts of the human anatomy, the K-means clustering approach is also effective for segmenting biological images. For instance, the MR picture of the human skull typically includes regions that indicate fat, soft tissue, bone, and background. Since there are 4 regions, K will also be 4. The algorithm's primary goal is to reduce the objective function [18]. In Fig. 6.3, comparison of segmentation maps, obtained using K-means clustering, against those obtained using conventional methodology has been shown [19].

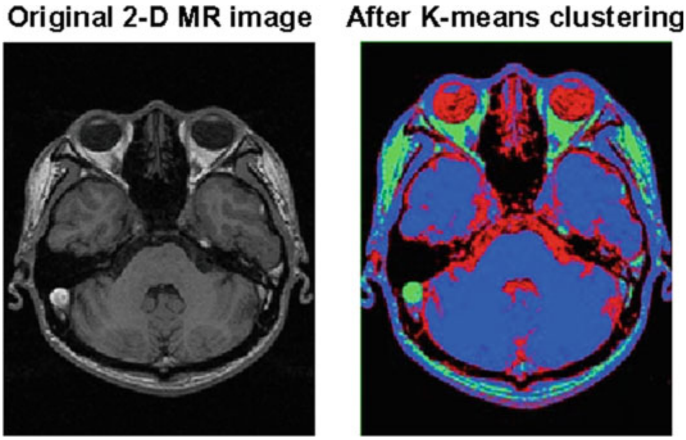


**Fig. 6.3** Comparison of segmentation maps using K-means clustering [19]

A straightforward validity metric based on the intra-cluster and inter-cluster distance metrics enables the automatic determination of the number of clusters. Producing all of the segmented images for 2 clusters up to K-max clusters, where K-max denotes the maximum number of clusters, is the basic technique. Then, by calculating the minimum value for our measure, it is determined which of their validity measures represents the best grouping.

The validity metric is put to the test using artificial photos whose number of clusters is known, as well as using real-world images.

With synthetic images, the validity measure produces a minimum value for the predicted number of clusters. For natural images, there is a propensity to choose fewer clusters, but this is because the inter-cluster distance is significantly bigger and has a much greater impact on the validity measure. They circumvent this by first seeking a local maximum for the validity measure and then, after the local maximum, determining the minimum value [20]. The least number of clusters that can be chosen using this modified technique is four. Natural color photos are anticipated to have more than two or three clusters, thus this is not a serious issue.

With the exception of the image with just two clusters—two clusters cannot be selected by this updated process—the modified rule still enables the optimal number of clusters to be chosen for the synthetic images. For all of the natural photos, the validity measure performed more consistently, leading to effective segmentation.

## 6.5 Conclusion

Medical science is a rapidly evolving field of study. Due to significant advancements in hardware and software, a growing amount of healthcare-related data is now easily accessible, including information from insurance providers, hospitals, pharmaceutical firms, patients, and users of personal health monitoring devices [21], 22. New medical treatments and machines are developed and introduced every day. Medical informatics is one of the contributing and supporting branches that has played a significant role in this evolution. Every day, the number of researchers working in the healthcare field increases. Funding for these scholars and their research is also expanding.

Lots of research has already been done in which the nature of medical data was studied, and it was found to be uncertain and fuzzy. Because of this, the classification of biomedical data remains a problem, where the hard clustering techniques are not very effective. The research in this domain has taken a shift from hard to soft computing techniques in biomedical science. Fuzzy logic is one of the techniques of soft computing. Fuzzy logic is specially designed for problems where the very nature of data is not clear, and is uncertain.

The long-standing disagreement between the medical and technical people, where the former claims that it is difficult to generalize about diseases and the latter forces them to create and generalize their conceptions, seems to be coming to an end. The main research areas of fuzzy logic appear to include ranking studies, grouping and classification studies, pattern recognition studies, performance comparison studies, and feature selection studies.

Traditional techniques suffer from flexibility due to the ambiguous nature of medical decision-making. By taking a look at the following, we can determine how useful fuzzy logic is in medicine:

- Flexibility is achieved by taking into account all conceivable values, including those that are ambiguous.
- Robustness: By making the process more resilient when compared to previous procedures, and by taking into account blurring boundaries.
- Efficiency: By utilizing more available data.

Fuzzy Logic, like many computer technologies, has drawbacks. Designing a Fuzzy Logic system or application takes more time and effort.

Fuzzy clustering, which assigns membership levels to data objects, allows things to belong to several clusters. Fuzzy clustering is more rational than hard clustering because a biomedical article may cover several medical disciplines and issues. An incremental technique is required for dealing with massive biomedical data sets. As a result, incremental fuzzy clustering is more effective.

## References

1. Barro, S., Marín, R. (eds.): Fuzzy Logic in Medicine, vol. 83. Springer, Berlin Eidelberg (2002)
2. Gürsel, G.: Healthcare, uncertainty, and fuzzy logic. Digit. Med. **2**, 101–112 (2016)
3. Singh, M., Kaur, K.: Clustering algorithm for genetic diversity. World Acad. Sci., Eng. Technol. **42** (2008)
4. Sadegh-Zadeh, K.: Fuzzy logic. In: Handbook of Analytic Philosophy of Medicine. Springer, The Netherlands, pp. 1055–110 (2015)
5. Beck, M., Melo, S.: Quality Management and Managerialism in Healthcare: A Critical Historical Survey. Palgrave Macmillan, Basingstoke, Hampshire (2014)
6. Torres, A., Nieto, J.J.: Fuzzy logic in medicine and bioinformatics. J Biomed Biotechnol **2006**, 91908 (2006)
7. Kwiatkowska, M., Michalik, K., Kielan, K.: Computational representation of medical concepts: a semiotic and fuzzy logic approach. In: Soft Computing in Humanities and Social Sciences, pp. 401–420. Springer, Berlin (2012)
8. Hudson, D.L., Cohen, M.E.: Uncertainty and complexity in personal health records. Conf. Proc. IEEE Eng. Med. Biol. Soc. **2010**, 6773–6776 (2010)
9. Shullman, A.: PACS/RIS/imaging. Radiology's golden age. This multi-site imaging firm improved patient care and its financial standing through RIS/PACS automation. Health Manag. Technol. **30**, 12–3, 24 (2009)
10. Fujita, H., Rudas, I.J., Fodor, J., Kurematsu, M., Hakura, J.: Fuzzy reasoning for medical diagnosis-based aggregation on different ontologies. In: 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 137–146. IEEE (2012)
11. Rai, B.K., Tyagi, A., Arora, B., Sharma, S.: Blockchain based Electronic Healthcare Record (EHR). In: Kumar, A., Mozar, S. (eds.) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7985-8_19
12. Liu, Y., Wan, X.: Information bottleneck based incremental fuzzy clustering for large biomedical data. J. Biomed. Inf. **62**, 48–58 (2016). pmid:27260783
13. Hammouda, K.M., Kamel, M.S.: Efficient phrase-based document indexing for web document clustering. IEEE Trans. Knowl. Data Eng. **16**, 1279–1296 (2004)
14. Interactive clustering: a comprehensive review. ACM Comput. Surv. **53**(1), Article 1
15. Liu, Y., Wu, S., Liu, Z., Chao, H.: A fuzzy co-clustering algorithm for biomedical data. PLoS ONE **12**(4), e0176536 (2017). https://doi.org/10.1371/journal.pone.0176536
16. Manikandan, R., Kumar, A., Gupta, D.: Chapter 5-Hybrid computational intelligence for healthcare and disease diagnosis. In: Hybrid Computational Intelligence for Pattern Analysis and Understanding, Hybrid Computational Intelligence, pp. 97–122. Academic (2020). ISBN 9780128186992, https://doi.org/10.1016/B978-0-12-818699-2.00006-8.
17. Yang, H., Pizzi, N.J.: Biomedical data classification using hierarchical clustering. In: Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No.04CH37513), vol. 4, pp. 1861–1864(2004). https://doi.org/10.1109/CCECE.2004.1347570
18. Ajala Funmilola, A., Oke, O.A., Adedeji, T.O., Alade, O.M., Adewusi, E.A.: Fuzzy k-c-means clustering algorithm for medical image segmentation. J. Inf. Eng. Appl. **2**(6) (2012). www.iiste.org. ISSN 2224–5782 (print). ISSN 2225–0506 (online)
19. Ng, H.P., Ong, S.H., Foong, K.W.C., Goh, P.S., Nowinski, W.L.: Medical image segmentation using K-means clustering and improved watershed algorithm. IEEE Southwest Symp. Image Anal. Interpret. **2006**, 61–65 (2006). https://doi.org/10.1109/SSIAI.2006.1633722
20. Ray, S., Turi, R.H.: Determination of number of clusters in K means. In: School of Computer Science and Software Engineering (1999)
21. Sharma, S., Kesarwani, A., Maheshwari, S., Rai, B.K.: Federated learning for data mining in healthcare. In: Yadav, S.P., Bhati, B.S., Mahato, D.P., Kumar, S. (eds.) Federated Learning for IoT Applications. EAI/Springer Innovations in Communication and Computing. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-85559-8_16

22. Solanki, T., Rai, B.K., Sharma, S.: Federated Learning using tensor flow. In: Yadav, S.P., Bhati, B.S., Mahato, D.P., Kumar, S. (eds.) Federated Learning for IoT Applications. EAI/Springer Innovations in Communication and Computing. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-85559-8_10
23. Chi, M.: Evolutionary Hierarchical Clustering Technique. Scientific Communication Avram I Ancu University (2001)
24. Lu, Y., Liang, L.: Hierarchical Clustering of Features on Categorical Data of Biomedical Applications. CAINE (2008)
25. Krishna, K.: Murty M "Genetic K-means algorithm." IEEE Trans. Syst. Man Cybern. B Cybern. **29**, 433–439 (1999)
26. Tajunisha, N., Saravanan, V.: Performance analysis of K-means with different initialization methods for high dimensional data. Int. J. Artif. Intell. Appl. **1**(4) (2010)
27. Wagsta, K., Cardie, C.: Constrained K Means Clustering Using Background Knowledge. Department of Computer Science Coenell University (2001)
28. Abbod, M.F., von Keyserlingk, D.G., Linkens, D.A., Mahfouf, M.: Survey of utilization of fuzzy technology in medicine and healthcare. Fuzzy Sets Syst **120**, 331–349 (2001)

# Part II
# Application of AI and Blockchain in Healthcare

# Chapter 7
# Applications of Machine Learning in Healthcare with a Case Study of Lung Cancer Diagnosis Through Deep Learning Approach

**Taskeen Zaidi and Bijjahalli Sadanandamurthy Sushma**

## 7.1 Introduction

The healthcare industry is offering supports to millions of patients and AI can be used in healthcare industry to make it more productive. The AI enabled system can perform tasks like decision making, solving complex problems, object detection, and many more. Machine learning is an area of AI which can be useful in the areas like marketing, finance, gaming, and others. The AI may be impactful in healthcare industry in future as the medical data is large. The AI based technology like deep learning and Machine learning will be helpful in organizing the data and keeping track of the data. Artificial Intelligence (AI) is a branch of science which deals with helping machines finds solutions to complex problems in a more human-like fashion. This generally involves borrowing characteristics from human intelligence, and applying them as algorithms in a computer friendly way. A more or less flexible or efficient approach can be taken care depending on the requirements established, which influences how artificial intelligent behavior appears. Artificial Intelligence is an umbrella term under Machine Learning which is a technique used to give machines the ability to learn like humans and analyze the human beings' learning journey by mapping it to how a machine can learn in the same way. A machine is trained to identify patterns and those patterns were used to predict some results based on algorithm. Machine learning involves training the machine using patterns, images, feedback, and more. If we try to summarize the way machine will learn, it can be put into three steps: Machines are not as capable as human beings; they will have to be given each step one by one to process the input to give us a workable output or prediction. There are namely four types of machine learning like Supervised

T. Zaidi (✉) · B. S. Sushma
Jain Deemed to be University, Bengaluru, India
e-mail: t.zaidi@jainuniversity.ac.in

B. S. Sushma
e-mail: bs.sushma@jainuniversity.ac.in

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_7

ML, Unsupervised ML, Reinforcement ML, and Semi Supervised ML approach. In supervised machine learning, we know about the data and the problem. Think of it as, "given a set of features x, we know the value of y," and so in supervised learning, we create a function that approximates results based on some set of data. There are two kinds of supervised learning: classification and regression. In unsupervised machine learning, our data is unlabeled. There are two forms of unsupervised machine learning: clustering and dimension reduction. In clustering, we learn more about data points as they are clustered, or grouped together. This allows learned models to understand a data set, detect anomalies, and assign relationships between points, often allowing users to develop new categories or features about the data set. In dimension reduction, we plot data points across different dimensions and feature sets to understand our data sets. This allows for techniques like feature selection or transformation.

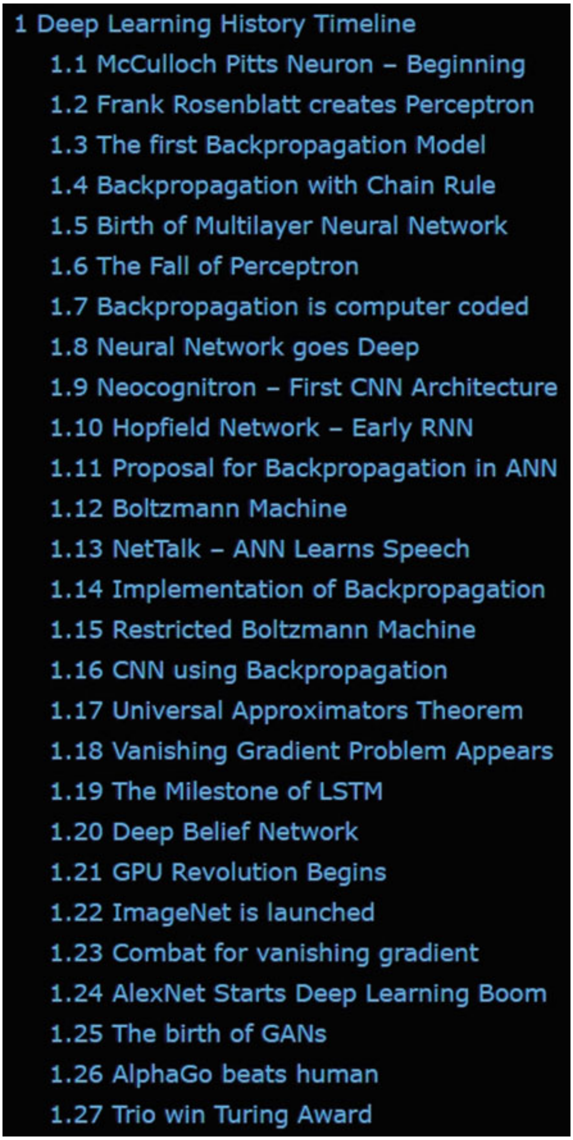
Reinforcement learning method aims at using patterns gathered from the interaction with the dataset to take actions that would maximize the reward or minimize the risk. Reinforcement learning algorithm continuously learns from the dataset in an iterative fashion. In the process, the agent learns from its experiences of the environment until it explores the full range of possible states.

Deep learning techniques are so popular because of their high computing performance. Deep learning can process a large number of features when dealing with unstructured data. Deep learning algorithms pass the data through many layers and each layer extracts some features and passes that to the next layer. Initial layers extract low-level features, and succeeding layers combine features to form a complete representation. The evolution of deep learning is shown in Fig. 7.1.

The machine learning algorithm helps in handling medical records and we can use machine learning algorithms to predict the disease risk in patients as well as helpful in diagnosing the disease. The complex ML algorithms help in development of health care industry by solving complex medical issues.

## 7.2 Related Work

Deep neural networks can be used in supervised learning, unsupervised learning, Reinforcement learning, as well as hybrid learning. Supervised learning uses labeled data to train the neural network. Unsupervised learning uses unlabeled data and learns the recurring patterns. Hybrid learning combines supervised and unsupervised methods to get a better result. Deep learning can be implemented using different architectures such as Convolutional Neural Networks, Recursive Neural Networks, Unsupervised Pre-trained Networks, and recurrent neural Networks. Few powerful training methods that can be applied to deep learning algorithms to reduce the training time and to optimize the model are Back propagation, stochastic gradient descent, learning rate decay, dropout, max-pooling, batch normalization, skip-gram, transfer learning. There are several applications of deep learning networks such as self-driving cars, Fake news detection, Natural language Processing, Google's Virtual



1 Deep Learning History Timeline
- 1.1 McCulloch Pitts Neuron – Beginning
- 1.2 Frank Rosenblatt creates Perceptron
- 1.3 The first Backpropagation Model
- 1.4 Backpropagation with Chain Rule
- 1.5 Birth of Multilayer Neural Network
- 1.6 The Fall of Perceptron
- 1.7 Backpropagation is computer coded
- 1.8 Neural Network goes Deep
- 1.9 Neocognitron – First CNN Architecture
- 1.10 Hopfield Network – Early RNN
- 1.11 Proposal for Backpropagation in ANN
- 1.12 Boltzmann Machine
- 1.13 NetTalk – ANN Learns Speech
- 1.14 Implementation of Backpropagation
- 1.15 Restricted Boltzmann Machine
- 1.16 CNN using Backpropagation
- 1.17 Universal Approximators Theorem
- 1.18 Vanishing Gradient Problem Appears
- 1.19 The Milestone of LSTM
- 1.20 Deep Belief Network
- 1.21 GPU Revolution Begins
- 1.22 ImageNet is launched
- 1.23 Combat for vanishing gradient
- 1.24 AlexNet Starts Deep Learning Boom
- 1.25 The birth of GANs
- 1.26 AlphaGo beats human
- 1.27 Trio win Turing Award

**Fig. 7.1** Evolution of Deep learning models. (https://machinelearningknowledge.ai/brief-history-of-deep-learning/)

Assistant, Fraud detection, Visual Recognition, healthcare, detecting developmental delay in children, adding sound to silent movies, automatic machine translation, text to image translation, image to image synthesis, automatic image recognition, Image colorization, earthquake prediction, market-rate forecasting, and news aggregation.

Cancer starts when cells in the body begin to grow out of control and lung cancer is leading cause of deaths in the current time. Lung cancers typically start in the cells lining the bronchi and parts of the lung such as the bronchioles or alveoli.

There are 2 main types of lung cancer: (i) Non-Small Cell Lung Cancer (NSCLC): About 80–85% of lung cancers are NSCLC. The main subtypes of NSCLC are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. These subtypes, which start from different types of lung cells, are grouped together as NSCLC because their treatment and prognoses (outlook) are often similar (ii) Small Cell Lung Cancer (SCLC): About 10–15% of all lung cancers are SCLC and it is sometimes called oat cell cancer. We can detect the lung cancer at the early stage by checking the presence of pulmonary nodules which may further develop to tumor in future. A new method using image analysis was proposed without the use of drugs for early detection of lung cancer [1]. The image processing techniques were used by the author using CT for early detection of lung cancer [2]. The authors have developed a Computer Aided Diagnosis (CAD) system for lung nodule detection on thoracic helical CT images [3]. The unsupervised learning approach is used by authors for cancer detection and cancer type analysis using gene expression data [4]. Deep learning is an improvement to Artificial neural networks which permits higher abstraction on data. The medical image analysis applying CNN and other deep learning methodologies on different applications [5]. The authors [6] have reviewed the deep learning methods in image registration, detection of cellular structures, tissue segmentation, CAD, etc. The authors discussed the research issues and future scope of deep learning approaches in detection of diseases. The Receiving Operating Curve (ROC), Random Forests (RFs), and Maximum Relevancy and Minimum Redundancy (mRMR) were used to capture the molecular signatures using machine learning methods for classification of small cell lung cancer (SCLC), lung adenocarcinoma (LADC), and squamous cell lung cancer (SQCLC) [7]. The statistical and machine learning techniques were applied to develop a CAD system for classifying the lung cancer. The system follows preprocessing phase, feature extraction phase, feature selection phase, and classification phase. The wavelet transform is used for feature extraction and two-step statistical technique is used for feature e selection. K-means is used for classification and Japanese Society of Radiological Technology's standard dataset of lung cancer is used for evaluating the system [8]. The authors have used curvelet transform and neural network for detection of lung cancer [9]. A CAD system, based on two-level artificial neural network (ANN) architecture is developed for lung nodule detection [10]. Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm were proposed by authors [11] for detection of lung cancer in early stages. A CAD classification method using computed tomography (CT) images of lungs through ANN developed for detection of lung cancer [12]. A diagnosis system of lung cancer detection based on deep learning is proposed in which the input data was generated from human urine using Gas Chromatography Mass Spectrometer (GC–MS) and the system is able to detect whether a person has lung cancer or not with 90% accuracy. This system is useful for pre and personal diagnosis [13]. A deep learning CNN is introduced by authors for nodule classification in CT images and the experimental results suggest that deep learning methods achieved better results in CAD application domain [14]. The feasibility of using deep learning algorithms for lung cancer diagnosis with case studies using Lung Image Database Consortium was done by authors [15]. The authors [17] analyzed that LIDC/IDRI database provide essential medical imaging research resource for CAD development and validation in clinical practice. A scalable tree boosting system XGBoost is created by authors [18] for handling sparse data. A sparse algorithm is proposed for handling sparse data for solving real world scale problems using optimum resources. A gradient descent boosting approach is developed based on fitting criterion. The algorithms are presented for least squares, least absolute deviation, and Huber-M loss functions for regression and classification [19–21]. The screening with low dose CT at early stages can be an effective approach to reduce the mortality from lung cancer [22]. The USPSTF evaluated the lung cancer by screening asymptomatic persons having average or high risk of lung cancer with the harm of screening tests [23, 24]. The concurrent CRT is preferred over sequential CRT in terms of response rate, survival rate, and progression rate [25]. NBIA software program funded by NCI was upgraded by TCIA to create archive of cancer images for collecting cancer imaging data. This resource is helpful for the researchers to publish their articles in this area using high quality data sets [26]. A network and training strategy on use of data augmentation is proposed [27]. A deep learning system based on multi stream multi scale convolutional network for which automatically classifies nodule types was presented [28]. The proposed deep learning system achieves good performance at classifying nodule type. A radiomic analysis of 440 features using image intensity, shape, and texture were extracted from CT data of 1019 patients having lung, head, or neck cancer helped in identifying more radiomic features related to cancer detection [29, 30]. Some easily computable textural features were proposed by authors [31] based on gray-tone spatial dependencies. The basic process of radiomics for evaluation of lung nodules predicting malignancy, histological subtyping, gene expression, and post-treatment prognosis is discussed [32]. The review study conducted by authors [33] proposed that texture features can predict patients with cancer. But the investigation on a single data set can cause inflation of type I errors. The authors have done a systematic review of type I error to analyze the associations between patient outcome and texture features derived using PET and CT [33]. A detail report on LUNGx challenge for computerized lung nodule classification and results were discussed [34]. Kaggle organized a Data Science Bowl (DSB) as an annual machine learning competition to build an automated system of forecasting lung cancer using CT scan[35]. A description of LUNA16 challenge is presented with results and the detection performance was also investigated and it was depicted that convolutional network can be used for automatic detection of pulmonary nodules in CT images [36]. The outcome of small lung nodules missed by (18)F-FDG PET/MRI was evaluated and it was depicted that small lung modules were benign [37]. A Wilcoxon Signed Generative Deep Learning (WS-GDL) method used by authors [38] based on machine learning technique for lung cancer detection. This study uses generator function and discriminator function for lung cancer diagnosis [38]. The authors have proposed an Optimizer Based Deep Neural Network (Op-DNN) to predict lung cancer patients survivability using regression technique [39].

## 7.3 Background

### *7.3.1 Convolutional Neural Network*

CNN is a deep learning method which takes an input image and then assigns various objects in the images to separate from the other. The architecture of convolutional network as collection of neurons in human brain and inspired by visual cortex. The CNN reduces the images size for further processing without losing the features. The CNN composed of multiple artificial neurons. The CNN can be used with deep learning as CNNs eliminate the manual feature extraction, CNN produces highly accuracy results, and CNN can be used to build pre-existing networks. The CNN application is composed of medical imaging, audio processing, stop sign detection, and synthetic data generation CNN can be used in image processing, Artificial Intelligence, deep learning, etc. CNN is able to execute different tasks and uses machine vision for image and video recognition as well as natural language processing for text classification. It is also called feed-forward neural network with 20–30 layers. It has 3–4 convolutional layers which help in recognizing hand written digits as well as it is able to distinguish human faces.

### *7.3.2 Deep Learning*

Deep learning is a sub field of machine learning that covers the structure and function of brain using ANNs. It consists of multiple layers of interconnected nodes build upon layer to refine and optimize the prediction and categorization of nodes. It is done by forward propagation. Then input layer performs data processing and output layer performs prediction or classification. The back propagation uses gradient descent for calculating and correcting errors. The deep learning applications include law enforcement, financial services, customer services and health care, etc.

### *7.3.3 Applications of Machine Learning*

Managing medical data: The machine learning is helpful in health care industry to perform medical diagnosis. The medical imaging and diagnosis will be helpful in developing deep learning models, MRI scanning, etc.

(i) The machine learning algorithms are very helpful in predicting medical diseases like heart attacks, diabetes, etc. The AI based devices are helpful in displaying warning whenever something is unusual. The apple watch is helpful in monitoring persons heart rate, sleep cycle, activity level, bp 24*7.

(ii) Machine learning is helpful in medical assistants also. The virtual nurses and assistants are helpful in monitoring patient health conditions which is helpful in treatment of disease and follow up visits with doctors.

(iii) The ML is very helpful in the decision making process in various fields. The ML algorithms may be helpful in identifying the customer needs and also helpful in predicting the risk related to the businesses. The surgical robots can minimize the errors and helpful in increasing the efficiency of surgeons. It is also helpful in performing surgeries with better flexibility.

(iv) The ML can be helpful for a user to get personalized treatment. The ML technologies can analyze the patient data and history to generate treatment reports. The ML algorithm is helpful in detecting and analyzing the errors in prescriptions also. The machine intelligence is helpful in inspecting patient's health records and correcting possible errors. The ML can be helpful in analyzing radiological images for predicting particular disease or disorder. The ML can be useful in analyzing

Retinal images for detecting the problems related to vision. The precision mechanism is one of the ML applications. It targets the specific requirement of person based on characteristics like genetic behavior.

### *7.3.4 Lung Cancer Causes*

In case lung blemishes at the start that may increase the chance of lung cancer. The denseness in the blemishes can be identified with CT. Lung Cancer Cell Testing Routes (LCST) may identify lung cancer and CT testing performed by radiologist for diagnosis. After CT scan, radiologist analyze the information in picture based on nodules modules and identifying the consequences based on detection velocity. The CAD methods are used for image processing and detection of latent lesions in health. The device CADx can be used in detection of lesions in the health. CADe method is also very efficient in detection of lesion with much better accuracy and lower interpretation time, with false positive velocity, low cost and it also determines the different size, space, place to avoid attacks.

Machine Learning: The machine learning is field of AI that provides an ability to learn and improve from experience and related to development of programs, the computer learns the data automatically without human intervention. It is a method of data analysis for identifying patterns and making decision without human intervention.

Supervised: This model is based on the uncertainty. It collects the set of input data and then converts to output and trains the model to generate the responses. It uses the classification and regression to build machine learning models.

Un-supervised: It finds the hidden patterns in data. It is used to draw interference from data. The clustering is a technique of unsupervised learning. The system is not able to identify the correct output.

Reinforcement: This model uses feedback learning method. The learning agent received awards for correct actions and penalty for incorrect actions. The robotic dog is an example of reinforcement learning.

## 7.4 Conclusion

A systematic review on lung cancer detection has been presented. Various methods were discussed with their significance and different machine learning approaches were also discussed. The authors have used many classifiers like SVM, M2P, Naïve–Byes, Neural networks, decision trees, etc. The literature review suggests that lung nodule classification is benign and malignant and deep learning with CNN is also used for image classification. The ML applications can be expandable independently to build hospitals in the future without physicians. The robotics was integrated with ML and AI to run the hospital. In the future robots may perform disease diagnoses and surgery. For example, the Mayo clinic is creating hospitals without doctors. The components are designed and testing is being performed on various components. The surgeons are also taking the help of robots to manage the surgical process.

**Acknowledgements** I would like to express my special thanks to Jain University for supporting me for writing this chapter.

## References

1. Palcic, B., Lam, S., Hung, J., MacAulay, C.: Detection and localization of early lung cancer by imaging techniques. CHEST J. **99**(3) 742–743 (1991)
2. Yamomoto, S., Jiang, H., Matsumoto, M., Tateno, Y., Iinuma, T., Matsumoto, T.: Image processing for computer-aided diagnosis of lung cancer by CT(LSCT). In: Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96, pp. 236–241 (1996). https://doi.org/10.1109/ACV.1996.572061
3. Gurcan, M.N., Sahiner, B., Petrick, N., Chan, H.P., Kazerooni, E.A., Cascade, P.N., Hadjiiski, L.: Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. Med. Phys. **29**(11), 2552–2558. https://doi.org/10.1118/1.1515762. PMID: 12462722
4. Fakoor, R., Ladhak, F., Nazi, A., Huber M.: Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning (2013)
5. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imag. **35**(5), 1153–1159 (2016)
6. Shen, D., Wu, G., Suk H.-I.: Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. (2017)
7. Cai, Z., et al.: Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molec. BioSyst. **11**(3), 791–800 (2015)
8. Al-Absi Hamada, R.H., Belhaouari Samir, B., Sulaiman, S.: A computer aided diagnosis system for lung cancer based on statistical and machine learning techniques. JCP **9**(2), 425–431 (2014)

9. Gupta, B., Tiwari, S.: Lung cancer detection using curvelet transform and neural network. Int. J. Comput. Appl. **86**(1) (2014)
10. Penedo, M.G., et al.: Computer-aided diagnosis: a neural-network-based approach to lung nodule detection. IEEE Trans. Med. Imag. **17**(6), 872–880 (1998)
11. Taher, F., Sammouda, R.: Lung cancer detection by using artificial neural network and fuzzy clustering methods. In: GCC Conference and Exhibition (GCC). IEEE (2011)
12. Kuruvilla, J., Gunavathi, K.: Lung cancer classification using neural networks for CT images. Comput. Methods Program. Biomed. **113**(1), 202–209 (2014)
13. Shimizu, R., et al.: Deep learning application trial to lung cancer diagnosis for medical sensor systems. In: 2016 International on SoC Design Conference (ISOCC). IEEE (2016)
14. Hua, K.-L., et al.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Therapy **8**, 2015–2022 (2014)
15. Sun, W., Zheng, B., Qian, W.: Computer aided lung cancer diagnosis with deep learning algorithms. In: SPIE Medical Imaging. International Society for Optics and Photonics (2016)
16. Armato, S.G., et al.: The lung image database consortium (LIDC) and image data-base resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
17. He, K., Zhang, X., Ren, S., Deep, S.J.: residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2016)
19. Friedman, J.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
20. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000)
21. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
22. Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Gareen, I.F., Gatsonis, C., Marcus, P.M., Sicks, J.D.: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**, 395–409 (2011)
23. Moyer, V.A.: U.S. preventive services task force. Screening for lung cancer: U.S. Preventive services task force recommendation statement. Ann. Int. Med. **160**, 330–338 (2014)
24. Lung Nodule Analysis (LUNA) Challenge. https://luna16.grand-challenge.org/description/
25. Zatloukal, P., et al.: Concurrent versus sequential chemoradiotherapy with cisplatin and vinorelbine in locally advanced non-small cell lung cancer: a randomized study. Lung Cancer **46**(1), 87–98 (2004)
26. Clark, K., et al.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**, 1045–1057 (2013)
27. Ronneberger, O., Fischer, P., Thomas Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351, pp. 234–241. Springer, LNCS (2015)
28. Ciompi, F., Chung, K., van Riel, S.J., et al.: Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci. Rep. **7**, 46479 (2017). [Crossref] [PubMed]
29. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**, 4006 (2014). [Crossref] [PubMed]
30. Lambin, P., Rios-Velazquez, E., Leijenaar, R., et al.: Extracting more information from medical images using advanced feature analysis. Eur. J. Cancer **48**, 441–446 (2012). [Crossref] [PubMed]
31. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. Syst. **3**, 610–621 (1973). [Crossref]
32. Wilson, R., Devaraj, A.: Radiomics of pulmonary nodules and lung cancer. Transl. Lung Cancer Res. **6**, 86–91 (2017). [Crossref] [PubMed]

33. Chalkidou, A., O'Doherty, M.J., Marsden, P.K.: False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One **10**, e0124165 (2015). [Crossref] [PubMed]
34. Armato, S.G., Drukker, K., Li, F., et al.: LUNGx challenge for computerized lung nodule classification. J. Med. Imaging (Bellingham) **3**, 044506 (2016). [Crossref] [PubMed]
35. Hammack, D.: Forecasting lung cancer diagnoses with deep learning. Available online: https://raw.githubusercontent.com/dhammack/DSB2017/master/dsb_2017_daniel_hammack.pdf
36. Setio, A.A., Traverso, A., de Bel, T., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. **42**, 1–13 (2017). [Crossref] [PubMed]
37. Sawicki, L.M., Grueneisen, J., Buchbender, C., Schaarschmidt, B.M., Gomez, B., Ruhlmann, V., Umutlu, L., Antoch, G., Heusch, P.: Evaluation of the outcome of lung nodules missed on 18F-FDG PET/MRI compared with 18F-FDG PET/CT in patients with known Malignancies. J. Nucl. Med. **57**(1), 15–20. https://doi.org/10.2967/jnumed.115.162966. Epub 2015 Oct 29. PMID: 26514173
38. Obulesu, O., Kallam, S., Dhiman, G., Patan, R., Kadiyala, R., Raparthi, Y., Kautish, S.: Adaptive Diagnosis of lung cancer by deep learning classification using wilcoxon gain and generator. J. Healthcare Eng. Article ID 5912051, 13 (2021). https://doi.org/10.1155/2021/5912051
39. Pradeep, K.R, Naveen, N.C.: A framework for lung cancer survivability prediction using optimized-deep neural network classification and regression technique. Int. J. Comput. Sci. Eng. 07(13), 57–66 (2019)

# Chapter 8
# Fetal Health Status Prediction During Labor and Delivery Based on Cardiotocogram Data Using Machine and Deep Learning

**Anu Singha, Joe Raj S. Noel, R. V. Adhikrishna, Nived Suthahar, S. Abinesh, and S. Jaya Sakthi Poorni**

## 8.1 Introduction

The happiest time in a woman's life is during her pregnancy. Given that she is carrying a child, a woman must actually take excellent care of her health during pregnancy. The fetus grows and develops during every three month period in the pregnancy process, this period is called as trimester. A number of laboratory tests are advised each trimester to track fetal growth and development. Since the pregnancy duration is about 9 months and in this period, there may be various reasons which may cause a problem which may lead to mortality in the newborn. Thus the probability of such incidents to take place must be minimized. One of the most significant tools to analyze the health of the fetal in the womb is by doing a Cardiotocography (CTG) which is profoundly used in evaluating the fetal heart beat rate (FHR) and uterine contraction (UC) of the fetus. This examination allows medical professionals to determine whether the fetus is healthy both before and after delivery [1]. For newborns' short- and long-term health outcomes, normal fetal growth is essential. The majority of morbidity and medical expenses in neonates born at term 2–4 are caused by infants at both ends of the birth weight distribution. The findings of CTG tests can be categorized into three groups by the International Federation of Gynecology and Obstetrics (FIGO): normal, suspect, or pathological [2]. The foundation of these classes is fetal heart rate (FHR) and uterine contractions (UC). FHR is the process of checking the condition of your fetus during labor and delivery by monitoring your fetus's heart rate with

A. Singha (✉)
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad, U.P., India
e-mail: anusingh5012@gmail.com

J. R. S. Noel · R. V. Adhikrishna · N. Suthahar · S. Abinesh · S. J. S. Poorni
Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_8

special equipment. UC is the tightening of uterine muscle fibers that occurs briefly and intermittently throughout pregnancy, and more regularly and forcefully during active labor.

According to the world health organization (WHO) report [3], every day in the year 2017, approximately 810 mothers died from preventable causes related to pregnancy and childbirth. Between 2000 and 2017, the maternal mortality ratio (MMR) which is a number of maternal deaths per 100,000 live births dropped by 38% worldwide. Sub-Saharan Africa and Southern Asia accounted for approximately 86% (254,000) of the estimated global maternal deaths in 2017.

Electronic fetal monitoring (EFM) is used in labor in order to detect fetuses at a risk of distress who might benefit from an emergency operative delivery (Caesarean or instrumental vaginal delivery), as mentioned in Fig. 8.1. However, visual interpretation is unreliable and ambiguous, the complex EFM graphs that continuously display the FHR and UCs remain poorly interpreted. A few classic EFM patterns which have been empirically identified and for certain EFM patterns where the disagreement in visual interpretation between experts reaches 100%. Computerized detection of such classic patterns, mimicking clinical visual assessment, is available commercially, but has not given any benefit in randomized clinical trials. The main goal of monitoring is the early detection of fetal hypoxia. The accurate analysis of cardiotocograms is required for further treatment. The fetal health assessment using machine learning and deep learning method from cardiotocogram data has been the center of attraction in literature [4–18]. Thus, the deep learning and machine learning concepts are significantly increasing the reliability and the accuracy in monitoring through the method of cardiotocography.

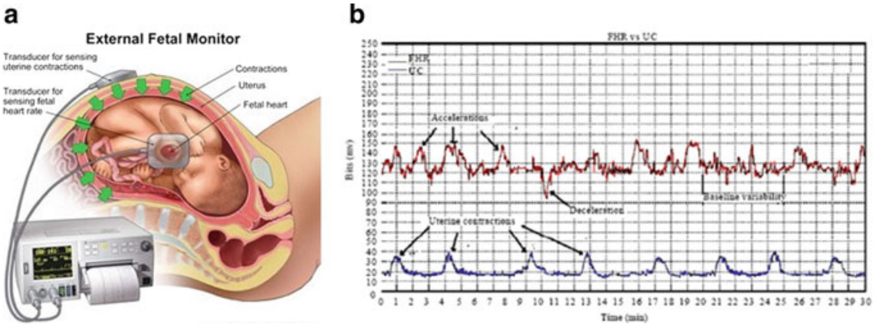


**Fig. 8.1** Brief visualization of cardiotocogram tool and signals

The contributions of the chapter are as follows:

- To understand the concepts of fetal health and the current research scenarios over machine learning and deep learning approaches, we have done a brief survey on the recent articles.
- We analyze the fetal health dataset information through several visualization techniques to pick right and meaningful data, and to pre-process those data before fit into training the machine learning and deep learning models.
- Implemented total 3 machine learning and 3 deep learning models for classification. These implementations were followed by extensive comparative analysis among three classes (Normal, Suspect, Pathological fetus) over these machine and deep learning models using metrics like precision, recall, $F_1$-score, and accuracy.

The rest of the chapter is organized as follows: Sect. 8.2 describes the related work, and Sect. 8.3 describes elaborately the machine and deep learning framework in this study. Section 8.4 the experimental results, discussions, and comparison with the state-of-the-art. Finally, the paper is concluded by Sect. 8.5.

## 8.2 Related Work

A machine learning approach that is support vector machine (SVM) has been used by several researchers to analyze CTG data. Cardiotocography was the main topic of Czabanski et al. [5] study on the bio-physical assessment of fetal state. The classification of FHR signals using a fuzzy approach is carried out as the initial stage. The proposed method then achieves a success rate in the second phase utilizing lagrangian support vector machines (LSVM). Using empirical mode decomposition and SVM, Krupa et al. [6] analyzed newborns' CTG data. On 90 randomly chosen data, classification into normal or at risk classifications could be made with 86% accuracy. Ocak [7] conducted a CTG analysis over UCI dataset of FHR-UC and suggested a model for estimating the fetal health using the SVM approach. Furthermore, the genetic algorithm used to remove the unnecessary features from the dataset which improved the classification performance. Batra et al. [8] studied a conjunction of the following several machine learning approaches to evaluate fetal distress. These are decision tree, support vector machine, random forest, neural networks, gradient boosting. The author achieved maximum accuracy of 99.25% which is higher than what was obtained in previous research. Kuhle et al. [9] were studied a data for 30,705 singleton infants born between 2009 and 2014 to mothers in Nova Scotia, Canada. The observation for 26 weeks was considered as predictor of the infant's size and gestational age. Only 7.9 and 13.5% of infants were SGA (primary outcomes were small) and LGA (large for gestational age), respectively and 48.6% of births were to primiparous women. The methods of logistic regression and other machine learning approaches were used to build the models. The accuracy of prediction for SGA and LGA based on maternal information is poor for primiparous women and fair for multiparous. Mehbodniya et al. [10] also conducted a research to deploy various machine learning algorithms to predict fetal health from the CTG data. It assesses influence of various factors measured through CTG to predict the health state of the fetus through algorithms like SVM, random forest, and k-nearest neighbors (KNN). The experimented results of the algorithms show that random forest performs better than its peers in terms of accuracy, precision, recall, and $F_1$-score. Similar work had seen in article [11] which was analyzed via seven machine learning algorithms. Classification is imperative in diagnosing the health of the fetus and new born specifically in critical cases such as uncertainty in CTG data [12]. The research at [12] aimed to classify the CTG data points into normal, suspicious, and pathologic using machine learning models through rough set. In the study, particle swarm optimization was used in pre-processing for selecting the relevant features, and rough set approximations were exploited in extracting the uncertain information from the data set. Akbulut et al. [13] developed a prediction system with assistive e-Health applications which both the pregnant women and practitioners can make use of 9 binary classification models such as averaged perceptron, boosted decision tree, bayes point machine, decision forest, and decision jungle. These models were trained with the private clinical dataset of 96 pregnant women which was obtained through maternal questionnaire and evaluations of 3 clinicians from RadyoEmar radiodiagnostics center in Istanbul, Turkey. The authors claimed that the proposed work aims to provide assistive services to pregnant women and clinicians via an online system consisting of a mobile side for patients, a web application side for their clinicians, and a prediction system.

The fetal mortality rate has been constantly rising, and also the genesis of machine and deep learning analysis in the healthcare domain has shown remarkable progress in diagnosis of various diseases. Deep learning is notably used in speech recognition and exists as a form of natural language processing unit. These concepts were adapted for medical data analysis in the past half-decade and have henceforth proven to be successful. Petrozziello et al. [1] analyzed the deep learning concept such as long short term memory (LSTM) and convolutional neural networks (CNN) to the electronic fetal monitoring (EFM) traces from 35,000 labors. The evaluation performance showed the sensitivity score of 42%. Yilmaz [4] studied a comparative assessment of fetal state through artificial neural network (ANN) models such as multilayer perception neural network (MLP-NN), probabilistic neural network (PNN), and generalized regression neural network (G-RNN). The experimented results showed that all the ANN models have high classification accuracy as MLP-NN is 90.35%, PNN is 92.15%, G-RNN is 91.86%. One of the most difficult tasks for the physicians is to acquire a quality fetal electrocardiogram (ECG) to analyze, manage, and plan according to the condition of the fetus in the womb. The fetal ECG signal can be acquired only after 25 weeks the fetus is developed in the womb. Jagannath et al. [14] utilized the deep learning techniques are used such as deep belief neural network (DBNN), convolutional neural network (CNN), back propagation neural network (BPNN) to execute the ECG synthesis. The evaluation results obtained are based on the signal-to-noise ratio (SNR) which were achieved 39.54 for DBNN, 39.732 for CNN, 39.454 for BPNN, respectively. The visual interpretation of the FHR signals by obstetricians varies due to the complexity and non-linearity nature of signal. Therefore, Zhao et al. [15] designed computer aided diagnosis systems based on an 8-layer deep CNN framework to predict the fetal acidemia. The FHR signals are pre-processed to the input 2-dimensional images. Unlike the conventional machine learning, the 2D CNN model can self-learn the features from the input data which representing the tremendous advantage over machine learning. The testing on the Open-access database achieved classification accuracy was about 98.34%. A. Petrozziello et al. [16] introduced a multi-model convolutional neural network (MCNN) and its consequence stacked MCNN to analyze fetal compromise during labor and delivery. MCNN enriched the prediction of cord acidemia at birth when compared with clinical practice.

## 8.3 Methodology

Figure 8.2 shows the systematic diagram of the workflow of this chapter. The data is collected from the dataset in kaggle repository, namely, fetal_health.csv. Several pre-processing methods are implemented like acquiring the dataset, finding the null values, splitting the dataset, and standardizing for scale of the features. The dataset is visualized with the help of plots and graphs which helps us to feature selection approach. After data preparation, the dataset is being split into train-test sets and fits into the desired machine and deep learning models. There are 3 machine learning models which are chosen, namely, decision tree, random forest, naïve bayes; and 2 deep learning concepts also chosen, namely, multilayer perceptron (MLP), recurrent neural network (RNN). In the testing phase, the trained models are tested via testing dataset and get executed into 3 classes such as normal, suspect, and pathological fetus. At last, performance evaluations are conducted in terms of precision, recall, $F_1$-score, and accuracy.

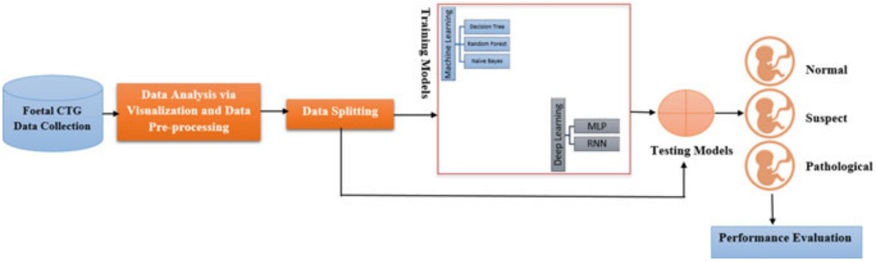


**Fig. 8.2** A schematic diagram of workflow of this chapter

## *8.3.1 Machine Learning Models*

In this chapter, we have implemented three well-known machine learning models.

**Algorithm 1**: Decision Tree

```
Step 1: Calculate the Information Gain of each feature.
Step 2: Considering that all rows don't belong to the same class, split the dataset **S** into subsets using the feature for which the Information Gain is maximum.
Step 3: Make a decision tree node using the feature with the maximum Information gain.
Step 4: If all rows belong to the same class, make the current node as a leaf node with the class as its label.
Step 5: Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.
```

### 8.3.1.1 Decision Tree [17]

Decision Tree is a supervised learning technique which having a structure that contains nodes and edges. Each node is either used to make a decision (known as decision node) or represents an outcome (known as leaf node). The algorithmic steps are shown in Algorithm 1. The step-by-step explanation of algorithm with sample dataset as follows.

*Step 1. Calculate the Information Gain of each feature.*

We'll be using a sample dataset of fetal_health. A preview of the entire dataset is shown below.

| Uterine contracrions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| Low | Low | Low | Normal |
| High | High | High | Pathological |
| High | High | Low | Normal |
| High | Low | High | Pathological |
| High | High | High | Pathological |
| Low | High | Low | Normal |
| High | Low | High | Pathological |
| High | Low | High | Pathological |
| Low | High | High | Pathological |
| High | High | Low | Pathological |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | High | Normal |
| High | High | Low | Normal |

Information Gain (IG) calculates the reduction in the entropy and measures how well a given feature separates or classifies the target classes. The feature with the highest IG is selected as the best one. In simple words, Entropy (E) is the measure of disorder and the Entropy of a dataset is the measure of disorder in the target feature of the dataset. In the case of binary classification (where the target column has only two types of classes) entropy is 0 if all values in the target column are homogenous (similar) and will be 1 if the target column has equal number values for both the classes. We denote our dataset as S, entropy is calculated as:

$$\text{Entropy (S)} = -\sum \mathrm{p_i} * \log_2(\mathrm{p_i});\quad \mathrm{i} = 1 \text{ to } \mathrm{n} \tag{8.1}$$

where n is the total number of classes in the target column (in our case $\mathrm{n} = 2$ i.e. Pathological and Normal), $\mathrm{p_i}$ is the probability of class 'i' or the ratio of 'number of rows with class i in the target column' to the 'total number of rows' in the dataset. Then, IG for a feature column A is calculated as:

$$\mathrm{IG(S, A)} = \mathrm{Entropy(S)} - \sum ((|\mathrm{S_v}|/|\mathrm{S}|) * \mathrm{Entropy(S_v)}) \tag{8.2}$$

where $\mathrm{S_v}$ is the set of rows in S for which the feature column A has value v, $|\mathrm{S_v}|$ is the number of rows in $\mathrm{S_v}$, and likewise |S| is the number of rows in S. Now, these concepts are implementing on our dataset. We'll calculate the IG for each of the features now, but for that, we first need to calculate the entropy of **S.** From the total of 14 rows in our dataset S, there are 8 rows with the target value PATHOLOGICAL and 6 rows with the target value NORMAL. The entropy of S is calculated as:

$$\text{Entropy (S)} = -(8/14) * \log_2(8/14) - (6/14) * \log_2(6/14) = 0.99$$

We now calculate the IG for each feature. In the uterine contractions (UC) feature, there are 8 rows having value HIGH and 6 rows having value LOW. As shown below, in the 8 rows with HIGH for Uterine contractions, there are 6 rows having target value PATHOLOGICAL and 2 rows having target value NORMAL.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| High | High | High | Pathological |
| High | High | Low | Normal |
| High | Low | High | Pathological |
| High | High | High | Pathological |
| High | Low | High | Pathological |
| High | Low | High | Pathological |
| High | High | Low | Pathological |
| High | High | Low | Normal |

As shown below, in the 6 rows with LOW, there are 2 rows having target value PATHOLOGICAL and 4 rows having target value NORMAL.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| Low | Low | Low | Normal |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | High | Normal |

The below content demonstrates the calculation of IG for UC.

$|S| = 14$.

For v = HIGH**,**

$|S_v| = 8$.

$\text{Entropy}(S_v) = -(6/8) * \log_2(6/8) - (2/8) * \log_2(2/8) = 0.81$.

For v = LOW,

$|S_v| = 6$.

$\text{Entropy}(S_v) = -(2/6) * \log_2(2/6) - (4/6) * \log_2(4/6) = 0.91$.

Expanding the summation in the IG formula:

IG(S**,** Uterine contractions) = Entropy(S) − (|SHIGH|/|S|) * Entropy(SHIGH) – (|SLOW|/|S|) ***** Entropy(SLOW).

∴ IG(S, Uterine contractions) = 0.99 − (8/14) * 0.81 − (6/14) * 0.91 = 0.13.

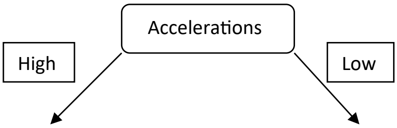
Next, we calculate the IG for the features 'Light Decelerations' and 'Accelerations'.

IG(S, Light decelerations) = 0.04.

IG(S, Accelerations) = 0.40.

*Steps 2 and 3: Considering that all rows don't belong to the same class, split the dataset S into subsets using the feature for which the Information Gain is maximum and make a decision tree node using the feature with the maximum Information gain.*

Since the feature Accelerations have the highest Information Gain it is used to create the root node. Hence, after this initial step our tree looks like this:



Next, from the remaining two unused features, namely, uterine contractions and light decelerations**,** we decide which one is the best for the left branch of Accelerations**.** Since the left branch of Accelerations denotes HIGH**,** we will work with the subset of the original data i.e. the set of rows having HIGH as the value in the Accelerations column. These 8 rows are shown below:

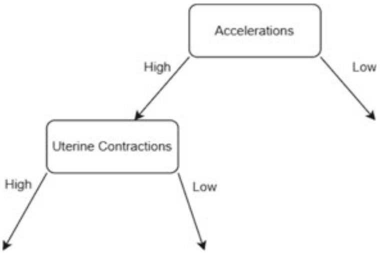
| Light decelerations | Accelerations | Fetal health |
|---|---|---|
| High | High | Pathological |
| Low | High | Pathological |
| High | High | Pathological |
| Low | High | Pathological |
| Low | High | Pathological |
| High | High | Pathological |
| High | High | Pathological |
| High | High | Normal |

Next, we calculate the IG for the features of uterine contractions and light decelerations using the subset **S**AH (**S**et **A**ccelerations **H**igh) which is shown above:

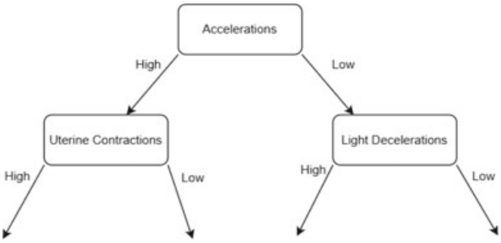
IG(**S**AH, Uterine contractions) = 0.20.

IG(**S**AH, Light decelerations) = 0.09.

IG of uterine contractions is greater than that of light decelerations, so we select uterine contractions as the left branch of Accelerations. Our tree now looks like this:



Next, we find the feature with the maximum IG for the right branch of Accelerations. But, since there is only one unused feature left we have no other choice but to make it the right branch of the root node. So, our tree now looks like this:



*Step 4: If all rows belong to the same class, make the current node as a leaf node with the class as its label.*

There are no more unused features, so we stop here and jump to the final step of creating the leaf nodes. For the left leaf node of uterine contractions, we see the subset of rows from the original data set that has Accelerations and uterine contractions both values as HIGH.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| High | High | High | Pathological |
| High | Normal | High | Pathological |
| High | High | High | Pathological |
| High | Normal | High | Pathological |
| High | Normal | High | Pathological |

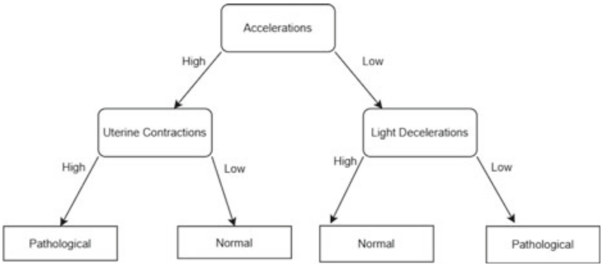
Since all the values in the target column are HIGH, we label the left leaf node as HIGH, but to make it more logical we label it Infected.

*Step 5: Repeating until there are no more subsets or all are leaf nodes at the end.*

Similarly, for the right node of Uterine Contractions we see the subset of rows from the original data set that have Accelerations value as HIGH and uterine contractions as LOW.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| Low | High | High | Pathological |
| Low | High | High | Normal |
| Low | High | High | Normal |

Here not all but most of the values are LOW, hence LOW or Normal becomes our right leaf node. Our tree, now, looks like this:



#### 8.3.1.2 Random Forest

Random Forest (RF) [18] is another standard machine learning approach that belongs to the supervised learning procedure. It is based on the concept of ensemble learning, which is a process of combining N decision trees are combined to generate the random forest, and then predictions are made for each tree that was produced in the first phase. The algorithmic steps are shown in Algorithm 2. The step-by-step explanation of algorithm with sample dataset as follows.

**Algorithm 2**: Random Forest

```
Step 1: Select random K data points from the training set.
Step 2: Build the decision trees associated with the selected data points (Subsets).
Step 3: Choose the number N for decision trees that you want to build.
Step 4: For new data points, find the predictions of each decision tree, and assign the new data
points to the category that wins the majority votes.
```

*Step 1. Select random K data points from the training set.*

The algorithm *s*elects a bunch of rows randomly. This process is called bootstrapping (random replacement). For our example, let's assume that it selects **k** records. We'll be using a sample dataset of fetal_health. A preview of the entire dataset is shown below.

| Uterine contracrions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| Low | Low | Low | Normal |
| High | High | High | Pathological |
| High | High | Low | Normal |
| High | Low | High | Pathological |
| High | High | High | Pathological |
| Low | High | Low | Normal |
| High | Low | High | Pathological |
| High | Low | High | Pathological |
| Low | High | High | Pathological |
| High | High | Low | Pathological |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | High | Normal |
| High | High | Low | Normal |

RF randomly selects a subset of features/columns. Here, for the sake of simplicity and for the example, we are choosing 3 random features i.e. Uterine contractions, Light decelerations, Accelerations.

Next, we have to select the root node.

$$\mathrm{IG(S,\ A)} = \mathrm{Entropy(S)} - \sum((|\mathrm{S_v}|/|\mathrm{S}|) * \mathrm{Entropy(S_v)})$$

where $\mathbf{S_v}$ is the set of rows in **S** for which the feature column **Uterine Contractions** has value **v**, **|$\mathbf{S_v}$|** is the number of rows in $\mathbf{S_v}$**,** and likewise **|S|** is the number of rows in **S**.

We now calculate the Information Gain for each feature:

**IG calculation for Uterine Contractions**

In this (Uterine contractions) feature there are **8** rows having value **HIGH** and **6** rows having value **LOW**. As shown below, in the **8** rows with **HIGH for** Uterine contractions, there are **6** rows having target value **Pathological** and **2** rows having target value **Normal**.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| High | High | High | Pathological |
| High | High | Low | Normal |
| High | Low | High | Pathological |
| High | High | High | Pathological |
| High | Low | High | Pathological |
| High | Low | High | Pathological |
| High | High | Low | Pathological |
| High | High | Low | Normal |

As shown below, in the **6** rows with **LOW**, there are **2** rows having target value **Pathological** and **4** rows having target value **Normal**.

| Uterine contractions | Light decelerations | Accelerations | Fetal health |
|---|---|---|---|
| Low | Low | Low | Normal |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | Low | Normal |
| Low | High | High | Pathological |
| Low | High | High | Normal |

*Step 2. Build the decision trees associated with the selected data points (subsets).*

The below content demonstrates the calculation of Information Gain for **Uterine Contractions**.

|S| = 14.

For v = HIGH,

$|\mathrm{S_v}| = 8$.

$\mathrm{Entropy(S_v)} = \text{-}(6/8) * \log_2(6/8)\text{-}(2/8) * \log_2 (2/8) = 0.81$.

For v = LOW,

$|S_v| = 6$.

Entropy($S_v$) = -(2/6) * $\log_2$(2/6)-(4/6) * $\log_2$(4/6) = 0.91.

**Expanding the summation in the IG formula**

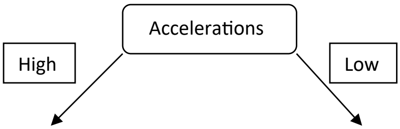
IG(S**,** Uterine contractions) = Entropy(S)-(|SHIGH|/|S|) * Entropy(SHIGH)–(|SLOW| / |S|) * Entropy(SLOW).

∴ IG(S, Uterine contractions) = 0.99-(8/14) * 0.81-(6/14) * 0.91 = 0.13.

Next, we calculate the **IG** for the features ‘**Light Decelerations**’ and ‘**Accelerations**’.

IG(S, Light decelerations) = 0.04.

IG(S, Accelerations) = 0.40.



Once the 3 random features are selected (in our example), the algorithm runs a splitting of the **k** record (from step 1) and does a quick calculation of the before and after values of a metric.

This metric could be either gini impurity or the entropy. Whichever of the random feature split gives the least combined gini impurity/entropy value, that feature is selected as the root node.

Now, we have to select the child nodes.

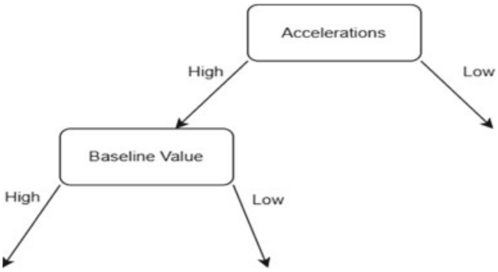
The algorithm performs the same process as in the previous root node selection and selects another set of 3 random features i.e. Baseline value, Prolonged decelerations, and Fetal Movement.

IG(S, Baseline Value) = 0.67.

IG(S, Prolonged decelerations) = 0.45.

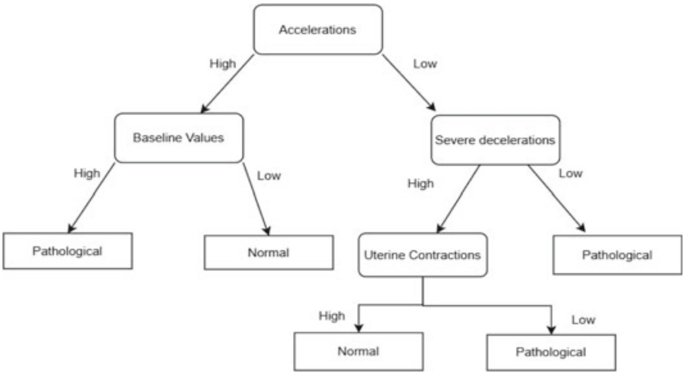
IG(S, Fetal Movement) = 0.27.

Hence, baseline value will go at the left side of the root node.



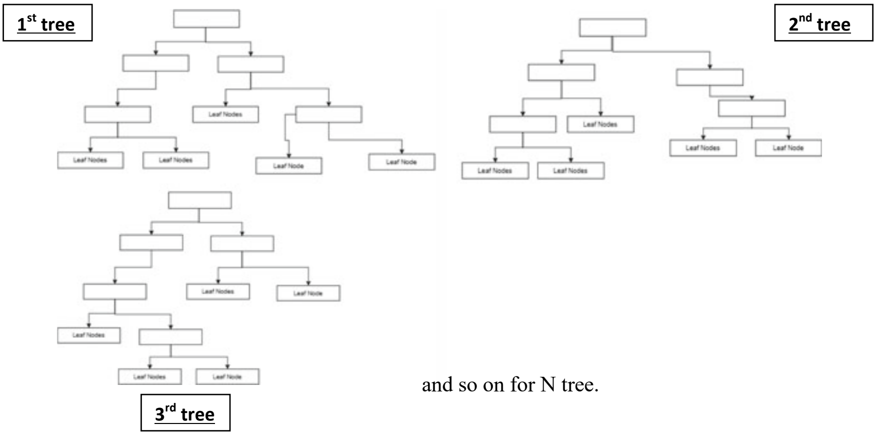
Based on the criteria (Gini/Entropy), it selects which feature will go into the next node/child node, and further splitting of the records happens here.

This process continues of selecting the random feature and splitting of the nodes happens till you have reached the number of splits specified by you i.e. (node = 5). You now have your first decision tree.



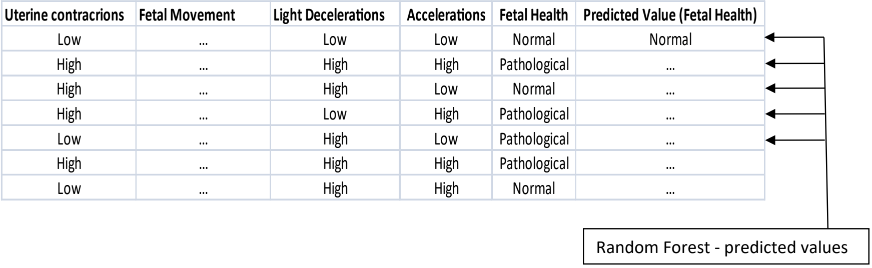
*Step 3. Choose the number N for decision trees that you want to build*

Algorithm goes back to your data and does steps 1–2 to create the N number of decision trees.



*Step 4. Find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.*

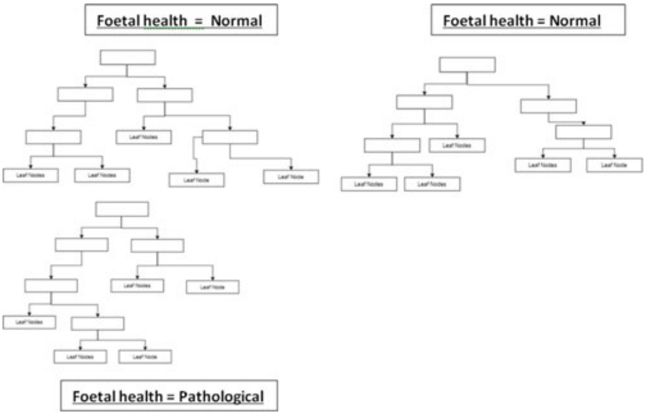
Once the default value of 100 trees is reached (i.e. you have 100 mini decision trees), the model is said to have completed its fit() process. Now let's predict the values in an unseen test dataset. For inference (more commonly referred to as predicting/scoring) the test data, the algorithm passes the record through each tree.



| Uterine contracrions | Fetal Movement | Light Decelerations | Accelerations | Fetal Health | Predicted Value (Fetal Health) |
|---|---|---|---|---|---|
| Low | ... | Low | Low | Normal | Normal |
| High | ... | High | High | Pathological | ... |
| High | ... | High | Low | Normal | ... |
| High | ... | Low | High | Pathological | ... |
| Low | ... | High | Low | Pathological | ... |
| High | ... | High | High | Pathological | ... |
| Low | ... | High | High | Normal | ... |

Random Forest - predicted values

The values from the record traverse through the mini tree based on the variables that each node represents and reaches a leaf node ultimately. Based on the predetermined value of the leaf node (during training) where this record ends up, that mini tree is assigned one prediction output. In the same manner, the same record goes through all the 100 mini decision trees and each of the 100 trees have a prediction output for that record as follows:



The final prediction value for this record is calculated by taking a simple voting of these 100 trees. Hence we get the predicted value of the fetal health by using random forest.

### 8.3.1.3 Naïve Bayes

Naïve Bayes classifiers [19] are a collection of classification methods based on Bayes' theorem. The algorithmic steps are shown in Algorithm 3. The step-by-step explanation of algorithm with sample dataset as follows.

**Algorithm 3**: Naïve Bayes Classifier

```
Step 1: Convert the given dataset into frequency tables.
Step 2: Generate Likelihood table by finding the probabilities of given features.
Step 3: Now, use Bayes theorem to calculate the posterior probability.
Step 4:
```

*Step 1. Convert the given dataset into frequency tables.*

Suppose we are taking a sample data from fetal health dataset.

| S. no. | Uterine contractions | Fetal health |
|---|---|---|
| 1 | 0.06 | Pathological |
| 2 | 0.09 | Pathological |
| 3 | 0.03 | Normal |
| 4 | 0.03 | Normal |

And the frequency table for the Uterine Contractions feature as.

| Frequency table | | Fetal health | |
|---|---|---|---|
| | | Pathological | Normal |
| Uterine contractions | 0.03 | 0 | 2 |
| | 0.06 | 1 | 0 |
| | 0.09 | 1 | 0 |

*Step 2. Generate likelihood table by finding the probabilities of given features..*

The likelihood table shows the probabilities as.

| Likelihood table | | Fetal health | | |
|---|---|---|---|---|
| | | Pathological | Normal | |
| | 0.03 | 0 | 2 | 2/4 = 0.5 |
| Uterine | 0.06 | 1 | 0 | 1/4 = 0.25 |
| Contractions | 0.09 | 1 | 0 | 1/4 = 0.25 |
| | | 2/4 = 0.5 | 2/4 = 0.5 | |

*Step 3. Use Bayes' theorem to calculate the posterior probability.*

Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

$$\text{Bayes' theorem}: \mathrm{P(A|B)} = \frac{\mathrm{P(B|A).P(A)}}{\mathrm{P(B)}}$$

where, P(A|B): Probability of hypothesis A on the observed event B. P(B|A): Probability of the evidence given that the probability of a hypothesis is true. P(A) is prior. P(B) is evidence. P(Pathological|Uterine Contraction) is Posterior probability. P(Uterine Contraction|Pathological) is Likelihood probability.

When Uterine Contractions (UC) = 0.03.

Pathological

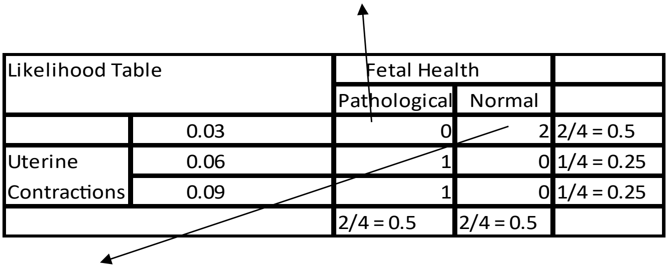
Posterior probability: P(Pathological|UC = 0.03) = $\frac{\mathrm{P(0.03|Pathological)*P(Pathological)}}{\mathrm{P(UC=0.03)}}$.

Likelihood probability: P (UC = 0.03 |Pathological) = 0/4.

Evidence: P(UC = 0.03) = 2/4 = 0.5

Prior: P(Pathological) = 0.5

Hence, Posterior probability: P(Pathological |UC = 0.03) = $\frac{0/4*0.5}{0.5} = 0$.



| Likelihood Table | | Fetal Health | | |
|---|---|---|---|---|
| | | Pathological | Normal | |
| | 0.03 | 0 | 2 | 2/4 = 0.5 |
| Uterine Contractions | 0.06 | 1 | 0 | 1/4 = 0.25 |
| | 0.09 | 1 | 0 | 1/4 = 0.25 |
| | | 2/4 = 0.5 | 2/4 = 0.5 | |

Normal

Posterior probability: P(Normal | UC = 0.03) = $\frac{\mathrm{P(0.03|Normal)*P(Normal)}}{\mathrm{P(UC=0.03)}}$.

Likelihood probability: P (UC = 0.03 |Normal) = 2/4 = 0.5

Evidence: P(UC = 0.03) = 2/4 = 0.5

Prior: P(Normal) = 0.5

Hence, Posterior probability: P(Normal |UC = 0.03) = $\frac{0.5*0.5}{0.5} = 0.5$

When uterine contraction is 0.03, the condition of the fetus is NORMAL because of higher posterior probability value.

Similarly, when Uterine Contractions (UC) = 0.06.

Pathological

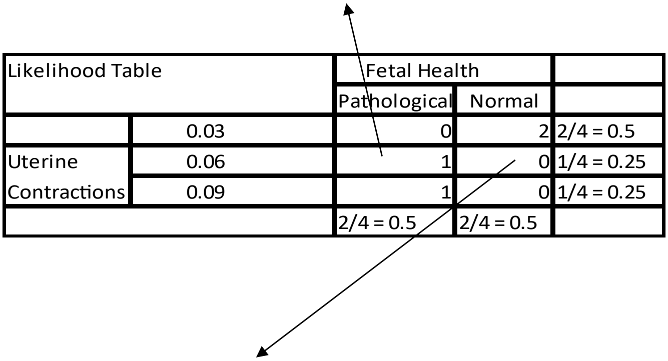
Posterior probability: P(Pathological | UC = 0.06) = $\frac{\text{P(0.06|Pathological)*P(Pathological)}}{\text{P(UC=0.06)}}$.

Likelihood probability: P (UC = 0.06 |Pathological) = 1/4 = 0.25.

Evidence: P(UC = 0.06) = 1/4 = 0.25.

Prior: P(Pathological) = 0.5

Hence, Posterior probability: P(Pathological | UC = 0.06) = $\frac{0.25*0.5}{0.25} = 0.5$



| Likelihood Table | | Fetal Health | | |
|---|---|---|---|---|
| | | Pathological | Normal | |
| | 0.03 | 0 | 2 | 2/4 = 0.5 |
| Uterine Contractions | 0.06 | 1 | 0 | 1/4 = 0.25 |
| | 0.09 | 1 | 0 | 1/4 = 0.25 |
| | | 2/4 = 0.5 | 2/4 = 0.5 | |

Normal

Posterior probability: P(Normal | UC = 0.06) = $\frac{\text{P(0.06|Normal)*P(Normal)}}{\text{P (UC=0.06)}}$.

Likelihood probability: P (UC = 0.06|Normal) = 0.

Evidence: P(UC = 0.06) = 0.25.

Prior: P(Normal) = 0.5

Hence, Posterior probability: P(Normal|UC = 0.06) = $\frac{0*0.5}{0.25} = 0$.

When uterine contraction is 0.06, the condition of the fetus is PATHOLOGICAL.

When Uterine Contractions (UC) = 0.09.

Pathological

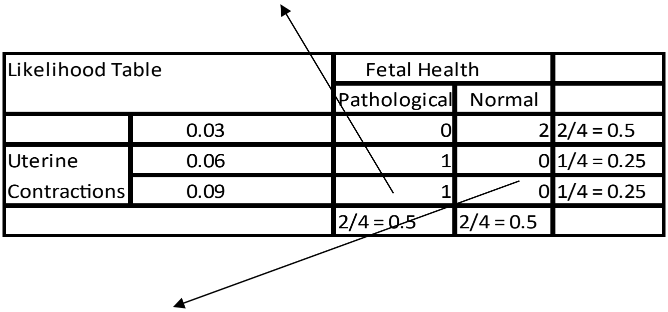
Posterior probability: P(Pathological|UC = 0.09) = $\frac{\text{P(0.09|Pathological)*P(Pathological)}}{\text{P(UC=0.09)}}$.

Likelihood probability: P (UC = 0.09 |Pathological) = 1/4 = 0.25.

Evidence: P(UC = 0.09) = 1/4 = 0.25.

Prior: P(Pathological) = 0.5

Hence, Posterior probability: P(Pathological|UC = 0.09) = $\frac{0.25*0.5}{0.25}$ = 0.5



| Likelihood Table | | Fetal Health | | |
|---|---|---|---|---|
| | | Pathological | Normal | |
| | 0.03 | 0 | 2 | 2/4 = 0.5 |
| Uterine<br>Contractions | 0.06 | 1 | 0 | 1/4 = 0.25 |
| | 0.09 | 1 | 0 | 1/4 = 0.25 |
| | | 2/4 = 0.5 | 2/4 = 0.5 | |

Normal

Posterior probability: P(Normal|UC = 0.09) = $\frac{\text{P(0.09|Normal)*P(Normal)}}{\text{P(UC=0.09)}}$.

Likelihood probability: P (UC = 0.09|Normal) = 0.

Evidence: P(UC = 0.09) = 0.25.

Prior: P(Normal) = 0.5

Hence, Posterior probability: P(Normal |UC = 0.09) = $\frac{0*0.5}{0.25}$ = 0.

When uterine contraction is 0.09, the condition of the fetus is PATHOLOGICAL.

## 8.3.2 *Deep Learning Models*

In this chapter, we have implemented two well-known deep learning approaches and tested three networks under each approach.

### 8.3.2.1 Multilayer Perceptron (MLP)

A supervised learning system called a multi-layer perceptron (MLP) trains on a dataset to learn a function. It can learn a non-linear function approximator for classification on given a set of features and a target. There may be one or more non-linear layers, known as hidden layers, between the input layer and the output. A single hidden layer MLP with scalar output is shown in Fig. 8.3. The input layer consists of a vector of predictor values as ($x_1 \ldots x_p$), and *L* neurons are used to represent a hidden layer. The values are distributed by the input layer to each of the neurons in the hidden layer. When the value from each input neuron reaches a neuron in the

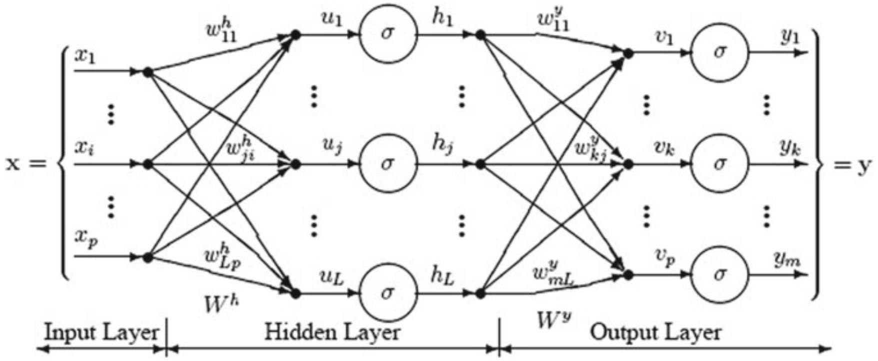


**Fig. 8.3** A simple network flow of MLP model

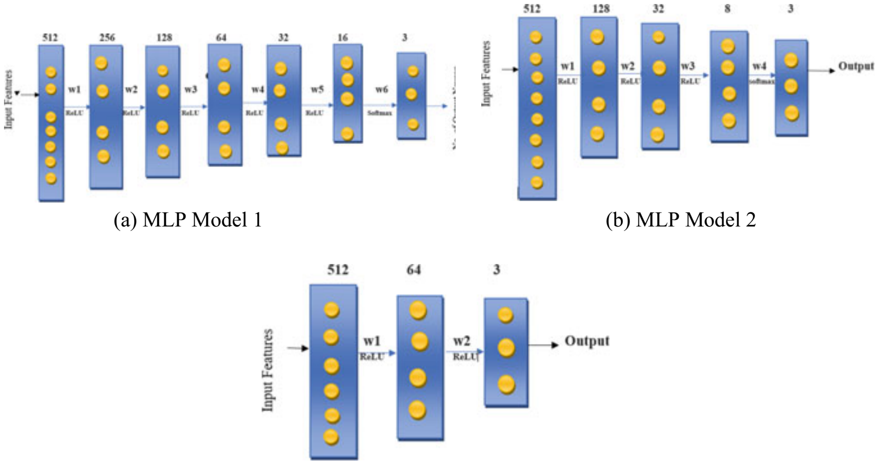
hidden layer, it is multiplied by a weight ($w_{ji}$), and the weighted values are added together to produce the combined weighted sum value $u_j$. The weighted sum ($u_j$) is passed through an activation (σ) function, which produces the hidden layer value $h_j$. These hidden layer's outputs are distributed to the output layer. When the value from each hidden layer neuron reaches a neuron in the output layer, it is multiplied by a weight ($w_{kj}$), and the weighted values are added together to produce a combined weighted sum value $v_j$. The weighted sum ($v_j$) is passed through again an activation ($\sigma$) function, which produces the value $y_k$. The y values are the network's outputs.

In this chapter, we have designed three MLP models as shown in Fig. 8.4. In model 1, there are about 7 hidden layers in which the neurons present in them are divided in the ratio of 2:1. Thus, the seven hidden layers having 512, 256, 128, 64, 32, 16, and 3 neurons, respectively. In model 2, there is a total of 6 hidden layers are presented and each one of them have their neurons divided in the ratio of 4:1. Thus, the six hidden layers having 512, 128, 32, 8, and 3 neurons, respectively. The third model hidden layers are ideally split into only 3 of them, each of having 512, 64, and 3 neurons, respectively.

#### 8.3.2.2 Recurrent Neural Network

A unique kind of artificial neural network called a recurrent neural network (RNN) which has designed to cope with data that contains sequences. Typically, feed forward neural networks are only suitable for independent data points. To include the dependencies between these data points, we must change the neural network if the data are organized in a sequence where each data point depends on the one before it. RNNs have the idea of 'memory', which enables them to store the states or details of earlier inputs to produce the subsequent output in the sequence.

In this chapter, we have designed three RNN models as shown in Fig. 8.5. Each input $x_i$ sequence is passed into its corresponding hidden RNN layer at time T. Each time-step, one input is presented to a recurrent neural network which then



(a) MLP Model 1

(b) MLP Model 2

(c) MLP Model 3

**Fig. 8.4** The proposed three MLP models

predicts one output and enters as input to next input sequence. An activation function ‘ReLU’/‘Softmax’ and regularization technique ‘Dropout’ applies and the predicted result is obtained as $y_j$. In model 1, five hidden RNN layers and 2 Dense or fully connected (FC) layers are present with 512, 256, 128, 64, 32, 16, and 3 neurons in the following order as mentioned in Fig. 5a. In model 2, there are 3 hidden RNN layers and 2 FC layers are presented where the number of neurons inside each layer are 512, 128, 32, 8, and 3, respectively. In model 3, there are 2 hidden RNN layers and one FC layer is presented with 512, 64, and 3 neurons.

## 8.4 Experimental Evaluation and Result Discussion

### *8.4.1 Brief Dataset Description*

The dataset is obtained from the Kaggle repository [20]. There are 21 features in the fetal health dataset. They are.

| | |
|---|---|
| **Baseline**—value FHR baseline (beats per minute) | **Accelerations**—Number of accelerations per second |
| **Fetal_movement**—Number of fetal movements per second | **Uterine_contractions**—Number of uterine contractions per second |
| **Light_decelerations**—Number of light decelerations per second | **Severe_decelerations**—Number of severe decelerations per second |

(continued)

(continued)

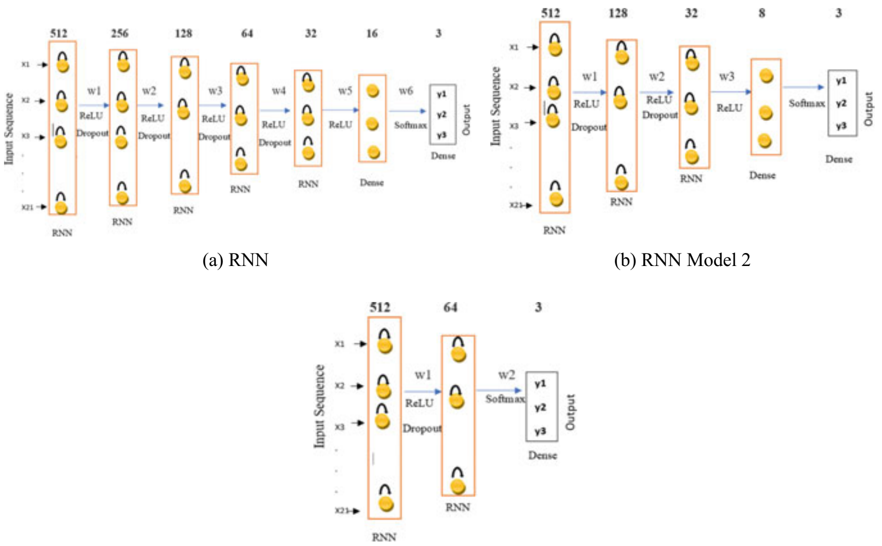
| **Baseline**—value FHR baseline (beats per minute) | **Accelerations**—Number of accelerations per second |
|---|---|
| **Prolonged_decelerations**—Number of prolonged decelerations per second | **ASTV**—Percentage of time with abnormal short-term variability |
| **Average ASTV**—Average value of short-term variability | **Percentage_of_time ALTV**—Percentage of time with abnormal long-term variability |
| **Mean ALTV**—Average value of long-term variability | **Histogram_width**—Width of the FHR histogram |
| **Histogram_min**—Minimum (low frequency) of the FHR histogram | **Histogram_max**—Maximum (high frequency) of the FHR histogram |
| **HNOP**—Number of histogram peaks | **HNOZ**—Number of zeros of the histogram |
| **Histogram_mode**—Fashion of the histogram | **Histogram_mean**—Mean of the histogram |
| **Histogram_median**—Median of the histogram | **Histogram_variance**—Variance of the histogram |
| **Histogram_tendency**—Tendency of the histogram | |

This dataset includes 2126 records of features from cardiotocogram tests that were divided into three classes by three expert obstetricians such as normal, suspect, and pathological. Normal class indicates that the fetus's health is in excellent condition and that no precautionary measures are required. Suspect class indicates that the fetus's health status may be unstable, and precautionary measures may be required. Pathological class indicates that the fetus's health is in mortal peril and that immediate precautions must be taken.

### *8.4.2 Performance Metrics*

Precision (Pre), Recall (Rec), $F_1$-score ($F_1$), Accuracy (Acc), and Support are employed as the assessment metric for result analysis. The calculation formulas are shown in Eqs. (8.3)–(8.7). Acc represents the ratio of the truly positive classified data. Among the total amount of classified data samples, Pre represents the ratio of truly positive predictive classified point in label data, and Rec represents the percentage of the total amount of data correctly classified in label data. $F_1$-score value is used to estimate the harmonic average of the Prec and Rec. Support is the number of actual occurrences of the class in the specified dataset. TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative respectively.

$$\text{Spec} = \frac{TN}{TN + FP} \tag{8.3}$$



(a) RNN

(b) RNN Model 2

(c) RNN Model 3

**Fig. 8.5** The proposed three RNN models

$$\text{Rec} = \frac{TP}{TP + FN} \tag{8.4}$$

$$\text{Pre} = \frac{TP}{TP + FP} \tag{8.5}$$

$$F_1 = \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{8.6}$$

$$\text{Support} = \sigma(\text{X} + \text{Y}) \div \text{Total} \tag{8.7}$$

### *8.4.3 Data Visualization and Data Pre-processing*

Any value derived from raw data cannot be used directly. Therefore, the data pre-processing is a process of taking raw data and converting it into usable information. Here, with the help of various plots, we have analyzed the given dataset. The process of manipulating, transforming, and visualizing data in order to derive meaningful insights from the results is known as data analysis. The following data visualization techniques have been used in our dataset. They are.

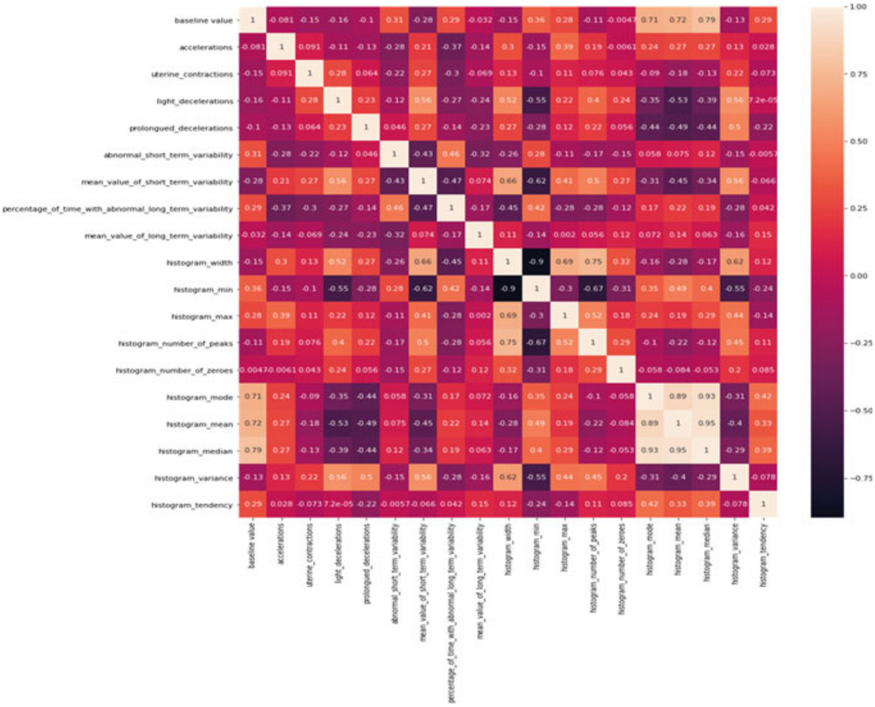


**Fig. 8.6** Correlation matrix among fetal data features

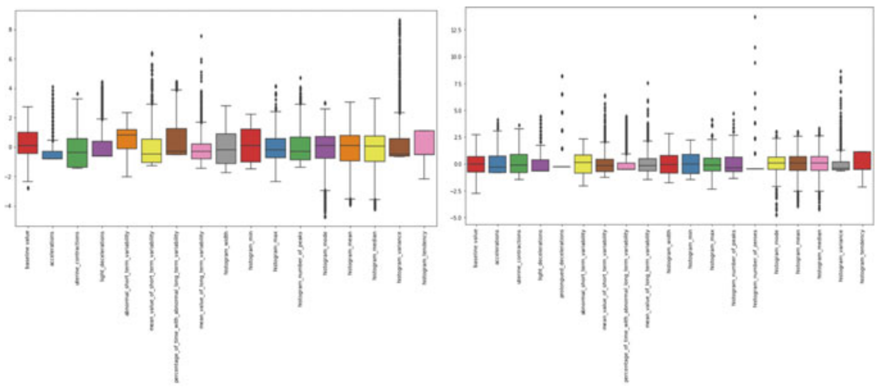
#### 8.4.3.1 Heat Map

A heat map represents these coefficients to visualize the strength of correlation among variables. It helps find features that are best for machine learning model building. The heat map transforms the correlation matrix into color coding as shown in Fig. 8.6.

We observe that the 'severe_deceleration' and 'fetal_movement' features have no impact on the dataset and their correlation is almost zero. Hence, these features are dropped for further experimental result evaluation.

#### 8.4.3.2 Box Plot

It is univariate in nature and summarizes the distribution of each attribute. It draws a line for the middle value i.e. median. Here, we observe that the features 'severe_deceleration', 'fetal_movement', 'light_decelerations', 'accelerations', and 'uterine contractions' show no distribution on the plot when compared to the other features. Hence, the dataset features are being normalization as shown in Fig. 8.7. This technique is also referred to as scaling. The Min–Max scaling method helps



(a)

(b)

**Fig. 8.7** Box plots of all the fetal data features. **a** Before Normalization **b** After Normalization

the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 and 1.

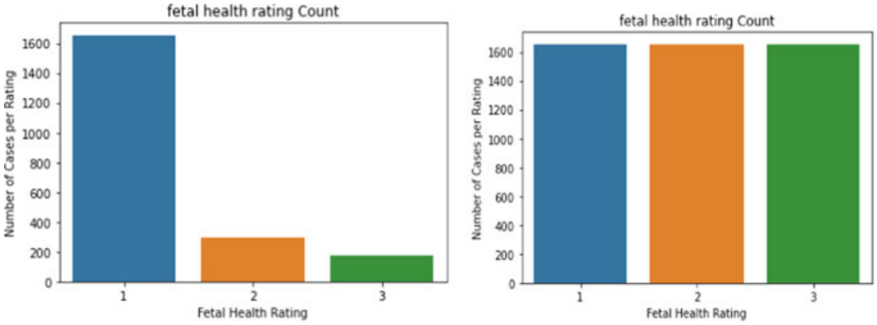
$$X_{\text{nomalized}} = (X - X_{\text{minimum}})/(X_{\text{maximum}} - X_{\text{minimum}}) \tag{8.8}$$

#### 8.4.3.3 Data Augmentation

Data augmentation is a process of artificially increasing the amount of data by generating new data points from existing data. This includes adding minor alterations to data to generate new data points in the latent space of original data to amplify the dataset. This is done by using the SMOTE (synthetic minority oversampling technique) approach. SMOTE is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. Fig. 8.8 shows the imbalance data among three fetal classes handles through data augmentation. As the count for 1: ‘Normal’ class is very high when compared to 2: ‘Suspect’ and 3: ‘Pathological’, we have used data augmentation technique to balance the data and keep it in a uniform manner.

### *8.4.4 Performance Evaluation—Machine Learning Models*

The confusion matrix is a matrix which has used to determine the performance of the classification models for our set of test data. It can only be determined if the true

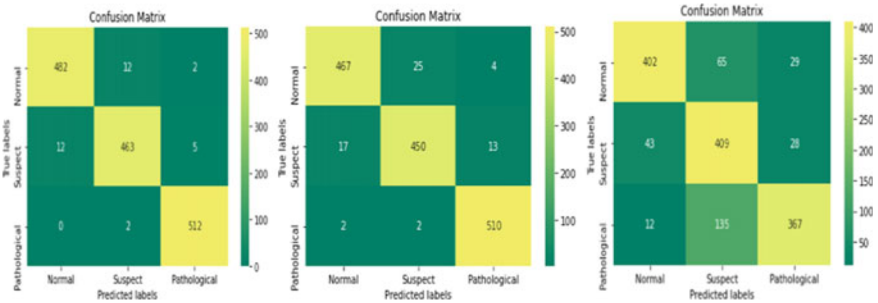


(a) (b)

**Fig. 8.8** Count plots of all the fetal Classes. **a** Before Data Augmentation **b** After Data Augmentation

values for test data are known. Figure 9a, b tell about random forest and decision tree respectively that the actual and predicted values for pathological is highest as compared to normal and suspect classes. Figure 9c tells about Naïve Bayes that the actual and predicted values for suspect class is highest as compared to normal and pathological.

Table 8.1 shows the results of classification where accuracy displays as average over all three classes. The recall, precision, and $F_1$-score display separately for the classes Normal (row 1), Suspect (row 2), and Pathological (row 3), respectively. The Support metrics show the number samples data uses for experiment result evaluation which can be noticeable through before and after augmentation number of samples. From the comparative point of view, the difference between the metrics values is clearly visible with data augmentation i.e. better performance. Overall the accuracies, random forest has shown the highest accuracy of 96.8 followed by decision tree with 93.98, and Naïve Bayes shows poorest performance of accuracy 77.77. As final



(a) Random Forest (b) Decision Tree (c) Naïve Bayes

**Fig. 8.9** Prediction for numbers of true and false classification of all the fetal classes

verdict, among the three machine learning models, the random forest is the best trained and tested model with and without data augmentation.

### *8.4.5 Performance Evaluation—Deep Learning Models*

We trained the proposed models for several epochs (applied early stop) on the training dataset from fetal_health.csv. From these training samples, the splits for training-validation has done at 0.8:0.2 ratio. The overfitting avoiding strategy has done through data augmentation. Throughout training, a batch size of 4 used, and the loss estimating is 'sparse_categorical_crossentrop'. It can be seen from Figs. 8.10 and 8.11 that when the iteration period is about 55 (maximum), the validation loss/accuracy stops decreasing/increasing which means of leading to the training termination. In fact, the models start to get relatively decent performance when training after the few epochs.

Figure 8.10a–c represents the training and validation loss of MLP models i.e. 1, 2, 3. The inference from the plot is that the model 1 is slightly overfit which shows testing loss about 0.54. The model 2 training and validation loss is inferring that the model is underfit which has very low performance with testing loss about 1.08. The model 3 seems to fit almost perfectly and it is considered as ideal fitting. However, this model seems to perform slightly higher loss than model 1. Figure 8.10d–f represents the training and validation loss of the RNN models. By examining the graph it is visible that the RNN models are also performing similar to MLP models but with lesser loss rate. RNN model 1 having test loss 0.34, RNN model 2 having test loss 10.75 (worst), and RNN model 3 having test loss 0f 0.41.

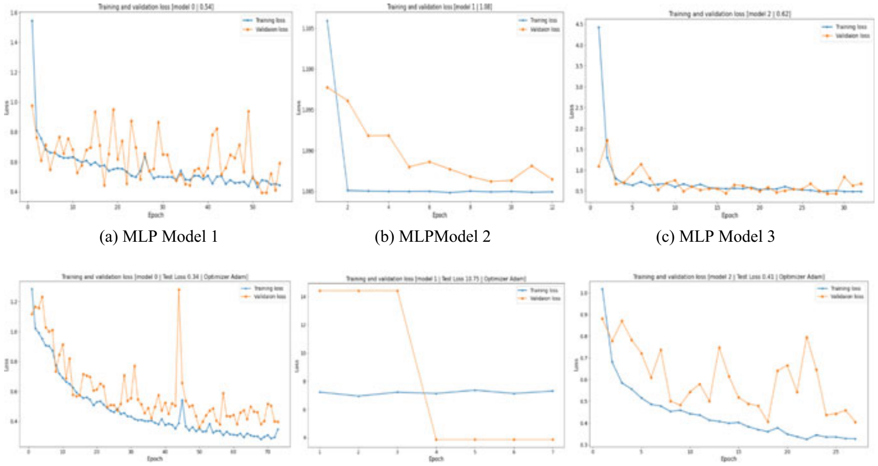
Table 8.2 shows the results of classification of MLP and RNN deep learning models over all three classes, 0: Normal, 1: Suspect, 2: Pathological, respectively. In case of MLP models, from the comparative point of view, the difference between the metrics values is not much visible even after data augmentation. Overall the performance metrics, MLP model 1 and MLP model 3 has shown the promising results. In case of RNN models, the difference between the metrics values increases noticeably after data augmentation. Overall the performance metrics, RNN model 1 and model 3 has shown the promising results but RNN Model 2 shown abnormal outcomes.

## 8.5 Conclusion

In this chapter, we have analyzed deeply the machine and deep learning models which classify the fetal health status based on the uterine contractions and fetal heart rate. These fetal heart rate and uterine contractions are collected data as a part of cardiotocogram (CTG). The obstetrician's visual assessment of the CTG data, however, could not be objective or correct. These learning models makes the job

**Table 8.1** Fetal classification results using machine learning. **a** Before Data Augmentation **b** After Data Augmentation

| (a) | | | | | | (b) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Without data augmentation | | | | | | With data augmentation | | | |
| Method | Accuracy | Precision | F1-Score | Recall | Support | Method | Accuracy | Precision | F1-Score | Recall | Support |
| | | 94 | 95 | 96 | 500 | | | 93 | 94 | 95 | 496 |
| DT | 91.8 | 76 | 73 | 71 | 85 | DT | 93.98 | 95 | 93 | 92 | 480 |
| | | 98 | 90 | 83 | 53 | | | 98 | 99 | 99 | 514 |
| | | 95 | 97 | 99 | 500 | | | 97 | 97 | 97 | 496 |
| RF | 93.88 | 89 | 81 | 74 | 85 | RF | 96.8 | 97 | 97 | 97 | 480 |
| | | 98 | 92 | 87 | 53 | | | 100 | 99 | 99 | 514 |
| | | 99 | 82 | 70 | 500 | | | 88 | 85 | 82 | 496 |
| NB | 72.17 | 34 | 50 | 94 | 85 | NB | 77.77 | 65 | 75 | 88 | 480 |
| | | 59 | 55 | 51 | 53 | | | 87 | 75 | 65 | 514 |



(a) MLP Model 1 (b) MLPModel 2 (c) MLP Model 3

(d) RNN Model 1 (e) RNN Model 2 (f) RNN Model 3

**Fig. 8.10** Training and validation losses of the generated three MLP and RNN models

**Table 8.2** Fetal classification results using deep learning

| (a) Before data augmentation | | | | | | (b) After data augmentation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MLP | Class | Pre | Rec | F1 | Support | MLP | Class | Pre | Rec | F1 | Support |
| Model 1 | 0 | 0.95 | 0.85 | 0.90 | 332 | Model 1 | 0 | 0.98 | 0.81 | 0.88 | 332 |
| | 1 | 0.49 | 0.73 | 0.59 | 59 | | 1 | 0.44 | 0.80 | 0.58 | 332 |
| | 2 | 0.68 | 0.86 | 0.76 | 35 | | 2 | 0.67 | 0.94 | 0.77 | 332 |
| Model 2 | 0 | 0.73 | 0.73 | 0.84 | 332 | Model 2 | 0 | 0.72 | 0.69 | 0.65 | 332 |
| | 1 | 0.33 | 0.92 | 0.53 | 59 | | 1 | 0.34 | 0.63 | 0.53 | 332 |
| | 2 | 0.66 | 0.74 | 0.78 | 35 | | 2 | 0.56 | 0.59 | 0.51 | 332 |
| Model 3 | 0 | 0.99 | 0.66 | 0.79 | 332 | Model 3 | 0 | 0.96 | 0.81 | 0.88 | 332 |
| | 1 | 0.35 | 0.95 | 0.51 | 59 | | 1 | 0.41 | 0.69 | 0.52 | 332 |
| | 2 | 0.67 | 0.89 | 0.77 | 35 | | 2 | 0.67 | 0.91 | 0.77 | 332 |
| *RNN* | *Class* | *Pre* | *Rec* | *F1* | *Support* | *RNN* | *Class* | *Pre* | *Rec* | *F1* | *Support* |
| Model 1 | 0 | 0.97 | 0.77 | 0.86 | 332 | Model 1 | 0 | 0.86 | 0.84 | 0.86 | 332 |
| | 1 | 0.43 | 0.92 | 0.58 | 59 | | 1 | 0.85 | 0.87 | 0.86 | 332 |
| | 2 | 0.55 | 0.60 | 0.58 | 35 | | 2 | 0.88 | 0.89 | 0.89 | 332 |
| Model 2 | 0 | 0.00 | 0.00 | 0.00 | 332 | Model 2 | 0 | 0.55 | 0.71 | 0.62 | 332 |
| | 1 | 0.14 | 0.97 | 0.25 | 59 | | 1 | 0.55 | 0.73 | 0.62 | 332 |
| | 2 | 0.45 | 0.37 | 0.41 | 35 | | 2 | 0.97 | 0.35 | 0.51 | 332 |
| Model 3 | 0 | 0.99 | 0.79 | 0.88 | 332 | Model 3 | 0 | 0.85 | 0.84 | 0.85 | 332 |
| | 1 | 0.51 | 0.83 | 0.63 | 59 | | 1 | 0.84 | 0.89 | 0.87 | 332 |
| | 2 | 0.53 | 0.97 | 0.69 | 35 | | 2 | 0.93 | 0.89 | 0.91 | 332 |

easier for the obstetricians to make decision and thus have a significant role in fetal health monitoring.

Data visualization is extremely important for this study of CTG data. Through the use of standard graphical representations such as heatmap, boxplot, and many more, it helped us analyze the data. This study also improved our understanding of unbalance data through the use of data augmentation. This enables us to artificially enlarge the dataset by adding new data points to the same dataset.

We have examined the effectiveness of deep learning and conventional machine learning models with varied dataset sizes and the number of target classes. If the dataset is tiny, we have discovered that conventional classifiers can train more effectively than deep learning classifiers. The performance of deep learning models improves as the dataset size grows.

## References

1. Petrozziello, A., Jordanov, I., Papageorghiou, A.T., Redman, C.W.G., Georgieva, A.: Deep learning for continuous electronic fetal monitoring in labor. In: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2018)
2. Ayres-de-Campos, D., Spong, C.Y., Chandraharan, E.: FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. Int. J. Gynecol. Obstet. **131**, 13–24 (2015)
3. World Health Organisation Official Website, Maternal Health page https://www.who.int/health-topics/maternal-health#tab=tab_1
4. Yilmaz, E.: Fetal state assessment from cardiotocogram data using artificial neural network. J. Med. Biol. Eng. **36**, 820–832 (2016)
5. Czabanski, R., Jezewski, J., Matonia, A., Jezewski, M.: Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. Exp. Syst. Appl. **39**, 11846–11860 (2012)
6. Krupa, N., Ali, M., Zahedi, E., Ahmed, S., Hassan, F.M.: Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine. Biomed. Eng. Online **10** (2011)
7. Ocak, H.: A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being. J. Med. Syst. **37** (2013)
8. Batra, A., Chandra, A., Matoria, V.: Cardiotocography analysis using conjunction of machine learning algorithms. In: Proceedings of International Conference on Machine Vision and Information Technology (CMVIT), pp. 1–6 (2017)
9. Kuhle, S., Maguire, B., Zhang, H., et al.: Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy Childbirth **18** (2018)
10. Mehbodniya, A., Jesu Prabhu, L.A., Webber J.L., et al.: Fetal health classification from cardiotocographic data using machinelearning. Exp. Syst. (2021)
11. Rahmayanti, N., Pradani, H., Pahlawan, M., Vinarti, R.: Comparison of machine learning algorithms to classify fetal health using cardiotocogram data. Procedia Comput Sci **197**, 162–171 (2022)
12. Kannan, E., Ravikumar, S., Anitha, A., Kumar, S.A.P., Vijayasarathy, M.: Analyzing uncertainty in cardiotocogram data for the prediction of fetal risks based on machine learning techniques using rough set. J. Ambient Intell. Hum. Comput. (2021)
13. Akbulut, A., Ertugrul, E., Topcu, V.: Fetal health status prediction based on maternal clinical history using machine learning techniques. Comput. Methods Programs Biomed. **163**, 87–100 (2018)

14. Jagannatha, D.J., Raveena Judie Dollya, D., Dinesh Peter, J.: Deep learning strategies for foetal electrocardiogram signal synthesis. Pattern Recognit. Lett. 136, 286–292 (2020)
15. Zhao, Z., Deng, Y., Zhang, Y., Zhang, Y., Zhang, X., Shao, L.: DeepFHR: intelligent prediction of fetal Acidemia using fetal heart rate signals based on convolutional neural network. BMC Med. Inf. Decis. Mak. **19** (2019)
16. Petrozziello, A., Redman, C.W.G., Papageorghiou, A.T., Jordanov, I., Geogieva, A.: Multimodal convolutional neural networks to detect fetel compromise during labor and delivery. IEEE Access 112026–112036 (2019)
17. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**, 1–37 (2008)
18. Breiman, L.: Random forests. J. Mach. Learn. **45**, 5–32 (2001)
19. Vikramkumar, Vijaykumar, B., Trilochan.: Bayes and Naive Bayes Classifier. arXiv:1404.0933
20. https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

# Chapter 9
# Blockchain and AI: Disruptive Digital Technologies in Designing the Potential Growth of Healthcare Industries

S. Uma

## 9.1 Introduction

Health care falls into the category of a large percentage of GDP in most developed countries. Yet, hospital costs are continuing to rise, along with inefficient practices and data breaches. Though disrupting technologies like artificial intelligence (AI) and Internet of Things (IoT) are used, existing solutions have several challenges in sharing lengthy electronic medical records, managing secure and transparent pharmaceutical product supply chains, and protecting and accessing genomic data. Significant issues in terms of interoperability, privacy, and supply chain traceability existed even before the outbreak of the COVID-19 pandemic [1]. AI-Powered Blockchain is the catch-phrase of the year and as this new technology slowly matures, it seems clear that from banking to supply chain logistics, it is ready for disruption. In the healthcare sector, there's a massive opportunity for the AI-integrated blockchain revolution to disrupt and lead a digital transformation. It is possible to leverage blockchain technology for a wide variety of things, including medical records, pharmaceutical supply chains, and payment distribution. With blockchain technology, patients' private keys are used to create database applications that are decentralized, robust, traceable, undeletable, and unchangeable [2].

S. Uma (✉)
Department of Computer Science and Engineering, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
e-mail: umakaruna19@yahoo.com

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_9

## 9.2 Review of AI in Health Care

AI refers to computer systems that simulate specific aspects of human intelligence such as learning, reasoning, and problem-solving. AI is a combination of several intelligent processes and behaviors generated by computational algorithms and models. With the drastic reduction in the price of storage devices, large volumes of data are stored in all kinds of transactions in all industrial sectors. The availability of this massive data has accelerated the progress in AI and the evolution of different types of machine learning algorithms fuelling research activities in all sectors. The developments in the fields of AI voice technology, Natural Language Processing (NLP), AI-based virtual assistants, and robotics have provided new and improved, powerful solutions in health care [3]. The learning abilities and the self-correcting abilities based on the feedback of AI can be used to improve its accuracy. AI is never a replacement for doctors. But AI can aid physicians in several ways in improving clinical decision-making or even replace certain functional areas where human judgment will not be accurate. Healthcare data has become increasingly available recently as big data analytic methods have developed. This has enabled the successful application of AI in health care. Using powerful AI techniques, relevant clinical questions can enable powerful AI techniques to extract relevant clinical information hidden in the massive amount of data, which in turn can support clinical decision-making [4]. AI in health care provides the following solutions for improving patient care [4].

- Provides updated information from various resources like journals, textbooks, and clinical practices.
- Reduces the diagnostic and therapeutic errors which cannot be avoided in human clinical practice.
- Extracts useful information from large patient populations to get insights to provide timely health risk alerts and also predict health outcomes.
- Early detection and diagnosis.
- Outcome prediction and prognosis evaluation.

## 9.3 Applications of AI in Health Care

The potential of AI is tremendous in revolutionizing the healthcare sector. Though technology has defined our lives, the impact of machine language and technology is dramatically transforming our lives across many spheres, but importantly never more than in the practice of medicine. Several questions such as those listed below flash into one's mind in applying AI to health care:

- How reliable are computers in making decisions about our health?
- What are the various possibilities by which AI can be used in health care?
- How can we rapidly analyze vast amounts of clinical data to diagnose disease?
- How do we identify the best treatment options and what will be the patient outcomes?

The top AI applications in health care will provide the answers to the above questions which are given below.

- Robot-assisted surgery
- Virtual nursing assistants
- Administrative workflow assistants
- Fraud detection
- Prescription error recognition
- Automated image diagnosis
- Cybersecurity
- Connected medical devices
- Identification of clinical trial participants
- Preliminary diagnosis and selection of optimal treatment strategies.

The value each of these AI applications will produce by 2026 is given in Fig. 9.1. In total, $150 billion is expected to be produced by applying AI in health care in these applications. The use of AI in health care is not a replacement for doctors, nurses, or clinical assistants. Rather, AI assists surgeons to make accurate and delicate motions.

Basically, our intelligence is what makes us human and AI is an extension of human intelligence to extract more powerful outcomes beyond the ability of human beings. The key benefits of using robotic assistants in operation theaters are reduced risk of infections, reduced pain, scarring, and blood loss, spending less time in the hospital, and recovering faster and getting back to daily routines faster. Most people google their symptoms, self-diagnose, and get scared. Most people take self-medication also. Using AI healthcare analytics, virtual nursing assistants can detect the worsening of chronic conditions, monitor medication intake, and schedule appointments to prevent problems. A multifunctional virtual nurse "Molly", an avatar, provides remote support for common medical conditions using the patient's weight, blood pressure, and other metrics from monitoring devices. Chatbots are also

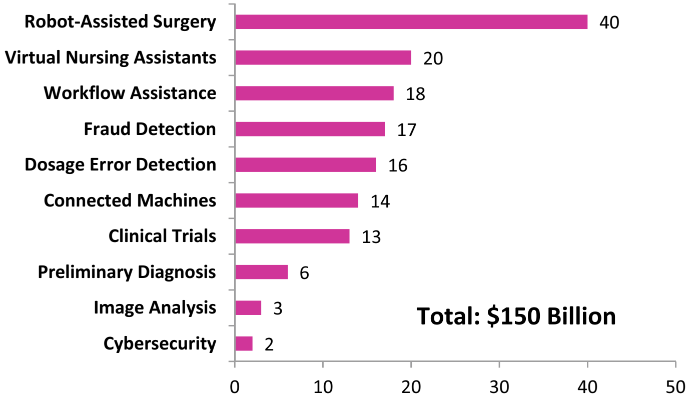


**Fig. 9.1** Top 10 Healthcare AI applications and the value they will produce by 2026 (in billion)

featured to discuss health issues privately and book an appointment with the consultants. Virtual nursing assistants provide 24 × 7 medical support, and monitoring and also provide quick medical solutions for common symptoms. Earlier, doctors had to spend minutes together to get the patient's complete medical history.

With AI and natural language processing and AI-powered workflow assistants, doctors are enabled to navigate, transcribe, and make more informed decisions using the medical records with voice commands more easily and speed up the consultancy time. Hence, a clinical specialist in a specific domain can address several patients more easily and consultation cost also gets reduced. AI-powered fraud detection systems are helpful in automating healthcare-related insurance services. Security of healthcare data and higher patient satisfaction are the outcomes of using AI. Artificial intelligence plays a vital role in removing the errors in prescriptions that occur due to the selection of wrong options from a drop-down list box. Based on the patient's history of health records, AI can identify such errors and increase the quality of care by preventing drug overdosing and health risks. AI reduces human error in diagnosing CT and MRI scans to improve the accuracy of treatment. It eliminates the threat to patient safety, protects the hospital's reputation, and cuts down the cost involved in patient data breaches. Connected medical devices are used to monitor the patient's health care periodically to provide informative insights from diverse data sources, and to prevent life-threatening risks which help to optimize and prioritize hospital workflows. Artificial intelligence in health care reduces data errors by 20%, costs by 30%, and 30% efficiency on time.

## 9.4 Review of Blockchain and AI in Health Care

Integrating with AI and other disruptive technologies, blockchain is a foundational digital technology used in health care as a replacement for traditional distributed database management systems [5].

When considering an opportunity to set up a robotic surgical program, hospitals should be ready to face organizational and technological challenges. For example, healthcare providers must ensure the precision of AI algorithms by training them on large amounts of reliable data and planning training programs for the hospital personnel to confidently work alongside intelligent machines [6]. Blockchain helps in overcoming a number of challenges met with the traditional practices followed in all sectors. The advantages of using blockchain are categorized as follows:

- Trust
- Decentralized structure
- Enhanced security and privacy
- Reduced costs
- Speed
- Data integrity assurance across multiple parties
- Visibility and traceability

- Immutability
- Individual control of data
- Tokenization
- Innovation
- Provide new operational efficiencies.

In the healthcare sector, blockchain technology is used for a variety of functionalities listed below [7].

- Acquiring, preserving, and maintenance of electronic health records
- Identifying the best diagnostic method
- liaising with healthcare professionals
- pharmaceutical supply chains
- health data analytics
- Patient consent management
- Drug traceability
- Secure electronic health records (ETRs)
- Micropayments incentivization
- Clinical trials data security.

## 9.5 Blockchain Framework

A blockchain is a chain of linked blocks that contain information. A group of researchers first described this technique in 1991, and it was originally intended to timestamp digital documents to make it impossible to backdate or corrupt them [8], similar to a notary public. It was mostly unused until Satoshi Nakamoto adapted it for use in Bitcoin until it became the digital currency Bitcoin in 2009.

Blockchains are distributed ledgers that are open for anyone to access. It becomes very difficult to change some data once it has been recorded in a blockchain. Each block contains some data, the hash of that block, and the hash of the previous block. In a blockchain, the type of data that is stored inside each block is different [9].

Say, for example, a blockchain has three blocks, and if someone tampers with the second block, the hash value of that block also changes. Due to this, the third block and all other following blocks will become invalid, since they can no longer store a valid hash of the previous block.

Though the advanced computation facilities permit very fast and calculate hundreds of thousands of hashes per second, effective tampering of a block could happen only if it is possible to recalculate the hashes of all other blocks to make the blockchain valid again [10]. Proof-of-work is a mechanism introduced by blockchain to mitigate this possibility. It slows down the creation of new blocks. For example, in bitcoins, it takes roughly 10 min to add a new block to the chain. To ensure the security of the system, the transaction details are stored in a distributed network rather than a centralized entity to manage the chain. Blockchain uses a peer-to-peer network and anyone can join it. Anyone who joins the network newly will get a full copy of the blockchain. The network node could verify that the blockchain is still in order.

Blockchain frameworks simplify the development, deployment, and integration of technically complex products [11]. Frameworks typically contain only the blockchain framework and its basic modules, and developers implement specific components based on them. This leads to a high development rate without compromising the final product's stability and performance. Some of the most widely used blockchain frameworks are Ethereum, Bitcoin, Hyperledger, Corda, EOS, IOTA, Ripple (XRP), Waves, NEM (XEM), and Quorum. While choosing a framework for an application, the following characteristics are taken into consideration [12].

- Completeness
- Performance
- Extendability
- Maturity
- Community.

## 9.6 Applications of Blockchain in Health Care

Blockchain is going to sweep away almost every domain and every industry and the pace will increase much more, and we need to equip ourselves with blockchain technology sooner or later. Combining artificial intelligence, machine learning, and blockchain is not a hype but really converging to bring in new paradigms in the way you are making solutions. There's a massive opportunity for the blockchain revolution to disrupt and lead a digital transformation in health care. From medical records to pharmaceutical supply chains, to smart contracts for payment distribution, there are plenty of opportunities to leverage this technology. Here are the three ways, how blockchain will change health care as we know it.

### *9.6.1 Health Records*

Electronic medical records are the backbone of every modern healthcare system. But the medical records grow longer and become more complex with each visit to the doctor. And since every hospital and every doctor's office has a different way of storing them, it's not always easy for healthcare providers to obtain them. There are already some companies out there like Patientory, Medibloc, and Medicalchain that aim to solve this problem [13]. The goal is to give patients authority over their entire medical history and to provide one-stop access to it for patients and physicians as well. Blockchain would not only simplify and make access more efficient, but inherently bring data security to the field as well.

### *9.6.2 Supply Chains*

The Pharmaceutical industry has one of the highest standards for product safety, security, and stability; it's ripe for disruption. For example, supply chain management with blockchain can be monitored securely and transparently. This greatly reduces time delays and human mistakes. It can also be used to monitor costs, labor, and even waste in emissions at every point in the supply chains. It can also be used to verify the authenticity of products by tracking them from their origin, combating the counterfeit drug market that costs 200 billion dollars in losses to the market annually [14]. Companies like Chronicled, Blockpharma, and Modum are already working towards more efficient blockchain logistic solutions [14]. Modum in particular works in compliance with EU laws that require proof that medicinal products have not been exposed to particular conditions, especially certain temperatures which may comprise their quality. Modum's solution was to develop a sensor that records environmental conditions while physical products are in transit and permanently record it on the blockchain.

### *9.6.3 Genomic Market*

Companies like EncrypGen and Nebula Genomics are building blockchain platforms to enable people to share genomic data safely and securely in a new emerging market. They bet that in the future, opportunities around personal genome sequencing will create a data market worth billions of dollars [14].

And what's the best technology to solve data security issues and to ensure that data gets from the source to its end-user without any middlemen? It is blockchain. These companies aim to use blockchain technology to enhance genomic data protection, enable buyers to efficiently acquire genomic data, and address the challenges of genomic big data. These companies are just a few of the dozens of startups that aim to use blockchain to disrupt health care. As it usually is, their marketing is great, they promise big, but the jury is still out on which ones will remain only promises, and which are the few that become the next big thing.

## 9.7 Metaverse

Metaverse is an integrated network of digital, 3D virtual worlds [15]. It is the future iteration of the Internet and a persistent, online 3D universe [16]. It is accessed with a virtual reality headset and the user's eye movements, voice commands, and feedback controllers are used to navigate the virtual world [17]. Metaverse users can socialize and engage in an unlimited variety of virtual experiences using digital avatars. It is a virtual world where we can live, work, travel, and play. Spending time online in Metaverse, a social utopia would be more fun and highly interactive. Meta, HSBC, Gucci, Coca-Cola, and J. P. Morgan are among the few companies involved in researching this fictional reality. By 2030, the metaverse economy is expected to be worth around $8-13trillion [18]. Though Internet is a distinct universe, it is 2-dimensional. But Metaverse, an artificially created universe, is 3-dimensional. It is infinite, persistent, self-sustaining, interoperable, and in real time. Metaverse gives the experience of being part of this 3D universe. Artificial intelligence, 3D reconstruction, virtual reality, augmented reality, 5G, Internet of Things, blockchain, and cryptocurrencies are the technologies used in Metaverse. Non-Fungible Tokens are tokens that exist in blockchain to acquire ownership of any digital asset.

Metaverse has created significant changes in the existing business processes in various dimensions of the working environment such as working in a virtual space remotely, organizing remote meetings, and interacting with colleagues in workrooms [19]. Metaverse is considered a game changer which will have a great impact on the global economy. It provides a new dimension for the job market, giving people access to jobs in more places no matter where they live. It has a positive impact on the environment, since mobility will be reduced and people will be less time stuck in traffic. Many business sectors are more enthusiastic about the metaverse ecosystem to build sector-specific projects in the metaverse which will provide new opportunities for generating income.

Healthcare, Virtual games, business, education, military, real estate, manufacturing, fashion, and social projects are some of the domains in which metaverse is used. Metaverse is helpful in addressing the challenges of working from remote places. It is an excellent tool for healthcare professionals, which permits patients who could not regularly visit doctors due to geographical limitations. Doctors could get greater insights into the health condition of their patients. Metaverse makes online games decentralized, economic, and more exciting. In the tourism industry, metaverse provides the experience of a virtual tour to any place and is more economical.

## 9.8 Metaverse for Health and Wellbeing

The use of augmented reality has emerged as a valuable tool for developing students' skills and knowledge [20]. A surgical assistive tool such as Microsoft Hololens assists surgeons in various surgical procedures. Such technologies are used to improve surgical precision and speed with the most common metaverse applications. Among the other benefits of AR headsets is the ability to view real-time patient data in addition to pre-operative pictures from MRI, CT, and 3D scans. As a result, patient's vital signs such as body temperature, respiration rate, heart rate, and blood pressure can be monitored through the metaverse.

Augmented reality can also be used to enhance vein identification in the metaverse. Thus, metaverse technologies are capable of solving the problem of finding a vein, especially in highly pigmented skin or blood vessels with tiny diameters. X-rays and CT scans, which make use of visual-based technology, are also appropriate candidates for transition to the metaverse. Medical practitioners and healthcare professionals can examine the inside of the body in a virtual world and figure out what the problem is [21].

Healthcare delivery entailed the physical interaction between the patient and physician as a means to diagnose the medical treatment. With the advent of telehealth, the patient-physician relationship slightly changed to digital means. As augmented and virtual reality got introduced, these paradigms are shifting to possibilities beyond imagination. In future, the metaverse might be the best way to get medical advice or go through a surgical procedure. It will be the best technology that will transform medical training and surgery to train medical staff. Trainees will get a close view of the surgeon's procedures enhanced with tactile haptic controls like intuitive surgical which is famous for its robot-assisted surgery systems. Surgeons from different geographical places can collaborate together in surgery with assistance from other consultants and experts during the surgery. The surgeon can see the vital images of the patient and navigate easily through multiple patient data. Surgeons will be empowered with virtual tools.

Yet there are major challenges in linking the real world and the virtual world. Since a physical examination of the patient is not needed always, for oral consultation telemedicine and consultation services are sufficient. In the metaverse, the doctor, patient visits, or consultations can be done virtually in a 3D clinic or virtual office. Similarly, digital therapeutic applications, cognitive therapy, psychiatric evaluations, and rehabilitations are the other medical applications of the metaverse.

## 9.9 Limoverse, The Blockchain and AI Revolution in Health Care

Limoverse is a global health and wellness ecosystem built on blockchain technology awarded as the best blockchain project in Crypto Expo Dubai 2022 [22]. It is the metaverse which provides a huge platform for health and wellness seekers and healthcare providers like doctors, nutrition experts, fitness and yoga therapists, and healthcare institutions. Limoverse is built on a number of disruptive technologies such as blockchain, smart contracts, artificial intelligence, machine learning, virtual reality, augmented reality, extended reality, and big data and distributed computing. In limoverse, people contribute to a world outside the metaverse where they connect, live, and get well [23]. It is a powerful real use case designed to attract millions of people with the blockchain-connected metaverse as its core. For example, consider a person carrying out research in a specific domain. Instead of listening to podcasts and non-specified content on the Internet which may or may not be useful for his research, he/she can personally connect with several experts and masters in the domain and bring new dimensions to the research. Thus, the time taken to search or read general information/irrelevant content will be avoided. This feature is very much useful in the medical field to carry out complicated surgeries.

Additionally, Limoverse features live sessions, augmented reality, and cutting-edge technology provide illumination of the mind. It offers its users the ability to monetize, decentralize, and gamify their personal wellness records through this network to health and wellness enthusiasts and practitioners [22]. Limo tokens will be awarded to the Masters, the native token of this blockchain. These are assets that can be earned by engaging meaningfully in Limoverse's ecosystem and by participating in it. Limo tokens are utility tokens with value within the ecosystem. Interaction and exchange within the Limoverse create, enhance, and multiply its value. Limoverse users can play fitness challenges based on wearable fitness trackers like Apple Watch, Fitbit, etc., and win Limo tokens in exchange. Limo tokens are powerful utility tokens powering this ecosystem and are tradable after listing in exchanges, an opportunity to create a huge crypto asset for entrepreneurs.

Limos can be earned by spending time, sharing expertise, and supporting the ecosystem. EPLIMO is at the core of this ecosystem, a highly comprehensive health data repository that is made up of genetic and metabolic analyses, stored in highly secure private blocks. Limo tokens can be earned even by sharing health data with credible universities and researchers. Users have absolute and final authority over how their genetic data is used in Limoverse, since it is built on a decentralized blockchain.

Limoverse, the globally acclaimed blockchain healthcare revolution, enables superhuman performance for its users by safely hacking their physiology and psychology with proven lifestyle variations in exercise, yoga, diet, meditation, supplements, etc. [24].

Epigenetic Personalized Lifestyle Modification (EPLIMO) is a Program which empowers people with personalized recommendations on nutrition, yoga, exercise, meditation, diets, sleep patterns, drug responses, and lifestyle using their Geno-Metabolic Analysis to take charge of their own health and wellbeing to live long, staying young and productive.

## 9.10 Impact of Blockchain and AI in Health Care

The healthcare industry is multifaceted and the integration of AI can be extended into hospital administration, treatment, diagnosis, medical imaging, personalized treatment, and remote care. The disruption of blockchain in health care is realized in terms of security, privacy, and interoperability of healthcare data. Supply chain maintenance of pharmaceutical products, fraud prevention, identifying, tracking drugs, and auditing the reports on licenses are some of the functionalities where blockchain is used in healthcare applications. Artificial intelligence in health care is used to increase the precision and accuracy of diagnosis, improves decision-making ability [25], introduces new drugs, and reduces the time for marketing the drugs. Improved data analysis capabilities, innovative treatment practices, higher success rates, and highly affordable drugs are the advantages of AI in health care. 55% of the applications in health care are expected to use blockchain technology by 2025. Also, global spending on health care is aimed to reach USD 18 trillion [26].

Similarly, AI helps doctors, scientists, and researchers to interpret, suggest, and review solutions for complex medical problems. Health care uses blockchain technology to eliminate redundant work, rework, and reconciliation. With AI, the right level of service intensity could be achieved by steering work in a predictive healthcare setting.

## 9.11 Future Prospects of Blockchain and AI in the Healthcare Ecosystem

Electronic Healthcare Records (EHR) help to maintain the security of patient information with the help of chained hashing [27]. "The unprecedented growth of augmented and virtual reality technologies has placed the digital health solutions on the cusp", says Dr Ben O. J. El Idrissi, COO and founder at Aimedis [28]. The Metaverse uses augmented reality and virtual reality along with AI and blockchains to create the virtual world. With Metaverse in health care, the benefits can be realized in the form of patient care, healthcare, and education sectors. Metaverse-based patient care will improve pre- and post-hospitalization care. Communicating with healthcare professionals from home may finally reduce the number of relapses after therapies and surgery and assist with treatment progress management with fewer costly and difficult office visits [28].

Additionally, in the Metaverse, patients and doctors will be able to control the long-term management of chronic diseases through regular reporting and follow-up sessions. Following up routinely in the Metaverse experience will significantly lower complications for patients and improve the long-term management of their conditions.

*Infrastructure stress is reduced in health care.* Through the adoption of the Metaverse, the healthcare industry will shift a large portion of its workflow to remote-based services. With the remote provision of the services, there will be less strain on infrastructure while giving patients better access to care. It will also strengthen our healthcare infrastructure, making it more able to meet the challenges of emergent situations like the COVID-19 pandemic.

*In the digital world, education has a crucial role to play in health care.* The participation of professionals and customers will be boosted significantly by exhibitions, seminars, and other informative events.

## 9.12 Conclusion

Researchers will have to address how this technology will be created in a way that is not only safe for the patient but also caters to the humanistic aspects of medicine, after all, health care is not just about treating the symptoms but additionally it's also about treating the person while artificial intelligence and blockchain hold incredible promise for health care. Moreover, it must be configured in a way that doesn't diminish the sensitivity of the patient-physician relationship that has long characterized health care when it comes to attracting an audience to AI and blockchain-enabled health care. With the proliferation of telehealth and mobile device integration, updated guidelines to accommodate the streaming online information that the metaverse and digital technologies provide are essential to protect the privacy of patients. Blockchain plays a vital role in privacy preservation, which enables patients to feel at ease to communicate and interact on a mass scale. High cost to enable the full potential of these technologies is required in addition to massive infrastructure like high-tech hardware, glasses, sensors, and uninterrupted 5G, and equipment to adhere to prescribed treatments. Thus, artificial intelligence and blockchain are emerging as immersive technologies with a big potential for optimizing patient care across the entire healthcare spectrum. Stakeholders need to pay close attention to the dimensions that need to transition to the advanced facilities discussed so far to create better expectations in health care.

## References

1. Blockchain healthcare and life sciences solutions (n.d.). www.ibm.com, https://www.ibm.com/blockchain/industries/healthcare
2. Thomas, L.: Blockchain Applications in Healthcare. News-Medical.net (2021). https://www.news-medical.net/health/Blockchain-Applications-in-Healthcare.aspx
3. Chen, M., Decary, M.: Artificial intelligence in healthcare: an essential guide for health leaders. Healthc. Manag. Forum **33**(1), 084047041987312 (2019). https://doi.org/10.1177/0840470419873123
4. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., Wang, Y.: Artificial intelligence in healthcare: past, present and future. Stroke Vascular Neurol. **2**(4), 230–243 (2017). https://doi.org/10.1136/svn-2017-000101
5. Ng, W.Y., Tan, T.-E., Movva, P.V.H., Fang, A.H.S., Yeo, K.-K., Ho, D., Foo, F.S.S., Xiao, Z., Sun, K., Wong, T.Y., Sia, A.T.-H., Ting, D.S.W.: Blockchain applications in health care for COVID-19 and beyond: a systematic review. Lancet Digital Health (2021). https://doi.org/10.1016/S2589-7500(21)00210-7
6. Victoria, S.: The Examples and Benefits of AI in Healthcare. ITRex (2021). https://itrexgroup.com/blog/examples-and-benefits-of-ai-in-healthcare/#
7. Gwyneth, I.: Why Blockchain is Important in 2021 and Beyond. 101 Blockchains (2021). https://101blockchains.com/why-blockchain-is-important/
8. Barakat, S., Al-Zagheer, H.: Blockchain tracking system of COVID-19 vaccination. Annals of the Romanian Society for Cell Biology 5059–5067 (2021)
9. IBM.: What is Blockchain Technology?—IBM Blockchain (n.d.). www.ibm.com, https://www.ibm.com/in-en/topics/what-is-blockchain

10. Limted, E.E.S.P.: Introduction to Cryptocurrency and Blockchain (n.d.). Edifypath.com. Accessed 24 June 2022, from https://edifypath.com/blog/post/global-outlook-of-blockchain
11. Eugene, T.: Best Blockchain Frameworks You Should Know About. Merehead (2021). https://merehead.com/blog/blockchain-frameworks-you-should-know-about/#:~:text=The%20blockchain%20frameworks%20are%20a
12. The top blockchain development frameworks for 2022. Fauna (2022). https://fauna.com/blog/top-blockchain-development-frameworks
13. Patil, R.A.: Blockchain and healthcare openings. Abhiguru (n.d.). Accessed 25 June 2022, from https://www.abhiguru.com/2020/11/blockchain-and-healthcare-openings.html
14. Role of Blockchain in Sustainable Development. GeeksforGeeks (2021). https://www.geeksforgeeks.org/role-of-blockchain-in-sustainable-development/
15. Robert, H.: What is the Metaverse: A Next Generation Virtual World. Dappradar (2022). https://dappradar.com/blog/what-is-the-metaverse?gclid=EAIaIQobChMIi-SizYG--AIVJJNmAh1VUgr6EAAYAiAAEgLgHvD_BwE
16. What Is the Metaverse? Binance Academy (2021). https://academy.binance.me/en/articles/what-is-the-metaverse?utm_campaign=googleadsxacademy&utm_source=googleadwords_int&utm_medium=cpc&ref=HDYAHEES&gclid=EAIaIQobChMIi-SizYG--AIVJJNmAh1VUgr6EAAYASAAEgL-8_D_BwE
17. Ma, A.: The conversation: What is the metaverse, and what can we do there? Daily Maverick (2022). https://www.dailymaverick.co.za/article/2022-06-13-what-is-the-metaverse-and-what-can-we-do-there/
18. Victor, C.: Is metaverse our future? Pixstory (2022). https://www.pixstory.com/story/is-metaverse-our-future/96980
19. Top Metaverse Use Cases and Applications.: PixelPlex (2022). https://pixelplex.io/blog/top-metaverse-use-cases/
20. Uma, S.: Latest Research Trends and Challenges of Computational Intelligence Using Artificial Intelligence and Augmented Reality: Intelligence and Sustainable Computing (2019). https://doi.org/10.1007/978-3-030-02674-5_3
21. Kaur, M.K., B.G.: Metaverse Technologies and Its Applications. Insights2Techinfo (2021). https://insights2techinfo.com/metaverse-technologies-and-their-applications/
22. Globally Acclaimed Blockchain Healthcare Revolution Limoverse is Launching in India. (2022). www.businesswireindia.com, https://www.businesswireindia.com/globally-acclaimed-blockchain-healthcare-revolution-limoverse-is-launching-in-india-78150.html
23. Impact Feature india today digital New, January 7, 2022UPDATED:, & Ist, 2022 11:59. (2022). Limoverse: A new metaverse for wellness enthusiasts built on blockchain. India Today. https://www.indiatoday.in/impact-feature/story/limoverse-a-new-metaverse-for-wellness-enthusiasts-built-on-blockchain-1897167-2022-01-07
24. Enter Limoverse, A Metaverse Everyone Can Benefit From (2022). www.outlookindia.com/, https://www.outlookindia.com/outlook-spotlight/enter-limoverse-a-metaverse-everyone-can-benefit-from-news-56214
25. Uma, S., Suganthi, J.: A smart and dynamic decision support system for nonlinear environments. In: Recent Advances in Intelligent Technologies and Information Systems. pp. 137–161. IGI Global (2015)
26. Revolutionizing Healthcare with Blockchain and Artificial Intelligence (2020). www.cbcamerica.org, https://www.cbcamerica.org/blockchain-insights/revolutionizing-healthcare-with-blockchain-and-artificial-intelligence#:~:text=Revolutionizing%20Healthcare%20with%20Blockchain%20and%20Artificial%20Intelligence
27. Rai, B.K.: Ephemeral pseudonym based de-identification system to reduce impact of inference attacks in healthcare information system. Health Serv. Outcomes Res. Method **2022**, 1–19 (2022). https://doi.org/10.1007/S10742-021-00268-2
28. Bhat, D.: What Role is Metaverse Playing in Reshaping the Healthcare Sector. Gulf Business (2022). https://gulfbusiness.com/what-role-is-metaverse-playing-in-reshaping-the-healthcare-sector/

29. Rai, B.K.: Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records **11**(1), 1–15 (2022). https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?https://doi.org/10.4018/IJRQEH.297076, https://doi.org/10.4018/IJRQEH.297076
30. Rai, B.K.: PcBEHR: patient-controlled blockchain enabled electronic health records for healthcare 4.0. Health Serv. Outcomes. Res. Method (2020). https://doi.org/10.1007/s10742-022-00279-7

# Chapter 10
# Recommendation Systems for Cancer Prognosis, Treatment and Wellness

**Harshita Bhargava, Snehal Gupta, Geetika Vyas, Amita Sharma, and Sreemoyee Chatterjee**

## 10.1 Introduction

Early stage diagnosis and effective treatment are critically acclaimed domains of research. Scientists and clinicians are experimenting with numerous approaches for early detection and medication. Much advancement has been made in this wellness journey. It is undeniable that computational tools have transformed the wellness process. Several severe diseases, such as cancer, diabetes, Parkinson disease, and HIV/AIDS, have been examined via computational methods in order to develop more sophisticated and faster treatments. Early identification and treatment of cancer is always valuable, however existing methods for early detection have low accuracy. In the majority of cases, cancer prognosis reaches a point where treatment is no longer possible.

Doctors and scientists are attempting to repurpose existing datasets of cancer patients in order to uncover patterns for disease occurrence. This pattern analysis is carried out using well-known computational techniques known as Machine Learning (ML). Artificial intelligence (AI) and machine learning (ML) are progressively gaining traction in everyday life, and are expected to have a significant impact on disease detection and treatment in the near future. They have paved the way for autonomous illness diagnosis tools by exploiting large data sets to face the future

H. Bhargava · G. Vyas · A. Sharma (✉)
Department of Computer Science & IT, IIS (deemed to be University), Jaipur, India
e-mail: amita.1983@iisuniv.ac.in

H. Bhargava
e-mail: harshita.bhargava@iisuniv.ac.in

S. Gupta · S. Chatterjee
Department of Biotechnology, IIS (deemed to be University), Jaipur, India
e-mail: snehalgupta31178@iisuniv.ac.in

S. Chatterjee
e-mail: Sreemoyee.Chatterjee@iisuniv.ac.in

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_10

problems of human disease identification at an early stage, particularly in cancer [1]. These techniques can find patterns and trends between cancer occurrences in massive datasets, as well as make predictions related to cancer consequences.

Another application of ML that has gained importance in cancer prognosis is customized medicine with special recommendations [2]. Patients' customized needs are catered by recommendation systems. This chapter provides a review of studies that use these methods to predict the prognosis of cancer. We also discuss the various types of recommendation algorithms and models that are used, the types of data that they integrate, and the overall performance of each given scheme in relation to breast cancer.

## 10.2 Cancer Diagnoses, Treatment and Rehabilitation

Cancer occurs as a consequence of the unrestrained growth of abnormal cells (tumor) that destroy other body tissues. Cancer cells replace healthy tissue with the tumor and spread to other tissues through the bloodstream. This is alluded to as *metastasis*. When there is a lump, a change in skin color, or organ enlargement, a physical examination is usually performed to diagnose cancer. A laboratory examination is the secondary method for affirming cancer. Imaging tests and biopsy are common methods of lab examination. Bone scan; Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) scan, Computerized Tomography (CT) scan, X-ray etc. are examples of imaging tests. Another important technique is biopsy, which involves removing tissue from a living body and examining it to verify the prevalence, aggravation, or extent of a disease.

There are various types of cancer, each with its own rate of inflammation and treatment regimen. All cancers, in general, have four stages: early stage, localized stage, regional spread, and distant spread. These stages describe the size and spread of solid tumors, as well as whether the cancer has spread to other parts of the body.

Cancer diagnosis and treatment has always been time-consuming, complicated, and painful. In order to alleviate these problems various computational strategies are being adopted for quicker and more effective diagnosis. Reports and data generated during certain diagnostic approaches have been combined with computational techniques to reduce the detection process. Urine and blood tests, for example, may assist your doctor in identifying abnormalities that may be caused by cancer. A common blood test known as a complete blood count, may reveal an unusual number or type of white blood cells in people with leukemia. Many studies used such datasets to create machine learning models that predict the likelihood of cancer. Noninvasive imaging tests, on the other hand, allow experts to examine internal organs(soft tissues), cartilages and bone. Some of the imaging tests used in cancer diagnosis include CT scans, bone scans, MRIs, PET scans, ultrasounds, and X-rays, which have contributed towards the development of massive data for computational analysis. Image classification, segmentation, and recognition have been used in such datasets to investigate cancer traces and grade.

Apart from this, biopsy procedures are also carried out to check the cell behavior. It is a surgical procedure in which a sample of tissue is removed by the experts. Later, a pathologist examines the tissue under a microscope and performs other tests to determine whether the tissue is cancerous. Pathology reports play an important role in cancer diagnosis and treatment planning. In most cases, a biopsy is the only way to confirm a cancer diagnosis. Fine needle aspiration (FNA) or endoscopy diagnoses are two methods for obtaining a biopsy sample. The doctor uses a needle to extract tissue or fluid during FNA. This technique is used for bone marrow aspirations, spinal taps, and some breast, prostate, and liver biopsies. Endoscopy involves the use of a thin, lighted tube called an endoscope by the doctor to examine areas inside the body. Endoscopes are inserted into the body's natural openings such as the mouth or anus. Histological images of infected tissues extracted in biopsy are also used in clinical studies to develop classification models.

During cancer treatment, patients frequently experience a variety of side effects, and their lifestyles change dramatically even after treatment. Physical, psychological, and cognitive issues are common side effects of treatment. These issues can make it more difficult to carry out everyday tasks or resume work. They may even have long-term consequences on their health. Cancer rehabilitation can assist with a variety of issues that arise during and after cancer therapy. Cancer rehabilitation encompasses a wide range of treatments aimed at improving a person's physical, emotional, spiritual, social, and economic well-being. Experts offered a variety of rehabilitation models to help people live healthier lives. However, multimodal, interdisciplinary rehabilitation is the ideal model of care, and it is best administered utilizing a prospective surveillance strategy, in which baseline measurements of performance and function are continuously assessed over time. Keeping this in mind, IoT-based technologies are being developed to continuously monitor patient health. Sensor-based gadgets, laboratory instruments, recommendation apps, modelling software, and a variety of other techniques have all helped to make comprehensive rehabilitation more successful and affordable. In a nutshell, computational tools aid cancer treatment at all levels [3].

Even it is instrumental in the progress toward personalized treatment. Even with cancer, methods that deliver customized therapy are being developed. Recommendation algorithms have aided in he development of these systems. Apps that provide recommendations are especially useful in the case of cancer rehabilitation. We've already talked about the different types of cancer and how to diagnose them. We shall cover recommendation/prescriptive methods for cancer in this chapter, with a focus on breast cancer.

## Breast Cancer

Breast cancer is the most frequent type of cancer in women all over the world [4]. Surviving rates, on the other hand, vary widely and are hopefully trending upward. The substantial shift in screening procedures, early diagnosis, and therapeutic advancements have all contributed to increased survival. We will briefly discuss breast cancer diagnosis, treatment, and rehabilitation in this section since it is critical to comprehend ideas linked to breast cancer before creating an effective recommendation system.

## Diagnosis

Breast cancer cells can be discovered in the breast tissue alone, in the axillary lymph nodes under the arm, or other parts of the body at the time of diagnosis. Breast cancer is classified into stages based on where it is discovered. Stages range from I to IV. Stage IV breast cancer, commonly known as metastatic breast cancer, is the cancer that has gone beyond the breast and axillary lymph nodes to other parts of the body [5].

Breast cancer is typically detected by screening or by a symptom (such as pain or a palpable tumor) that prompts a diagnostic examination. Most physicians advise women to undergo monthly breast self-examination (BSE) to get acquainted with their general anatomy and to empower themselves in terms of their healthcare. The cancer diagnosis is divided into basic three stages viz., (1) screening, (2) pathological evaluation, and (3) imaging and staging. The screening includes self- and clinical breast examination, mammography, ultrasound, and MRI along with physical symptoms which may indicate the presence of cancer. Then in the case of pathological evaluation, the diseased tissue is typically collected in clinical practice through fine-needle aspiration, core biopsy, or surgical excision [6]. The imaging and the staging part includes the generation of image using either Mammography or an MRI or an ultrasound. Early diagnosis of non-palpable tumors is one of the most significant advancements in breast cancer treatment which could be attained through mammography. Mammogram is carried out in women that have palpable masses or any other symptoms which are indication of breast cancer. Diagnostic mammograms shows major areas like a particular area of breast tissue viewed by focal compression or magnification images. Although mammography has always been the primary diagnostic imaging method for breast cancer, magnetic resonance imaging (MRI) has emerged as an essential modality in the identification, evaluation, staging, and therapy of breast cancer in a subset of patients. In high-risk women, screening MRI is more sensitive but less specific for cancer detection. MRI is useful in screening high-risk individuals, those whose breast augmentation limits effective screening by mammography and patients with equivocal findings on other imaging modalities [7]. Whole breast ultrasonography may help clinicians to screen breast cancers that standard mammography may not detect, particularly in dense breasts in where mammographic sensitivity is lower.

Multiple physicians from various disciplines often use distinct subsets of biomarkers and multiple clinical criteria to determine a cancer prognosis, such as the patient's age and general health, the location, and kind of cancer, as well as the tumor's grade and size.

## Treatment

The treatment for breast cancer includes surgery, radiation, and chemotherapy. Surgical intervention is still the major method of local and regional breast cancer therapy [8]. The varied surgical approaches that are in practice for treating breast tumors are mastectomy alone or with reconstruction which can be either primary or delayed, or breast conserving therapy (BCT), with or without the utilization of oncoplastic techniques. Another surgical option is a mastectomy that is no longer a simple surgery. It can be conducted as a traditional mastectomy, which is typically a viable choice of therapy for a particular set of patients since it is an outpatient operation with a short recovery and low risk of complications. The other surgical treatment for breast cancer includes the removal of the axilla in the early stage breast cancer. The axillary node status is critical in determining the fate of individuals with early stage breast cancer. The usual surgical strategy to invasive breast cancer was complete axillary lymph node dissection (ALND), or excision of level I and II axillary nodes. This surgery is now reserved for patients who have clinically positive lymph nodes verified by needle biopsy at the time of the first examination, or for patients who have clinically negative axilla assessed by ultrasonography and discovered to have a suspicious node, which is confirmed by needle biopsy. A sentinel lymph node (SLN) biopsy can be done safely at the time of mastectomy or lumpectomy in patients with a clinically and radiologically negative axilla, sparing them the morbidity associated with ALND. SLND enables a thorough histological study of lymph nodes, both with and without metastases. A comprehensive approach to breast cancer treatment has been critical to recent breakthroughs in the disease's management. One of these approaches is the adjuvant systemic therapy. The goal of this therapy is to enhance the disease-free survival (DFS) and overall survival (OS) rates which is related to treating BC with local treatments alone (surgery and/or radiation). Chemotherapy, endocrine treatment, and tissue- targeted medicines improve the efficacy of definitive local therapy (surgery, radiation therapy, or both), significantly lowering cancer recurrence and disease-specific mortality [9].

### Rehabilitation

According to the World Health Organization, rehabilitation is a process that enables patients to achieve and maintain optimal physical, intellectual, psychological, social, and spiritual functioning [10]. More emphasis should be placed on how cancer survivors cope with the results of their frequently extensive therapy, as well as how their quality of life evolves after treatment is over. While a cure for breast cancer remains the primary and most important goal for treatment, how the cure is accomplished is becoming increasingly crucial. At the same time, it must be acknowledged that breast cancer might reoccur for up to 20 years. As a result, the rehabilitation process includes not only dealing with the effects of completed treatments, such as surgery and radiotherapy, but also managing side effects from ongoing treatments, as well as assisting women in overcoming the psychological consequences of the cancer diagnosis, such as persistent fear of recurrence, a higher than normal rate of depression, and social implications, such as job loss.

Because breast cancer is the most frequent disease diagnosed in women, with relatively good overall survival rates, a comprehensive approach to survivorship care that focuses on restoring the physical function of women living with breast cancer is required.

## 10.3 Applications of Computer Based System in Cancer Study

Recently, Artificial intelligence and Internet of Thing (IOT) have become quite popular in devising better supportive tools in handling cancer cases. Artificial intelligence (AI) has taken the globe by storm. It has been reverberating around the world since the 1960s and continues to be a game changer in every industry. Medicine is no exception; AI has proved to provide a potential platform in medicine and health-care, in fact, oncology is working hard to comprehend the complicated algorithms at the root of cancer. Machine learning which is a subset of AI has dragged the attention of clinical researchers towards large scale ML algorithms. With the aid of ML, the computer has gained the ability to learn from large pharmaceutical data present at industrial scale along with its application in drug discovery in a short span of time and which is less costly. AI is advancing at a breakneck pace. Clinical oncology research is increasingly focusing on decoding the molecular genesis of cancer by studying the complicated biological architecture of cancer cell proliferation. It also aimed to process millions of relevant cases in big data and computational biology in order to address the current global scenario of increasing cancer mortality. Furthermore, the use of AI in clinical decision-making is thought to boost the likelihood of early disease prediction and diagnosis through Next Generation Sequencing (NGS) and high- resolution imaging techniques. It would also lead to the development of novel biomarkers for cancer detection, the development of novel tailored medications, and the delivery of prospective treatment options by creating large datasets and employing specialist bioinformatics tools.

Radiology has been a major part of the health-care system especially in the case of cancer and other cancer-related consequences. Radiologists are expected to be more technologically savvy as compared to any other medical professional. They are continually on the cutting edge of adopting digital medical imaging information. AI could detect abnormal data at a look, demonstrating a high sensitivity rate when compared to other conventional technologies[]. When considering traditional computer assisted detection (CAD) systems, they can only inform about the presence or absence of characters in the image under study, however AI-based systems collects all visible and non-visible image elements to produce more accurate results.

AI-based CAD (recommendation) systems for breast cancer detection on various modalities such as mammography, ultrasound, MRI, and biopsy histopathological images have gained popularity in recent years. Pre-processing, Segmentation, Feature extraction, and Classification are the four stages that make up the foundation of CAD systems. These stages are made up of various algorithms that help with cancer prognosis. In this chapter, we have discussed classification and recognition techniques used in hybrid recommendation models. Breast cancer diagnosis can be viewed as a classification problem in machine learning, with the outcome indicating which cancer class it belongs to or presence of cancer or strain presence in sample.

## 10.4 Recommendation Systems: History and Introduction

Recommendation systems are a widely known application of machine learning. Their integrity and usage in our daily lives can be visualized by the rise of web services like youtube, Netflix, Amazon, and many more. It was in 1979, that recommender systems were first talked about. A computer-based library system called Grundy implemented this concept and recommended books to the user. In recent times, recommender systems have proliferated after NetFlix and Amazon implemented the concept. Today recommender systems have an implementation in almost all facets of our life ranging from movie recommendations to e-commerce, e-government, e-learning, and also acting as a gateway to disease prediction and treatment [11].

Recommender systems are tools that are designed in such a way that considering a variety of factors, they recommend relevant suggestions to their users. These recommendations help users in identifying the correct information, products, or services. The underlying concept is that it has machine learning algorithms that filter information and data analytics techniques for making predictions pertaining to rating or user preference. They deal with a large volume of data in order to find out the most accurate match between users and items. Their basic principle of operation is to find patterns in data that can be implicit or explicit.

Working of Recommendation Systems:

Recommendation items work by combining data and machine learning technology. Data has a very critical aspect and is the basic building block for the whole system. The efficiency and effectiveness of recommendations depend majorly on the type of data being supplied to the tool. Data collection, storage, analysis, and filtering are the four steps in the Recommendation Systems Process.

Types of Recommendation Systems:

The three major types of recommendation systems based on the filtering technique used are Collaborative recommendation systems, Content-based recommendation systems, and hybrid recommendation systems. Table 10.1 includes their brief description [12–14].

Because of their growing popularity, recommendation systems are rapidly changing the medical scenario. They are bringing a revolution in the medical field through effective monitoring and analytics of medical records, as they are intelligent decision support systems. Their accurate disease diagnosis and recommendation of effective treatment promise faster recovery. The domain of cancer is also not untouched by its growing popularity. In the next section, existing work related to recommendation systems for the study of cancer disease is taken up.

**Table 10.1** Types of recommendation systems

| | |
|---|---|
| Collaborative filtering method | Such systems gather and analyze data based on users' behavior. It works on metadata based on its knowledge about the user and makes limited recommendations. The collaborative filtering technique can be User-User collaborative filtering or Item-Item collaborative filtering |
| Content-based filtering methods | In these systems, product description keywords and the profile of the user's choices are the base for filtering relevant information. It works upon transactional data and limits its recommendations up to its knowledge about user-profiles and product descriptions |
| Hybrid recommendation systems | Such systems use content-based and collaborative filtering simultaneously and recommend a broader range to customers. Being hybrid in nature, they work on metadata as well as transactional data and provide more accurate recommendations in comparison to other recommender systems. Commonly used approaches by these systems are listed in Fig. 10.1 |

## 10.5 Recommendation Systems\Algorithms for Cancer Study

### *10.5.1 Predicting Cancer Drug Response Using a Recommender System*

The availability of genomic databases such as Cancer Cell-Line Encyclopedia (CCLE) and the collaborative Genomics of Drug Sensitivity in Cancer (GDSC) have proved to be a useful resource for studying and predicting patient specific drug responses. With the same idea in mind, this study used the concept of collaborative filtering technique from the ecommerce domain while mapping the unseen cell lines as users and drugs as items. As a first step in developing the recommender system: CaDRReS ,the cell line features are extracted based on the gene expression data.

In this study, the authors used the popular matrix factorization from the latent factor model category expressing the cell line specific drug response as the dot product of two vectors. These two vectors include drug vector and the cell line vector which depict the respective latent vectors and their dot product depicts the interaction between drugs and cell lines. The dot product was termed as the "pharmacogenomics space" which captured the interactions between drugs and cell lines. The resultant learned space was also used to infer drug-pathway associations, find drug mechanisms and cell line subtypes [14].The advantage of the developed model was its ability to predict even for newer datasets. As future work the authors suggested the use of mutations and interactome networks along with the genomic information into the existing model for robust predictions.

Weighted Approach - Takes the outputs from each of the models and combines the result in the form of static weight which doesn't change across the training and testing set.

Switching Hybrid Approach - Adds an additional layer to the model, to select the suitable model to be used. This system is sensitive to the strengths and weaknesses of the constituent recommendation model.

Mixed Hybrid Approach - The recommendation system inputs different sets of candidates to the recommendation model and combines the predictions to produce the resulting recommendation.

Feature Combination Hybrid Approach - A virtually contributing recommendation model is added to the system, this additional feature-engineering contributes to the original user profile dataset.

Feature Augmentati on Hybrid Approach - A contributing recommendation model generates ratings or classifications of the user profile or items to be used in the main model for producing final recommendations.

Cascade Hybrid Approach - The main system produces the primary results; the secondary system, defined in the hierarchy, is used to resolve issues pertaining to the primary results.

Meta-Level Hybrid Approach - A contributing model provides datasets to the main model which replaces the original datasets.

**Fig. 10.1** Commonly used approaches used to build hybrid recommender systems

### *10.5.2 Recommender System for Breast Cancer Patients*

Breast cancer is one of the major diseases that account for high mortality rate amongst women all over the world.In order to address such an adverse situation recommender systems can act as an information filtering tool for the patients.This idea was translated by the authors [15] by developing a recommender system for suggesting personalized articles for breast cancer patients.These articles may be related to treatments, lifestyle, associated risks and behavioral concerns.A hybrid recommendation system named Breast Cancer Recommender System (BCRS) was developed to deliver all the information aspects related to cancer, through a single system.The hybrid system used a switching approach between the collaborative and content based techniques to handle the cold start issue. The articles are recommended on the basis of user profile similarity with the target user. If the target article has been already read by the neighboring users then their rating is predicted on the basis of ratings provided by these similar users.In case the target article has not been read by the neighboring users then the rating is predicted on the basis of the articles read by neighboring users from the same document category. The drawback was that the system could not be evaluated using real cancer patients.

### *10.5.3 Personal Health Information Recommender for Empowering Cancer Patients*

Patient empowerment is a new buzzword in today's world wherein, the patients search for health related information using the internet. The major challenge is that the authenticity/reliability of the varied information sources cannot be assured. Hence the user needs to filter out the relevant information on the basis of one's experience and discretion. In order to facilitate this search task the authors developed a Personal Health Information Recommender (PHIR) system. The system integrates (a) PHIR search engine for searching health related data based on the patients' medical profile and preferences (b) Semantic annotator app for the experts/doctors/health professionals to add validated text/articles/multimedia content to the system. The main advantage of the developed system was that even if two patients are similar in terms of ailment and treatment history then also they might get a different piece of information based on their individual preferences and contexts. Search engine feature with the addon recommendation ability to match the patient's profile the relevant/validated information from the health experts marks the uniqueness of this system [16].

### *10.5.4 Gene Based Recommendation Algorithm to Recommend Genes for Cancer Patients*

The open source biological databases have encouraged the medical practitioners to study the diseases on the basis of genetic factors as opposed to the traditional symptoms based approaches.In line with this approach a gene based collaborative filtering algorithm [GeneCF] was proposed to assist the doctors to find the list of genes with respect to each patient for effective diagnosis [17]. A new evaluation metric named gene precision coverage was used to assess the effectiveness along with a cancer staining procedure of the genes from the Human Protein Atlas to verify the obtained results. AMP, SAA1, S100P, SPP1 and CY2A7 and AFP were the six genes which were listed out of which only S100P, SPP1 and CY2A7 and AFP were identified as cancer staining. In order to decide the number of genes to be recommended an experimental design was proposed which was further validated from RNA expression of both healthy and liver cancer patients. The results were further validated against the genes and samples from cancer staining of patients.Also Gene Precision (GP), Gene Precision (GP), Gene Coverage (GC) and Gene Precision

Coverage (GPC) were proposed as performance indicators with respect to gene based data. These indicators were modified version of coverage and precision metrics used in evaluating recommendation systems.This algorithm used a two step process: (1) finding the gene interest (Gi) matrix for each patient using the following formula:

$$\left.\begin{aligned}
&\text{Gene expression (Ge): Ge} = 2^{\text{LOG}}\\
&\text{Gene singlemode network (G): G} = \text{Ge} * \text{Ge}'\\
&\text{In the Gene similarity matrix (Ja)} : ja_{ij} = \frac{g_{ij}}{g_{ii} + g_{jj} - g_{ij}}\\
&\text{Gene interest (Gi): Gi} = \text{Ja} * \text{Ge}\\
&\text{Gene Rank (GR): GR} = \text{sort (Gi)}
\end{aligned}\right\} \Rightarrow \text{Gene Interest}$$

(2) Finding the minimum no of genes to recommend with maximum precision and coverage

$$\left.\begin{aligned}
&\text{Gene Precision (GP): GP} = \frac{\sum_u |R(u) \cap T(u)|}{N}\\
&\text{Gene Coverage (GC): GC} = 1 - \frac{S}{M}
\end{aligned}\right\} \Rightarrow \text{GPC} = \sqrt{K * \frac{\text{GP} * \text{GC}}{\text{N}}}$$

The above use-cases justify the use of different recommender systems/algorithms for cancer based studies. These include predicting drug response for unseen cell lines, identifying cancer staining genes to help doctors for effective diagnosis, empowering cancer patients to search validated information from experts/doctors, predicting relevant articles to patients related to treatments, lifestyle, associated risks and behavioral concerns. The main advantage with such systems/algorithms is the ability to generate relevant and useful recommendations as per the patient's profile. The main drawback is that they employ only a single data source in a particular format in contrast to a varied data sources of different formats and types.

## 10.6 Recommendation System with Blended Approach for Breast Cancer Diagnosis-BC Recommender

In this case study, we would explain a recommender model for breast cancer study specifically with a blended approach. While designing the recommender model several data sources have been considered from different studies and arranged in a layered architecture to form layers with rules. Each model comprises of different machine learning models and takes different kinds of input as per the condition required. This architecture tries to blend different conditions with data and provides layer by layer recommendations.

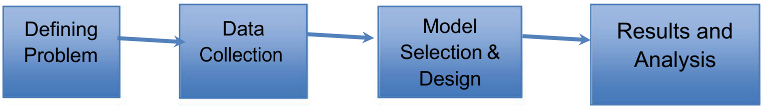


**Fig. 10.2** Methodology for designing breast cancer recommendation system- BC recommender

### *10.6.1 Methodology*

This research has a four stage methodology as discussed in Fig. 10.2 .

(i) Problem Definition and Objectives:

Current systems are based on clinical verifications with computational approaches. Only difficulty with such systems is that they are very specific and take only particular kind of data. To be more precise in diagnosis and treatment of this deadly disease, we need systems to consider all aspects of the disease like its symptoms, lifestyles, allergens, genetics etc. Therefore, there is a need for a blended approach in the recommendation system. The goal of this study was to collect secondary data from various levels of cancer testing and create a recommendation model by combining data and machine learning models. We are attempting to include symptoms and reports in this system; in the future, we will incorporate lifestyles and genetics into this system by utilizing secondary datasets.

(ii) Data Collection:

A review of the literature revealed a wide range of breast cancer data. It combines clinical test results, pathology reports, genetics, and lifestyle information. Table 10.2 discusses the datasets derived from various sources (Fig. 10.3 and Table 10.3).

From these datasets 4 datasets (1, 3, 4 and 6) have been selected for designing the recommendation model (Fig. 10.4).

(iii) Model Selection and Design Model Selection:

After data selection, suitable classification models were identified. Table 10.4 discusses data, selected model and reason of selection.

Layered architecture:

As we know, breast cancer screening is an important strategy to allow for early detection and ensure a greater probability of having a good outcome in treatment, hence in this research, a layered architecture with blending of data has been proposed. Table 10.5 discusses the layers, purpose & prerequisite, dataset & model, recommendation, feedback in the proposed model.

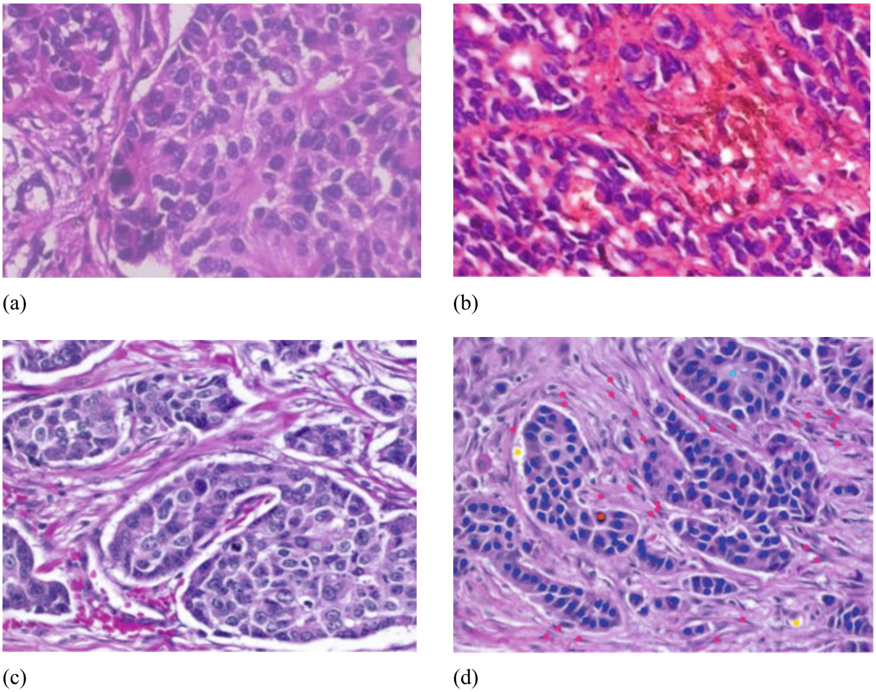
**Table 10.2** Datasets for cancer study

| S.No | Sources | Sample collection | Information present in data | Nature of data & Size |
|---|---|---|---|---|
| 1 | https://www.bmccancer.biomedcentral.com/track/pdf/10.1186/s12885-017-3877-1.pdf [18] | Clinical report (biomarker of breast cancer) | Blood analyses—notably, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI) | • Numeric and categorical<br>• Size 166 (both patients with and without cancer) |
| 2 | https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric | Clinical profiles | Patient age with cancer type, surgery, cellularit y, chemotherapy, Pam50 + Claudin-low subtype, ER measurements | • Numeric and categorical<br>• 2,509 breast cancer patients |
| 3 | https://www.archive.ics.uci.edu/ml/datasets/breast%2Bcancer%2Bwisconsin%2B(Diagnostic) [19] | Fine needle aspirate (FN A) | 30 real-valued features of each cell nucleus including radius,texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension | – Real values<br>– Size 569 |
| 4 | https://www.web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ [20] | Histopathology images | Four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenona (TA); and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC) | – Images<br>– 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples |

(continued)

**Table 10.2** (continued)

| S.No | Sources | Sample collection | Information present in data | Nature of data & Size |
|---|---|---|---|---|
| 5 | https://www.sciencedirect.com/science/article/pii/S2352914820300757 [21] | Histopathol ogy images | Grading IDC based on mitotic score | 922 images related to 124 patients with IDC |
| 6 | https://www.bmcresnotes.biomedcentral.com/articles/10.1186/s13104-019-4121-7 [20] | Histopathol ogy images | BreCaHAD, The dataset currently contains four malignant tumors (breast cancer): ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and tubular carcinoma (TC). These annotations are mitosis, apoptosis, tumor nuclei, non-tumor nuclei, tubule, and non-tubule | 162 breast cancer histopathology images |
| 7 | https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida [22] | Mammogra phy scans | Character of background tissue, Class of abnormality, Severity of abnormality, x, y image-coordinates of centre of abnormality | Grayscale image file saved in the portable gray map nearly 325 files |
| 8 | https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida [22] | Curated Microarray Database (CuMiDa) genetic data | The gene expression levels of 54,676 genes (columns) from 151 samples (rows) with 6 classes | Numeric values Size 151 |



(a) (b) (c) (d)

**Fig. 10.3** Histopathology slide for ductal carcinoma cancer. **a** Histopathology slide for lobular carcinoma. **b** Non annotated histopathology image of cancer. **c** Annotated histopathology image of cancer

**Table 10.3** Model selection for different datasets

| Data Sources | Model | Reason |
|---|---|---|
| Table 10.3 (1) | Ensemble | Data set is a combination of numeric and categorical data. In such cases, ensemble has high performance and is quite robust |
| Table 10.3 (3) | Ensemble | |
| Table 10.3 (4) | CNN | Image data sets are processed using CNN which perform well with medical images |
| Table 10.3 (6) | Mask R-CNN | This data set is masked with specific parts of infection. Problem here is image recognition and in such cases Mask CNN is suitable |

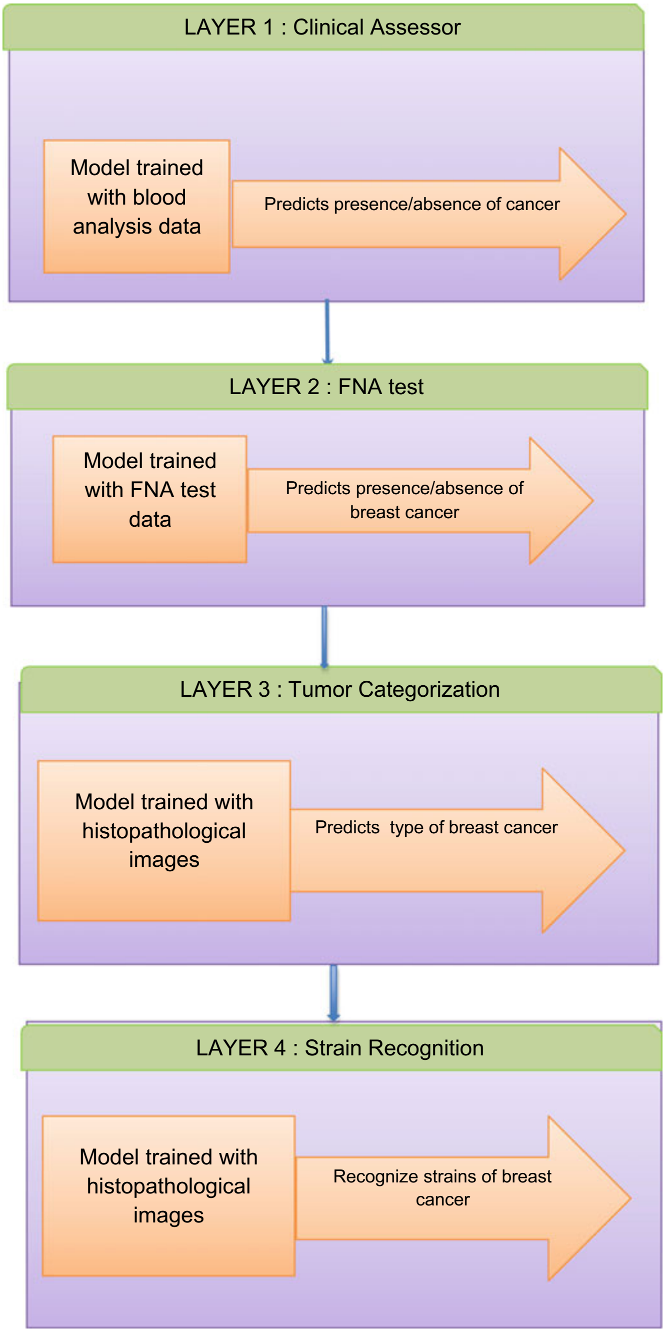
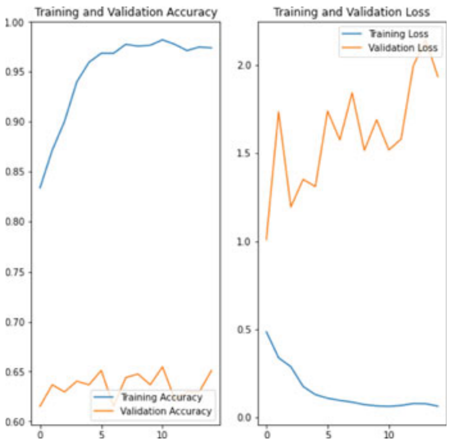

LAYER 1 : Clinical Assessor
Model trained with blood analysis data
Predicts presence/absence of cancer
LAYER 2 : FNA test
Model trained with FNA test data
Predicts presence/absence of breast cancer
LAYER 3 : Tumor Categorization
Model trained with histopathological images
Predicts type of breast cancer
LAYER 4 : Strain Recognition
Model trained with histopathological images
Recognize strains of breast cancer

**Fig. 10.4** Working model of proposed BCRecommender

**Table 10.4** Layer structure of blended recommendation system-BCRecomender

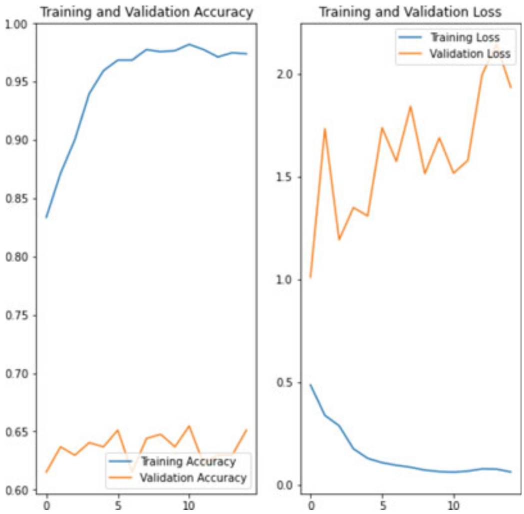
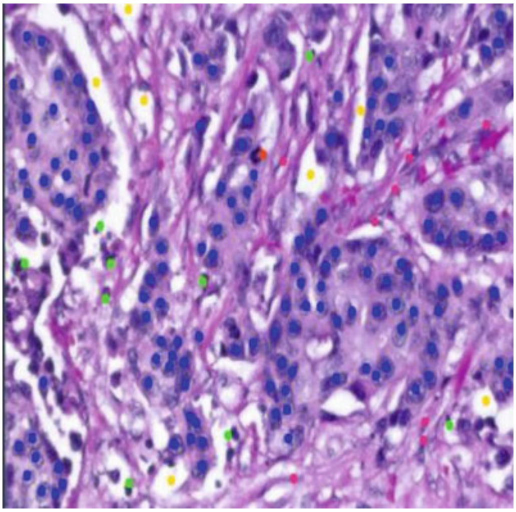
| Layer no | Name | Purpose and prerequisite | Dataset and model | Recommendation | Feedback |
|---|---|---|---|---|---|
| 1 | Clinical Assessor | – Examines the presence of cancer-blood sample | Dataset: Table 10.3 (1)-Train/Test:70:3 0<br>Model: Ensemble | – Possibility of cancer<br>– If possibility is true, recommended for physical examination followed by FNA test | Outcomes with sample stored in database after verification from expert |
| 2 | FNA Test | – Examine the type of tumor<br>– digitized image sample | Dataset: 569 Table 10.3 (3)<br>Train/Test:70:3 0<br>Model: Ensemble | – Possibility of lump or node to be cancerous<br>– If possibility is true, recommended for examination of type of cancer and treatment | – Sample values with recommendati on stored in database after verification from expert |
| 3 | Tumor Categorization | – Examine the type of tumor | Actual Dataset: 2,013 from Table 10.3 (4) | – Possibility of particular type of cancer | – Sample images with recommendati |
| | | – Digitized image sample | Selected images:1390 of malignant types (consist of 200X images only)<br>Train/Test: 70:30<br>Model: CNN | – Annotate sample for better recognition | On stored in database after verification from expert |
| 4 | Strain Recognition | – Recognize infected part among: Mitosis, Apoptosis, Tumor nuclei, Non-tumor nuclei, Tubule and Non-tubule | Dataset: 162 Table 10.3 (7)<br>Train/Test: 162 with number of annotations: 23,549<br>Model: Mask R-CNN | – Strain recognition in different parts | – Sample annotated images with recommendati on stored in database after verification from expert |

**Table 10.5** Performance of each layer in recommendation system

| Layers | Model performance | | | | |
|---|---|---|---|---|---|
| 1 | Ensemble classifier | Accuracy (mean) | Recall | Precision | F1-score |
| | Bagging | 61.06 | 73.0 | 73.0 | 74.0 |
| | Adaboost | 56.72 | 73.0 | 75.0 | 74.0 |
| | Voting classifier | 53.49 | 74.0 | 74.0 | 74.0 |
| 2 | Ensemble classifier | Accuracy (mean) | Recall | Precision | F1-score |
| | Bagging | 97.52 | 94.0 | 94.0 | 94.0 |
| | Adaboost | 95.60 | 95.0 | 95.0 | 95.0 |
| | Voting classifier | 55.56 | 96.0 | 96.0 | 96.0 |
| 3 | CNN | Training loss | Training accuracy | Validation loss | Validation accuracy |
| | Model before augmentation | 0.4311 | 83.18 | 0.9860 | 66.91 |
| | |  | | | |
| | Model after augmentation | Training loss | Training accuracy | Validation loss | Validation accuracy |
| | | 0.0631 | 97.39 | 1.9344 | 65.11 |

(continued)

**Table 10.5** (continued)

| Layers | Model performance | |
|---|---|---|
| | |  |
| 4 | Mask RCNN<br>(resnet) | <br>Ground Truth |

(continued)

**Table 10.5** (continued)

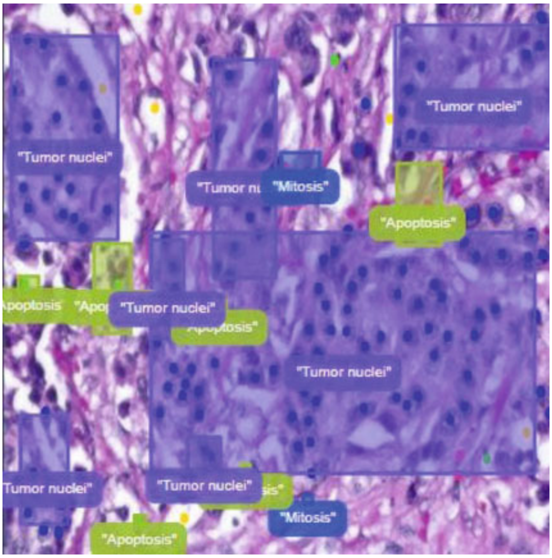
| Layers | Model performance | |
|---|---|---|
| | |   "Tumor nuclei"<br>"Tumor nuclei"<br>"Tumor nu<br>"Mitosis"<br>"Apoptosis"<br>poptosis<br>"Apo<br>"Tumor nuclei"<br>poptosis"<br>"Tumor nuclei"<br>Tumor nuclei"<br>"Tumor nuclei"<br>sis"<br>"Mitosis"<br>"Apoptosis"<br>Evaluation check points |

**Table 10.6** Observations on BC recommender

| Layers | Utility | Constraints | Enhancement needed |
|---|---|---|---|
| 1 | Support to analyze clinical data at initial phase of cancer | Training dataset is limited and imbalanced | Dataset can be expanded and balanced for better training |
| 2 | FNT test analysis based on existing data | Training dataset is imbalanced and numeric | • Balanced dataset<br>• Image dataset can be combined for better prediction |
| 3 | Support in Recognizing cancer type | • Model was trained based on 4 types of cancer<br>• 200X images were considered | • Images from other types of cancer can be combined to improve recommendations<br>• Images of 400X<br>Configuration can also be considered for evaluation |
| 4 | Support strain detection | • Type of strains and images are limited | • Images from other types of cancer can be considered<br>• Strain annotated images can be increased for better analysis |

In the first layer, the dataset contains data from routine blood analyses—notably, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age and Body Mass Index (BMI)—all of which can be used to predict the presence of breast cancer. This is an entry layer which will try to screen cancer based on clinical data. The model used is an ensemble, trained on already existing data as discussed in Table(5). The outcome of this model is the likelihood of presence (1)/absence(0) of breast cancer. During recommendation executing original data of a new patient is provided to this model which will generate binary results (presence or absence). If the outcome is presence, then we move to the second layer and also record the case in the training dataset of layer 1. The system also recommends to go for pathology tests as the model shows likelihood of presence of cancer. In the case of absence no further test data would be taken into consideration.

In layer 2, a dataset from a fine needle aspirate (FNA) test was considered to design the model. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. FNA dataset is used to train the ensemble model of this layer. The FNA of target is then collected from FNA test after converting digital images into numeric values. The outcome of this model will again confirm the presence and absence of breast cancer. If target values reveal the presence, then the target would be suggested to go for further test and shifted to layer 3.

In layer 3, histopathological images are considered to train the CNN model. The dataset currently contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenona (TA); and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC).The dataset is taken from BreakHis, Breast Cancer Histopathological Image Classification (BreakHis) which is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples (700 × 460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). 200X images are considered for model fit.

In this system, we have considered a malignant tumor dataset with 200X images (i.e. 1390 total images) for training our model. The reason for this selection is to keep the system fast and focus towards cancer type detection. The outcome of this model helps to classify the type of cancer based on images. The recommendation will be type of cancer PC, DC, LC or MC. If the input image is different from these classes, then this system will not give correct results.

In layer 4, an image recognition model is created using breast cancer histopathology images named BreCaHAD which is publicly available to the biomedical imaging community. The images were obtained from archived surgical pathology example cases which have been archived for teaching purposes. The hematoxylin and eosin (H&E) stained histological images are annotated or marked by a pathologist as either mitosis, apoptosis, tumor nuclei, non-tumor nuclei, tubule, and non-tubule. This model has not been trained for papillary carcinoma (PC), as BreCaHAD does not consist of the annotated data of such cases. This model will provide a marked part by highlighting the image regions with rectangle boxes.

Though the system is its infancy, we will use MYSQL database for importing the data from the dataset and thereby storing the generated recommendations. These recommendations can be used as input to the feedback module wherein the experts can verify the results (Fig. 10.4).

iv) Results and Analysis

## 10.7 Observations and Discussion

The BCRecomender system designed in this research is unique and tries to provide continuous assistance for faster cancer detection. As the system still is in its initial phase, there are certain constraints that need to be handled to make this system more flexible and extensive. This section discusses each layer with their advantages and limitations. Table 10.7 describes the detailed observations.

## 10.8 Challenges and Future Work

The proposed BCRecommender is an effort to address the challenges related to the early detection of breast cancer, thereby providing necessary medications and rehabilitation. One of the major challenges is the imbalance of the datasets which result in biased predictions. Secondly the deep learning model in the layered architecture of the recommender system requires a large amount of image data for training thereby affecting the accuracy of predictions. Most of the datasets either lack this in number or quality. As a part of future work we would consider other cancer datasets including gene expression or metabolomics data.

## References

1. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Futur. Healthc. J. **6**(2), 94 (2019)
2. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)
3. Stout, N.L., Santa Mina, D., Lyons, K.D., Robb, K., Silver, J.K.: A systematic review of rehabilitation and exercise recommendations in oncology guidelines. CA: Cancer J. Clin. **71**(2), 149–175 (2021)
4. Ghoncheh, M., Pournamdar, Z., Salehiniya, H.: Incidence and mortality and epidemiology of breast cancer in the world. Asian Pac. J. Cancer Prev. **17**(sup3), 43–46 (2016)
5. Lim, B., Hortobagyi, G.N.: Current challenges of metastatic breast cancer. Cancer Metastasis Rev. **35**(4), 495–514 (2016)

6. Stewart, C.J.R., Coldewey, J., Stewart, I.S.: Comparison of fine needle aspiration cytology and needle core biopsy in the diagnosis of radiologically detected abdominal lesions. J. Clin. Pathol. **55**(2), 93–97 (2002)
7. Bicchierai, G., Di Naro, F., De Benedetto, D., Cozzi, D., Pradella, S., Miele, V., Nori, J.: A review of breast imaging for timely diagnosis of disease. Int. J. Environ. Res. Public Health **18**(11), 5509 (2021)
8. McDonald, E.S., Clark, A.S., Tchou, J., Zhang, P., Freedman, G.M.: Clinical diagnosis and management of breast cancer. J. Nucl. Med. **57**(Supplement 1), 9S-16S (2016)
9. Shah, R., Rosso, K., Nathanson, S.D.: Pathogenesis, prevention, diagnosis and treatment of breast cancer. World J. Clin. Oncol. **5**(3), 283–298 (2014). https://doi.org/10.5306/wjco.v5.i3.283
10. World Health Organization.: Promoting mental health: Concepts, emerging evidence, practice: Summary report. World Health Organization (2004)
11. Nagarnaik, P., Thomas, A.: Survey on recommendation system methods. In: 2015 2nd International Conference on Electronics and Communication Systems (ICECS). IEEE (2015)
12. Lucero-Álvarez, C., et al.: Literature review on information filtering methods in recommendation systems. In: 2021 Mexican International Conference on Computer Science (ENC). IEEE (2021)
13. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: principles, methods and evaluation. Egypt. Inform. J. **16**(3), 261–273 (2015)
14. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: The Adaptive Web. Springer, Berlin, pp. 325–341 (2007)
15. Suphavilai, C., Bertrand, D., Nagarajan, N.: Predicting cancer drug response using a recommender system. Bioinformatics (2018). https://doi.org/10.1093/bioinformatics/bty452
16. Kanimozhi, G., Shanmugavadivu, P., Rani, M.M.S.: Machine learning-based recommender system for breast cancer prognosis. Recomm. Syst. Mach. Learn. Artif. Intell. 121–140 (2020). https://doi.org/10.1002/9781119711582.ch7
17. Iatraki, G., Kondylakis, H., Koumakis, L., Chatzimina, M., Kazantzaki, E., Marias, K., Tsiknakis, M.: Personal health information recommender: implementing a tool for the empowerment of cancer patients. Ecancermedicalscience **12** (2018)
18. Hu, J., Sharma, S., Gao, Z., Chang, V.: Gene-based collaborative filtering using recommender system. Comput. Electr. Eng. **1**(65), 332–341 (2018Jan)
19. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., Caramelo, F.: Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer **18**(1), 1–8 (2018)
20. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7), 1455–1462 (2015)
21. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007)
22. Bolhasani, H., Amjadi, E., Tabatabaeian, M., Jassbi, S.J.: A histopathological image dataset for grading breast invasive ductal carcinomas. Inform. Med. Unlocked **19**, 100341 (2020)
23. Aksac, A., Demetrick, D.J., Ozyer, T., Alhajj, R.: BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis. BMC. Res. Notes **12**(1), 1–3 (2019)
24. Feltes, B.C., Chandelier, E.B., Grisci, B.I., Dorn, M.: Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. J. Comput. Biol. **26**(4), 376–386 (2019)

# Chapter 11
# Real-Time Data Mining-Based Cancer Disease Classification Using KEGG Gene Dataset

**Parvathala Balakesava Reddy, Ahmed J. Obaid, V. Sivakumar Reddy, and K. Saikumar**

## 11.1 Introduction

Data Mining is a method for searching through a vast quantity of data to extract the important information. It allows for the pattern recognition and analysis of big datasets with the use of the software. Documents withdrawal may be used in a variety of fields, including research and science. So, by using data mining, businesses may simply learn about their consumers and adopt a variety of valuable ideas that are relevant to a wide range of business operations. As a result, business people will be able to better use their existing resources and make better business judgments. This method's main purpose is to shift through a large volume of data in search of useful information. By using this technique, one may discover relevant examples and fresh knowledge for a group or individual in need of it. Devices like these are used by humans. Data mining is generally capable of handling computer processing, data storage, and data collecting. Data is partitioned and future occurrences are estimated using mathematical programmers or algorithms that are optimized. Insights gained via data mining include:

P. B. Reddy
VNR Vignana Jyothi Institute of Engineering Technology, Vignana Jyothi Nagar, Hyderabad, India

A. J. Obaid (✉)
Faculty of Computer Science and Mathematics, University of Kufa, Kufa, Iraq
e-mail: ahmedj.aljanaby@uokufa.eu.iq

V. S. Reddy
Department of CSE, Malla Reddy University Hyderabad, Hyderabad, India

K. Saikumar
Department of ECE, Koneru Lakshmaiah Education Foundation, Green Fields, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_11

- Automatic pattern estimates may be provided based on behavior analysis and trends.
- The estimate is based only on probabilities.
- Decision-oriented information is most likely to be created.
- Research and analysis should focus on databases and massive data sets.
- The clustering process is dependent on identifying and locating physically accessible documented groupings.

Database size and query complexity are key factors in determining the best results. To build a strong and successful system, more data is required. Execution and organization both make use of this technique [1].

For more crucial inquiries and a large number of questions, query complexity is necessary. So a standard system is needed to meet all of these objectives. Marketing, mathematics, engineering, and science are just a few of the fields where data mining methods have had a significant impact. The distributor may use data mining to determine the customer's purchase range so that it can provide advances depending on the individual's purchasing history. With the aid of remark or guarantee cards, the distributor or retailer may mine segment information to produce items and improvements that communicate with exterior client partitions. Oncology is the branch of medicine that deals with diseases caused by the growth of abnormal cells. It may affect different sections of the body by attacking or spreading. There will be cancer in proportion to tumors that are benign and do not grow or disperse. Cancer may be divided into four broad categories. It's cancer that may be classified as either malignant or non-malignant. Carcinomas will have the most impact on cancer development when it comes to these factors. The skin, lungs, breasts, and other organs may all be affected by cancer. Tumors of the lymphocytes are known as lymphomas. Leukemia is a cancerous blood tumor. It is possible to accurately identify cancer using the association rule mining technique (ARM). Flexible statements like if/then will be accessible in this rule mining and will greatly aid in determining the relationships between different free social databases or other information repositories. Because they operate with numerical datasets, computer-based algorithms are almost always scientific. It is appropriate to use this technique when dealing with numeric data since it simply requires simple counting [2].

Infections, medicines, and synthetic compounds are all included in KEGG's collection of information on genomes and organic processes. Uses of the KEGG include bioinformatics examination and training, reenactment in systems biology, and translational inquiry in medication development. Our approach to drug discovery is undergoing a radical shift because of the developments in proteomics. Proteomics, particularly in the context of examining bodily fluid samples, will be described. Proteomics is the study of proteins in an organic sample, which involves the accurate partitioning, identification, and depiction of proteins. It is possible to identify alterations in the articulation of proteins that may be linked to organ poisoning by comparing ill samples to healthy samples. When analyzing tissues with clinical pathologies, a strange connection may exist between the degree to which particular features have an impact and the relative abundance of the appropriate proteins inside the tissue that is measured using proteomics. The serum is the term used to describe the liquid and clear portion of blood that does not play a significant role in clotting. Blood plasma lacking fibrinogens might be defined as such. Fibrinogens are the proteins found in blood plasma that are converted into fibrin blood clots [3].

## 11.2 Literature Survey

Ong et al. [4] for effective early detection and clinical diagnosis, the identification of reliable cancer biomarkers is essential. Integrative microarray data analysis is one method for finding cancer biomarkers based on gene expression. Microarrays are a very effective high-throughput technique that enable the comprehensive study of human genes together with related biological data. However, more research is required to enhance the predictive capacity, repeatability, and interpretability of the identified gene biomarkers, making them suitable for clinical use. Using the informed top-k class associative rule (iTCAR) method, we propose a unified framework for discovering cancer-related candidate genes. We present iTCAR, a refined associative classification system that integrates microarray data with biological knowledge from gene ontology, KEGG pathways, and protein-protein interactions to provide informative class associative rules. Ranking and selecting class associative rules for creating trustworthy classifiers using a new measure of interestingness. With an average classification accuracy of over 90% and an average area under the curve of over 0.80%, the testing results show a high prediction of iTCAR. The use of functional enrichment analysis and the retrieval of pertinent cancer concepts also considerably improve iTCAR's credibility and interpretability. These positive results show that the proposed method has great potential for identifying candidate genes that may be further investigated as biomarkers for cancer illnesses [4].

Approaches that analyze biochemical routes represented by functional units are rapidly replacing the extra conventional methods that concentrate on individual genes in microarray data analysis because of the greater insight they provide into gene expression and illness connections [5]. Genotypic functional modules may be related to disease manifestations through known molecular interactions, but effective procedures to make this connection are still in their infancy. In the first part of this essay, we talk about methods that use sets of genes to perform illness classification tasks that ultimately try to establish causal relationships between gene expression patterns and clinical outcomes. Following this, we provide a mathematical programming model using hyper-box concepts to classify diseases based on pathways. The molecular patterns associated with the diseased pathway are examined with the association rules retrieved from the model. We argue that using gene sets that correspond to disease-relevant pathways is a promising way to discover expression-to-phenotype relations in disease classification and demonstrate how hyper-box classification can be used to evaluate the predictive power of functional pathways and identify the effect of individual genes in this setting [5].

Association rules mining is an effective technique for identifying previously unknown relationships between genes within a biological dataset [6]. While prior methods often only work for a single biological data set, a single minimum support cutoff may be used globally, i.e. across all gene sets/item sets. By combining gene expression, methylation, and protein-protein interaction profiles, this work offers a dynamic threshold-based FP-growth rule mining approach that makes use of weighted shortest distance to find novel correlations between pairs of genes in multitier data sets. To do this, we introduce three new thresholds—distance-based variable/dynamic supports (DVS), distance-based variable confidences (DVC), and distance-based variable lifts (DVL) for each rule—and incorporate co-expression, co-methylation, and protein-protein interactions into the multi-omics data set. Three state-of-the-art multiple threshold tests were used to develop the proposed procedure. In the suggested approach, the DVS, DVC, and DVL values are computed separately for each rule, and then the support, confidence, and lift off the evolved rule are compared to those of the individual rules to ensure they are more than or equal to those of the individual rules. Only if these three conditions are true is a rule regarded as a resultant rule. The key advantage of the proposed technique over comparable state-of-the-art approaches is that it accounts for the quantitative and interaction relevance among all paired genes associated with each rule. Additionally, the proposed method generates fewer rules, needs less running time, and provides superior biological significance for the resultant top-ranking rules compared to prior methods [6].

Gene set enrichment analysis is already a standard practice for analyzing high-throughput gene expression data, although evaluation of enrichment strategies is still in its infancy and often ad hoc [7]. To make up for the lack of appropriate gold standards, the majority of evaluations, on the other hand, are based on selectively chosen datasets and biological reasoning about the importance of enriched gene sets. Using predetermined criteria for generalizability, gene set prioritizing, and important process identification, we provide an adaptable framework for repeatable benchmarking of enrichment approaches. This structure contains a selected collection of 75 expression datasets focusing on 42 human disorders. Each dataset has a precompiled GO/KEGG relevance rating for the related ailment, and the compendium includes microarray and RNA-seq measurements [7].

Especially in cancer research, microarrays have generated vast amounts of genetic data that may be used for transcriptase analysis [8]. The gold standard for statistical analysis is a comparison of each transcript or gene in a set of cancer samples to the corresponding gene in a set of matched control samples. Using a rule-based methodology, “association rule mining” seeks previously unseen sets of objects. Discovering these causal relationships among the transcripts is useful. Based on the weighted similarity scores and an association rule-based learning system, we provide two novel rule-based similarity measures: the weighted rank-based Jaccard and Cosine measures. In conclusion, we provide a unique computational methodology for identifying dense gene co-expression modules (ConGEM s). The created condensed marker set, which includes both simple and complicated markers, is determined by the condensed gene sets that match the preceding or subsequent rules of the resulting modules. Our research showed that these indicators were supported by gene ontology and pathway annotations from databases like KEGG (Kyoto Encyclopedia of Genes and Genomes), as well as the published literature. The expression levels of these genes were predicted to be variable using an empirical Bayes test. Then, a recently developed algorithm known as RANWAR was used to the problem of identifying association rules. We used clustering to compute the integrated similarity scores of these rule-based similarity measures between each rule pair, which allowed us to identify the co-expressed rule modules. We validated our method using gene expression data for lung squamous cell carcinoma and genome methylation data for uterine cervical carcinogenesis. Our proposed method for identifying gene modules produced much better results than previous methods. In conclusion, the rule-based method we proposed is useful for examining biomarker modules using transcriptase data. Go to the whole text [8]

Xu et al. [9] despite the growing popularity of traditional Chinese medicine (TCM) as an adjunctive treatment for (GC) gastric cancer, large-scale data supporting the use of deep learning in healthcare settings is currently in short supply. Herbal prescriptions based on clinical studies conducted during the previous three decades [1990–2020] were retrieved from a public database using the search phrases "Gastric cancer or gastric malignancy" and "Traditional Chinese Medicine." Extracts from qualifying studies were analyzed for their prescribing patterns of herbs using association rules mining (ARM). We used computational prediction and deep machine learning to look at potential GC treatments that might be useful for the population at large. Network pharmacology was used to learn about the mechanism of action of the recommended medication, and in vivo and in vitro testing confirmed the findings [9].

Long et al. [10] among all cancers, lung cancer has the highest fatality rate. This research used bioinformatics analysis to look for genes that were differentially expressed (DEGs) and pathways that were enriched in lung cancer patients in the hopes of finding new diagnostic and therapeutic targets for this disease. Valid microarray data from 31 pairs of matched lung cancer and normal tissue samples were retrieved from the Gene Expression Omnibus database (GSE19804). Using significance analysis of the gene expression profile, 1,970 DEGs that were significantly enriched in biological processes were discovered between cancer tissues and normal tissues. Toll-like receptor pathway was shown to be significant among the 77 KEGG pathways linked to lung cancer that was discovered using Gene Ontology enrichment analysis using the KEGG database. After extracting 1,770 nodes and 10,667 edges from a protein-protein interaction network, we found that the 10 genes with the greatest degrees, of hub centrality and betweenness, played the most significant roles in lung cancer. Lung cancer was shown to have close ties to the 'chemokine signaling route,' the 'cell cycle,' and the 'pathways in cancer' as shown by protein-protein interaction modules. In conclusion, our knowledge of lung cancer's development is aided by the DEGs identified, particularly the hub genes, and certain genes (including advanced glycosylation end product-specific receptor and epidermal growth factor receptor) may be used as candidate target molecules in the diagnosis, monitoring, and treatment of lung cancer [10].

Jiang et al. [11] connect metabolism, redox biology, and illnesses like cancer in a process called ferroptosis, which is a kind of programmed cell death. The goal of this study was to develop a gene prediction signature for pancreatic cancer (PCa) based on Ferro ptosis by using data from the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTE) (GTEx). To create a risk score, researchers first identified 14 genes associated with ferroptosis that may have prognostic significance. We divided the patients into two categories: those with high and those with low-risk ratings. Gene Expression Omnibus (GEO) and International Cancer Genome Consortium data were used to verify its accuracy (ICGC). Patients with lower risk ratings showed better overall survival (OS) as seen by the Kaplan-Meier survival curves (P 0.0001). The ROC area for 12, 18, and 24 months was all around 0.8. Proteins involved in immune infiltration and immune checkpoint blockade were shown to have a substantial connection with the signature in an examination of the immunological state (ICB). In addition, q-rtPCR and the Human Protein Atlas were used to confirm the critical gene expression levels (HPA). The signature as a whole helps determine which individuals with PCa will survive treatment. Since the signature correlates with immunological features, it has the potential to enhance the success of customized immunotherapy [11].

It is believed that about 200 million individuals throughout the globe are infected with the Hepatitis C Virus, which is the infectious agent responsible for the most severe type of chronic liver disease [12]. There is still a lot that is unclear about the way through which HCV pathogenesis occurs. A better knowledge of the HCV process may be attained by investigating the interactions that occur between HCV and human proteins [12].

Data mining is a useful tool for discovering patterns in the enormous volumes of data that are now at our fingertips. Large volumes of data may be processed using this technology. Structures that can be readily updated will be created from this data in the future. Data may be simply processed without any difficulty. Data mining is governed by databases, and the administration of databases is used to conduct the data mining. Data mining as a concept was first introduced to computer science in 1990. Data mining includes fields like KDD (Knowledge Discovery in Databases), PA (Predictive Analytics) and DS (Data Science). Cooperative and repeated data mining is really what it is all about. This technique's primary goal is to mine large amounts of data for groupings, structures, variations, and anomalies. This approach should provide results that are accurate, long-lasting, precise, and above all, simple to comprehend. Blood clots separate serum from the rest of the blood. Unclotted blood contains both platelets and red and white blood cells, unlike plasma, which only contains red and white blood cells. When producing cheese, the watery liquid that separates from the curd is known as "whey." The name "serum" is derived from this Latin word. In terms of skincare, a serum is a simple solution that contains high-quality ingredients. They're used as a last step before saturation, just after purging. For example, anti-aging serums may be applied to the face to address specific skin issues.

There are various proteins in serum; however they are not involved in blood coagulation. Antibodies, exogenous chemicals, electrolytes, and hormones are among the proteins that make up this list. There are no serum samples available for platelets, red platelets, clotting factors, and white platelets. The study of serum is known as serology. It is possible to utilize the serum in conventional assays like blood composition. In a few cases, such as clinical studies identifying the restorative index of a medication competitor, the process to predict the creation of various particles is more advantageous. If the blood test is allowed to coagulate for long enough, a serum will be produced. Serum does not include platelets, clotting factors, or any of the other components of the platelet clot. In the subject of serology, the study of serum is emphasized. Analytical tests employ serum in the same way that blood is used. It is possible that measuring uplifting directory of pharmaceutical rival in scientific trial might benefit from an estimation of the concentration of various particles. This would be one application in which such estimation would be valuable. It is permissible for a blood test to clot in order to get serum. After centrifuging the sample to remove the coagulation and platelets, serum is obtained from the fluid supernatant. Because the antibodies that are produced as a result of a successful recovery are powerful soldiers against the microorganism, the serum of recovered patients may be used as a biopharmaceutical in the treatment of those who are suffering from the same ailment. Immunotherapy is one way of putting it.

Protein electrophoresis may also benefit from the lack of fibrinogen, which can lead to false results. Fetal bovine serum (FBS) is a rich source of growth factors and is often included in media for eukaryotic cell culture that are designed for this purpose. Supporting young, immature bacteria was traditionally done using fontal bovine serum (FBS) and the cytokine leukemia inhibitory factor, but worries about bunch-to-clump variations in FBS have led to the usage of serum substitutes. As an immediate and direct result of this, the cancer feature association rules miming (CFARM) and KEGG may be detected significantly more quickly.

## 11.3 CFARM-KEGG Architecture

Anatomical items (such as characteristics, proteins, and nanoparticles) are mapped to sub-atomic association, response and connection systems by a process known as "KEGG mapping". KEGG modules, KEGG route maps, and BRITE hierarchy. It's a set action to generate another set, not merely a progression technique. While it was first intended to develop routes that were only relevant for the study of animals, the goal has always been to generate pathways that are relevant to the study of all organisms. It's an important function for the KEGG planning set activity to increase the KEGG's knowledge and understanding in this way. As a result, it played a critical role in the integration and translation of large-scale datasets from high-throughput discoveries. Pathway mapping, bride mapping, and module mapping are all critical components of the KEGG mapping process.

Research shows that those who consume more protein are more likely to die from cancer than those who consume less protein. However, for those over 65, a modest protein intake may actually be beneficial and help prevent deterioration, according to experts. Preclinical research shows that reducing protein intake alone is insufficient to prevent cancer growth; altering the protein's structure also has an effect on the tumor. Another way to look at it is that the protein's source acts in a manner similar to that of reducing the total amount in slowing the growth of prostate and breast tumors. A hybrid approach to the selection of features might be useful for an informative collection that contains tens of thousands of different variables. After applying the channel model, which ranks the features based on the mutual data given by each feature and each class, we pick the k features that are the most relevant to each class based on the results of this ranking. According to the Correlation feature selection (CFS) Subset hypothesis, the predictive characteristics ought to have had an exclusive relationship with the aim class and the least relevance to any other indicator qualities.

If the prediction test is repeated with each perception removed, the result is called a "jackknife estimator," for short. For each forecast, the inclination and change in measurement may be calculated. It is more difficult to cross-examine information that has been related to other perceptions since forgetting a perception does not eliminate all of the associated data. Both preparation and forecasting are necessary in order to use this strategy. Based on performance criteria throughout the estimate process, which includes information preparation, feature selection, and grouping, classifier models are assessed using jack-knife cross-evaluation tests When preparing the data, The two types of lung cancer, small cell and non-small cell, need separate gene sets and it is crucial to identify them. The option of hybrid features gives you the ability to handle all of them. When it comes to choosing this feature, the principle of incremental feature selection will have to be taken into account. Models are then created that differ significantly from calculation-based prediction approaches helpful for locating lung tumors. Physicochemical and biological characteristics of proteins may be studied using these models. Class values and characteristics were found to be linked by the gain ratio criteria based on data gains and information entropy values. It is hoped that subset assessors who use incremental element selection in conjunction with this Bayesian method would be able to discern between lung cancer tumors.

Hybrid feature selection is the next step in this procedure as shown in Fig. 11.1. In order to get them to commit to rating the examples under the many objective categories in this selection, features like feature location were critical. Features selection algorithms will focus on placing features in accordance with their chosen value, leaving it up to clients to decide on a limiting essential. Correlation Subset Evaluators were then used as a follow-up to the positioning feature determination algorithms in order to determine a small or insignificant set of features that were suitable for the class but not necessarily connected to each other. There were two methods used to determine the best features: ranking and subset evaluation. This was dubbed the hybrid feature selection method since it incorporated both methods. Proteins that are essential to the cell's genomic integrity and also serve as a useful hotspot for drug design can be identified through the precise classification of small cell lung cancer and

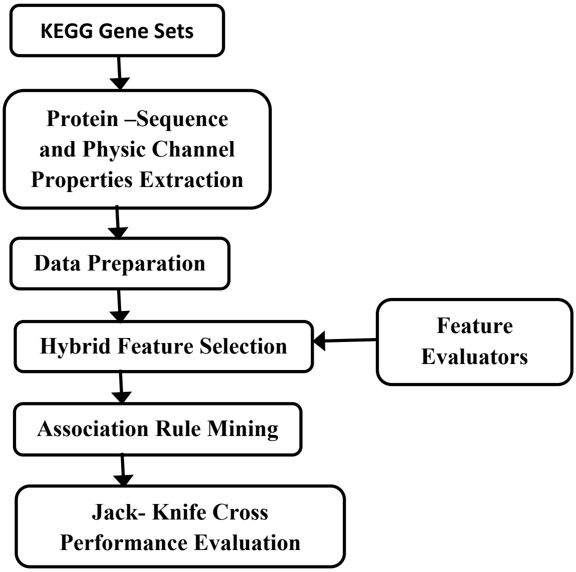

KEGG Gene Sets
Protein –Sequence and Physic Channel Properties Extraction
Data Preparation
Hybrid Feature Selection
Feature Evaluators
Association Rule Mining
Jack- Knife Cross Performance Evaluation

**Fig. 11.1** CFARM-KEGG for protein sequence-based cancer categorization

non-small cell lung cancer oncogenic genes, which are based on auxiliary properties and physical properties, respectively.

$$P\left(\frac{x}{y}\right) = p(x/y)p(y) \Big/ \int p(x/y)p(y)dy$$

## 11.4 Results

This approach uses a negligible and perfect arrangement of characteristics to identify the different types of lung tumors so that they may be used in analytical practice and the construction of medicine. Then we sorted the features using the Gain Ratio criteria, the Information Gain basis, and the Symmetric Uncertainty criterion. Then, we used the Correlation Feature Subset evaluator with a hunt end limit of 5 and the Best First Search method to find the smallest subset of features that had a robust correlation to the target class while having the least relationship to each other. The resulting subset of element features included 39 individual characteristics. Lung tumor classes may be classified in this manner for utilization in analytical practice and medicine structure by arranging characteristics in a negligible and optimal manner. This was accomplished via the use of Incremental Feature Selection (IFS), which ultimately led to us acquiring the feature set that was the most perfect out of a total of 35 feature sets (which included a subset of Gain Ratio and CFS). Then, using the

**Table 11.1** Accuracy in classification: several classes

| Selecting features in a hybrid manner | Feature | To classify with an algorithm | Effectiveness of the JKCV (%) |
|---|---|---|---|
| Profit Ratio + CFS Subset | 35 | CFARM-KEGG | 89.3 |
| Subset of symmetric uncertainty + the CFS | 31 | J48 | 87.2 |
| Theoretical advancement + CFS subgroup | 33 | The use of Bayesian networks | 89.4 |

Gain Ratio meter, the Information Gain foundation, and the Symmetric Uncertainty, we ordered the characteristics. Then, we used the Best First Search technique and the Correlation Feature Subset evaluator with a search end limit of 5 to identify the fewest features with the highest correlation to the goal class and the fewest connections to one another. As a direct consequence, a subset with 39 unique feature items was created. The results of using incremental feature selection are shown in Table 11.1. (IFS). Accuracy for the whole set of 35 characteristics (including the subset of Gain Ratio + CFS) was determined to be 89.3% using the Jack-Knife cross-check (JKCV). Moreover, 33 features are included in the results, leading to an 86.4% accuracy rate when the CFS subset was used for data collection. Using Jack-Knife cross-evaluation (JKCV) with 89.3% accuracy and similarly, information gain with a CFS subset of 33 features and 86.4% accuracy as shown in Table 11.1; all of these evaluation methods are accurate.

Here, 7 grouping models that may be used to sort cancer tumors into categories depending on the attributes they contain. When doing hybrid feature selection on the dataset, we can see the results of our work in developing clustering methods. Table 11.2 shows the classifications that may be used to group evaluation findings. There was no new insight to be gained from this study's results, since their presenting accuracy was quite poor, as seen by the categorized ones. Conversations about the data and its consequences are introduced in this section. A comparison was made between seven different clustering algorithms, including the Expectation–Maximization (EM) Algorithm, the COBWEB Clustering Framework, the Hierarchical Clustering Framework (HC), K-Means Clustering, the Nearest First Clustering Framework (NFCF), and the Density-Based Clustering Framework (DBCF). Cobweb automatically allocated the number of clusters; however the client may specify the optimal number of groups in all other computations. When considering the large number of features for bunching, a few computations performed better than others, but the presentation of the hybrid feature selection datasets degraded with time. Figure 11.2 shows the amount of characteristics that may be selected using the hybrid feature selection approaches. Using CFS, the gain ratio has a greater number of characteristics than the other approaches . According to Fig. 11.3, comparison of the hybrid feature selection methods based on the proportion of correct jack-knife assessments. It demonstrates that the gain ratio using CFS is more accurate than other methods. Areas in which performance evaluation methodologies and limits are discussed are included. For pre- and post-Hybrid feature selection (HFS), the connection between different parameters of cluster evaluation accuracy (CEA) in % may be shown in Fig. 11.4.

**Table 11.2** Assessing Clustering Classes using the CFARM-KEGG System

| S. No | Clustering Model | Classes to Cluster Evaluation Accuracy (%) | |
|---|---|---|---|
| | | Selection for Pre-Hybrid characteristics (%) | Post-hybrid characteristic Selection (%) |
| 1 | K-Means | 53.0973 | 51.3274 |
| 2 | It uses an algorithm known as E-M | 51.8942 | 49.8542 |
| 3 | Clustering According to Density | 52.1867 | 50.7842 |
| 4 | The Filtered Clustering Method | 54.706 | 50.1782 |
| 5 | COB WEB | 3.7945 | 5.2605 |
| 6 | Distance-First Cluster Analysis | 47.8521 | 45.9527 |
| 7 | The Clustering Hierarchy | 50.8963% | 50.6534% |

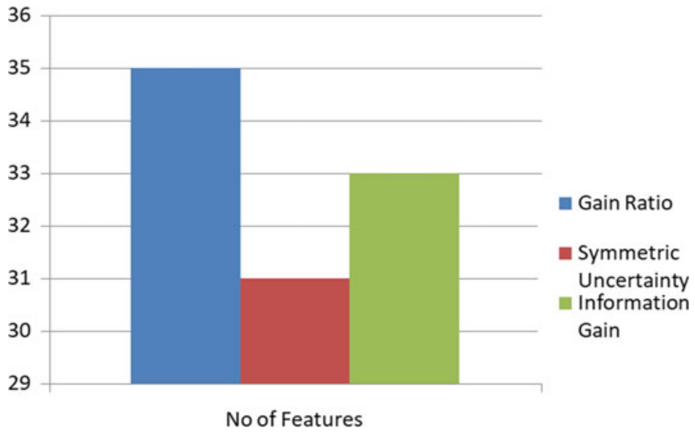


**Fig. 11.2** Scale of CFARM-KEGG characteristics

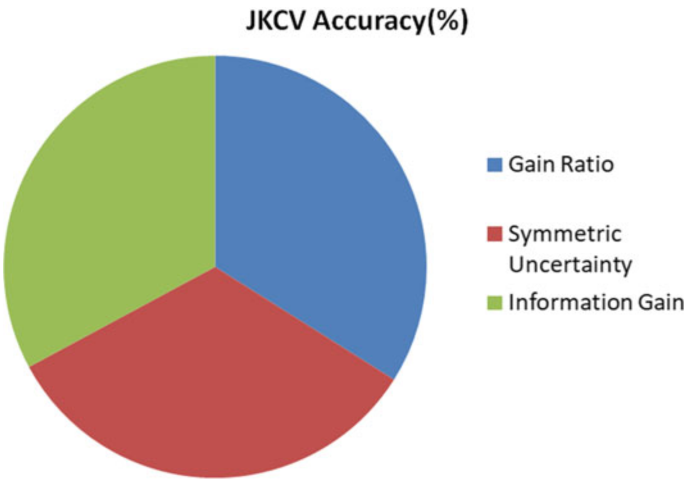


**Fig. 11.3** Precision of different CFAME-KEGG Methods in terms of the JKCV

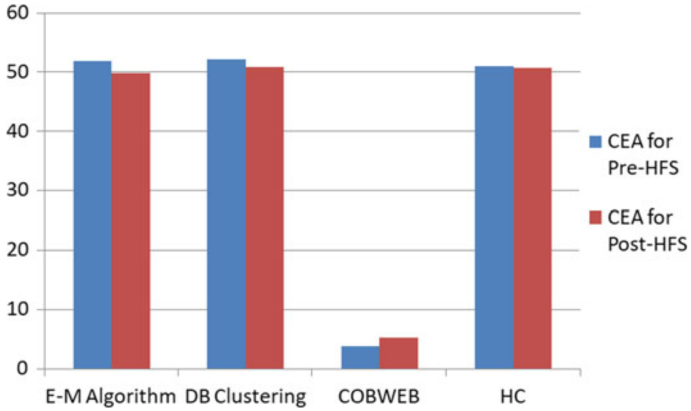


**Fig. 11.4** Analysis of CEA's Efficiency in CFAME-KEGG Methods

### Proposed Algorithm

**Algorithm: 1**

```
M data is then subjected to the following categorization tests: $T_a \ldots T_J$
Step 1: Zero
Step 2: Bayes net
Step 3: Naïve Bayes
Step 4: Logistic
Step 5: Simple Logistic
Step 6: Method based on many layers of perceptron
Step 7: Networks Based on the Radial Basis Function
Step 8: Sequential Minimal Optimization
Step 9: Classification via Clustering
Step 10: End of algorithm
```

## 11.5 Conclusion

Unlike malignant tumors, benign tumors do not spread or grow in size according to their presence in the body. Understanding how oncogenic tumors differ from one another is essential for both diagnosing and treating illness. Oncology is the branch of medicine that deals with diseases caused by the growth of abnormal cells. It has the potential to assault or spread too many sections of the body. The cancer was caused by physiochemical factors. The microarray-based research is the best way to define these characteristics. An effective computational approach has been adopted in this study to predict the kind of cancer tumor in the body. AI algorithms are often used to numerical datasets, which necessitates that they be scientific in nature. The Kyoto Encyclopedia of Genes and Genomes contains a wide variety of genome management databases, organic processes, illnesses, pharmaceuticals, and synthetic compounds (KEGG). Using the method of association rule mining, it is feasible to properly diagnose cancer (ARM). This rule mining made use of flexible expressions such as if/then, which were a big help in figuring out the connections between a variety of allowed public databanks and extra data repositories. In bioinformatics, it serves as a testing ground and classroom tool, as well as the memorization of knowledge pertaining to genomics, met genomics, and metabolomics. KEGG gene sets and association rule mining were used to identify several cancer-related diseases in this study.

## References

1. Schulte-Sasse, R., Budach, S., Hnisz, D., Marsico, A.: Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat. Mach. Intell. **3**(6), 513–526 (2021)
2. Li, C., Xu, J.: Feature selection with the Fisher score followed by the Maximal Clique Centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma. Sci. Rep. **9**(1), 1–11 (2019)
3. Mallik, S., Zhao, Z.: Towards integrated oncogenic marker recognition through mutual information-based statistically significant feature extraction: an association rule mining based study on cancer expression and methylation profiles. Quant. Biol. **5**(4), 302–327 (2017)
4. Ong, H.F., Mustapha, N., Hamdan, H., Rosli, R., Mustapha, A.: Informative top-k class associative rule for cancer biomarker discovery on microarray data. Expert Syst. Appl. **146**, 113169 (2020)
5. Yang, L., Ainali, C., Kittas, A., Nestle, F.O., Papageorgiou, L.G., Tsoka, S.: Pathway-level disease data mining through hyper-box principles. Math. Biosci. **260**, 25–34 (2015)
6. Mallik, S., Bhadra, T., Mukherji, A.: DTFP-growth: dynamic threshold-based FP-growth rule mining algorithm through integrating gene expression, methylation, and protein-protein interaction profiles. IEEE Trans. Nanobiosci. **17**(2), 117–125 (2018)
7. Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Waldron, L.: Toward a gold standard for benchmarking gene set enrichment analysis. Brief.S Bioinform. **22**(1), 545–556 (2021)
8. Mallik, S., Zhao, Z.: ConGEMs: Condensed gene co-expression module discovery through rule-based clustering and its application to carcinogenesis. Genes **9**(1), 7 (2017)
9. Xu, X., Chen, Y., Zhang, X., Zhang, R., Chen, X., Liu, S., Sun, Q.: Modular characteristics and the mechanism of Chinese medicine's treatment of gastric cancer: a data mining and pharmacology-based identification. Ann. Transl. Med. **9**(24) (2021)
10. Long, T., Liu, Z., Zhou, X., Yu, S., Tian, H., Bao, Y.: Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. Mol. Med. Rep. **19**(3), 2029–2040 (2019)
11. Jiang, P., Yang, F., Zou, C., Bao, T., Wu, M., Yang, D., Bu, S.: The construction and analysis of a ferroptosis-related gene prognostic signature for pancreatic cancer. Aging (Albany NY) **13**(7), 10396 (2021)
12. Indhumathy, M., Nabhan, A.R., Arumugam, S.: A weighted association rule mining method for predicting HCV-human protein interactions. Curr. Bioinform. **13**(1), 73–84 (2018)

# Chapter 12
# Solution Architecting on Remote Medical Monitoring with AWS Cloud and IoT

**Parul Dubey and Arvind Kumar Tiwari**

## 12.1 Introduction

A technical definition of cloud computing in healthcare is “the storage, administration, and processing of patient data on distant servers that are accessible via the internet.” On a practical level, cloud computing is defined as “the storing, administration, and processing of patient data on distant servers that are accessible via the internet.” Another option is to store files on a personal computer at the company’s office, or to use a data centre with servers on-site.

It is possible to customize cloud storage, which enables hospitals and health-care providers to store enormous volumes of information on a secure network of computers that can be accessed from any location.

Cloud-based medical techniques are becoming more popular among health care practitioners in the United States as a consequence of the Electronic Medical Records (EMR) requirement, which was enacted in 2012 [1].

The ageing of the population has a significant impact on the health-care industry. While it is true that many older people suffer from long-term medical issues, such as diabetes or high blood pressure, they also need regular care. However, although this may be the case for some, it is not the case for others. As the senior population continues to rise, hospitals may find themselves unable to meet the demands of an ever-increasing patient population in the foreseeable future. A new approach to senior care is necessary in order to enhance and simplify the everyday lives of health-care

P. Dubey (✉)
Department of Computer Science Engineering, Dr. C. V. Raman University, Kota, Chhattisgarh, India
e-mail: dubeyparul29@gmail.com

Department of Artificial Intelligence, G H Raisoni College of Engineering, Nagpur, Maharashtra, India

A. K. Tiwari
Dr. C. V. Raman University, Kota, Chhattisgarh, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_12

workers, while also assisting the elderly in preserving their health and independence while reducing the demand on national resources.

Patients with limited mobility, who reside in distant areas, or who suffer from a chronic condition are the topic of this article, which analyses the use of remote patient monitoring systems. It is possible to observe delocalized patient health data in real time, which is a more accurate depiction.

## 12.2 Literature Review

When it comes to health care applications, cellular networks have grown increasingly prevalent [2]. Doctor-to-doctor contact is becoming increasingly significant in India, thanks to the ongoing growth of mobile phone networks in rural locales that are remote from metropolitan areas. [3] Mobile technology is gradually complementing the capabilities of portable devices, namely as smart phones and personal digital assistants (PDAs), which have the potential to replace technology-based alternatives while satisfying the mobility demands of patients and medical practitioners. With these advantages in mind, it is becoming more viable to design unique technologies for poor regions that may improve the distribution of services and reduce the challenges produced by health-care delivery systems [4].

Cloud computing is an intriguing technology that makes use of software, infrastructure, and the whole computer platform as a service in order to provide a service to the public. Cloud computing, with the exception of traditional web hosting companies, provides services that are based on a pay-as-you-go model. According to this approach [5], clients pay only for the resources they utilise over time, rather than for the resources they purchase in advance. Using cloud computing, patients may find it easier to locate and keep track of their medical conditions on their health records [6].

The area of patient data security inside healthcare systems, on the other hand, has made significant strides in recent years. For example, Nacha and Pattra [7] developed a cloud-based, mobile, and secure solution for their customers. WSN and cloud computing were used to develop a comprehensive healthcare system, which included the deployment of Ciphertext Policy-ABE on the basis of WSN and cloud computing. This was solved in full by developing a DRM (Data Relationship Management) service that ensures the integrity of data transfers and gives a roadmap for how any future cloud-based healthcare system may minimize stress while simultaneously enhancing security. As a result of this trend, Blockchain technology is becoming more popular for enhancing security [8]. It combines decentralized databases with encryption, a combination that has previously shown to be beneficial in other industries, such as the bitcoin industry.

On the other hand, cloud computing, as opposed to on-premises computing, is based on services rather than applications [9]. This strategy delivers virtual resources in the form of a service that can be evaluated and charged for. Because of these benefits, it is now simpler than ever before to come up with innovative ideas. Applications that are innovative and original and have the potential to make an impact in developing markets are encouraged. It is critical that the health-care system continually searches out innovative methods of providing treatment and resolving difficulties. As a result, there has been an increase in security and privacy risks. Cloud computing has the potential to be beneficial in a variety of industries, including healthcare. The usage of computer-based solutions has the potential to provide several benefits. It can reduce healthcare costs while simultaneously enhancing patient satisfaction [10].

The services in issue are provided by Amazon Web Services, which is the firm in charge of doing so (AWS). This service, which serves as an ideal example of genuine cloud computing, protects the security, integrity, and availability of client data while also offering high-quality cloud services. Because AWS offers the necessary resources for application development and deployment, an application may be created and deployed on AWS at a cheap cost. The realisation that it is impossible to provide high-quality service to all of their consumers has prompted business owners to concentrate their efforts on a certain geographic region at a time. This problem has been resolved due to Amazon Web Services, which enables companies all around the globe to provide a better customer experience [11] to their customers. The services in issue are provided by Amazon Web Services, which is the firm in charge of doing so (AWS) Because it is so realistic, it serves as an excellent representation of what cloud computing is all about. Cloud computing services, as well as data security, accuracy, and availability guarantees, are made available to customers [2]. It is has a free tier facility available also.

The Internet of Things (IoT) has a bright future in the healthcare business, and the present view is quite positive. It is also well-liked for its capacity to detect and quantify a variety of different items. Narrowband Internet of Things (IoT) is a low-power variation of the Internet of Things (IoT). It is a popular option in the healthcare business because of its low power consumption [12]. The Internet of Things (IoT) may be used in healthcare in a variety of ways. Because it has been standardized, LTE is completely compatible with the Internet of Things. As a consequence, in recent years, the Internet of Things has shown to be an excellent match for healthcare-related applications. The most serious threats to the Internet of Things arise from security measures and other system-related difficulties. In healthcare, low-power wide-area deployments have the potential to be a top choice provided the concerns and impediments that have been raised about it are addressed effectively. These are the most significant obstacles and challenges that the NBIoT regime is now confronted with at this point in time. Additionally, researchers provide some recommendations for overcoming these concerns.

Machine learning models and systems in healthcare are the topic of another study [13], which covers definitions, intricacies, difficulties, and requirements for interpretable and explainable models and systems. The use of interpretable machine learning models in healthcare has a variety of applications, and researchers investigate the most effective methods of implementing them. The process of identifying the most appropriate interpretable machine learning approach for a given healthcare challenge is also discussed, as is the landscape of recent advancements in the field of model interpretability in healthcare.

Machine learning methods applied to electronic health records (EHR) may allow for the improvement of patient risk score systems, the prediction of sickness onset, and the streamlining of hospital operations. Although EHR-derived data-driven statistical models are still in their infancy, the variety and richness they bring open up a whole new area of inquiry. In another article [14], they present an overview of clinical applications of machine learning and explain the advantages that machine learning has over more traditional ways of data analysis. Machine learning research and practice are confronted with a variety of challenges, which they described in this paper. They also discussed how machine learning will have a significant impact on the future of healthcare and healthcare delivery in the next years.

## 12.3 Internet of Things (IoT)

The Internet of Things (IoT) is a wireless communication and interaction platform that allows electronic-equipped devices to communicate and interact with the rest of the world via the use of the Internet (IoT). The Internet of Things (IoT) is expected to have a significant influence on the way people live and work in the near future within a few years of its introduction. The Internet of Things (IoT) has had a substantial influence on a variety of fields, including medicine, energy, gene therapy, agriculture, smart cities, and smart homes, to name a few examples [15]. More than 9 billion "Things" are now estimated to be present on the internet, according to current estimates (physical commodities). Within a very short period of time, this number is expected to increase to a mind-boggling 20 billion individuals.

In order to form the Industrial Internet of Things (IIoT), devices, apps, and technical systems must all have the same fundamental components. The architecture of the Internet of Things is well shown in Fig. 12.1.

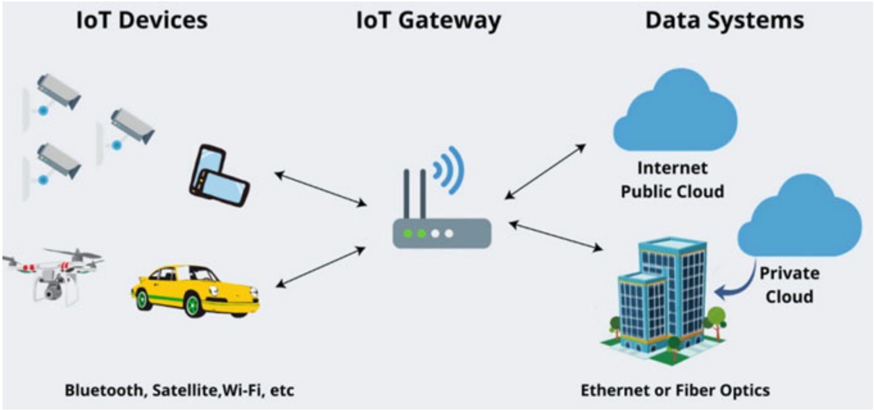


**Fig. 12.1** The Internet of Things (IoT) architecture

Following the increased bandwidth and falling hardware prices brought about by the Internet of Things, various industries are seeing rapid transformation. These are just a few of the industries where the Internet of Things is finding new applications these days, including healthcare, construction, government and insurance. Massive corporations and financial institutions are all increasing their expenditures in information technology as a result of these advancements.

- Gadgets and sensors that are intelligent

The initial layer of a protocol stack is comprised of the connectivity between devices and sensors. Smart sensors are continually collecting and transmitting data from their immediate surrounds, which they refer to as their immediate environment, according to experts. The most sophisticated semiconductor technology available today makes it feasible to manufacture compact smart sensors that may be used in a broad variety of different applications.

A variety of sensors and devices are employed to obtain information about their local environment, ranging from a simple temperature measurement to a comprehensive live video feed. The phrase "sensor/device" may be used to describe a sensor that is a component of a larger gadget that performs functions other than just sensing the surrounding environment. With sensors (such as a camera and an accelerometer), phones are more than just sensors, since they can perform a broad range of functions with the help of their sensors.

In order to relocate data, we need to set up a cloud computing service. Cellular, satellite, WiFi, Bluetooth, low-power wide area networks (LPWAN), a gateway, or a router are examples of techniques for connecting sensors to the cloud, as are cellular and satellite networks. In order to make the best selection, it's necessary to consider elements such as power consumption as well as range and bandwidth. There are a variety of methods to connect to the Internet of Things, but they are all geared at delivering data to a cloud-based storage service.

- Gateway

The Internet of Things Gateway is responsible for bidirectional data flow across a wide range of networks and protocols. Network gateways are used to interpret network protocols in order to assure the compatibility of linked devices and sensors. Before transmitting data to the next level, gateways may do local pre-processing on data acquired from hundreds of sensors before transferring it to the next level. In certain circumstances, the compatibility of the TCP/IP protocol may necessitate this. Higher-order encryption methods, in conjunction with Internet of Things gateways, increase the security of networks and the data they transfer, according to the Federal Bureau of Investigation. Intermediary protection, which exists between the cloud and the devices, protects against harmful assaults and unauthorized access.

- Cloud

It is expected that the Internet of Things will create massive volumes of data, and that managing this data would be a significant difficulty. The Internet of Things makes it possible to gather and analyse data in real time, as well as to store and manage data in the cloud. Customers have the ability to access this information from any place, enabling companies and services to make critical decisions when the situation calls for them to do so. The Internet of Things (IoT) platform enables high-performance computer networks to analyze billions of devices and govern traffic in a short amount of time utilizing a single platform. The popularity of this product is increasing. DBMSs (distributed database management systems) are widely used in the Internet of Things cloud environment (DBMSs). Predictive analytics and data storage are made possible by the interconnection of billions of devices in a cloud-based infrastructure. Companies use this data to improve the quality of their goods and services in the future, as well as to develop a new business model based on the information they have gathered. In order to avoid problems in the future, companies use this data to improve the quality of their goods and services and to develop a new business model based on the information they have gathered.

## 12.4 Cloud Healthcare Management

In this article, we'll look at three different cloud computing models. Each model serves a particular purpose in the stack of computing.

### *12.4.1 Infrastructure as a Service (IaaS)*

A shorthand way to describe Infrastructure as a Service is that it is the foundation of cloud computing. Customers of cloud computing services have access to network resources, computers (virtual or dedicated hardware), and computer storage space for a monthly or annual fee. It's the closest thing we have now in terms of computing resources. When you choose with an IaaS provider, you have the most flexibility and control over your IT infrastructure (IT).

### *12.4.2 Platform as a Service (PaaS)*

Rather than focusing on infrastructure management, firms may focus on developing, deploying, and administering their applications (typically hardware and operating systems). When businesses don't have to worry about collecting the resources they need, planning for capacity, or updating and patching your application as often, simply have more time to concentrate on other things.

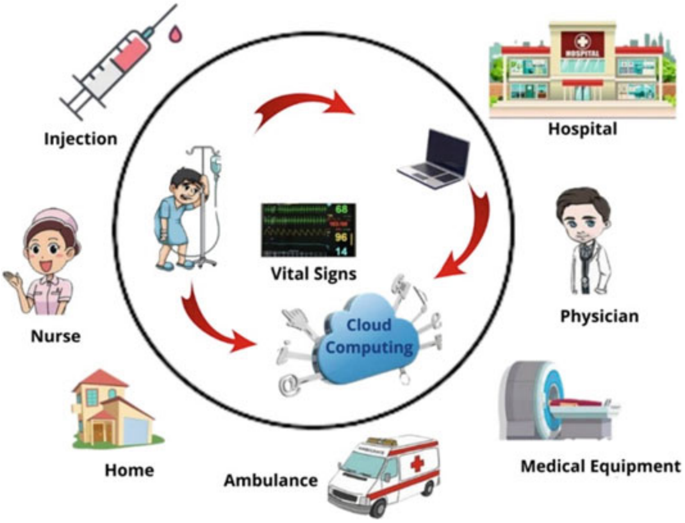


**Fig. 12.2** Cloud healthcare management service

### 12.4.3 Software as a Service (SaaS)

In certain cases, the service provider oversees and maintains the completed product for customers (SaaS). SaaS refers to the software that users interact with. Instead of worrying about the maintenance or administration of a cloud-based SaaS solution, consider how to use it. To respond to emails through the internet, a SaaS provider that manages the infrastructure and operating systems may be employed. Using web-based email services, one may safeguard messages without having to deal with the email system's design or operating system.

Figure 12.2 depicts the Cloud healthcare management service. Using this graphic, people can see how cloud computing takes care of all the little details. This may be gained at a lower cost, resulting in improved service and patient care as well as increased security and availability.

### 12.4.4 Deployment Models

Let's have a look at some of the different Cloud Computing Deployment Models.

### 12.4.5 Cloud

All code and data are stored and executed on the cloud when creating a cloud-based application. Either a cloud-native application was created from the ground up or an on-premises application was moved to the cloud. With the help of these services, it is feasible to build cloud-based applications since they offer abstraction from the administrative needs of the cloud's infrastructure [16].

### 12.4.6 Hybrid

Using a hybrid deployment technique, it is feasible to integrate cloud-based resources with those that are not kept in the cloud. The type of mixed deployment described above [16] is the most common. It's possible to build and improve an on-premises architecture, and cloud resources can be connected with internal systems.

### 12.4.7 On-Premises

When virtualization and resource management technologies are leveraged to deploy resources on-premises, they are referred to as "private clouds" in various circles [16]. On-premises installations remain popular despite the fact that they lack many of the benefits of cloud technology. This is owing to the dedicated resources that they can supply. Application management and virtualization technologies are utilized in an attempt to optimize resource utilization in this deployment strategy, which is, in the majority of situations, the same as that of traditional information technology infrastructure. Therefore, making advantage of cloud computing might be partial or total (Table 12.1).

This is consumers' option and comfort to decide on which model to pick.

**Table 12.1** Comparison of traditional and cloud based approaches

| Dimension | Cloud based | Traditional approach |
|---|---|---|
| Capacity | Unlimited | Limited |
| Containers | Cheap Rented Storage | Local Storage |
| Availability | 24/7 over internet | Limited |
| Synergy | Real-time | Not Real-time |
| Expenditure | Pay peruse | Upfront cost and maintenance |
| Scalability | No limits | Limited |
| Accessibility | Anywhere, anytime | Limited |

## 12.5 AWS

AWS stands for Amazon Web Services, is a market lead for Cloud service providers at present. Today, most businesses have access to streaming data from a number of sources, including website click stream data and telemetry from Internet of Things (IoT) devices, among other things. The ability to manage streaming data in real time has always been a challenge so long as there hasn't been any on-the-fly processing. Despite the fact that data insights might be received days or even weeks after the information has been gathered, they can be deemed meaningless if they are discovered at the wrong moment.

Streaming technologies, which can handle millions to tens of thousands of events per hour, enable real-time data processing and evaluation. Because of this, real-time data is being integrated into company processes in an attempt to better serve customers, improve the customer experience, increase productivity, and generate new ideas.

For real-time patient health monitoring, we have suggested an AWS-based architecture based on the above diagram. Figure 12.3 shows the architecture in details.

AWS IoT provides cloud services in order to connect the Internet of Things devices to other devices and AWS cloud services. One may link the IoT devices to AWS-based apps by using the device software provided by AWS IoT. If the devices are capable of communicating with AWS IoT, AWS IoT can connect them to AWS's cloud services.

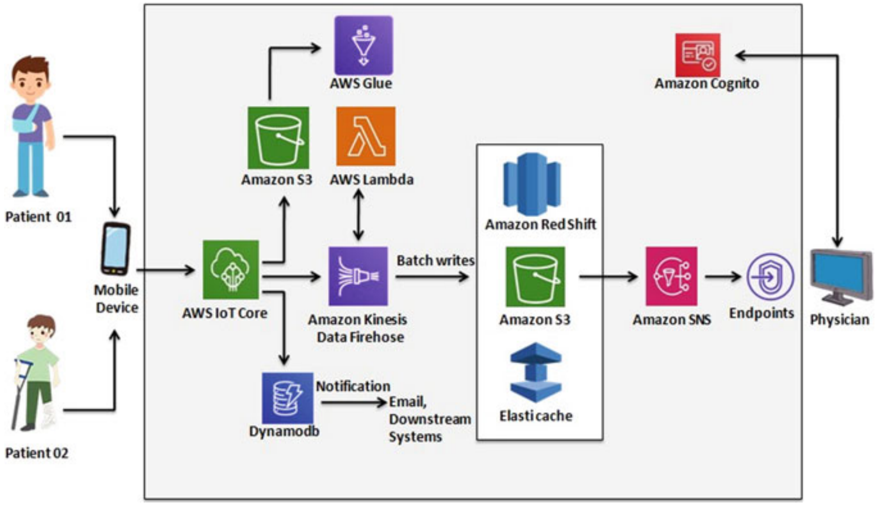


Fig. 12.3 AWS based architecture for health monitoring

- AWS IoT-Amazon Web Services (AWS) Internet of Things (IoT) allows customers to construct their solution using a range of cutting-edge technologies from a single source.The MQTT and MQTT over WSS protocols are supported by the AWS IoT Core message broker, which enables for the publication and subscribing of messages [17]. It is also compatible with devices and clients that broadcast messages using the HTTPS protocol, which is a secure version of HTTP.

It is possible to connect and control wireless LoRaWAN devices using the Internet of Things Core for LoRaWAN (Low Power Long Range Wide Area Network). It is possible to bypass the burden of setting up and maintaining your own LoRaWAN Network Server on-premises by using AWS IoT Core for LoRaWAN services (LNS).

- Amazon S3-Storage as a service (S3) is a kind of object storage that offers industry-leading scalability, as well as data availability and security. Among the many applications in which Amazon S3 may be used to store and preserve any amount of data are data lakes, websites, mobile apps, backup and restore, archiving, business applications, IoT devices, big data analytics, and other types of big data storage [18]. S3 provides administration solutions to assist clients in optimization, managing, and arranging access to their data in line with their specific commercial, organizational, and regulatory requirements.

Amazon S3 is a cloud-based service that allows users to store data in the form of objects in buckets in the Amazon data centre. An object is comprised of files, as well as any metadata that distinguishes them from one another. A bucket is a container that is used to store items.

First and foremost, one must create a bucket in Amazon S3 with a bucket name and an AWS Region associated with it. The data is subsequently uploaded to that bucket as objects in Amazon S3, which is a storage service. A key is a unique identifier that is assigned to each item in the bucket .S3 provides a diverse selection of choices that may be customized to meet our specific requirements. S3 Versioning allows users to save several versions of an item in the same bucket, which makes it possible to recover items that have been accidentally deleted or overwritten.

Individuals with access permissions to private buckets and their contents will only be able to see the contents of those buckets. Bucket policies, AWS IAM policies, access control lists (ACLs), and S3 Access Points may all be used to govern who has access to your S3 buckets and resources.

- AWS Glue-The discovery, preparation, and aggregation of data for analytics, machine learning, and software are all made easier with AWS Glue, a serverless interface solution provided by Amazon Web Services. With AWS Glue, you don't have to wait months to begin analyzing the data since it comes pre-loaded with all of the tools business need to get started right now. Data integration is the process of bringing together different types of data in analytics, machine learning, and application development. The data identification and extraction from multiple sources, data enrichment, cleaning and normalization as well as merging and loading and organization in databases, data warehouses or lakes are just a few of the operations that must be completed to achieve success. All of this is a part of the process of integrating large amounts of data. Frequently, these duties are carried out by a number of distinct persons, each of whom employs a number of different tools [19].

AWS Glue's visual and code-based interfaces make it easier to integrate data from several sources. Users will be able to find data more rapidly as a consequence of the AWS Glue Data Catalog. AWS Glue Studio simplifies the development, execution, and monitoring of ETL (extract, transform, and load) operations for data engineers and ETL developers. Without having to write a single line of code, AWS Glue DataBrew lets data analysts to enrich, clean, and standardize data visually. Application developers may mix and replicate data from several data stores using the standard Structured Query Language when utilizing AWS Glue Elastic Views, which is provided by AWS Glue.

- Kinesis Data Firehose—It is a live ETL solution that combines with the Kinesis Data Firehose platform. Using this strategy, users may store and analyze online streaming data in the fastest and most convenient manner possible. So streaming data can be ingested, converted, and evaluated in real time using your existing advanced analytics and dashboards in AWS S3, AWS Redshift, Amazon OpenSearch Service, Splunk, and other cloud-based services. Solution that is completely controlled and that automatically changes to match your data traffic without the need for any continuing administration. Additionally, data may be compressed and encrypted before being loaded, which reduces the amount of storage space needed at the end destination while also increasing security and reliability.

Among the services supported by Kinesis Data Firehose are S3, Amazon Redshift, Amazon OpenSearch, and Splunk. Kinesis Data Firehose handles all of the infrastructure, data storage, networking, and configuration requirements for collecting and loading your data. Maintaining hardware and software, as well as provisioning them, is done automatically, so we don't have to be concerned about anything. Aside from that, Kinesis Data Firehose may be scaled without the requirement for any involvement from the developer or the incurring of additional charges. Kinesis Data Firehose is used to duplicate data across three locations in an AWS Region in a synchronous manner, assuring high availability and long-term durability as the data travels to its ultimate destinations.

- Lambda-It is possible to run the code using Amazon Web Services (AWS) Lambda without needing to set up or maintain a server instance yourself. Only the time the code is running is charged; otherwise, there is no charge. With Lambda, it is possible to run code for any kind of application or backend service without the need for further administration. All we have to do in order for that code to be executed and scaled is submit it to Lambda for execution and scaling. Depending on your needs, AWS services may activate your code, or you may call it directly from a web or mobile application.

Serverless computing allows us to develop and run applications and services without needing to worry about maintaining a server infrastructure. Serverless computing allows administrators to outsource server management to Amazon Web Services (AWS). AWS Lambda is at the core of serverless computing, enabling you to run your code without the need to maintain or provision any servers.

- DynamoDB-It is a NoSQL database, provides support for Amazon Key-value and document data formats. Petabytes of data may be stored in DynamoDB, and it can handle millions of read/write requests per second, making it an excellent choice for serverless applications of the modern day. The DynamoDB database system is a better match for high-performance, internet-scale applications than traditional relational databases.

DynamoDB is a key-value and document database that has the capacity to accommodate tables of almost any size due to its horizontal scalability. With this setup, there are more than 10 trillion requests each day, with peaks of more than 20 million searches per second, and over a petabyte of storage is used each second.

- SNS-It is possible to send messages from one app to another and from one app to a person using Amazon Simple Notification Service (Amazon SNS), which is a managed service [20]. Using the A2A pub/sub capabilities, distributed systems, microservices, and serverless applications that are triggered by events may interact with one another in real time. SNS topics enable message publishers to send their messages to a diverse range of subscriber systems, including Amazon SQS queues and AWS Lambda functions, as well as HTTPS endpoints and the Amazon Kinesis Data Firehose. As a consequence of the A2P function, users will be able to send messages to a large number of individuals at the same time through SMS and push notifications on smartphones.

Cognito-This service allows users to quickly and easily incorporate user registration and authentication into existing mobile and web-based apps. Additional to this, Amazon Cognito gives temporary security credentials for accessing the app's backend resources in AWS or any service behind the AWS CloudTrail. Amazon Cognito allows us to leverage external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, and Amazon), or even your own identity provider to authenticate your users [21]. Additionally, by using Amazon Cognito, it is able to sync user data over several devices, ensuring that their app experience remains constant regardless of whether they switch devices or upgrade to a new device. Because device is connected to the internet, the app may be able to store data locally, allowing your applications to continue to function even if the user's device is removed from the internet. Instead of worrying about how to build, secure, and scale a solution to handle user management, authentication, and sync across platforms and devices, you can utilise Amazon Cognito to focus on providing exceptional app experiences to your customers instead.

## 12.6 Conclusion

In this chapter, cloud computing and healthcare are inextricably linked with one another. According to the proposed design, Amazon Web Services (AWS) is used to link patients with doctors online, as well as to boost productivity and minimise response time in this industry. The suggested system collects data via the use of an IoTcore. The rest is taken care of by AWS, which offers a wide range of tools and services such as DynamoDB, SNS, S3, glue for analytics, and so on and so forth.

## References

1. Dhilawala, A.: 9 Key Benefits of Cloud Computing In Healthcare-Galen Data. Galen Data (2019). https://www.galendata.com/9-benefits-cloud-computing-healthcare/
2. Blake, H.: Innovation in practice: mobile phone technology in patient care. Br. J. Community Nurs. **13**(4): 160, 2–5 (2008)
3. Bali, S., Singh, A.1.: Mobile phone consultation for community health care in rural north India. J. Ielemedicine Ielecare **13**(8), 421–424 (2007)
4. Perera, I.: J. Health Inform. Ctries. **3**(2) (2009)
5. Chauhan, R., Kumar, A.: Cloud computing for improved healthcare: Techniques, potential and challenges. In: 2013 E-Health and Bioengineering Conference (EHB), pp. 1–4. IEEE (2013)
6. Grogan, 1.: EHRs and information availability: are you at risk? Health Manag. Technol. **27**(5), 16 (2006)
7. Chondamrongkul, N., Chondamrongkul, P.: Secure mobile cloud architecturefor healthcare application. Int. J. Futur. Comput. Commun. **6**(3) (2017). From http://www.ijfcc.org/vol6/493-F062.pdf
8. Azhar, M., Laxman, M.: Secured health monitoring system in mobilecloud computing. Int. J. Comput. Trends Technol. (IJCTT) **13**(3) (2014)
9. Buyya, R., Yeo, C.S., Venugopal, S.: Market-oriented Cloud computing: vision, hype, and reality for delivering IT services as computing utilities. In: Proceedings of IEEE/ACM Grid Conference, pp. 50–57 (2008)
10. Muir, E.: Challenges of cloud computing in healthcare integration (2011). Accessed from http://www.zdnet.com/news/challenges-of-cloudcomputing-in-healthcare-integration/6266971. Accessed 23 July 2013
11. Bouslama, A., Laaziz, Y., Tali, A., Eddabbah, M.: AWS and IoT for real-time remote medical monitoring. Int. J. Intell. Enterp **6**(2–4), 369–381 (2019)
12. Anand, S., Routray, S.K.: Issues and challenges in healthcare narrowband IoT. In: 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 486–489. IEEE (2017)
13. Ahmad, M.A., Eckert, C., Teredesai, A.: Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 559–560 (2018).
14. Callahan, A., Shah, N.H.: Machine learning in healthcare. In: Key Advances in Clinical Informatics, pp. 279–291. Academic (2017)
15. Introduction to Internet of Things (IoT)|Set 1—GeeksforGeeks (2018). GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-internet-of-things-iot-set-1/
16. Types of Cloud Computing. (n.d.). Amazon Web ServicesInc. https://aws.amazon.com/types-of-cloud-computing/
17. What Is AWS IoT?—AWS IoT Core. (n.d.). What is AWS IoT?—AWS IoT Core. https://docs.aws.amazon.com/iot/latest/developerguide/what-is-aws-iot.html

18. What Is Amazon S3?—Amazon Simple Storage Service. (n.d.). What is Amazon S3?—Amazon Simple Storage Service. https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
19. AWS Glue|Serverless Data Integration Service | Amazon Web Services. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/glue/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc
20. Amazon Simple Notification Service (SNS)|Messaging Service | AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/sns/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc
21. FAQs|Amazon Cognito|Amazon Web Services (AWS). (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/cognito/faqs/
22. Traoré, M., Yamamoto, S.: Healthcare cloud ecosystem risk analysis and modeling: a fair approach—a case study of Arterys™ on AWS. In: 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 841–844. IEEE (2018)
23. Bracci, F., Corradi, A., Foschini, L.: Database security management for healthcare SaaS in the Amazon AWS Cloud. In: 2012 IEEE Symposium on Computers and Communications (ISCC), pp. 000812–000819. IEEE (2012)
24. Wong, A., Benedict, N.J., Lohr, B.R., Pizon, A.F., Kane-Gill, S.L.: Management of benzodiazepine-resistant alcohol withdrawal across a healthcare system: benzodiazepine dose-escalation with or without propofol. Drug Alcohol Dependence **154**, 296–299 (2015)
25. Doukas, C., Pliakas, T., Maglogiannis, I.: Mobile healthcare information management utilizing Cloud Computing and Android OS. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 1037–1040. IEEE (2010)
26. Healthcare Solutions|Healthcare & Life Sciences | AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/health/healthcare/solutions/
27. Sha M.M., Rahamathulla, M.P.: Cloud-based Healthcare data management Framework. KSII Trans. Internet Inf. Syst. (TIIS) **14**(3), 1014–1025 (2020)
28. Singh, I., Kumar, D., Khatri, S.K.: Improving the efficiency of e-healthcare system based on cloud. In: 2019 Amity International Conference on Artificial Intelligence (AICAI), pp. 930–933. IEEE (2019)
29. Modi, K.J., Kapadia, N.: Securing healthcare information over cloud using hybrid approach. In Progress in Advanced Computing and Intelligent Engineering, pp. 63–74. Springer, Singapore (2019)
30. Mitropoulos, S., Veletsos, A.: A categorization of cloud-based services and their security analysis in the healthcare sector. In: 2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), pp. 1–8. IEEE (2020)
31. Kashani, M.H., Madanipour, M., Nikravan, M., Asghari, P., Mahdipour, E.: A systematic review of IoT in healthcare: applications, techniques, and trends. J. Netw. Comput. Appl. **192**, 103164 (2021)
32. Chacko, A., Hayajneh, T.: Security and privacy issues with IoT in healthcare. EAI Endorsed Trans. Pervasive Health Technol. **4**(14) (2018)
33. Bharadwaj, H.K., Agarwal, A., Chamola, V., Lakkaniga, N.R., Hassija, V., Guizani, M., Sikdar, B.: A review on the role of machine learning in enabling IoT based healthcare applications. IEEE Access **9**, 38859–38890 (2021)
34. Elhoseny, M., Ramírez-González, G., Abu-Elnasr, O.M., Shawkat, S.A., Arunkumar, N., Farouk, A.: Secure medical data transmission model for IoT-based healthcare systems. IEEE Access **6**, 20596–20608 (2018)
35. Li, J., Cai, J., Khan, F., Rehman, A.U., Balasubramaniam, V., Sun, J., Venu, P.: A secured framework for sdn-based edge computing in IOT-enabled healthcare system. IEEE Access **8**, 135479–135490 (2020)
36. Mahmud, R., Koch, F.L., Buyya, R.: Cloud-fog interoperability in IoT-enabled healthcare solutions. In Proceedings of the 19th International Conference on Distributed Computing and Networking. pp. 1–10 (2018)

# Chapter 13
# A Domain Oriented Framework for Prediction of Diabetes Disease and Classification of Diet Using Machine Learning Techniques

**Salliah Shafi Bhat and Gufran Ahmad Ansari**

## 13.1 Introduction

Diabetes is a chronic disease. If this disease is untreated for a long-term then it can cause damage to the kidney, heart, eyes, nerves and blood vessels due to high blood sugar levels over age. Due to the shortage of insulin produced by the pancreas diabetics require extra insulin therapy to maintain their blood glucose level whereas too much insulin reduces blood glucose levels dangerously (hypoglycemia) [1]. The most common treatment for diabetes is insulin regimen in which patients take insulin dose to control fasting blood glucose levels and insulin boluses around mealtimes to quickly reduce the impact of carbohydrate intake or an insulin pump that provides a continuous insulin infusion [2]. To avoid diabetes the insulin pumps simultaneously normal daily insulin rates and meal that represents the collection initiated by the user consuming food intake. Insulin is injected in the fatty tissue right beneath the epidermis in treatments [3]. This can be done using a continuous glucose monitor (CGM) installed in the subcutaneous tissue or manually using fingerprick measures multiple times each day. Measuring blood sugar levels is required in addition to external insulin therapy. This can be done with a continuously blood glucose concentration (CGM) installed in the epidermis or manually multiple times each day using dipstick measures [4]. Furthermore, diabetes Patients will build a treatment plan with their physician based on their specific needs. Patients will self-administer insulin dosages based on their treatment plan and self-measured blood

S. S. Bhat (✉) · G. A. Ansari
Faculty of Science, Dr. Vishwanath Karad MIT World Peace University, Pune 411 038, India
e-mail: Salliahshafi678@gmail.com

G. A. Ansari
e-mail: gufran.ansari@mitwpu.edu.in

S. S. Bhat
Department of Computer Applications, B. S. Abdur Rahman Crescent Institute of Science and Technology, Chennai 48, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_13

sugar level. Artificial drug treatment seeks to maintain blood sugar levels between 70 and 180 mg/dl in the norm glycemic range [5]. Increased heart rate, psychological confusion and unconsciousness are among symptoms of hypoglycemia. Diabetes can cause memory loss if it occurs frequently enough, because of a phenomenon known as hypoglycemia unaware hypoglycemic symptoms can be difficult to identify in many diabetic patients [6]. Physical activity has numerous health effects and is therefore commonly recommended for diabetics [7]. Exercise on the other hand has a considerable impact on glucose metabolism in diabetics [8]. Moreover enhanced insulin awareness effects are long-term activities that have a harmful effect on patients daily routines. Since an automated blood glucose control system can be greatly useful to diabetics numerous studies have been conducted to develop algorithms for this goal [9]. It is influenced by a variety of variables including dietary consumption, active insulin and hormone changes. Furthermore, many BG strategies are difficult to apply because they either require extensive manual adjustment for individual patients or assume the availability of an accurate mathematical formula of the patient's BG dynamics [10]. Machine learning algorithms have recently been popular because they can learn and gain intelligence from a significant amount of data generated by the technological innovations [11]. Artificial neural networks (ANN) for example are a powerful tool that may be used to create a more personalized model of BG activities by simulating how the nervous system functions in a direct way. It was created and studied in the fields of control theory and information theory as well as being used in a number of applications such as diabetes. Blindness, excessive blood pressure, heart failure, and kidney failure are all common side effects. According to the American Diabetes Association's 2021 statistics report, 3 million persons had diabetes in 2019. Nearly 2 Crore 30 Lakhs individuals were diagnosed from this population [12]. The remaining people were not even given a diagnosis. As a result diabetes detection and prediction is a major problem that can be addressed using Machine Learning Techniques. According to data analysis provided by the World Health Organization 42 million people worldwide suffer from diabetes. This population comprises together not high-income of developing nations but also developing and underdeveloped countries. Type-1 Diabetes Mellitus and Type-2Diabetes Mellitus are its main type.

**Type-1 Diabetes Mellitus**: It is caused due to the production of less insulin than the body requires through a process known as "glucose Diabetes mellitus". It require supplemental insulin to compensate for the pancreas decreased insulin production and low blood sugar levels in the human body define its characteristics.

**Type-2 Diabetes Mellitus**: It is defined by the body's resistance to insulin, since the body's cells react to insulin differently than they are normally expected to. This could eventually result in the body having no insulin. This is also known as "adult beginning Diabetes" or "non-insulin dependent Diabetes mellitus". Individuals with a high bodyweight or who do not live an active life are more likely to get this disease. Sugar levels in a normal human being range from 70 to 99 mg/dl. Only when a person's overnight glucose level exceeds 126 mg/dl he or she is diagnosed as diabetic. An individual with a glucose concentration of 100–125 mg/dl is known as prediabetic in medicine. Obesity, hypertension, dyslipidemia, arteriosclerosis and other diseases are often associated with Type-2 diabetes which is more common in middle-aged and elderly people. Insulin resistance is a feature of Type-2 diabetes. Another essential fact is that Diabetes is the major cause of death. Type-2 diabetes, on the other hand can be discovered early and properly managed. Doctors must be able to acknowledge possible cases quickly. But our focus will be on Diabetes Mellitus.

## *13.1.1 Machine Learning*

The research field that deals with the way in which the machine adapts to changes is called Machine Learning. The term "Artificial Intelligence" is similar to the word Machine Learning for many researchers since the probability of studying is the main feature of an object and is known as Intelligent in the best possible way. Design electronic systems that would adjust and learn from their experiences is the purpose of Machine Learning. Mitchell provides a more comprehensive and systematic description of Machine Learning with regard to certain classes of the tasks T and performance measures P, a programming language is set to learn from the knowledge. We have to find the potential to find an answer with the emergence of Machine Learning methods. Authors have to build a model using a data collection that can be determined whether or not the patient has Diabetes mellitus. In addition early disease prediction leads to the treatment of patients until it becomes crucial. Machine learning is capable of removing the secret insights from a vast volume of information linked to Diabetes Mellitus research and it has an important role more than ever before. The goal of this study is used to build a method that can estimate the level of Diabetes mellitus risk to patients with greater precision. This is a study focusing on designing a structure on several classifications namely SVM, KNN, NB, DT and Model of integrated algorithm.

### 13.1.1.1 Supervised Learning for Predictive Models

Predictive models are made using supervised Learning Techniques. A predictive model utilizes other values in the dataset to predict null values. The supervised Learning Techniques consist of set of intake data and output data to build a model to predict the response to a new dataset in a real manner, Bayesian Method, ANN are all examples of supervised learning. Machine learning is exploding with these techniques [13].

### 13.1.1.2 Unsupervised Learning for Predictive Models

Unsupervised Learning is used to create descriptive models. We have known set of inputs in this model but the outcome is unknown. On data information unsupervised learning is most commonly used. Clustering algorithms such as k-Means clustering and k-Medians clustering are included in this method [14].

#### 13.1.1.3 Semi-supervised Learning for Predictive Models

On the training dataset the semi-supervised learning method uses both labeled and unlabeled data. Semi Supervised Learning includes techniques such as classification and regression. Regression techniques such as logistic regression and linear regression are examples [15].

So, the main objective of this chapter is to focus on the Framework for Diabetic Prediction and Diet Using Machine Learning Techniques. The proposed approach is also based on Data Exploring and cleaning. The focus is on building a predictive model using Machine Learning algorithms for Diabetes prediction. The classifications have been done using the XGBoost-Nearest Neighbor, Decision tree and Random Forest.

The remaining chapter is structured as follows: Sect. 13.2 is about Literature Survey. Section 13.3 is about the Framework for Diabetic Patients and Diet Using Machine Learning Prediction and Sect. 13.4 includes Machine Learning Classifications. Section 13.5 Discusses about the Algorithms for Food Recommendation and diabetes prediction using Machine Learning. Section 13.6 discusses about the Experimental set up. Section 13.7 contains the result and Analysis and finally Sect. 13.8 is an elaborate discussion followed by Conclusion and Future work.

## 13.2 Literature Survey

A lot of research is being conducted in the domain of diabetic diagnosis. As a result some of the relevant literature is offered below. Diabetes Mellitus (DM) is a serious public health issue and its prevalence is on the rise. Many classification algorithms have been used in this field in the aim of classifying patients or forecasting their future states. This section will provide an overview of these works. In recent times, plenty of diabetes prediction methods have been proposed and published. Researchers proposed a Machine Learning framework in which they implemented Linear Discriminant Analysis (LDA), Naive Bayes (NB), Gaussian Process Classification (GPC), Support Vector Machine (SVM), Artificial Neural Network (ANN), AdaBoost (ADB) and Decision Tree (DT) [16–24]. They also carried out an investigation on outlier rejection and filling missing values in order to improve the ML performance. Several algorithms, including the traditional Machine Learning method such as Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression and others have recently been used to predict diabetes. Principal Component Analysis (PCA) and neuro fuzzy inference were used to distinguish diabetics from healthy people [25, 26]. The role of AdaBoost and bagging ensemble ML methods for classifying Diabetes Mellitus and as diabetic or non-diabetic based on Diabetes risk factors, using the J48 decision tree as the basis [27]. The result of the research showed that Ada boost Machine Learning ensemble technique outperforms bagging and a J48 decision tree in terms of performance. The research study reveals a variety of reports about Diabetes and the Diabetes epidemic in the United States. With 30,383 people with diabetes the data storage established a wide structured care system in the region [28]. To test the data sets they used a classification or regression tree method. Diabetes was analyzed in India and it was discovered that the total prevalence of diabetic adults is 72.96%. The authors also suggest implementing a long-term study to show the significance of modifying risk factors for diabetes development and prevention [29]. Machine Learning showed that the learning frequency did not predict changes in HbA1c levels [30].The status of insulin therapy for schoolchildren with Type-1 diabetes over previous decade was examined and it was discovered that 143 Patients were diagnosed at less than five years of age in the past decade (2011–2020).During the school years their methods of frequent insulin treatment and episodes of extreme high blood sugar were examined. In the year 2018 a Diabetes study was conducted in India to examine the clinical management and regulation of Diabetes, as well as modifiable cardiovascular risk factors in Patients with Type-2 Diabetes providing specialist care [31]. A total of 900 training candidates lead from across India were invited to participate in the research. This is due in part to the seriousness of Diabetes in these patients that are seen by Diabetes specialists however, public understanding and implementation of recommended schedule must be improved [32]. The researchers wanted to learn more about glycemic control, diabetic guidance and complications in old aged with Type-2 diabetes. Blood glucose regulation and hypoglycemia-related factors of 2312 children and adolescents were engaged in a cross-sectional clinic-based sample (aged 18 years; 45% males).The Swedish National Diabetes Register (NDR) received data on cigarette smoking in order to research the trend in the ratio of diabetic patients who smoke as well as the links between smoking, serum lipids and hyperlipidemia. The research smoking was common among Diabetic Patients, especially among young male Type-1 diabetes and middle-aged Type-1 and Type-2 diabetes patients and that they should be targeted for smoking cessation campaigns. Independent of other study characteristics, smoking was linked to impaired glycemic regulation and hyperlipidemia [33]. In an actual translation of the lifestyle intervention, an investigation on weight loss target among participants participating in an adapted Diabetes Prevention Program (DPP) found that those who met their goal of losing weight were more likely to have closely tracked their food intake and increased their physical activity substantially both in a drug concentration relation. The research discusses the importance of helping patients in treatment outcomes in developing and sustaining dietary personality and increasing levels of physical activity [34]. The findings show the importance of facilitating individuals in lifestyle programs to begin and sustain diet and lifestyle self-monitoring and higher participation in sports, Normal body movement [35]. LR, K Nearest Neighbors, Support Vector Machine, Gradient Boost, Decision Tree, MLP, Random Forests and Naive Bayes were the eight Machine Learning methods that were analyzed. Among all the algorithms the KNN has the accurate results.

## 13.3 Framework for Diabetic Prediction and Diet Using Machine Learning

A Framework has now been proposed and well explained in detail with a focus on Machine Learning based prediction. In addition, this part provides insight into the Diabetic Patient population's data extracted from UCI Machine Learning repository.[1] The proposed Machine Learning based Framework for Diabetes Mellitus prediction as shown in Fig. 13.1 motivates the future Machine Learning based disease prediction of Diabetes and diet. The author proposed a Framework in which data is collected; after that the author has done the preprocessing and then extracted the value from Machine Learning Algorithms. Then we apply Machine Learning algorithms and find results and analysis. Then the author has finally done the prediction to predict the diabetes.

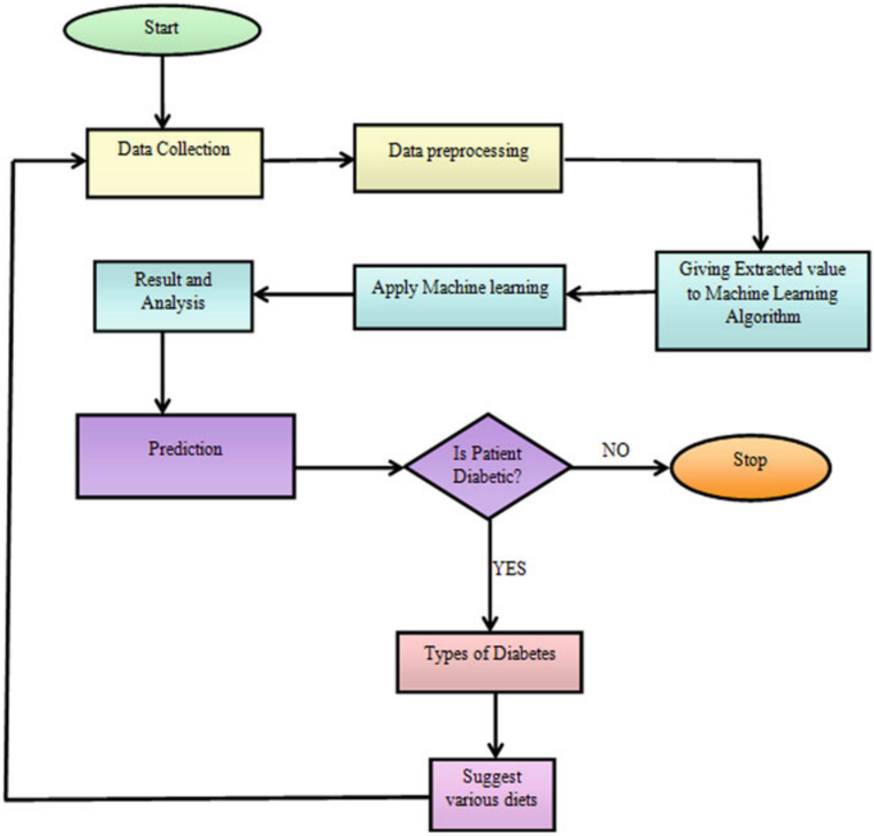


**Fig. 13.1** Framework for diabetes prediction and diet

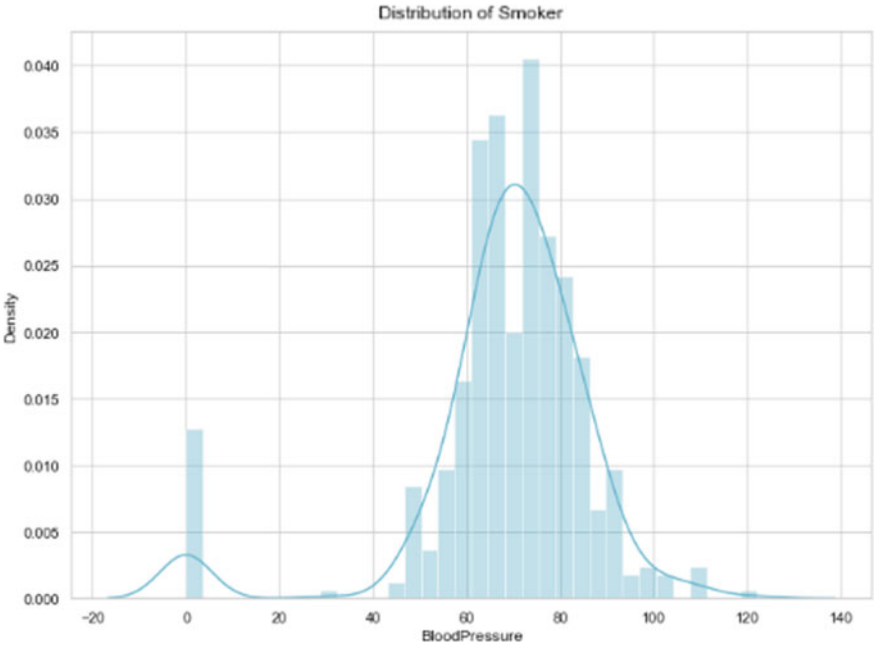


**Fig. 13.2** Distribution of high blood pressure and smokers

### 13.3.1 Data Set Collection

The data used for this study is data collected from the UCI machine learning repository.[1] These affect health data collection and analysis in order to investigate patterns and trends which aids in forecasting and valuating outcomes. It consists of 41 instances and 41 attributes which are Diabetes, High Blood Pressure, Cholcheck, HighChol, BMI, Smoker, Fruits, Veggies, Gender, Age, Blood Pressure, Skin thickness, Insulin and Outcomes. Figure 13.3 gives the length distribution of Pima Indian Data set (PIDD). The PIDD are labeled for their corresponding classes. Here the people with high blood pressure tend to be smokers as well. This shows there is a strong positive correlation between the people who smoke and the ones who have high blood pressure (Table 13.1).

### 13.3.2 Data Preprocessing

The most essential factor of this research is data preprocessing. Insufficient data and different existing values are common in data from health-care services which might lead to ambiguity in data analysis. Data preprocessing is performed on the dataset in order to increase the quality and viability of the information obtained after

**Table 13.1** Depicting patient characteristics

| #N | Feature name | Description | Feature units | Feature value |
|---|---|---|---|---|
| 1 | Age | Person's age | Number | 1, 2, 3…… |
| 2 | Skin thickness | Skin fold thickness of the triceps (mm) | Categorical | 1, 2, 3…… |
| 3 | BMI | Body mass index of a person | Number | 1, 2, 3…… |
| 4 | Blood pressure | Whether the person is having blood pressure or not | Number | 1, 2, 3, … |
| 5 | Insulin | Whether the person needs insulin or not | Number | 1, 2, 3…… |
| 6 | Diabetes | Whether the person is having Diabetes or not | Number | 1, 2, 3…… |
| 7 | High chol | Whether the person is having high cholesterol level high or low | Number | 1, 2, 3…… |
| 8 | Smoker | Whether the person active or passive smoker | Number | 1, 2, 3…. |
| 9 | Chol check | Whether the person checks cholesterol or not | Number | 1, 2, 3…… |
| 10 | Fruits | Whether the person eats Fruits or not | Number | 1, 2, 3…… |
| 11 | Gender | Whether the person is male or female | Categorical | 1, 2, 3…… |
| 12 | High blood pressure | Whether the person is having high blood pressure or not | Categorical | 1, 2, 3…… |
| 13 | Veggies | Whether the person eats vegetables or not | Categorical | 1, 2, 3… |
| 14 | Outcomes | It is a class variable of 0&1 | Categorical | 1, 2, 3…. |

the processing method. The value is a fundamental basis for precise outcome and accurate forecasting when using Machine Learning Techniques on a dataset. Data preprocessing is conducted in two stages for the dataset utilized in this study. The Process of removing data is known as data preprocessing.

### *13.3.3 Data Distribution*

Some data have a normal distribution in their distribution for instance Age, Skin Thickness, BMI, Blood pressure, Insulin, HighChol, Diabetes, HighBloodPressure, Smoker, Chol Check, Fruits, Gender, Veggies, outcome this sort of exponential and depicted in Fig. 13.3.

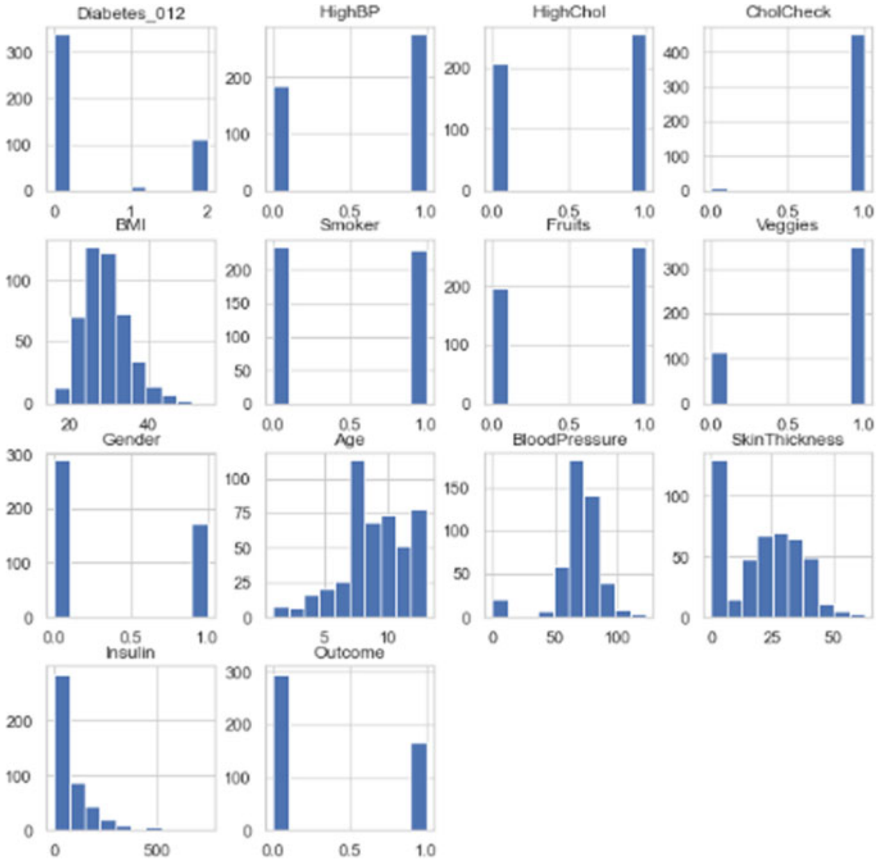


Fig. 13.3 Shows the data distribution visualization

### *13.3.4 Data Exploring and Cleaning*

In replacing the null values, analysis and purification is based on a descriptive analysis distribution of variables used as data to our system. The goal is to get statistical analysis that describes the distribution's tendency, variance and shape. Table 13.2 shows the results of a data analysis of the dataset. The main goal of this stage is to get a description of the collection and count number of times missing values. Researchers choose to build the pairs plot diagram to analyze the distribution of input parameters, relation, and influence on the outcome variable (diabetic or non-diabetic). Figure 13.4 shows the histogram and scatter plot. The x-axis is the horizontal axis. The x-axis indicates the values of input variables whilst the y-axis indicates the values of output variables. This depicts the distribution of the variables in each dataset.

**Table 13.2** Result of statistical analysis

| | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Diabetes | 461.00 | 0.50 | 0.86 | 0.00 | 0.00 | 0.00 | 1.00 | 2.00 |
| High blood pressure | 461.00 | 0.60 | 0.49 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| High chol | 461.00 | 0.55 | 0.50 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| Chol check | 461.00 | 0.98 | 0.14 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| BMI | 461.00 | 28.84 | 5.95 | 16.00 | 24.00 | 28.00 | 32.00 | 55.00 |
| Smoker | 461.00 | 0.49 | 0.50 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Fruits | 461.00 | 0.57 | 0.49 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| Veggies | 461.00 | 0.75 | 0.43 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Gender | 461.00 | 0.38 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |
| Age | 461.00 | 8.88 | 2.63 | 1.00 | 7.00 | 9.00 | 11.00 | 13.00 |
| Blood pressure | 461.00 | 68.67 | 18.94 | 0.00 | 64.00 | 70.00 | 78.00 | 122.00 |
| Skin thickness | 461.00 | 20.67 | 15.55 | 0.00 | 0.00 | 23.00 | 32.00 | 63.00 |
| Insulin | 461.00 | 79.49 | 114.80 | 0.00 | 0.00 | 37.00 | 122.00 | 744.00 |
| Outcomes | 461.00 | 0.36 | 0.48 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 |

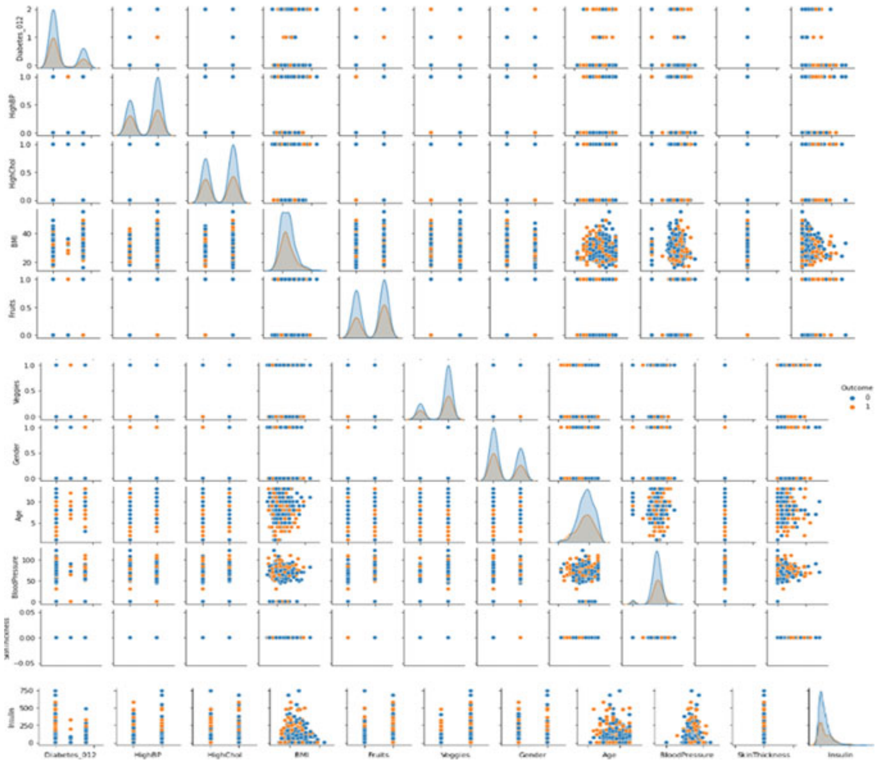
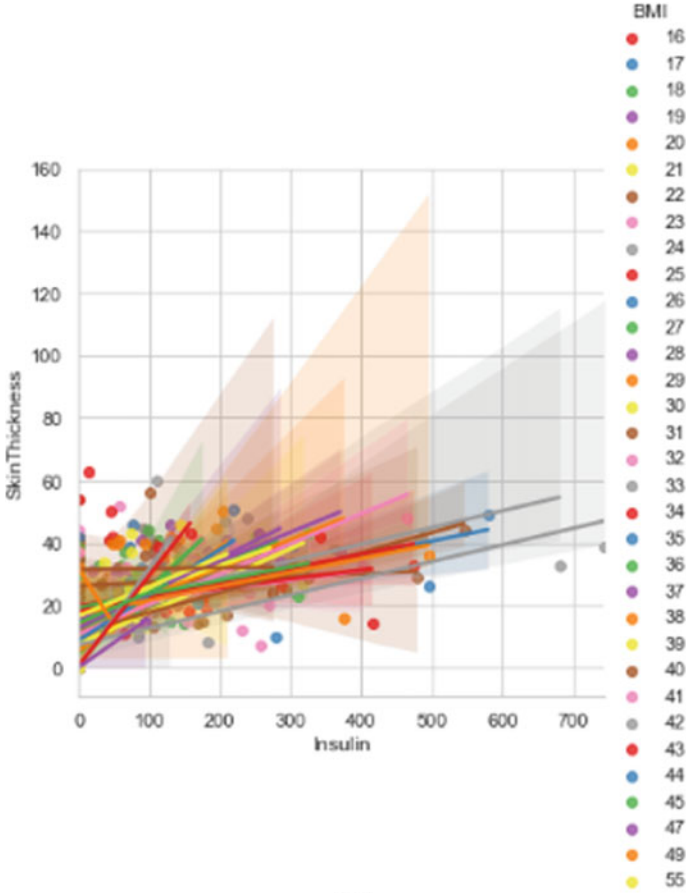


**Fig. 13.4** Pair plot diagram drawn on diabetes data sets



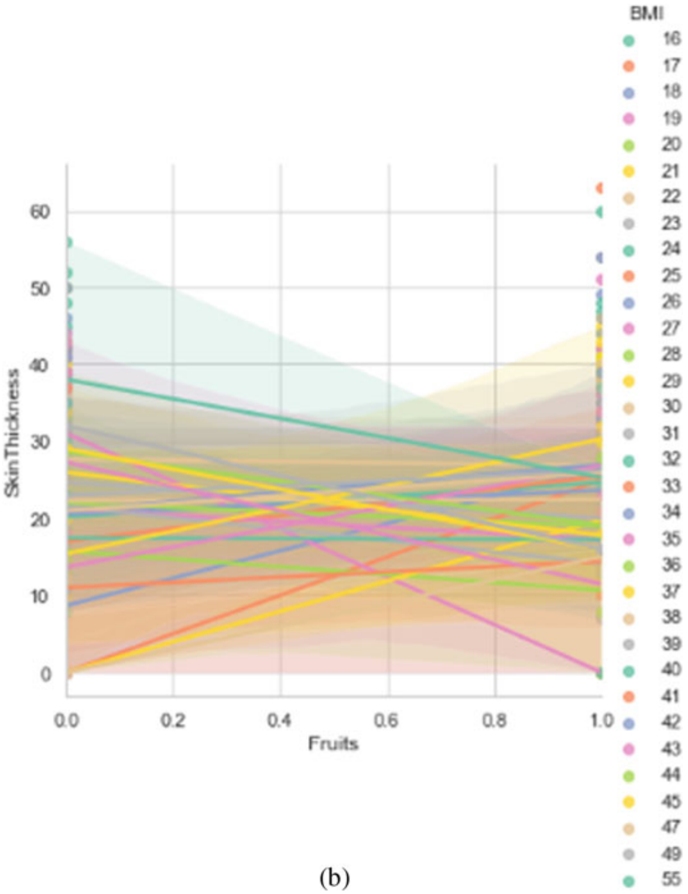
(a)

**Fig. 13.5** Seaborn plot

## 13.4 Machine Learning Classification

On the basis of the scope of research the author set out to employ four of the most recognized algorithms for estimating such as XG Boost, K-Nearest Neighbor, Random Forest, and Decision Tree. Algorithms will create the model for our dataset based on the problem. In this chapter four algorithms were picked for the data set which is given underneath:

(i) **XG Boost (XGB)**: XGBoost is a convincing distributed Machine Learning platform for scaling tree boosting methods as well as an efficient and easy deployment of the Gradient Boosted Trees algorithm. The classier is well configured and has the responsibility to solve in a distributed environment for a rapid parallel tree structure. It combines a single node with tens of

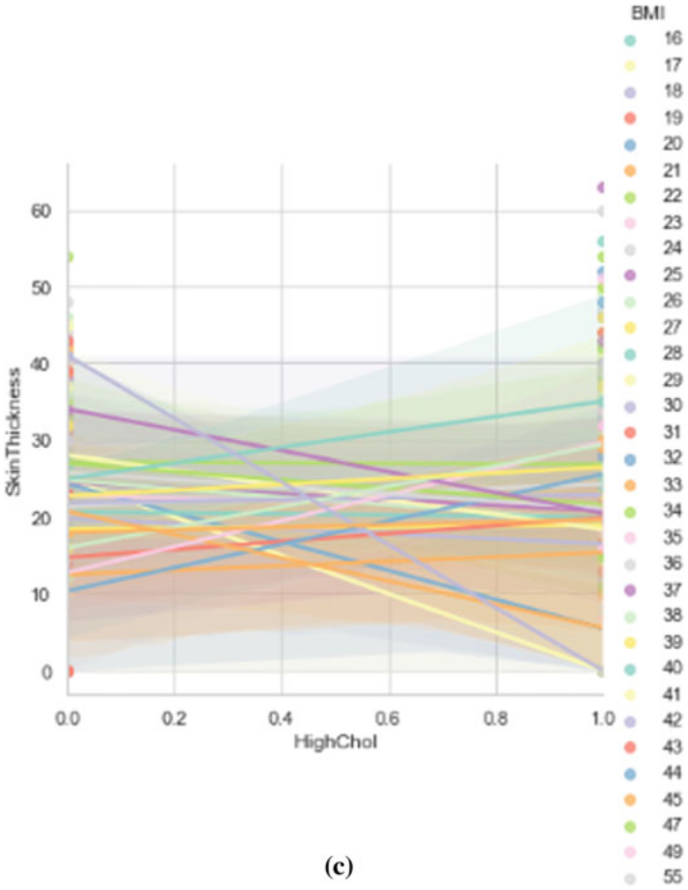


(b)

**Fig. 13.5** (continued)

millions of samples and billions of distributed software samples, allowing it to expand to record levels [36]. XGBoost is a Machine Learning method that has recently controlled Kaggle competitions for unstructured or structured data. Boosting is a high-speed and high-performance implementation of gradient boosted decision trees.

(ii) **K-Nearest Neighbor (KNN)**: The K Nearest Neighbor Algorithm is an easy supervised machine learning model that trains on the complete dataset. This is typically applied to classification models. Whenever unknown data needs to be predicted, it searches the complete training dataset for the K-most similar instances, and only the data with the most similar instances is returned as prediction (that is it classifies the data points based on how its neighbors are classified).This is frequently used in technology application to find similar objects. The letter K in KNN stands for the number of nearest neighbors, or



(c)

**Fig. 13.5** (continued)

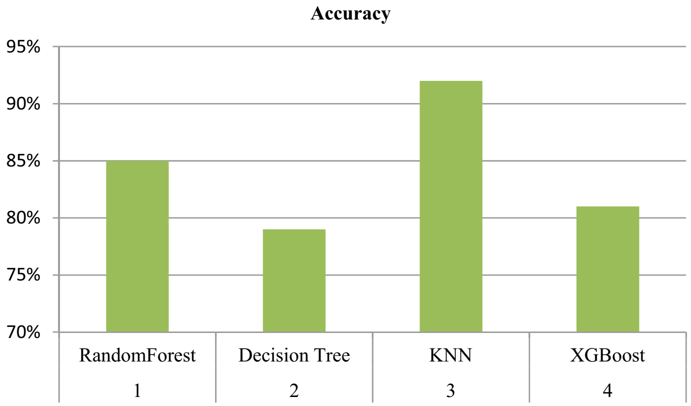


**Fig. 13.6** Comparison and accuracy of various algorithms

voting class, in the fresh or testing data [37]. This algorithm is based on resemblance of features. Parameter tuning is the process of selecting the appropriate value for K, which can improve accuracy. Choosing the correct number for K is critical since a value that is too low causes noise, while a value that is too large cause's capacity or processing concerns. The square root of n is a typical method of determining the proper value for K. (where n is the total number of data points) The k-nearest neighbors (KNN) technique is a supervised machine learning algorithm that is simple to develop and may be used to address both classification or regression problems.

(iii) **Random Forest (RF)**: It is a group of models that operate together as an ensemble, as the title implies. The knowledge of the people is a core idea in RF; each model forecasts an outcome, and the majority decides in the end. It has been shown to be effective in the literature for diabetic prediction [38]. The RF classifier iterates B times by picking samples with replacement by fitting a tree to the training examples given a collection of training examples X × 1, × 2,…, xm and their respective targets Y y1, y2,…, ym. The -e training method is made up of the steps shown in Eq. (13.1)

$$f = \frac{1}{b}\sum_{b=1}^{n} Fb(X1) \tag{13.1}$$

Sample n training instances from X and Y with replacement for b 1…B.
On Xb and Yb, train a classification tree fb.

(iv) **Decision Tree (DT)**: It is one of the most common classifications for machine learning. Algorithm that imputes meaning much of the value of the single level algorithm Time is especially suitable for an ensemble system in improving that is one of the factors for boosting.

Algorithm for Decision Tree for healthcare.

Step 1: Using the dataset S and all of its D and C properties train and prune a DT.

Step 2:Let D' and C' represent the sets of discrete and continuous attributes remaining present in the network, respectively, and S' represents the set of data samples correctly classified by the trimmed network.

Step 3: Using both discrete and continuous C' features, create a decision tree.

Step 4: Ri is created for each rule.

Step 5: Create the tree's root node.

Step 6: Return the individual tree Root, with label = + , if all examples are positive.

Step 7: Return the single-node tree Root, with label = −, if all examples are negative.

Step 8: Return the single node tree Root with label the most common value of the target attribute in the instances if the number of predicting attributes is empty.

Step 9: Otherwise, start with A the Attribute that best represents a classifier. Root = A is a Decision Tree attribute.

Step 10: For each of A's potential values.

Step 11: Below Root, add a new tree branch that corresponds to the test A = vi.

Step 12: Let Examples (vi) be the subset of examples for which vi is the value.

Step 13: If Examples (vi) is empty, create a leaf node with label = least appropriate target value in the examples below this new branch.

Step 14: End.

Step 15: Get back to the beginning (Return root).

Step 16: If RI > si is supported and RI > s2 is an error.

Step 17: Let Si represent the set of data samples that satisfy the condition of rule RI, Di the set of attribute selection, and Ci the set of attribute values that do not best fit the needs of rule Ri.

Step 18: Otherwise stop.

## 13.5 Algorithms for Food Recommendation to Diabetic Patients by Using Machine Learning

Patients can be given two dietary recommendations depending on the length of their activities. People should only consume at the beginning of short physical activities according to the method. The amount of carbohydrate (CHO) is predicted based on the Multi-Layer perception network's prediction of the BG outcome listed. Long physical activities require spreading food intake throughout the activity to keep blood sugar levels stable.

### *13.5.1 Diet Recommendations for Diabetic Patients*

Dietary recommendations are essential in preventing the development of diabetes. Patients with life-threatening diseases, such as high blood pressure or Diabetes, must obey the rules. Diabetes prevention requires a careful diet. The patients food will be recommended based on the area discovered and the outcome of the prediction. Diet database is used to get food items for the diet plan. Dietary management is important in the treatment of diabetes disease since it slows the course of the disease. To avoid kidney failure people with diabetes and high blood pressure should follow a very strict diet.

### *13.5.2 Characteristics of Diabetes*

**Drug**: Oral medicines in the form of medicines help in the management of symptoms. Levels of blood sugar in patients whose bodies are still producing small portion of insulin (the majority of people with Type-2Diabetes) People with Diabetes (Type-1) are often prescribed drugs. As well as ideas for relevant dietary changes increases and exercising on a daily basis. So many of these medicines to achieve maximum blood circulation are often used in combination carbohydrate management.

**Diet**: Nutrition process is important for diabetic patients. What they feed, what they eat. This supports the patient in avoiding very high or low blood sugar levels. Meal planning entails selecting healthy foods and consuming the required amount of food at the appropriate time. The patient must consult with his doctor and a nutritionist on a regular basis to decide how much protein, calcium, and carbohydrates are needed in his diet. Meal preparations for the patient should be personalized to his everyday routine and habits.

**Weight Loss**: It is indeed one of the most integrated disease treatments. Weight loss increases insulin resistance, which helps to keep blood sugar levels under control.

**Insulin**: Most diabetic patients use insulin to treat their disease in various ways.

**Cessation for Smoking**: One of several causes of Diabetes is smoking. Cigarette increases the harm caused to the body by Diabetes, which leads to heart stiffening. High blood pressure is made more likely by smoking.

**Exercise**: In support of Diabetes prevention, exercise is important. Weight and blood sugar levels can be controlled by balancing diet, exercise, and medicine (if needed). Exercise improves the body's insulin use, which aids in Diabetes management. Exercise also results in the loss of excess body fat and weight management.

### *13.5.3 Fruit Consumption and Diabetes Prevention*

Many statistical research shows that taking fruit on a daily basis helps to prevent diabetes. People who eat 400 g of fruit each day have a 40% lower risk of diabetes over a 5- to ten year period. The advantages of fruit and vegetable intake can be noticed up to 200–300 g. per day. The amount of benefit varies depending on the fruit. A mixture of diverse fruits on the other hand is more healthy than a single fruit kind.

### *13.5.4 How Does Fruit Consumption Help to Avoid Diabetes*

The favorable effect of fruits in diabetes care can be associated with a number of variables. Fruit consumption lowers caloric intake while increasing fiber consumption resulting in a lower risk of weight gain. The consequences of high-energy diets are being gradually reduced. Bioactive chemicals found in fruits (such as vitamin C, carotenoids and chromogenic acid) have been demonstrated to lower the incidence of diabetes.

### Type of Fruits

The benefits are the same for all fruits; however some fruits are more advantageous than others depending on the caloric content vs. fiber content. Apples, pears, strawberries, grapes, bananas, oranges, peaches, plums, and cherries are all good for diabetes prevention.

### Fruit consumption schedule

Fruit is not digested or processed correctly to reap the advantages, therefore eating it immediately after a meal is not a smart idea. Fruit should be consumed first thing in the morning, with a glass of water, or an hour before or two hours after a meal.

### Flavors and fruit juices that have been processed

Fruit and juices that have been processed or canned are low in fiber and rich in sugar. Consistent usage of these increases the risk of obesity and diabetes. Fruit juice drinking causes a quick rise in blood sugar in diabetics as well as chronic hunger.

### Consumption of fruits by diabetics

Fruits have an impact on diabetes due to two components: carbohydrate and fiber content. Carbohydrates can raise blood sugar levels on the other hand fiber can help those with diabetes. Fiber delays the absorption of meals, lowering blood glucose levels. Fiber also increases a sense of fullness and lowers hunger which aids in weight loss. The effect of carbohydrate content in food on blood glucose levels is measured using two parameters—glycemic index (how quickly it raises blood sugar) and insulin levels (how much it elevates blood sugar) (how long blood glucose remains high). In diabetes, fruits with a low glycemic index and glycemic load are favored. The following fruits are recommended as given below.

**Apple**: This fruit has been examined the most for diabetes prevention. It's also one of the healthiest fruits for people who have diabetes. Apple skin includes fiber which slows sugar absorptions as well as polyphenol chemicals which aid in insulin release from the pancreas. Both of these variables aid in diabetes management. Apples are also high in antioxidants which aid in the prevention of diabetes problems.

**Pear**: Pears are high in fiber, potassium, and antioxidants, and have a low sugar level. It is one of the most beneficial fruits for diabetics.

**Cherries**: Cherries are recommended in diabetes because they have one of the lowest diabetic indices and are high in antioxidants.

### Fruits to Avoid

Pineapple, watermelon and pumpkin are high-glycemic fruits that should be avoided or consumed in small amounts with precaution. Fruit juices, cooked, processed fruit, dried fruits and flavored fruit items should be avoided because they contain a lot of sugar and cause levels to rise.

## 13.6 Experimental Setup

The Machine Learning Techniques were implemented in a window operating system environment using the Python programming language on a ThinkPad Laptop System Corei5 with 8 GB of RAM and a 2.8 GHz processor (laptop) speed. On the python notebook all of the necessary libraries were installed and used for data analysis, including analysis and model creation. The Fig. 13.5a above illustrates the relationship between three variables BMI, Insulin and Skin thickness. The data in the Figure illustrates the level of BMI at various insulin and Skin thickness points. Also the BMI points are majorly concentrated around the lines and the shaded area. The Inference is clear that the Skin Thickness is directly proportional to the Insulin and BMI. The Fig. 13.5b above demonstrates the relationship between three variables BMI, Fruits and Skin thickness. The data in the Fig illustrates the level of BMI at various fruits consumption level and Skin thickness points. Also the BMI points are majorly concentrated around the lines and the shaded area. The Inference is clear that the Skin Thickness is directly proportional to the Fruits Consumption level and BMI. Figure 13.5c above demonstrates the relationship between three variables BMI, High Chol and Skin thickness. The data in the Fig illustrates the level of BMI at various HighChol and Skin thickness points. Also the BMI points are majorly concentrated around the lines and the shaded area. The Inference is clear that the Skin Thickness is directly proportional to the HighChol and BMI.

## 13.7 Result and Analysis

PIMA an Indian Diabetes dataset is being used for this study's analysis. It is made up of fourteen distinct attributes and forty one instances. The experiment was carried out with the help of the simple python programming language using Jupiter Notebook. Machine Learning Classifications like Decision Tree, Random Forest, XGBoost, and KNN. It is also used to predict diabetes and diet in early stages as shown in Table 13.3. As described in Fig. 13.6, the Measure Performance model is based on accuracy.

**Table 13.3** Shows the predictive accuracy of algorithms

| S. no. | Algorithms | Accuracy (%) |
|---|---|---|
| 1 | Random forest | 85 |
| 2 | Decision tree | 79 |
| 3 | KNN | 92 |
| 4 | XGBoost | 81 |

## 13.8 Conclusion and Future Work

Diabetes is a reality metabolic condition that affects every organ in the body. Missing and partial values are common in established health data which leads to erroneous diabetes diagnosis decisions. The purpose of this research was to create a useful predictive model for better diabetes diagnosis and diet. In this chapter, authors have done the classification using XGBoost, KNN, Random Forest and Decision tree for diabetes people successful prediction of Diabetes Miletus Diseases using the method on PimaIndian Data set. Although we developed and created a Machine Learning method for diabetic illness predictions that has created a strong relationship in the field of medical science for the accurate identification of varied medical data. Further the future plan is to build a classification model and create a Location-based dataset from real medical data to successfully forecast diabetes complications in Type-1 and Type-2 diabetes. In addition to that parameters such as heredity, Smoking habit, exercise and diet sheet should be examined for a better diabetes Prediction is one of best choice for research enhancements.

## References

1. Samy, A.L., Hairi, N.N., Low, W.Y.: psychosocial stress, sleep deprivation, and its impact on type II Diabetes mellitus: policies, guidelines, and initiatives from Malaysia. FASEB Bio Adv. **3**(8), 593–600 (2021)
2. Zavitsanou, S., Massa, J., Deshpande, S., Pinsker, J.E., Church, M.M., Andre, C., Eisenberg, D.M.: The effect of two types of pasta versus white rice on postprandial blood glucose levels in adults with Type-1 Diabetes: A randomized crossover trial. Diabetes Technol. Ther. **21**(9), 485–492 (2019)
3. Perkins, B.A., Sherr, J.L., Mathieu, C.: Type-1 Diabetes glycemic management: Insulin therapy, glucose monitoring, and automation. Science **373**(6554), 522–527 (2021)
4. Bodington, R., Kassianides, X., &Bhandari, S.: Point-of-care testing technologies for the home in chronic kidney disease: a narrative review. Clin. Kidney J. (2021)
5. Powers, M.A., Bardsley, J.K., Cypress, M., Funnell, M.M., Harms, D., Hess-Fischl, A., Uelmen, S.: Diabetes self-management education and support in adults with Type-2 Diabetes: a consensus report of the American diabetes association, the association of diabetes care and education specialists, the academy of nutrition and dietetics, the American academy of family physicians, the american academy of PAs, the American Association of nurse practitioners, and the american pharmacists association. Diabetes Care **43**(7), 1636–1649 (2020)
6. Colberg, S.R., Sigal, R.J., Yardley, J.E., Riddell, M.C., Dunstan, D.W., Dempsey, P.C., Tate, D.F.: Physical activity/exercise and Diabetes: a position statement of the American Diabetes Association. Diabetes Care **39**(11), 2065–2079 (2016)
7. Merjaneh, L., Hasan, S., Kasim, N., Ode, K.L.: The role of modulators in cystic fibrosis related diabetes. J. Clin. Transl. Endocrinol. **27**, 100286 (2022)
8. Singh, S.P., Prakash, T., Singh, V.P., Babu, M.G.: Analytic hierarchy process based automatic generation control of multi-area interconnected power system using Jaya algorithm. Eng. Appl. Artif. Intell. **60**, 35–44 (2017)
9. Patra, A.K., Mishra, A.K., Rout, P.K.: Backstopping model predictive controller for blood glucose regulation in type-I Diabetes patient. IETE J. Res. **66**(3), 326–340 (2020)

10. Ghazal, T.M., Hasan, M.K., Alshurideh, M.T., Alzoubi, H.M., Ahmad, M., Akbar, S.S., Akour, I.A.: IoT for smart cities: machine learning approaches in smart healthcare—a review. Future Internet **13**(8), 218 (2021)
11. Ghosh, A., Nundy, S., Mallick, T.K.: How India is dealing with COVID-19 pandemic. Sens. Int. **1**, 100021 (2020)
12. Everett, J.A.: The 12 item social and economic conservatism scale (SECS). PloS One **8**(12), e82131 (2013)
13. Wohlrab, P., Boehme, S., Kaun, C., Wojta, J., Spittler, A., Saleh, L., Tretter, V.: Ropivacaine activates multiple proapoptotic and inflammatory signaling pathways that might subsume to trigger epidural-related maternal fever. Anesth. Analg. **130**(2), 321–331 (2020)
14. Prakash, V.J., Nithya, D.L.: A Survey on Semi-Supervised Learning Techniques (2014). arXiv:1402.4645
15. Lee, S., Zhou, J., Wong, W.T., Liu, T., Wu, W.K., Wong, I.C.K., Tse, G.: Glycemic and lipid variability for predicting complications and mortality in Diabetes mellitus using machine learning. BMC Endocr. Disord. **21**(1), 1–15 (2021)
16. Srivastava, R., Dwivedi, R.K.: A survey on diabetes mellitus prediction using machine learning algorithms. In: ICT Systems and Sustainability, pp. 437–480. Springer, Singapore (2022)
17. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York (2005)
18. Cover, T.M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. (3), 326–334 (1965)
19. Boughton, W., WANG, Z.: Not so naive Bayes: aggregating one-dependence estimators. Mach. Learn **58**(1), 5–24 (2005)
20. Brahim-Belhouari, S., Bermak, A.: Gaussian process for no stationary time series prediction. Comput. Stat. Data Anal. **47**(4), 705–712 (2004)
21. Assad, A., Bouferguene, A.: Data mining algorithms for water main condition prediction—comparative analysis. J. Water Resour. Plan. Manag. **148**(2), 04021101 (2022)
22. Ahuja, R., Dixit, P., Banga, A., Sharma, S.C.: Classification algorithms for predicting diabetes mellitus: a comparative analysis. In: Pervasive Healthcare, pp. 233–253. Springer, Cham (2022)
23. Kégl, B.: The return of AdaBoost. MH: multi-class Hamming trees (2013). arXiv:1312.6086
24. Jenhani, I., Amor, N., Elouedi.: Decision trees as possibility classifiers. Int. J. Approx. Reason. 784–807 (2008)
25. Pal, M., Parija, S., Panda, G.: Improved prediction of diabetes mellitus using machine learning based approach. In: 2021 2nd International Conference on Range Technology (ICORT), pp. 1–6. IEEE (2021)
26. Bhat, S.S., Ansari, G.A.: Predictions of diabetes and diet recommendation system for diabetic patients using machine learning techniques. In: 2021 2nd International Conference for Emerging Technology (INCET), pp. 1–5. IEEE (2021)
27. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance analysis of data mining classification techniques to predict Diabetes. Procedia Comput. Sci. **82**, 115–121 (2016)
28. Shafi, S., Ansari, G.A.: Early Prediction of Diabetes Disease and Classification of Algorithms Using Machine Learning Approach (2021). Available at SSRN 3852590
29. https://en.wikipedia.org/wiki/Diabetes_in_India(29-12-2021).
30. Sigurdardottir, A.K., Jonsdottir, H.: Outcomes of educational interventions in Type-2Diabetes: WEKA data-mining analysis. Patient Educ. Couns. **67**(1–2), 21–31 (2007)
31. Khawaja, N., Abu-Shennar, J., Saleh, M., Dahbour, S.S., Khader, Y.S., Ajlouni, K.M.: The prevalence and risk factors of peripheral neuropathy among Patients with Type-2Diabetes mellitus the case of Jordan. Dialectol. Metab. Syndr. **10**(1), 1–10
32. Saini, P., Ahuja, R.: A review for predicting the diabetes mellitus using different techniques and methods. In: Proceedings of International Conference on Data Science and Applications. Springer, pp. 425–440 (2022)
33. Nilsson, P.M., Gudbjörnsdottir, S., Eliasson, B., Cederholm, J.: Steering committee of the swedish national diabetes register. Smoking is associated with increased HbA1c values and micro albuminuria in Patients with Diabetes–data from the National Diabetes Register in Sweden. DiabetesMetab **30**(3), 261–268 (2004)

34. Harwell, T.S., Vander wood, K.K., Hall, T.O., Butcher, M.K., Helgerson, S.D.: Factors associated with achieving a weight loss goal among participants in an adapted Diabetes Prevention Program. Prim. Care Diabetes **5**(2), 125–129 (2011)
35. Verdezoto, N., Grönvall, E.: On preventive blood pressure self-monitoring at home. Cogn. Technol. Work **18**(2), 267–285 (2016)
36. Onan, A.: Mining opinions from instructor evaluation reviews: a deep learning approach. Comput. Appl. Eng. Educ. **28**(1), 117–138 (2020)
37. Wang, Y., Zhang, L., Niu, M., Li, R., Tu, R., Liu, X., & Wang, C, (2021). Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type-2Diabetes Mellitus Using Machine Learning: Cohort Study *Frontiers in public health*, *9*, and 96.
38. Bhat, S.S., Selvam, V., Ansari, G.A., Ansari, M.D., Rahman, M.H.: Prevalence and Early prediction of diabetes using machine learning in North Kashmir: a case study of district bandipora. Comput. Intell. Neurosci. (2022)

# Chapter 14
# An Accurate Swine Flu Prediction and Early Prediction Using Data Mining Technique

**Srinivas Kolli, Ahmed J. Obaid, K. Saikumar, and V. Sivakumar Reddy**

## 14.1 Introduction

Emerging technologies for Healthcare information and it's use became an important area of research [1–4]. In the United States, the swine flu ranks high on both the list of the most prevalent infectious illnesses and the list of the diseases with the highest transmission rates. For the time being, we'll be using a Naive Bayes classifier to compute this sickness, which will help us save money and time at the examination office. In order to help restorative administrations experts make smart clinical selections, how can we transform accommodating data? This is the motivating force for this piece. A. the avian influenza pandemic Swine illness is a respiratory condition caused by influenza viruses, which weakens lungs of pigs and causes coughing, wheezing, nasal discharges, and drowsiness in the animals. The infection may then pass to humans. It is possible that swine flu outbreaks might alter in order to be more easily transmitted to humans. Pig flu or swine flu, crowd infection, or pig infection is an ailment caused by a variety of swine infectious illnesses, including swine flu. This season's infection social affair pollutions harm is any damage caused by pigs being endemic for this season's infection contamination. $H_1N_1$ and $H_1N_2$ flu A subtypes, as well as $H_3N_1$, $H_3N_2$, and $H_2N_3$, were all recognised in 2009 as flu a subtypes.

S. Kolli
Department of Information Technology, VNR Vignana Jyothi Institute of Engineering Technology, Hyderabad, India

A. J. Obaid (✉)
Faculty of Computer Science and Mathematics, University of Kufa, Kufa, Iraq
e-mail: ahmedj.aljanaby@uokufa.edu.iq

K. Saikumar
Department of ECE, Koneru Lakshmaiah Education Foundation, Green Fields, India

V. S. Reddy
Department of CSE, Malla Reddy University, Hyderabad, India

B. K. Rai et al. (eds.), *AI and Blockchain in Healthcare*, Advanced Technologies and Societal Change, https://doi.org/10.1007/978-981-99-0377-1_14

The spread of swine flu across the globe starts with pig populations. Contagious contaminant transmission from pigs to humans is unusual and does not always tend toward human flu; antibodies in human blood are sometimes all that's needed. Zoonotic swine flu occurs when the virus spreads via the air and infects people.

People who interact with pigs on a regular basis are at an increased risk of contracting swine flu. Testing for infection subtypes became possible in the mid-twentieth century, enabling the proper termination of transmission to people. The Best 50 of these broadcasts need to be confirmed now and in the near future. To transmit a swine infection, one must inhale or consume anything that is infected with swine flu or swine pneumonia; it is not spread by eating cooked pork. Infection a $H_3N_2$v can't be avoided by the most forward-thinking swine flu contamination. Swine flu is characterised by fever, sore throat, runny nose, headache, cold, lethargy, and nausea when it is detected.

*Statistics of Swine flu*

One of the main reasons for 2009's H1N1-caused swine flu epidemic was the creation and distribution of an infectious illness. A novel strain of swine flu, infection A ($H_3N_2$) v, was identified in 2011 as a new form of the illness. Even though only a small number of individuals (mostly children) were initially infected during the 2012–2013 flu season, specialists from the U.S. Spots for illness management and avoidance (SMA) pronounced large numbers of people to be infected.

Those infected with swine flu are now responsible for the deaths of around 1500 family members in India. For this reason, it's unlikely that $H_3N_2$v will have an impact on clearing amounts for family members. Unfortunately, a contaminant with the unfortunate designation of $H_3N_2$ (notice there is no "v" in its name) has also been detected. However, this flu strain may be one of the most prevalent. $H_3N_2$v was previously associated with the virus. The larger majority of the population will not be claiming this season's cold illness when the final tally is tallied. A disease has a similar structure to the $H_1N_1$ virus; each kind requires a different H in the same way that of an n structure.

Masters need assistance in dealing with pc abilities throughout the computational evaluation for the transmission of disease. Some unravel the complicated web of claims about the rise of pandemic contamination. Infections like as swine flu, cholera, and jaundice may spread in particular trademark states. Swine influenza and its associated factors need constant assistance, which is provided in this study by several data mining approaches. In order to show how the k mean and Google graph would look in the future, a pressing check was made. In addition to helping with a Google representation, these clumping estimates control the domain of intelligent swine influenza research. The true blue notion about the observation of swine spoilage is mapped out for Google's research. These packs provide a quick diagram of investigations that have been focused on a certain location. Preliminary calculations from the prior data records are utilised as a basis for the new data collection standards.

Such a large percentage of the data documents the disasters caused by swine contamination in a given area as well as the natural domain factors immediately surrounding the incident and after it. It is necessary to recognise the standard count of temperature, stickiness, and twist velocity in addition to any prospective scenarios of swine contamination. When it comes to such recognitions, data records from the past need to be employed as an educational tool. There is a strong correlation between how C4. 5 weighs and the credulous Bayes classifier. With regard to the fundamentally observed investigations, the reasonable outcome shows a true blue weight inconsistency of investigational findings. It monitors new domains that may be affected by the zone in order to prevent them from claiming a significant number of investigations. In addition, this conclusion tells individuals who may be claiming tainting in a new area the hard road ahead of them. In order to achieve those standard basic parameters employed to image the sea of swine influenza, the national atmosphere figure's web co-operations would be used as Choice trees are counted as part of an overall inspection. In the same way as finding a profit and a true count for investigations from those key frameworks need aggravated subject, Nave Bayes classifiers require an aggravated subject as well.

## 14.2 Literature Survey

Thakkar et al. the healthcare sector produces reams of data that are not being mined effectively or used to their full potential. Unfortunately, the potential of uncovering these unseen connections and patterns is frequently wasted. There is, however, ongoing research in medical diagnostics that uses patient data to make predictions about conditions including heart disease, lung cancer, and other cancers. Using the gathered data on Swine Flu, our study focuses on this element of medical diagnosis. As a result of this study, a prototype of an intelligent swine flu prediction programme has been constructed (ISWPS). Nave Bayes classifier was used to divide swine flu patients into three groups: least possible, probable or most probable. There are 17 swine flu symptoms that we utilised; we compiled 110 symptom sets from different medical facilities and doctors. Through the use of ISWPS, we were able to increase our precision to 63.33%. The JAVA platform is used for its implementation [5].

Information in health data may be greatly aided by opinion mining. Many studies have shown that public tweets may be used to monitor illness spread. Twitter posts in other languages were not included in much of the research. Infectious diseases like influenza are a major problem all around the globe right now. In this article, an existing new process for detecting the spread of the flu using Arabic tweets in Arab nations was done by employing machine learning strategies. We believe this work to be the first to use Arabic tweets to investigate the spread of illness. The aim of this effort was to collect, categorise, filter, and analyse Arabic tweets mentioning influenza. Classifiers such as Naive Bayes, Support Vector Machines, Decision Trees, and K-Nearest Neighbours were used to evaluate the method's efficacy and performance. Across all three studies, Nave Bayes (90.06%) and K-Nearest Neighbor (86.43%) performed best as classifiers in terms of accuracy [6].

Alsmadi et al. the recently identified coronavirus Covid-19 causes an infectious condition. First noticed in Wuhan, Hubei Province, where a number of patients had pneumonia-like symptoms. Since there is no vaccination available and the pandemic is spreading rapidly, health care providers need assistance in assessing their patients' conditions. Data mining methods are the key to finding a remedy for this issue Anticipating recoveries is important for countries working to contain the virus, and these forecasts can help public health experts keep track of COVID-19 positive citizens, improve doctors' ability to predict the general perception of the course of events over time, and assess patients at early risk using approaches based on new data. In this article we describe supervised learning with special reference to the COVID-19 Corona Virus India dataset, which contains data from 3,799 patients and was used to classify the COVID-19 patient data into two groups: recovered and died. The patient dataset has been subjected to a variety of classification strategies, including decision tree (DT), support vector machine (SVM), logistic regression (LR), random forest (RF), k-nearest neighbours (KNN), naïve bayes (NB), and ANN models, with the best method selected using hold-out or cross-validation to evaluate accuracy, precision, and recall [7].

Nouira et al. the number of people who have access to the internet continues to rise; we are now in a position to put a numerical value on contemporary social phenomena. Health care providers have a critical challenge in keeping tabs on the spread of the current epidemic. To assess the present global influenza outbreak, we suggest FluSpider, a novel concept for digital surveillance based on monitoring the number of people who visit certain websites. Our approach is precise because it employs Big Data technology and Massive Data Mining techniques. Based on our findings, the FluSpider system has the potential to tackle the challenge of monitoring the spread of influenza-like illnesses in real time, two weeks ahead of the response time of conventional centres [8].

Using auto regressive integrated moving average and intervention time series analysis [9], this research shows how to separate the effects of swine flu on tourism in Brunei from the continuing effects of the 2008 global financial crisis. We employed an auto regressive integrated moving average model for the virus, a model for the global financial crisis that impacted the number of tourist arrivals, and a model for intervention time series analysis to assess the impacts of the swine flu in the first year after the epidemic. By examining the model coefficients in a time series intervention, we can observe that the swine flu and the global financial crisis have had a catastrophic impact on travel and tourism. It was estimated that Brunei lost around 30,000 visitors (or 15% of its total) and B$15 million in the first year after the swine flu epidemic [9].

Twitter is a free micro blogging and social networking website where users may communicate with one another via brief messages (called "tweets") of up to 140 characters in length. More than 190 million people use the service on a regular basis, and it handles over 55 million tweets per day. The Twitter stream, which is the sum of the thoughts and reactions of Twitter users, has a plethora of information regarding news and geopolitical events. Twitter's stream contains a wealth of information that, if collected efficiently, might be valuable for monitoring and even predicting user behaviour. In this research, we analyse how data from the Twitter stream may be used to (1) monitor the public's developing opinion on $H_1N_1$ (swine flu) and (2) monitor and quantify the disease's real spread. In addition, we show that Twitter may be a useful barometer of public concern about health crises. Findings demonstrate Twitter conversation may be used to estimate the prevalence of influenza-like illnesses, and that these estimates closely mirror the levels of sickness that have been officially reported [10].

Peng et al. the prevalence of COVID-19 was shown to be associated with data from internet search engines like Google Trends, but only in a subset of nations. We want to build a model using data from a selected number of nations that may be used to forecast the global pandemic warning level. We looked at Google Trends data for the phrases "coronavirus," "pneumonia," and "six COVID symptoms" to see how its popularity has changed over time and where it is most popular. The World Health Organization was queried for the daily COVID-19 incidence across 202 countries from 10 January to 23 April 2020. There are now three distinct "alarm states." Machine learning algorithms were trained using data collected across 20 nations over the course of 10 weeks. The characteristics were chosen based on their relevance and association. Thereafter, 2,830 samples from 202 nations were used to evaluate the model [11].

Biswas et al. the medical sciences might benefit greatly from using the problem-solving technique known as Case Based Reasoning (CBR). Nonetheless, one of the crucial technologies for the achievement of CBR is the acceptable appraisal of case similarity. When making predictions about the spread of swine flu, inductive learning methods like weighted closest neighbour (w-NN) are considerations given to the problem of feature weighting in the context of similarity. Using w-NN and extensibility in similarity, a novel case categorization system for swine flu prediction has been created. The unique case categorization method draws upon a mountain of healthcare industry data that has not been adequately mined and is thus not being used to its full potential. The suggested model is also compared to a previously proposed model, and the empirical findings show that our approach has a lot of promise by producing a result that is 85.6% correct, which is better than the previously proposed model [12].

This paper explores Techno social Predictive Analytics (TPA) and related methods for Web "data mining" [13], wherein users' posts and queries are gathered to create coherent representations of real-time health events, using Social Web ("Web 2.0") tools like blogs, micro-blogging, and social networking sites. This article presents an introduction of widely used Social Web tools like as mashups and aggregators, explores their origins as a method of gathering insight into the overall health of communities, and traces their stratospheric growth as an open architecture of participation for the masses. Epidemiological data, such as flu outbreaks, may be visualised in a clear, location-specific fashion by using a variety of health-related tools, which are discussed and demonstrated in detail [13].

Since the COVID19 epidemic, mosquito-borne illnesses have emerged and reappeared in many parts of the world [14]. The latest research in text mining for infectious diseases has the potential to speedily provide both explicit and implicit connections between textual information. Because there is so much unstructured or semi-structured text data available online with rich content of information from this domain, there has been a proliferation of studies in recent years aiming to address problems in this field, such as disease-related knowledge discovery, disease surveillance, early detection systems, etc. Unfortunately, we were unable to locate a comprehensive review of text mining in the area of mosquito-borne diseases. Here, we provide a thorough review of the latest research on text mining tactics for protecting against mosquito-borne diseases. In this article, we analyse the current trends in this area, including the corpus sources, technologies, uses, and difficulties encountered by the research, before looking forward to potential future developments. Here, we report a bibliometric study of the 294 scholarly papers on the topic of text mining in mosquito-borne illnesses that appeared in Scopus and PubMed between 2016 and 2021. After initial screening, the publications were further whittled down and examined according to the methods employed to analyse the text in relation to mosquito-borne illnesses. Using a corpus of 158 articles, we were able to determine that 27 of them made use of edition withdrawal to address mosquito-borne illnesses. Bulk of these publications focused on over 40% of those surveyed were concerned about Zika, 32% about Dengue, and 29% about Malaria, coverage for Chikungunya, Yellow Fever, and West Nile Fever ranged from very low to nonexistent [14].

Nagaraj et al. when it comes to data, the nonprofit sector gathers a humongous quantity that goes unused. Disclosure of these hidden representations and connections is often wasted. However, there is active research toward a therapeutic goal that can foresee infections of the heart, lungs, and other tumours by analysing patient records. Our research is predicated on this medical dead end by knowledge map of the available data on swine flu. This research has opened up a new frontier in the field of Intelligent Swine Infection Prediction modelling. We implemented the DLSC Classifier (Dynamic Learning split classifier). Data mining plays a significant role in health care disease prediction. Since the patient database isn't getting any better, we've started a project to identify Swine fever; the most widely dispersed infectious disease in the globe. The swine flu is a respiratory sickness, and there are a plethora of diagnostic procedures that the patient must through in order to be properly diagnosed. High-powered data mining algorithms are helping us find a solution to this problem [15].

Ali et al. the heart is one of the most important organs in the body, and heart illness is a prominent cause of mortality, so it seems sense that doctors would try to offer the most likely diagnosis when faced with such a condition. Experts in the area of heart illness forecasting may benefit from a number of novel hybrid approaches developed by researchers, which improve a number of machine learning methods. This research presents a technique known as Convolution Neural Network Gate Recurrent Unit (CNN GRU). One of the main goals of this method is to develop a more effective machine learning approach for predicting heart disease. In order to isolate the most important features of the dataset, we use feature selection methods like Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). Multiple machine learning methods using the same set of characteristics were used to evaluate the suggested method. A technique called "K-fold" cross-validation was used to improve precision. According to the findings, the (CNN GRU) method outperformed the competition with an accuracy of 94.5% [16].

Olson et al. the time, location, and intensity of outbreaks may be determined by tracking the onset and progression of clinical episodes of influenza-like illness (ILI). Increases in computing power and data storage capacity have made feasible the automatic collecting of massive amounts of electronic data, allowing for more immediate assessments. Google Flu Trends (GFT) is one example of an innovative monitoring system that uses data from internet search queries to track the spread of influenza-like sickness (ILI). We contrasted weekly public health monitoring with search query data and analysed the reliability of GFT over the last decade to characterise the onset and progression of seasonal and pandemic influenza at the national (United States), regional (Mid-Atlantic), and local (New York City) levels (2003–2013). To provide just two instances, the original and improved GFT models failed miserably at the regional, national, and international levels to predict the first wave of the 2009 influenza A/H1N1 pandemic, and the 2012/2013 season's A/H3N2 epidemic was also significantly overstated in severity. The results were the same for both the 2008 and 2009 GFT versions. Both models underperformed the gold standard due to differences in the seasonality, regional heterogeneity, and age distribution of the epidemics between the timeframes of GFT model-fitting and prospective use, as well as probable modifications in internet search behaviour [17].

After analysing over 500 million tweets over a period of eight months, we found that by monitoring a small number of flu-related terms, we were able to predict future influenza rates with great accuracy, correlating with national health data 95% of the time. We next examine how well this method stands up against false positives for keywords and suggest adding a document categorization step to the process to weed out such misleading signals. In simulated false alarm studies, we discover that our document classifier can cut mistake rates by more than half, however further work is required to build approaches that are resilient in the face of very high noise [18].

The development of an epidemic illness may be mitigated and a response strategy developed if people can keep tabs on the spread of something like seasonal or pandemic flu [19]. In particular, this surveillance is useful for the early identification and geographical localization of an epidemic. This tracking is often done via a variety of means, such as keeping track of how often doctors' offices are visited. We provide a method that analyses the content of social networking sites like Twitter to determine the illness frequency in a given population. Our technique involves reading through thousands of tweets every day to look for references to flu symptoms and then using that data to generate a statistical score that may be used to track the spread of the virus. We put it through 24 weeks of testing in the UK during the 2009 H1N1 flu pandemic. When comparing our flu-score to statistics from the Health Protection Agency, we find a linear correlation of above 95% on average. This approach provides low-cost and up-to-date data on the status of an epidemic since it utilises data that is entirely separate from what is typically utilised for such purposes and may be used at frequent intervals [19].

Achrekar et al. the public health community places a high priority on mitigating the effects of pandemics like H1N1 and seasonal influenza outbreaks. Early detection is crucial for controlling epidemics, according to the literature. The CDC's standard method involves gathering information on the spread of influenza-like illnesses (ILIs) from a sample of "sentinel" medical facilities. The time lag between a patient's diagnosis and its inclusion in aggregate ILI reports is typically between two and four weeks. This research introduces the Social Network Enabled Flu Trends (SNEFT) architecture, which tracks and predicts the appearance and spread of an influenza pandemic in a community by monitoring tweets mentioning flu indicators. Our analysis of data from 2009 and 2010 shows that the number of influenza-like illness (ILI) cases recorded by the CDC is substantially associated with the volume of tweets about the virus. We also develop auto-regression models to estimate the prevalence of ILI in a given area. The models foretell the proportion of visits to "sentinel" doctors in subsequent weeks that may be attributed to ILI, based on data gathered and released by the CDC. We do experiments using historical CDC data on models with and without Twitter data metrics, and find that the latter significantly improves the models' prediction accuracy. So, Twitter information gives a real-time evaluation of ILI activity [20].

Venna et al. we present a novel data-driven machine learning method to influenza forecasting based on long short-term memory (LSTM). Some of the unique elements of the procedure are as follows: Two recent breakthroughs have significantly impacted this field: A method to capture the influence of external variables, such as proximity to geographic features and climatic features like humidity, temperature, precipitation, and sun exposure, has been introduced, and the Long Short-Term Memory (LSTM) method has been developed to capture the temporal dynamics of seasonal flu. The suggested model is tested on two public datasets using two state-of-the-art techniques. Our proposed strategy outperforms the current gold standard and most popular methods for predicting the spread of influenza. The results suggest a promising direction for improving influenza forecasting via the use of data-driven forecasting methods and the incorporation of spatio-temporal and environmental factors [21].

AlAmoodi et al. [22] this study compares and contrasts many AI techniques for identifying and categorising medical images of the 2019 coronavirus outbreak (COVID-19). Relevant research was uncovered by searching the five highly regarded databases of IEEE Xplore, Web of Science, PubMed, Science Direct, and Scopus. After screening the 36 studies collected via a series of filtering and scanning processes based on inclusion/exclusion criteria, only 11 studies were found to be suitable. Using taxonomy, the eleven papers were split into two groups: reviews and original research. Then, the scholarly literature on the topic was critically reviewed and analysed in depth to reveal the obstacles and gaps that needed filling. Using the COVID-19 image database, no papers were found that evaluated and benchmarked the performance of various artificial intelligence algorithms used for classification tasks (binary, multi-class, multi-labelled, and hierarchical classifications). It is possible that evaluation and benchmarking may face future challenges due to the use of several assessment criteria within each categorization assignment, trade-offs between criteria, and the relevance of these criteria [22].

## 14.3 Existing Methods

Representation as a number in this paper, we're focusing on the most important swine flu studies and directing you to the Google graphic. Swine Flu investigations conducted at various social gatherings are checked using the Fig. 14.1 tally. Set theory is used to describe the data and the many outcomes [23].

$$S = \{I,\ O,\ F\}$$

Set S is comprised of the data sources, limitations, and the specific yields that they produce, all of which are shown below in a manner resembling set theory [24].

(1) I = D, P, S y is the input.

Dataset = di, where di is a plan for all patient records.

P = Weather Parameter = Si is a method of Swine Flu side effects.

In response to the sy = sy question: The Pi Climate Parameters Strategy

(2) F = {K, D, B}

K = Mean

D = Decision Tree

B = Bayesian Theorem.

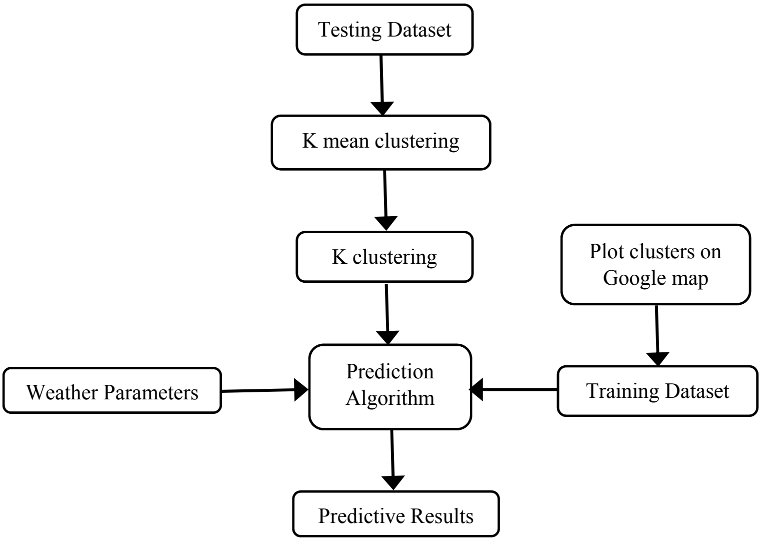


**Fig. 14.1** System architecture

*Existing Architecture*

Once patients have been thoroughly reviewed, the data in the records shows the main findings of the swine flu inquiry. The patient's progress may be gauged by the indicators he or she is displaying. On Google Maps, the precise groups of essential research are framed and shown. By this point, the total includes the actual results of the investigations. Working of the structure contains an indisputable subset lighted up as follows [25, 26].

**Testing Data set**

An introduction data record for those patients who require help dealing with magical side effects of being infected with swine influenza will serve as the testing dataset. Those datasets that k infers are urgently needed will be an obligation. The vast amount of data records related to swine flu investigations that K expects to find will have a high connection with the social event estimates that K intends [27].

**Planning Data set**

Swine Flu observation areas may be plotted on Google Maps using the "K" Clusters feature, which focuses on a particular social issue and plots it on Google's map [28].

It is used to map the yearning to continue running by using past information accounts as a course of action dataset, which is used to monitor really accurate checks of investigations from the present gathered evidence [29].

**Perceptive Investigation**

Using the present patient outline as a guide, we prepare a dataset using a decision tree computation and a Naive Bayes classifier to track real inspections. Current situation subsystems are used to investigate the potential Swine Flu effect on the local region and its environs.

**Proposed Algorithm**

Algorithm: feature selection procedure

[N is the total number of characteristics included in the pattern's issue description, and A is the pattern's attribute.]

Step: 1 K-Nearest Neighbor (K-NN) is used for the categorization.

(1) The criteria for evaluating the classifier's efficacy are set in stone. The proportion of false positives to total classifications is a good indicator of how well it performs. The attribute with the highest recognition rate is chosen as the criteria.
(2) When running the classifier on the training set one attribute at a time, the five-folded cross-validation technique is called upon several times. Therefore, the classifier will only employ attribute $A_1$ in the beginning, followed by $A_2$ and so on up to $A_N$. We keep track of the recognition and mistake rates for each individual characteristic. We choose the attribute Ai, where 1 to N is the range of possible values, based on the criteria that were created in step 2. This ensures that the classifier will have the greatest possible performance.

(3) The classifier is used in cross validation using two characteristics at once:

$$A_i A_1, A_i A_2, \ldots, A_{i-1} A_{i+1}, \ldots, A_i A_n$$

The top performing pair is chosen by the classifier. To illustrate, let's say the optimal pair is $A_I A_J$ for $1 \leq i \leq N$ and $1 \leq j \leq N$.

(4) The classifier is used in the cross-validation process, and three characteristics are used simultaneously: $A_i A_j A_1$, $A_i A_j A_2$, and so on.

(5) Therefore, the list of chosen qualities expands by one attribute at each stage of the process. This procedure is repeated until the performance of the classifier is deemed adequate.

## 14.4 Conclusion

When it comes to disease transmission research, this literature shows how a data mining technique may be employed in an effective way. The Swine Flu indication on Google's layout shows that the batching count k signifies is anticipated to design the sufferers. The desired estimates are used to get a real check on people's imaginations and possible danger. When dealing with the many opposing atmospheric factors, a similar examination of the ordinary assessment tree computation, for example, C4.5 contains a strong producing gathering regulation but settles on a tough choice tree structure. It performs better than any assessment tree estimate in the case of an event of changing atmospheric parameter than the Bayesian Theorem, which is an original probabilistic computation.

## 14.5 Future Work

By applying the DLSC Classifier, I am extending this expectation model in my research to find seasonal infection-affected persons with issue regions (Dynamic Learning and managed classifier). With the use of the dynamic learning model, I've been able to identify which persons are most likely to get infected with the flu this season based on their symptoms. As far as we are aware, this model is the only one that accounts for the identification of individually significant cases of the disease in a dataset. Straight hotspot disclosure scalability may now be achieved with new algorithmic modifications (i.e., fragmenting persons and regions differentiating evidence). A cost model for the suggested computations is also shown to show that our proposed algorithmic enhancements are correct. Examining obligations, in particular, include these:

Dynamic division allows us to deal with sub-edge hotspots in the dataset in a manner that is consistent with the suggested method.

New algorithmic enhancements, including an area channel and pruning computation, are also included to increase the adaptability of direct hotspot detection with dynamic division, which includes an area channel.

We give a cost study and scientific proof of the master's estimates' correctness.

A comparison of the finding outcomes under powerful division with those found in related research, such as area recognition, will be presented in two contextual investigations.

## References

1. Rai, B.K.: Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records. Int. J. Reliab. Qual. E-Healthc. (IJRQEH) **11**(1), 1–15 (2022)
2. Rai, B.K.: Ephemeral pseudonym based de-identification system to reduce impact of inference attacks in healthcare information system. Health Serv. Outcomes Res. Methodol. 1–19 (2022)
3. Rai, B.K.: PcBEHR: patient-controlled blockchain enabled electronic health records for healthcare 4.0. Health Serv. Outcomes Res. Methodol. 1–23 (2022)
4. Rai, B.K., Tyagi, A., Arora, B., Sharma, S.: Blockchain based Electronic Healthcare Record (EHR). In: ICCCE 2021 Lecture Notes in Electrical Engineering, vol. 828. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7985-8_19
5. Thakkar, B.A., Hasan, M.I., Desai, M.A.: Health care decision support system for swine flu prediction using naïve bayes classifier. In: 2010 International Conference on Advances in Recent Technologies in Communication and Computing, pp. 101–105. IEEE (2010)
6. Baker, Q.B., Shatnawi, F., Rawashdeh, S., Al-Smadi, M., Jararweh, Y.: Detecting epidemic diseases using sentiment analysis of arabic tweets. J. Univers. Comput. Sci. **26**(1), 50–70 (2020)
7. Alsmadi, T., Alqudah, N., Najadat, H.: Prediction of Covid-19 patients states using Data mining techniques. In: 2021 International Conference on Information Technology (ICIT), pp. 251–256. IEEE (2021)
8. Nouira, K., Njima, N.B.: FluSpider as a new vision of digital influenza surveillance system: based on Big Data technologies and Massive Data Mining techniques. In: 2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA), pp. 1–10. IEEE (2020)
9. Haque, T.H., Haque, M.O.: The swine flu and its impacts on tourism in Brunei. J. Hosp. Tour. Manag. **36**, 92–101 (2018)
10. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of Twitter to track levels of disease activity and public concern in the US during the influenza A $H_1N_1$ pandemic. PLoS ONE **6**(5), e19467 (2011)
11. Peng, Y., Li, C., Rong, Y., Chen, X., Chen, H.: Retrospective analysis of the accuracy of predicting the alert level of COVID-19 in 202 countries using Google Trends and machine learning. J. Glob. Health **10**(2) (2022)
12. Biswas, S.K., Sinha, N., Baruah, B., Purkayastha, B.: Intelligent decision support system of swine flu prediction using novel case classification algorithm. Int. J. Knowl. Eng. Data Min. **3**(1), 1–19 (2014)
13. Boulos, M.N.K., Sanfilippo, A.P., Corley, C.D., Wheeler, S.: Social Web mining and exploitation for serious applications: technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Comput. Methods Programs Biomed. **100**(1), 16–23 (2010)
14. Ong, S.Q., Pauzi, M.B.M., Gan, K.H.: Text mining in mosquito-borne disease: a systematic review. Acta Tropica 106447 (2022)

15. Nagaraj, P., Prasad, A.K.: Survey on swine flu prediction. Int. J. Manag. Technol. Eng. **9**, 937–941 (2019)
16. Ali, A.A., Hassan, H.S., Anwar, E.M.: Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique. In: 2020 12th International Conference on Electrical Engineering (ICEENG), pp. 145–150. IEEE (2020)
17. Olson, D.R., Konty, K.J., Paladini, M., Viboud, C., Simonsen, L.: Reassessing Google Flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput. Biol. **9**(10), e1003256 (2013)
18. Culotta, A.: Detecting Influenza Outbreaks by Analyzing Twitter Messages (2010). arXiv:1007.4748.
19. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: 2010 2nd International Workshop on Cognitive Information Processing, pp. 411–416. IEEE (2010)
20. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707. IEEE (2011)
21. Venna, S.R., Tavanaei, A., Gottumukkala, R.N., Raghavan, V.V., Maida, A.S., Nichols, S.: A novel data-driven model for real-time influenza forecasting. IEEE Access **7**, 7691–7701 (2018)
22. Albahri, A.S., Hamid, R.A., Al-qays, Z.T., Zaidan, A.A., Zaidan, B.B., Albahri, A.O., Madhloom, H.T.: Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. J. Med. Syst. **44**(7), 1–11 (2020)
23. Saikumar, K., Rajesh, V., Babu, B.S.: Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal 39(1), 31–42 (2022). https://doi.org/10.18280/ts.390104
24. Kailasam, S., Achanta, S.D.M., Rama Koteswara Rao, P., Vatambeti, R., Kayam, S.: An IoT-based agriculture maintenance using pervasive computing with machine learning technique. Int. J. Intell. Comput. Cybern. **15**(2), 184–197 (2022)
25. Saikumar, K., Rajesh, V.: A machine intelligence technique for predicting cardiovascular disease (CVD) using radiology dataset. Int. J. Syst. Assur. Eng. Manag. (2022). https://doi.org/10.1007/s13198-022-01681-7
26. Nagendram, S., Nag, M.S.R.K., Ahammad, S.H., Satish, K., Saikumar, K.: Analysis for the system recommended books that are fetched from the available dataset. In: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1801–1804. IEEE (2022)
27. Shravani, C., Krishna, G.R., Bollam, H.L., Vatambeti, R., Saikumar, K.: A novel approach for implementing conventional LBIST by high execution microprocessors. In: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 804–809. IEEE (2022)
28. Kiran, K.U., Srikanth, D., Nair, P.S., Ahammad, S.H., Saikumar, K.: Dimensionality reduction procedure for bigdata in machine learning techniques. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 836–840. IEEE (2022)
29. Saikumar, K., Rajesh, V.: A novel implementation heart diagnosis system based on random forest machine learning technique. Int. J. Pharm. Res. **12**, 3904–3916 (2020)