Artificial Intelligence and Machine Learning Techniques Using Omics Data for Cancer Diagnosis and Treatment

2

Priyanka Gawade, Sutanu Nandi, Chandrakala Meena, and Ram Rup Sarkar

Abstract

Cancer is a heterogeneous disease concerning molecular, functional and clinical behaviour, and poses a challenge for timely detection and treatment. Early detection and prognosis of cancer type may facilitate refined clinical management of cancer treatment. Recent technological development, such as next-generation sequencing, generated a large number of omics datasets in cancer genomics. The genome-wide biological information, such as cancer driver mutations, aberrantly methylated regions, gene, and miRNA expression profiles, is helpful for predicting the cancer onset, subtypes, and treatment response and is valuable for improving diagnosis and therapeutic and clinical decisions. In this context, machine learning (ML) algorithms and artificial intelligence have been beneficial and essential for the better accuracy of cancer-related predictions. Here, we mainly focus on research based on these omics data, paying close attention to machine learning methods. We summarize various kinds of omics data and different ML algorithms effective in cancer prediction. We also highlighted the

P. Gawade
Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, Pune, Maharashtra, India

Bioinformatics Centre, Savitribai Phule Pune University, Pune, Maharashtra, India

S. Nandi · R. R. Sarkar (✉)
Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, Pune, Maharashtra, India

Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh, India
e-mail: rr.sarkar@ncl.res.in

C. Meena
Chemical Engineering and Process Development, CSIR-National Chemical Laboratory, Pune, Maharashtra, India

S. Singh (ed.), *Systems Biomedicine Approaches in Cancer Research*,
https://doi.org/10.1007/978-981-19-1953-4_2

applications of the ML algorithm on genomic information in cancer, including cancer classification, therapy response, survival, metastasis, and biomarker identification. Further we discussed the novel approaches in machine learning for improving cancer prediction. These data-driven approaches can potentially provide a new solution for enhancing the precise treatment of cancer.

2.1 Introduction

Cancer shows significant disease burden globally due to its high prevalence and death rate. It occurs due to the development of atypical cells that divide in an uncontrolled manner. A central feature of cancer malignancy is metastasis. In the metastasis stage, cancerous cells leave their pre-neoplastic lesions, enter the bloodstream, disseminate throughout the body, and acclimate to new cellular surroundings in a secondary site, ultimately destroying the normal body tissue (Kang and Pantel 2013; Welch and Hurst 2019). These abilities of cancer cells viz., dissemination and invasion, eventually prove fatal to the host.

Cancer is a multistep and progressive disease in which gene expression alters because of the accretion of numerous genetic and epigenetic aberrations within a genome. The genomic complexity of cancer cells arises due to intrinsic factors and/or extrinsic factors that cause gross-scale abnormalities, i.e., variation in chromosome numbers (including aneuploidy and whole-genome duplication) (Hasty and Montagna 2014). Also, small-scale/local changes, i.e., genome rearrangements (consist of gene amplification, deletions, and non-reciprocal translocations), occurs due to causative agents and are responsible for genomic complexity. In addition to this, aberrant alterations in genes encoding epigenetic players that control epigenetic mechanisms are also responsible for increasing the complexity of cancer by causing the inappropriate onset (initiation/inhibition) of genetic expressions and promoting tumorigenesis. The epigenetic changes modify DNA (via methylation), histones (by post-translational modifications PTMs, namely methylation, acetylation, and phosphorylation, etc.), and non-coding RNAs (small and long ncRNAs) regulations and nucleosome remodeling, to form a regulatory system that controls accessibility between DNA elements and histones/non-coding RNAs (Ilango et al. 2020; Lu et al. 2020). The epigenetic players that participate in these modifications are susceptible to extrinsic factors, and changes caused by these players are reversible. These genetic and epigenetic alterations are often found in two kinds of genes, namely, proto-oncogenes and tumor suppressor genes. The activation changes like gain-of-function mutations and hypomethylation converts the proto-oncogenes into oncogenes (OGs), which are overactive positive cell cycle regulators responsible for cell survival, growth, and division, ultimately leading to cancer progression. The changes like loss-of-function mutations, epigenetic silencing like hypermethylation, proteasomal degradation by ubiquitination, and abnormal cellular localization of tumor suppressor genes (TSGs) leads to their inactivation. As a consequence of this, tumor development occurs due to elimination of negative regulatory proteins that

usually restrict cell growth (by apoptosis or activating DNA repair and cell cycle checkpoint) (Wang et al. 2018; Kontomanolis et al. 2020). Numerous studies have been done for identifying the genetic and epigenetic changes in different cancer types. For example, glioblastoma is associated with genetic alterations in the number of tumor suppressors, viz., PTEN, TP53, PIK3R1, NF1, RB1, and oncogenes, i.e., EGFR, PIK3CA, and IDH1 (Zhang et al. 2019a). The other tumor suppressor genes such as BRCA1/2, P53, PTEN, ATM, Rb, LKB, Nm23, P16, and oncogenes like HER2, c-MYC, and ERBB2, MYC, PIK3CA are very frequently mutated in breast cancer (Oliveira et al. 2005; Perera and Bardeesy 2012).

The complexity of cancer is further enhanced due to *tumor heterogeneity* that can occur during cancer evolution. Tumor heterogeneity is of two types as follows: (a) *Intra-tumor heterogeneity*, in which subsets of cancer cells within a tumor of a single patient possess discrete phenotypic and molecular characteristics and (b) *Inter-tumor heterogeneity*, which comprises tumor genotype variations among tumors of the same histological type between different patients (Meacham and Morrison 2013). This heterogeneity can arise from genetic, epigenetic, transcriptomic, or phenotypic changes (McQuerry et al. 2017). Genomic-level studies of *tumor heterogeneity* showed that cells in a tumor are highly diverse, spatio-temporally by analyzing their genetic variations like single-nucleotide variants (SNV), insertion–deletion mutations (indels), and copy number variation (CNV) (Murtaza et al. 2015; Li et al. 2017). Several studies provided information on *epigenetic heterogeneity* by inspecting DNA methylome and micro-RNA (miRNA) pools (Liu et al. 2018; Dietz et al. 2019; Wang et al. 2019; Guo et al. 2019; Alfardus et al. 2021). Studies of *tumor heterogeneity* at the transcriptome level revealed variation in the gene expression pattern of particular pathways like cell cycle, MAPK signaling pathway, immune/complement system pathways, and biological programs, namely hypoxia and epithelial–mesenchymal transition (EMT) (Patel et al. 2014; Zhang et al. 2016). Some studies supported the proteomic heterogeneity of tumors, but it is less prominent than genomic and transcriptomic heterogeneity (Ahmed et al. 2016; Sood et al. 2016). *Tumor heterogeneity* also includes heterogeneity of the tumor microenvironment (consists of endothelial cells, fibroblasts, adipocytes, immune cells, mesenchymal stroma/stem-like cells, and extracellular matrix), that sends physical and chemical signals to tumor cells and influences epigenetic machinery (Hass et al. 2020). Such a dynamic and highly variable nature of cancer hinders diagnosis and prognosis and leads to treatment resistance, relapse, and eventually death (Dagogo-Jack and Shaw 2018; Marusyk et al. 2020). Hence, understanding the mechanism of cancer development at different biological levels and early prediction of cancer may help in designing better therapeutic strategies.

2.2 Omics Data in Cancer Research

Advancements of high-throughput sequencing (or next-generation sequencing, NGS) techniques and the availability of omics data provide genome-wide measurements of genomic features (including genetic variants, DNA methylation,

and transcripts etc.) at various levels and resulted in remarkable progress in cancer research. Several databases are available online which provide free access to genomic information related to cancer. Among these, a few popular and important resources are listed in Table 2.1.

In order to properly analyze various kinds of omics data and to perform exploratory analysis, several computational tools are freely available online (see Table 2.1). Integration of different omics data will decode interrelationships between these features and their functions. This holistic approach seems to be promising to understand cancer development, recurrence, therapy response, and patient survival. The subsequent sections will discuss different types of omics data produced by various high-throughput sequencing approaches.

2.2.1 Genomic Data

Genomic information helps to unravel functional information present in DNA sequences.

2.2.1.1 Genomic Variation Data

Genetic variation is an alteration in the nucleotide order of DNA sequences that occur either due to mutation or genetic recombination. It can be grouped into following classes on the basis of size: (1) Small-scale sequence variation (<1 kb) consists of single-nucleotide variants (SNV), single nucleotide insertions/deletions (indels), etc., (2) Large-scale structural variation includes copy number variations (CNV) (loss or gain) and chromosomal rearrangement (genomic inversions, translocations) (Cardoso et al. 2015). SNVs are the most prevalent variants and can be present in different genomic locations: (1) protein-coding sequences, (2) non-coding regions like splice sites, promoters, ribosome binding sites, etc. Indels cause frameshift mutations within a coding region, whereas chromosomal rearrangements affect the spatial organization of chromosomes and cause nuclear reorganization. This kind of genomic variation is a fundamental constituent of genomics data and provides an opportunity to explore associations between genes, tissues, individuals, and phenotypes. DNA-sequencing (DNA-seq) techniques have been used to study genomic alterations, which include whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted massively parallel sequencing (TS) (Lightbody et al. 2019). WGS technique analyzes entire genomes and allows investigation of changes within coding and regulatory sites (Meienberg et al. 2016). It offers identification of CNVs, chromosomal rearrangements, and other structural variations that may be missed by targeted sequencing. This technique provides global insight into novel genomic changes in cancer samples as it gives the base-pair resolution of complete cancer genome in a single run (Zhao et al. 2019). Yates et al. (2017) performed WGS of primary as well as metastatic tumor samples of breast cancer and observed that cell clones causing metastasis or relapse migrate late from the primary tumors; however, they constantly gain alterations, mostly in the same biological process as the primary tumor. Another WGS study of

Table 2.1 Resources (data repositories and analysis tools) for cancer genomics study

Database/tools	Features	Link	References
The Cancer Genome Atlas (TCGA)	Exhaustive data repository of genomic, epigenomic data of cancer and control samples	https://portal.gdc.cancer.gov/	Tomczak et al. (2015)
Gene Expression Omnibus (GEO)	Public repository of genomic and proteomic data from array- and sequencing-based techniques	https://www.ncbi.nlm.nih.gov/geo/	Barrett et al. (2012)
International Cancer Genome Consortium (ICGC)	Data portal consists of somatic mutations and molecular data of major tumor types for competent visualization and analysis	https://dcc.icgc.org/	Zhang et al. (2019b)
Database of DNA methylation and gene expression in human cancer (MethHC)	DNA methylomes and mRNA/microRNA expression database; provides clinical and genomic variation data; multiplicity of information present	https://awi.cuhk.edu.cn/~MethHC/methhc_2020/php/index.php	Huang et al. (2021)
The database of human DNA methylation and cancer (MethyCancer)	Database comprises of DNA methylation data, cancer-related gene and mutations; also provides an efficient visualization tool, MethyView	http://methycancer.psych.ac.cn/	He et al. (2007)
Chinese Glioma Genome Atlas (CGGA)	Database contains mRNA/miRNA expression profiles and DNA methylation data of brain tumors from Chinese cohorts	http://www.cgga.org.cn/	Zhao et al. (2021)
UCSC Xena	Graphical viewer for gene- and genomic-coordinate across multiple data types of tumors	http://xena.ucsc.edu/	Goldman et al. (2020)
cBioPortal	Data portals provide genetic alterations across samples, genes, and pathways by analyzing multi-omics cancer data	https://www.cbioportal.org/	Gao et al. (2013)
SomamiR	Comprehensive resource for somatic and germline alterations in miRNA and their target sites in cancer	https://compbio.uthsc.edu/SomamiR/	Bhattacharya et al. (2013)
Database of Epigenetic Modifiers (dbEM)	Data resource for genomic information of epigenetic modifiers in cancer and healthy samples	https://webs.iiitd.edu.in/raghava/dbem/	Nanda et al. (2016)

glioblastoma (GBM) tumors identified novel non-coding constraint mutations for GBM-associated genes (Sakthikumar et al. 2020). In contrast to WGS, targeted sequencing approaches examine specific genomic regions of interest for the detection of rare variants and include WES and TS. WES covers coding genomic portions (i.e., genes and their flanking regions) to find out disease-causing variants in these portions (Gupta et al. 2017; Mueller et al. 2018). Mainly, WES is useful for identifying indels and SNV/SNPs inside the genome's coding sites. TS technique is helpful when prior information of disease is available and performed on particular locations of the genome (Davis et al. 2021). Recently, Weigelt et al. (2018) performed WES of breast tumors and TS of 410 breast cancer genes to investigate the somatic changes and the phenotypic characteristics associated with breast cancer which is originated from ataxia–telangiectasia (ATM) germline mutation. Garrett et al. (2020) carried out WES study of GBM tumor samples to analyze their genetic profile and correlated this information with drug treatment response to develop personalized treatments against GBM. Targeted sequencing was used to identify somatic mutations and CNV alterations in 30 genes which are most frequently altered in gliomas in order to detect biomarkers associated with the long-term survival of GBM patients (Cantero et al. 2018).

2.2.2 Epigenomic Data

Epigenomic information is useful to map the dynamic state of the genome in order to elucidate phenotypic characteristics observed via gene expression studies.

2.2.2.1 DNA Methylation Data

DNA methylation process is an epigenetic mechanism which incorporates a methyl (CH_3) group into the cytosine residue of DNA via the action of DNA methyltransferase enzymes. It controls gene expression and chromatin remodeling by influencing the interactions of DNA with histone or specific transcription factors. Whole-genome bisulfite-sequencing (WGBS) is a high-throughput technique used to quantify genome-wide DNA methylation. It provides a higher resolution to allele-specific DNA methylation as compared to DNA methylation assays and DNA microarrays. This technique allows identification of differentially methylated positions (DMPs) and differentially methylated regions (DMRs) which are the genomic positions/regions having distinct of DNA methylation levels in various biological circumstances (Wu et al. 2015). These DMPs and DMRs in disease conditions are useful for the development of potential epigenetic biomarkers which may help in early detection and diagnosis. The methylation changes in circulating DNA of metastatic breast cancer were studied using WGBS and found 21 DNA hypermethylation hotspots that could be potential blood-based biomarkers (Legendre et al. 2015). Bam et al. (2021) analyzed the global methylation status of both tumor-infiltrating and blood CD4+ T-cell from glioblastoma patients. The study found that the epigenetic modifications in tumor-infiltrating helper T-cells are affected by tumor cells.

2.2.2.2 Histone Modification Data

Chromosomal DNA tightly wraps around histone proteins and forms a chromatin structure in the nucleus. The post-translational modifications (PTMs) of histone proteins are crucial in chromatin remodelling which influence transcription. There are two mechanisms by which histone modifications exert their effect: (1) by directly altering overall chromatin structure either over short or long distances and (2) regulating (either positively or negatively) the binding of histone modifiers (Bannister and Kouzarides 2011). The detection of various histone modifications enables a greater understanding of epigenetic regulation and leads to the development of therapeutic strategies against histone-modifying enzymes. Chromatin immunoprecipitation-sequencing (i.e., ChIP-seq), an effective method for detecting DNA, targets for histone modifications as well as for transcription factors (TFs) at genomic scale with base-pair resolution (O'Geen et al. 2011). It identifies differences in the histone modification patterns which help in understanding epigenetic mechanisms that regulate various biological processes in diseases and thus a powerful tool to analyze chromatin structure and gene expression. ChIP-seq data also reveals how the genome is organized and the functional domains across the entire genome which aid in predicting and validating a set of large, non-coding RNAs. Xi et al. (2018) used the ChIP-seq technique to profile the distributions of 8 key histone modifications (i.e., H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3 and H3K79me2) across 13 breast cancer cell lines and from the epigenetic landscape of 5 molecular subtypes of breast cancer defined subtypes-specific key chromatin signatures to determine potential biomarkers. ChIP-seq analysis of histone H3 Lys27 acetylation (H3K27ac) revealed that alteration in the metabolite acetyl-CoA stimulates site-specific regulation of H3K27ac through which acetyl-CoA impacts the expression of distinct sets of genes associated with malignant phenotypes of glioblastoma, i.e., cell adhesion and migration (Lee et al. 2018).

2.2.3 Transcriptomics Data

Gene expression data are useful to obtain information on the abundance of complete sets of RNA transcripts that are produced by the genome within a biological sample simultaneously.

2.2.3.1 Transcript Profiling Data

The RNA molecules are used to form proteins that serve a crucial part of the cell. Thus, RNA expression reveals active transcription of cell and core activities in cells and tissue under specific conditions. Different types of RNA molecules present in eukaryotic cells play different biological functions like: mRNAs—carry the genetic blueprint from a cell's DNA to its ribosomes to make protein; microRNAs (miRNAs)—involve in gene silencing by repressing translation; long ncRNAs (lncRNAs)— involve in regulating chromatin function, modulating mRNA translation and also interfere with signalling pathways by acting as decoys, scaffolds or

enhancer RNAs. RNA sequencing (RNA-seq) technique is useful to study expression level of transcripts under particular conditions, namely, different environmental conditions, disease scenarios, and therapeutics exposure etc. Analysis of transcriptome data reveals which genes are activated or silent in cells/tissue (qualitative information) and to what extent genes are expressed (quantitative information) (Wang et al. 2009). RNA-seq methods provide information on differentially expressed genes to detect both known and novel transcripts. The profiling of mRNA molecules can be done using several RNA-seq assays viz., mRNA-seq, single-cell RNA-seq (scRNA-seq), strand-specific RNA-seq, ultra-low input RNA-seq and isoform sequencing (Iso-seq). However, the small RNA-seq technique is useful for expression profiling of small non-coding RNAs (like miRNA, siRNA, and piRNA). Total RNA-seq technique provides genome-wide expression data of both coding and non-coding RNAs. For example, using total RNA-seq technology, miRNA associated with metastatic breast cancer response to systemic treatment was identified based on miRNA count (Martinez-Gutierrez et al. 2019). Gao et al. (2021) showed that circular RNA (circRNA)-encoded unique E-cadherin variant *circ-E-Cad* (C-E-Cad) activates oncogenic EGFR signalling by directly binding to it and contributes to glioma stem cell tumorigenicity. Recently, Ren et al. (2021) discussed usage of scRNA-seq technology in breast cancer heterogeneity, metastasis, drug resistance, and prognosis and highlighted the importance of scRNA-seq for development of better treatment strategies.

High-throughput technologies generate massive amount of omics data which present a challenge due to its high dimension and redundancy. There is still a gap in understanding of these data that are often publicly and freely available. The traditional simplex classification algorithms are not suitable to handle large data sets as they contain a small sample size and large gene count. In this scenario, machine learning-based methods provide an excellent tool for analyzing such large and complex data, thus promoting clinical diagnosis and precision medicine against cancer.

2.3 Machine Learning Approaches

Nowadays, machine learning (ML), a subset of artificial intelligence (AI), is extensively applied in growing areas of healthcare, like medical imaging and gene expression pattern analysis, etc. and is extremely useful for high-dimensional data analysis and prediction. It is a data driven approach, which handles large datasets and automatically learns inherent patterns in the data that are useful to make decisions for new sets of data (Witten and Frank 2000). These characteristics make ML a suitable approach to design effective strategies for cancer diagnosis and treatment. Recent developments in ML models have indicated pronounced potential in preclinical conditions. The following sections give details of machine learning algorithms and their applications in cancer research.

The terminologies used in machine learning are mentioned below:

Dataset: It is a matrix containing features from which the machine learns and class label/target to predict. Each column in the matrix represents a feature or target, whereas each row represents an instance/observation. An initial dataset from which the model learns any relationships between features and targets during model training is called as training dataset. However, testing dataset is a subset of data which is not provided during model training but is useful for unbiased model evaluation by comparing predictions with the true value of the dataset.

Instance: An observation or data point is denoted as instance.

Feature/Attribute/Variable: This describes instances by measurable values and acts as input for prediction.

Target/Class Label: A value of an observation that a machine learns to predict is called as target or class label. For example, molecular subtype identification of breast cancer is a multi-classification task. Here, four class labels, i.e., luminal A, luminal B, HER2, and triple negative, are available.

Cross-Validation (*CV*): It is a technique that uses a subset of the original dataset for model training and utilizes other subset for model evaluation. This is generally useful to reduce model overfitting during training time. This method generates a fixed number of subset (fold) of data and performs the analysis for each subset. Further, it averages the final error estimate. Types of cross-validation methods are mentioned below.

a) k-Fold Cross-Validation: The *k*-fold cross-validation method performs random splitting on the original dataset to generate k equal size subsets and uses $(k - 1)$ subsets for training. For the testing purpose, it uses one subset.

b) Leave One Out Cross-Validation: The "leave one out cross-validation" method selects one instance from the original dataset for testing and the remaining instances for model training. The iteration is performed for each instance, and the final outcome is the average of results obtained from each iteration.

c) Bootstrap Cross-Validation: In this method, the complete original dataset is used for model training with sample replacement technique, and the remaining instances are used for model testing.

Machine learning algorithms are mainly grouped into following three categories on the basis of the availability of class labels/targets (Kotsiantis et al. 2007).

1. *Supervised Learning Algorithms*: It uses known targets during training and a model learn the relationship between features and targets. This information can be used for predicting unknown instances.
2. *Unsupervised Learning Algorithms*: Targets for unsupervised machine learning algorithms are unknown. It is used to find hidden structures/patterns or groups of similar samples during training the model. For clustering and pattern detection in biological research, these algorithms are mostly applied and also, useful for identification of gene signature in cancer and survival prediction.
3. *Semi-Supervised Learning Algorithms*: In this case, limited class labels are available, and thus, both labeled and unlabeled data are used during model building to improve accuracy. These algorithms are self-learning and show great potential in cancer prediction problems.

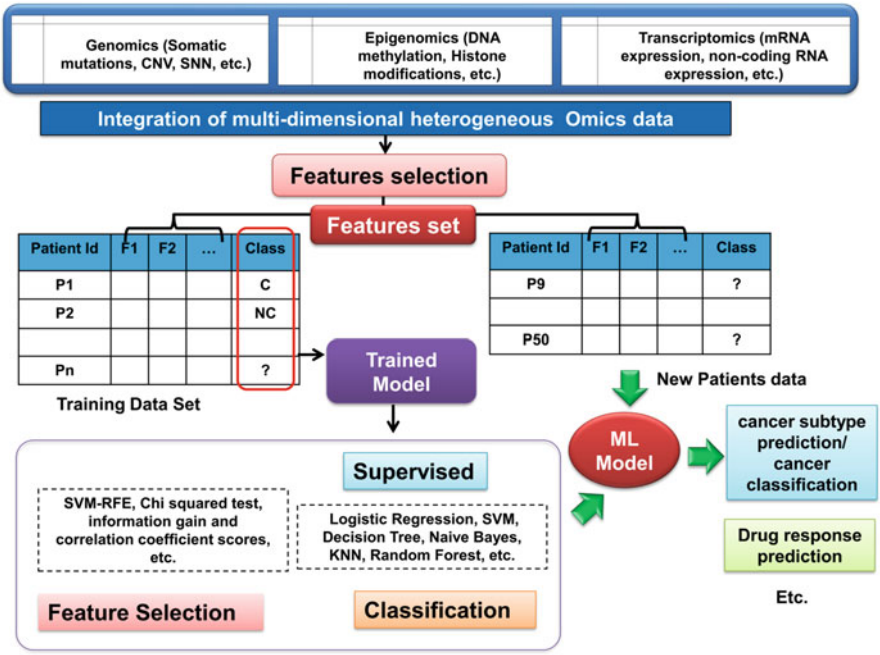

Fig. 2.1 Schematic overview of a machine learning workflow for cancer prediction using multi-omics data. Cancer patient's omics data, i.e., genomics, epigenomics, and transcriptomics, can act as input for machine learning models. Once a model is trained, it can be used to make predictions like cancer classification and drug response etc. for new patient's data

Depending on the nature of target values, above-discussed ML algorithms are further divided into either classification or regression type. The classification algorithms are used to predict categories of new instances by training the input dataset. However, regression algorithms learn from input datasets and then predict the outcome for continuous values. The schematic representation of machine learning workflow in the case of cancer prediction is given in Fig. 2.1.

The commonly used methods during model construction for improving model performance are discussed below.

2.3.1 Feature Selection Methods

Post-genomics era generated a large amount of transcriptomic, mutational, copy number variation (CNV), DNA methylation, histone modification, and miRNA expression data from various high-throughput techniques when applied on cancer cell lines or patients. These different types of data act as features in machine learning models and hold predictive power. To improve prediction accuracy, the feature selection method selects relevant features and removes irrelevant features present

in the original dataset without changing its original value. This method is very important when a dataset contains a large number of features. In such cases, there is no need to give every feature to the algorithm but only important ones for model prediction. This step will make the algorithm to perform fast, and will decrease the model complexity, increase model accuracy, reduce overfitting and simplify interpretation.

The feature selection is mainly grouped into three classes, namely, filter, wrapper, and embedded (Hira and Gillies 2015).

Filter Method: This feature selection algorithm employs some ranking over features that decide the importance of each feature for prediction. In this way, it selects best N features without depending on any ML algorithms. This method is used as pre-processing step. Few examples of this method are Pearson's correlation, *t* test, variance thresholds, information gain (IG), and Bayesian networks.

Wrapper Method: This method selects the best N features using machine learning classifiers. It uses forward selection or backward elimination or bi-directional elimination techniques to decide which features to retain or remove.

Embedded Method: It combines the filter and wrapper techniques to check feature importance. This method is useful for avoiding the overfitting. The gradient boosting machine (GBM), ridge regression, recursive feature elimination (RFE), and LASSO are few examples of embedded feature selection algorithms.

2.3.2 Dimension Reduction Methods

In model construction, the feature selection method selects a subset of relevant features, resulting in a reduction in the dataset's dimension. It retains a subset of original features. However, in the case of high-dimensional data (i.e., 100 or 1000 features), the dimension reduction approach is used to reduce the high number of features into low numbers by transforming the original values. The implementation of this method will reduce computational time and provide quick visualization. Few commonly used dimension reduction techniques are, principal component analysis (PCA) (Pearson 1901), metric dimensional scaling (MDS) (Torgerson 1952), and t-distributed stochastic neighbor embedding (t-SNE) (Hinton and Roweis 2002). The high-dimensional data in the biological area like high-throughput gene expression data, can be analyzed using the above techniques, and some of them are discussed below.

Principal Component Analysis (Pearson 1901): This method uses the orthogonal transformation process to convert instances of correlated features into a group of linearly uncorrelated features. In this way, it reduces the dimension of the dataset with the most negligible information loss, and newly formed features are known as principal components. If data are nonlinear, kernel PCA is beneficial with nonlinear kernel mapping. PCA works well on the dataset which shows the Gaussian distribution.

Metric Dimensional Scaling (Torgerson 1952): This statistical method uses data that contains dissimilarities among pairs of instances. MDS denotes these

dissimilarities as distances among instances and obtain low dimension data points from the high-dimensional dataset by keeping pairwise distances the same.

2.3.3 Overview of Machine Learning Algorithms

2.3.3.1 Supervised Machine Learning Algorithms

Supervised machine learning algorithms have been used for cancer diagnosis and prognosis. Different supervised ML algorithms are available to analyze multi-omics data with categorical and quantitative variables in cancer research and build prediction models. Omics data of individual cancer patient at variety of molecular levels can also be used with these classifiers to develop personalized predictions; they are also useful for personalized predictions models. A detailed description of some of the supervised ML algorithms useful in cancer prediction/prognosis is given below.

Support Vector Machine (SVM)

SVM is a commonly applied supervised machine learning algorithm that searches hyperplane with maximal separation from each data class. Vapnik first described such a kind of classifier to classify data classes using only a hyperplane (Cortes and Vapnik 1995). The general principle of SVM is presented in Fig. 2.2a. SVM uses a multidimensional function known as kernel to transform input data points from the feature space to target space so as to differentiate complex real-life datasets. The classification, as well as regression problems, can be solved using SVM. The proper selection of kernel functions and their parameters significantly helps to improve the model performance.

The following function describes SVM:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i^+ + \xi_i^-\right)$$

$$s.t.\begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i^+ \\ y_i - f(x_i) \le -\varepsilon - \xi_i^- \\ \xi_i^+, \xi_i^- \ge 0 \end{cases},$$

where f, y, and ε represent prediction, actual class label, and free threshold parameter, respectively. The constant C is a coefficient of adjustment between the margin of separation and error on the hyper-plane. The ξ_i^+ and ξ_i^- parameters representing slack variables for error calculation.

Naive Bayes (NB)

Naive Bayes, another supervised ML algorithm (Rish 2001), is a probabilistic method based on Bayes' law. It assumes that a particular feature in a class is independent of another feature in the same class and each feature is equally

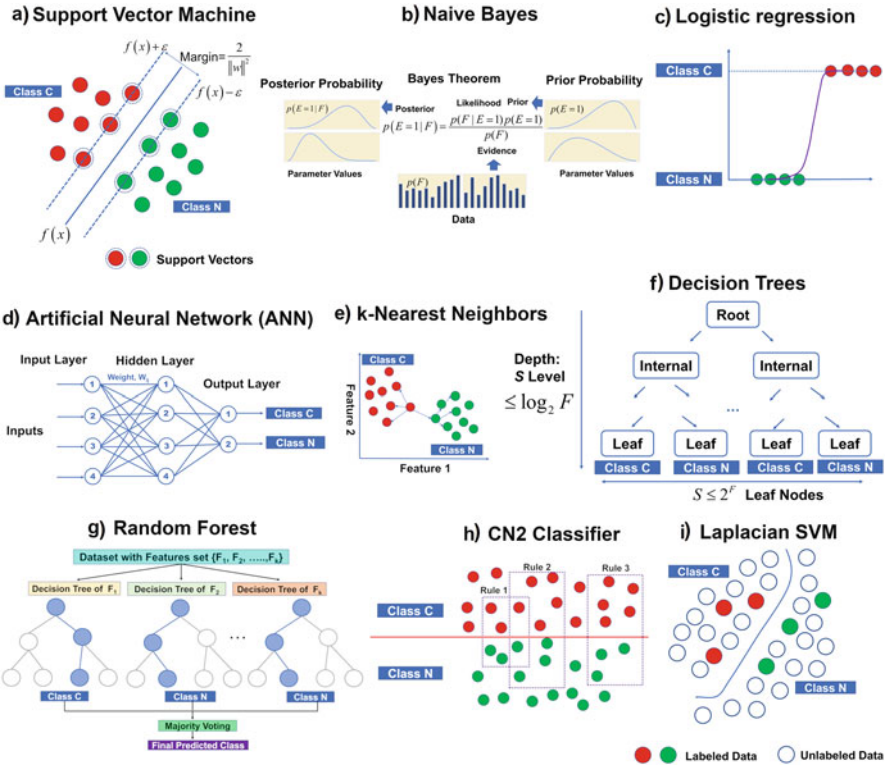

Fig. 2.2 The basic principles of different types of machine learning algorithms. (**a**) Support vector machine (SVM), (**b**) Naive Bayes, (**c**) logistic regression, (**d**) artificial neural network (ANN), (**e**) *k*-nearest neighbors (KNN), (**f**) decision tree, (**g**) random forest, (**h**) CN2, and (**i**) Laplacian SVM

contributing to target class. See Fig. 2.2b. This algorithm is used for classification purposes.

Naïve Bayes classifier can be defined as follows:

$$p(E=1|F) = \frac{p(E=1)\prod_{i=1}^{n} p(f_i|E=1)}{p(F)},$$

where, $F = (f_1, f_2, \ldots . f_n)$ denotes all the features, $p(E=1)$ is obtained from a training set and known as target class prior probability, $p(F)$ is the feature prior probability, $p(f_i|E=1)$ is likelihood that is probability of feature given target, and $p(E=1|F)$ is the posterior probability of target class given feature.

The below function finds class with maximum probability:

$$\text{classify}(f_1, f_2, \ldots . f_n) = \arg\max_{E=1,0} p(E) \prod_{i=1}^{n} p(f_i|E).$$

Logistic Regression Classifier

Logistic regression is a classification algorithm utilized for probability prediction of target class by logistic function (refer Fig. 2.2c). To use this algorithm, the target class must be categorical, and multi-collinearity should not be in features. This classifier helps to detect the best fitting model so as to represent the association between features and the target class.

The logistic function with the feature set $F = \{f_1, f_2, \ldots \ldots \ldots \ldots, f_n\}$ is

$$p(E = 1|F) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 f_1 + \ldots . \beta_n f_n)}}.$$

Artificial Neural Networks (ANNs)

ANN (Hagan et al. 1997), also known as a neural network or simulated neural network (SNN), simulates behaviour of the human nervous system. This computational network comprises numerous interconnected layers (i.e., multi-layer perceptron) that learn (without any programming), generalize training data, and give output from complex data. It mainly contains three layers, namely, (1) input layer, only one input layer in which input data are fed, (2) hidden layers, one or more hidden layers in which processing takes place to derive results based on the weighted sum of connections, and (3) output layer, demonstrating the results. Each layer consists of multiple processing units called nodes, which possess an "activation function" that converts input signal to output signal. Model performance gets affected by the number of nodes and hidden layers. Figure 2.2d shows the computational scheme of ANN.

ANN's objective function is as follows:

$$\arg\min_{w} E(w) = \frac{1}{2} \sum_{i=1}^{m} (N(w, x_i) - y_i)^2,$$

where x, w, and y represents input vector, weight between nodes, and target vector, respectively. ANN algorithms are useful for prediction, classification, regression, and pattern recognition.

k-Nearest Neighbors (KNNs)

The *k*-nearest neighbors are a distance-based algorithm as it first finds all the closest points around new unknown data point and calculates the distance between them to determine the class of new data points (Aha et al. 1991) (shown in Fig. 2.2e). The number of closest points near new unknown data points is denoted as "*k*" symbol, and fine-tuning of this value improves the model performance. This method helps to

solve the classification task by considering the majority of votes, while for the regression problem, KNN takes the mean for all the closest points.

Decision Trees (DTs)

Decision tree, a supervised machine learning algorithm, is a tree-structured classifier that continuously divides the data based on specific parameters. This classifier starts with the root node (i.e., entire dataset), which further expands based on features into a number of branches (represent decision rule) and finally forms leaf nodes (viz., final outcome) (Breiman et al. 2017). Figure 2.2f illustrates the decision tree classifier. Decision tree has two types as follows: (1) classification tree (for categorical class variable) and (2) regression tree (for continuous class variable).

The following measures in a decision tree are used to check the impurity of a node t:

$$\text{Entropy}(t) = -\sum_{i\in(0,\,1)} p(i|t)\log_2 p(i|t)$$

$$\text{Gini}(t) = 1 - \sum_{i\in(0,\,1)} [p(i|t)]^2$$

$$\text{Classification error} = 1 - \max_i [p(i|t)].$$

The gain ratio is as follows:

$$= \frac{I(\text{parent}) - \sum_{i=1}^{n} \frac{N(\text{child}_i)}{N} I(\text{child}_i)}{-\sum_{i=1}^{n} p(\text{child}_i)\log_2 p(\text{child}_i)}.$$

In the decision tree, the gain ratio is used to measure the goodness of a node's split. This measure decides which feature in a tree should be the parent node and which should be set down after being split as a child node.

Random Forest (RF)

As the name suggests, the random forest comprises multiple decision trees and can provide more accurate predictions by combining all of them (Fig. 2.2g). However, this algorithm solves the problem of overfitting associated with the decision tree. Each decision tree in a random forest makes prediction of a class, and the class with highest number of hits will be the prediction of model. If an optimal classifier is unfeasible, a random forest classifier is especially helpful (Ditterrich 1997; Breiman 2001). This classifier applies bagging and feature randomness to generate uncorrelated trees in a forest, ultimately giving a more accurate and stable result. It is used to solve classification and regression problems.

CN2 Classifier

The CN2 algorithm induces classification rules "if. . .then.." from data using entropy (Clark and Niblett 1989). This classifier is used only for classification purposes and works well with imperfect/noisy training data. Figure 2.2h gives the schematic representation of this classifier.

The advantages and disadvantages of above supervised ML algorithms are given in Table 2.2.

2.3.3.2 Semi-Supervised Classifier

Semi-supervised ML algorithms use a combination of supervised learning on a small amount of labeled data and unsupervised learning on large amount of unlabeled data (Chapelle et al. 2009). This approach is applicable when a large number of labeled data is not available and overcomes the drawbacks of supervised (i.e., require sufficient labels and costly process) and unsupervised (i.e., limited range of applications) algorithms. This algorithm works on the basis of any of these three assumptions, viz., (1) continuity assumption, data points around each other belong to the same class; (2) cluster assumption, data can be split into distinct clusters and data points in the same cluster tend to share class; and (3) the manifold assumption, assumes that data points are present on the manifold of lower dimensions than input space. The manifold assumption is useful in condition where data points may locate in high dimensions, and is very difficult to map data points in those dimensions. Semi-supervised classifier includes Laplacian SVM, generative models, and transductive SVM.

Laplacian SVM

Laplacian support vector machine (LapSVM) is based on a support vector machine algorithm and obeys manifold regularization (Belkin et al. 2006). This is a graph-based approach in which nodes are formed from labeled and unlabeled data. The KNN algorithm is employed to compute edge weight to define similarity between data points in a graph. Through this procedure unlabeled nodes can be labeled by transferring the information of labeled data points to other nodes. See Fig. 2.2i for pictorial representation of LapSVM.

LapSVM solves the following optimization problem.

$$\arg\min_{f \in H_k} \frac{1}{n_\mathrm{l}} \sum_{i=1}^{n_i} |1 - y_i f(x_i)|_+ + \lambda_\mathrm{a} \|f\|_K^2 + \frac{\lambda_\mathrm{b}}{(n_\mathrm{l} + n_\mathrm{u})^2} \times f^\mathrm{T} L f,$$

where $\|f\|_K^2$, n_l, n_u are a regularization function for smoothness, number of labeled data points, number of unlabeled data points, respectively, and λ_a, λ_b are hyperparameters.

$$\text{loss function} = |1 - y_i f(x_i)|_+ = \max(0, 1 - yf(x)),$$

Table 2.2 Advantages and disadvantages of different supervised ML algorithms

Algorithm name	Advantages	Disadvantages
Support Vector Machine (SVM)	• High prediction accuracy • Handle high-dimensional space • Generalized well with small amount of data • Less prone to condition of overfitting • Less influence of outliers	• Extensive memory required for optimization • Not appropriate for large datasets • High time complexity • Selection of proper kernel function is challenging • Difficult to fine-tune some hyper-parameters
Naive Bayes	• Its implementation is easy and simple • Computationally very fast • If conditional independence assumption holds, it quickly generates outcomes • Works well with categorical and continuous data	• The conditional independence assumption does not always hold in the complex biological problems • Not suitable for imbalanced data • Shows decrease in performance with increase in sample size of dataset
Logistic regression classifier	• Simplest algorithm to use • Very fast • Do not suffer from overfitting in case of low-dimensional dataset • Very efficient for linearly-separable dataset	• Causes model overfitting on high-dimensional dataset • Shows decrease in performance with increase in number of samples and features in dataset • Not suitable for non-linear data • Sensitive to outliers
Artificial Neural Network (ANN)	• Robust to noise • Shows good fault tolerance • Works well on complex nonlinear association among dependent and independent features • Able to perform parallel processing	• Training performance increases with increase in training dataset • Unexplained functioning • The algorithm may be stuck into local minima • Hardware dependent • Long training time is required • Difficult to determine network structure • Suffers from overfitting
k-Nearest neighbors	• Implementation is simple and easy • Fast, as no training time is require • Highly reserved for local information • Versatile as it performs classification, regression, and search tasks • Analytically tractable	• Huge storage space requires • Different values of *k* give different outcomes • Takes long computation time for large dataset • Larger *k* values increase the time complexity • Sensitive to noisy data and outliers • Difficult to work with high dimensional data • Standardization and normalization steps require
Decision Trees	• Simple, easy to understand and interpret • Can handle irrelevant features and nonlinear associations	• Small changes affect stability of decision tree structure • Suffer from overfitting without proper tree pruning

(continued)

Table 2.2 (continued)

Algorithm name	Advantages	Disadvantages
	• Not sensitive to missing values • Runs fast • No normalization and scaling require • Data preparation takes less efforts	• Stuck in local minima • Not suitable for regression problem and prediction of continuous values • Difficult to find optimal decision tree
Random Forest	• High predictive performance • Works well with both classification and regression problems • Efficiently handles large datasets • Reduces overfitting and variance • Easy to understand model predictions	• Complex thus requires more computational power and resources • Suffers from overfitting for noisy datasets • Takes more time than decision tree
CN2 classifier	• Implementation is simple and easy to understand • Can handle irrelevant features and nonlinear relationships • Works fast	• In case of large number of features, difficult to define rules for training datasets

$$\sum_{i,j=1}^{n} W_{ij}\left(f(x_i) - f\left(x_j\right)\right)^2 = f^T L f,$$

W_{ij}, is the edge weights in the graph,
Laplacian operator, $L = D - W$.

This classifier is less vulnerable to overfitting, robustness to noise and outliers, and has high prediction power and good generalization ability with small labeled data. However, it does not work well with large number of data points because it needs high memory to construct a graph and is time-consuming.

2.3.4 Model Performance Evaluation

The major part of building an effective ML model is evaluation of model's performance. ML requires evaluation metrics for selecting the best model. Following are the primary building blocks of several evaluation metrics, formed from confusion matrix, which is obtained from actual and predicted class labels:

True Positive (*TP*): It represents an outcome in which positive samples are accurately predicted as positive by the model.

True Negative (*TN*): It represents an outcome in which negative samples are accurately predicted as negative samples by the model.

False Positive (*FP*): It represents an outcome in which negative samples are incorrectly predicted as positive by the model.

False Negative (*FN*): It represents an outcome in which positive samples are wrongly predicted as negative samples by the model.

Based on these 4 outcomes, model performance metrics are given below.

True-Positive Rate (*TPR*) (*Also Known as Sensitivity*): The probability that positive samples will predict positive.

$$\text{True Positive Rate (TPR) or Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{TPR} \in [0, 1].$$

False-Positive Rate (*FPR*): The probability that negative samples will predict positive.

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{FPR} \in [0, 1].$$

Precision: It estimates positive sample predictions that are genuinely from the positive class label.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Precision} \in [0, 1].$$

Recall: It estimates positive sample predictions from all actual positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Recall} \in [0, 1].$$

F-Measure: It balances precision and recalls both together to provide a single score.

$$F\text{-measure} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{(\text{Precision} + \text{Recall})}, \quad F\text{-measure} \in [0, 1].$$

Accuracy: It gives the total correct predictions (TP + TN) made by the model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad \text{Accuracy} \in [0, 1].$$

The Area Under the Receiver Operating Characteristic Curve (auROC): It shows whether the model is capable of correctly discriminating between class labels.

$$\text{auROC} \in [0, 1].$$

Matthews Correlation Coefficient (*MCC*): It computes the correlation between actual and predicted labels and is calculated by the following formula.

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}, \quad \text{MCC} \in [-1, 1].$$

Sufficient labeled data should be available to get statistically significant measures. The above performance metrics are useful to evaluate various ML algorithms to check how good the model is in predicting the outcome.

To implement different machine learning algorithms, various functions are available in scikit-learn (Python) (Pedregosa et al. 2011; Kramer 2016), e1071 (R) (Dimitriadou et al. 2008; Meyer et al. 2019), and Weka (Java) (Witten et al. 1999; Dimov et al. 2007).

2.4 Application of AI and Machine Learning Techniques in Cancer

As mentioned previously, cancer is a heterogeneous disease and shows distinct molecular as well as phenotypic characteristics within a tumor. This heterogeneous nature of the tumor poses a challenge for successful treatment and recovery. Different machine learning and artificial intelligence techniques have been effectively applied to combat the disease.

2.4.1 Cancer Classification

In order to determine the proper treatment regime and to reduce cancer-related mortality, correct classification of cancer is needed. RNA-seq provides genome-wide gene expression data that can be useful to determine cancer types and unravel cancer subtypes, indicating a profound impact on cancer prediction/diagnosis.

However, gene expression data have several limitations like small sample sizes, large number of genes, and presence of some uninformative genes. All these factors decrease classification performance. This indicates the need for filtration and feature selection steps before model building. With the stringent threshold, these two steps ensure that only informative and sufficiently differentially expressed genes between the target classes can be used in building the classifiers. Various supervised and unsupervised algorithms are developed using gene expression data for cancer classification purposes. For instance, Flynn et al. (2018) identified primary site of 33 cancers and the molecular subtype of 11 cancers by applying several machine learning approaches like diagonal linear discriminant analysis (DLDA), KNN, RF, and SVM on gene expression profiles from the TCGA. The gene expression data of breast cancer provide an information based on which ML methods classified the disease into triple-negative breast cancer (TNBC) and non triple-negative breast cancer (non-TNBC) (Wu and Hicks 2021). The authors evaluated four different classification algorithms, namely SVM, KNN, NB and DT and found that SVM was able to divide TNBC and non-TNBC with less errors compared to others. Zhang et al. (2020) classified glioblastoma subtypes using SVM and RF with methylation data.

2.4.2 Anti Cancer Drug Response Prediction

The complexity of the tumor and its microenvironment lead to partial or no response to anti cancer drugs. Therefore, finding the relationship between drug response and molecular features of cancer cells or their microenvironment will be helpful in the identification of novel diagnostic/predictive biomarkers and evaluating drug response to guide personalized medicine. Miranda et al. (2021) used DNA methylation profiles at global scale from several cancer cell lines in the Genomics of Drug Sensitivity in Cancer (GDSC) database to predict eight anti cancer drug's cytotoxic responses by machine learning algorithms. Here, authors used RF, SVM, gradient boosting machines, and KNN for both classification and regression. The predictions made by the RF classifier were significantly correlated with Temozolomide drug responses for low-grade gliomas. Bomane et al. (2019) assessed ML algorithms, namely RF, XGB, LGBM, logistic regression LR, classification and regression tree (CART) on six molecular profiles (CPG and CGI DNA methylation, mRNA expression, miRNA profiles, isomiR expression, CNV) of breast tumors to predict paclitaxel response. A study found that DNA methylation and miRNA profiles out of six molecular profiles were the most informative overall.

2.4.3 Survival Prediction

Survival is the time during which a patient survives after disease diagnosis. Survival analysis is crucial in cancer patient management because of *tumor heterogeneity*. Integration of multi omics data and ML algorithms holds promise for improving the

survival of cancer patients. Mitchel et al. (2019) developed ML workflow using decision-level integration of multi omics tumor data to predict the overall survival of breast cancer patients. This study predicted the survival with an accuracy of 85% and area under the curve (AUC) of 87% with multi omics data and identified best integrated classification combination as methylation, miRNA, and gene expression. Recently, an auto-encoder was used to integrate and reduce the dimensions of pancreatic cancer patients' microRNA expression and DNA methylation data (Baek and Lee 2020). Machine learning models like SVM, RF, and LR and L2 regularized logistic regression were implemented to combine the clonal expansion of DNA mutations and multi omics data to predict cancer recurrence and survival within five years. This study revealed that mutated genes with low cellular prevalence (CP) values (i.e., mutated in smaller clones) were not significantly associated with recurrence and survival. However, the topmost CP value genes which usually mutated in the initial stages of tumor development were significantly related to poor prognosis in pancreatic cancer.

2.4.4 Metastasis Prediction

Cancer metastasis contributes to cancer-related mortality. Early prediction of it can improve prognosis. Most of the time, the metastasis prediction models use gene expression data. Recently, miRNA expression levels and DNA methylation patterns have also been explored for metastasis prediction. Tuo et al. (2018) used the SVM-based classifier on gene expression profiles to predict whether the breast cancer samples were metastatic or non metastatic, and prediction accuracy was evaluated by training and validating the model on TCGA data, an independent dataset. The mRNA- and miRNA-specific classifiers were used to differentiate cross-cancer tissue samples as primary or metastatic (Lee et al. 2019a). This study used three classification algorithms (LASSO, RF, and SVM) with bootstrap cross-validation method to determine how accurately mRNA and miRNA biomarkers can classify metastasis.

2.4.5 Biomarker Prediction

Cancer biomarkers may evaluate the risk of cancer development or progression in a specific tissue or therapeutic response. Thus, to decide appropriate therapy for cancer patients, the identification of cancer biomarkers is essential and may be useful for patients' survival. Tabl et al. (2019) used a machine learning multi class approach, the one-versus-rest technique to identify potential biomarkers which can increase breast cancer patients' survival. In this study, gene expression profiles of cancer patients who received different treatments like surgery, hormone therapy, and radiotherapy were used and also considered their status as living or deceased. The classifiers, namely random forest, SVM, and Naive Bayes, were implemented and found that random forest outperformed the others and showed a better classification

power for the hierarchical model. In another study, microRNA expression data were used for validating clinically selected miRNAs as breast cancer biomarkers using several machine learning classifiers (Rehman et al. 2019). The feature selection methods like information gain (IG), chi-squared (CHI2) and least absolute shrinkage and selection operation (LASSO) were implemented to rank the miRNAs by their importance and concluded that not all miRNAs carry equal weightage to act as a cancer biomarker, even among those clinically selected ones.

2.5 Conclusion and Future Directions

In the upcoming decade, advancement in artificial intelligence and proper implementation of machine learning methods in cancer genomics will reveal crucial aspects in oncology. Recent studies showed that the utilization of diverse omics data and their combination enhanced the cancer prediction performance of the machine learning models. Several challenges still exist, like data collections, pre processing, and storage. The collaboration between clinicians and bioinformaticians will be beneficial to get organized or structured data from various sources and at numerous scales. Recently, a study combined multi omics data with drug data to predict overall survival and different subtypes (at pathological, histological, and molecular levels) of glioma patients (Saurabh et al. 2020). This kind of comprehensive analysis can be helpful for doctors and clinicians in early diagnosis as well as in deciding the correct and personalized therapeutic strategies for individual cancer patients. In recent years, several studies integrated genomics data with pathological image data for identifying distinct cellular subtypes, prognostic biomarkers, mutational status, therapeutic strategies, and clinical outcomes. Applying a deep learning classifier on point mutation, copy number alteration, and gene expression data, Qu et al. (2021) predicted driver mutations and signaling pathways activity from histopathological whole slide images (WSI) of breast carcinoma patients. Recently, a new approach, radiogenomics, has emerged in the area of personalized medicine which integrates genetic and radiomic data for monitoring genetic variations in patients through medical images and can act as a better substitute for painful mediation (Shui et al. 2020; Gullo et al. 2020). Lee et al. (2019b) used machine learning classifiers on quantitative radiomic data of glioblastoma patients obtained from magnetic resonance images (MRI) and targeted sequencing of the IDH1 gene to predict its mutation status from images. The study showed a good (~80%) predictive power for IDH1 mutation by training 31 features of MRI and also observed good (i.e., 66.3–83.4%) accuracy through validation on an external set (Lee et al. 2019b). Such an integrative machine learning approach has an important role in improving diagnosis and prognosis. Nowadays, model explainability is gaining importance in the machine learning field as it explains the backend process of the model prediction, i.e., which particular features mainly contribute to the model prediction. The success of ML solutions may provide a handful of clinically relevant tools for cancer patient's treatment and management.

Acknowledgements CM acknowledges the Department of Science and Technology (DST) for the support as an Inspire Faculty (Award No. IFA19-PH248).

References

Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6(1):37–66

Ahmed N, Greening D, Samardzija C, Escalona RM, Chen M, Findlay JK et al (2016) Unique proteome signature of post-chemotherapy ovarian cancer ascites-derived tumor cells. Sci Rep 6(1):1–13

Alfardus H, de los Angeles Estevez-Cebrero CM, Rowlinson J, Aboalmaaly A, Lourdusamy A, Abdelrazig S et al (2021) Intratumour heterogeneity in microRNAs expression regulates glioblastoma metabolism. Sci Rep 11(1):1–14

Baek B, Lee H (2020) Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data. Sci Rep 10(1):1–11

Bam M, Chintala S, Fetcko K, Williamsen BC, Siraj S, Liu S et al (2021) Genome wide DNA methylation landscape reveals glioblastoma's influence on epigenetic changes in tumor infiltrating CD4+ T cells. Oncotarget 12(10):967

Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell Res 21(3):381–395

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M et al (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41(D1):D991–D9D5

Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7(11):2399–2434

Bhattacharya A, Ziebarth JD, Cui Y (2013) SomamiR: a database for somatic mutations impacting microRNA function in cancer. Nucleic Acids Res 41(D1):D977–DD82

Bomane A, Gonçalves A, Ballester PJ (2019) Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. Front Genet 10: 1041

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees. Routledge, New York

Cantero D, Rodríguez de Lope Á, Moreno De La Presa R, Sepúlveda JM, Borrás JM, Castresana JS et al (2018) Molecular study of long-term survivors of glioblastoma by gene-targeted next-generation sequencing. J Neuropathol Exp Neurol 77(8):710–716

Cardoso JG, Andersen MR, Herrgård MJ, Sonnenschein N (2015) Analysis of genetic variation and potential applications in genome-scale metabolic modeling. Front Bioeng Biotechnol 3:13

Chapelle O, Scholkopf B, Zien A (2009) Semi-supervised learning (Chapelle, O. et al., Eds.; 2006) [book reviews]. IEEE Trans Neural Netw 20(3):542

Clark P, Niblett T (1989) The CN2 induction algorithm. Mach Learn 3(4):261–283

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Dagogo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15(2):81–94

Davis AR, Stone SL, Oran AR, Sussman RT, Bhattacharyya S, Morrissette JJ et al (2021) Targeted massively parallel sequencing of mature lymphoid neoplasms: assessment of empirical application and diagnostic utility in routine clinical practice. Mod Pathol 34(5):904–921

Dietz S, Lifshitz A, Kazdal D, Harms A, Endris V, Winter H et al (2019) Global DNA methylation reflects spatial heterogeneity and molecular evolution of lung adenocarcinomas. Int J Cancer 144(5):1061–1072

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A (2008) Misc functions of the Department of Statistics (e1071), TU Wien. R Pack 1:5–24

Dimov R, Feld M, Kipp DM, Ndiaye DA, Heckmann DD. Weka: Practical machine learning tools and techniques with java implementations. AI tools Seminar University of Saarland WS University of Waikato, Hamilton 2007;6(07)

Ditterrich T (1997) Machine learning research: four current direction. Artif Intell Mag 4:97–136

Flynn WF, Namburi S, Paisie CA, Reddi HV, Li S, Karuturi RKM et al (2018) Pan-cancer machine learning predictors of primary site of origin and molecular subtype. bioRxiv 2018:333914

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6(269):pl1–ppl

Gao X, Xia X, Li F, Zhang M, Zhou H, Wu X et al (2021) Circular RNA-encoded oncogenic E-cadherin variant promotes glioblastoma tumorigenicity through activation of EGFR–STAT3 signalling. Nat Cell Biol 23(3):278–291

Garrett A-M, Lastakchi S, McConville C (2020) The personalisation of glioblastoma treatment using whole exome sequencing: a pilot study. Gene 11(2):173

Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A et al (2020) Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 38(6):675–678

Gullo RL, Daimiel I, Morris EA, Pinker K (2020) Combining molecular and imaging metrics in cancer: radiogenomics. Insights Imaging 11(1):1–17

Guo M, Peng Y, Gao A, Du C, Herman JG (2019) Epigenetic heterogeneity in cancer. Biomark Res 7(1):1–19

Gupta S, Chatterjee S, Mukherjee A, Mutsuddi M (2017) Whole exome sequencing: uncovering causal genetic variants for ocular diseases. Exp Eye Res 164:139–150

Hagan MT, Demuth HB, Beale M (1997) Neural network design. PWS Publishing, Boston

Hass R, von der Ohe J, Ungefroren H (2020) Impact of the tumor microenvironment on tumor heterogeneity and consequences for cancer cell plasticity and stemness. Cancer 12(12):3716

Hasty P, Montagna C (2014) Chromosomal rearrangements in cancer: detection and potential causal mechanisms. Mol Cell Oncol 1(1):e29904

He X, Chang S, Zhang J, Zhao Q, Xiang H, Kusonmano K et al (2007) MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res 36:D836–DD41

Hinton G, Roweis ST (2002) Stochastic neighbor embedding. NIPS, Toronto

Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform 2015:198363

Huang H-Y, Li J, Tang Y, Huang Y-X, Chen Y-G, Xie Y-Y et al (2021) MethHC 2.0: Information repository of DNA methylation and gene expression in human cancer. Nucleic Acids Res 49 (D1):D1268–D1D75

Ilango S, Paital B, Jayachandran P, Padma PR, Nirmaladevi R (2020) Epigenetic alterations in cancer. Front Biosci 25(1):1058–1109

Kang Y, Pantel K (2013) Tumor cell dissemination: emerging biological insights from animal models and cancer patients. Cancer Cell 23(5):573–581

Kontomanolis EN, Koutras A, Syllaios A, Schizas D, Mastoraki A, Garmpis N et al (2020) Role of oncogenes and tumor-suppressor genes in carcinogenesis: a review. Anticancer Res 40(11): 6009–6015

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160(1):3–24

Kramer O (2016) Scikit-learn. Machine learning for evolution strategies. Springer, Cham, pp 45–53

Lee JV, Berry CT, Kim K, Sen P, Kim T, Carrer A et al (2018) Acetyl-CoA promotes glioblastoma cell adhesion and migration through Ca^{2+}-NFAT signaling. Genes Dev 32(7–8):497–511

Lee SC, Quinn A, Nguyen T, Venkatesh S, Quinn TP (2019a) A cross-cancer metastasis signature in the microRNA–mRNA axis of paired tissue samples. Mol Biol Rep 46(6):5919–5930

Lee MH, Kim J, Kim S-T, Shin H-M, You H-J, Choi JW et al (2019b) Prediction of IDH1 mutation status in glioblastoma using machine learning technique based on quantitative radiomic data. World Neurosurg 125:e688–ee96

Legendre C, Gooden GC, Johnson K, Martinez RA, Liang WS, Salhia B (2015) Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. Clin Epigenetics 7(1):1–10

Li C, Wu S, Yang Z, Zhang X, Zheng Q, Lin L et al (2017) Single-cell exome sequencing identifies mutations in KCP, LOC440040, and LOC440563 as drivers in renal cell carcinoma stem cells. Cell Res 27(4):590–593

Lightbody G, Haberland V, Browne F, Taggart L, Zheng H, Parkes E et al (2019) Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. Brief Bioinform 20(5):1795–1811

Liu Y, Huang R, Liu Y, Song W, Wang Y, Yang Y et al (2018) Insights from multidimensional analyses of the pan-cancer DNA methylome heterogeneity and the uncanonical CpG–gene associations. Int J Cancer 143(11):2814–2827

Lu Y, Chan Y-T, Tan H-Y, Li S, Wang N, Feng Y (2020) Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy. Mol Cancer 19(1):1–16

Martinez-Gutierrez AD, Catalan OM, Vázquez-Romo R, Porras Reyes FI, Alvarado-Miranda A, Lara Medina F et al (2019) miRNA profile obtained by next-generation sequencing in metastatic breast cancer patients is able to predict the response to systemic treatments. Int J Mol Med 44(4): 1267–1280

Marusyk A, Janiszewska M, Polyak K (2020) Intratumor heterogeneity: the Rosetta stone of therapy resistance. Cancer Cell 37(4):471–484

McQuerry JA, Chang JT, Bowtell DD, Cohen A, Bild AH (2017) Mechanisms and clinical implications of tumor heterogeneity and convergence on recurrent phenotypes. J Mol Med 95(11):1167–1178

Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. Nature 501(7467):328–337

Meienberg J, Bruggmann R, Oexle K, Matyas G (2016) Clinical sequencing: is WGS the better WES? Hum Genet 135(3):359–362

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C et al (2019) Package 'e1071'. R J 2019:1071

Miranda SP, Baião FA, Fleck JL, Piccolo SR (2021) Predicting drug sensitivity of cancer cells based on DNA methylation levels. PLoS One 16(9):e0238757

Mitchel J, Chatlin K, Tong L, Wang MD (2019) A translational pipeline for overall survival prediction of breast cancer patients by decision-level integration of multi-omics data. In: 2019 IEEE International conference on bioinformatics and biomedicine (BIBM). IEEE, Piscataway, NJ

Mueller JJ, Schlappe BA, Kumar R, Olvera N, Dao F, Abu-Rustum N et al (2018) Massively parallel sequencing analysis of mucinous ovarian carcinomas: genomic profiling and differential diagnoses. Gynecol Oncol 150(1):127–135

Murtaza M, Dawson S-J, Pogrebniak K, Rueda OM, Provenzano E, Grant J et al (2015) Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. Nat Commun 6(1):1–6

Nanda JS, Kumar R, Raghava GP (2016) dbEM: a database of epigenetic modifiers curated from cancerous and normal genomes. Sci Rep 6(1):1–6

O'Geen H, Echipare L, Farnham PJ (2011) Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Epigenetics protocols. Springer, Cham, pp 265–286

Oliveira AM, Ross JS, Fletcher JA (2005) Tumor suppressor genes in breast cancer: the gatekeepers and the caretakers. Pathol Patterns Rev 124(suppl_1):S16–S28

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H et al (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344(6190): 1396–1401

Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Perera RM, Bardeesy N (2012) On oncogenes and tumor suppressor genes in the mammary gland. Cold Spring Harb Perspect Biol 4(6):a013466
Qu H, Zhou M, Yan Z, Wang H, Rustgi VK, Zhang S et al (2021) Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. NPJ Precis Oncol 5(1):1–11
Rehman O, Zhuang H, Muhamed Ali A, Ibrahim A, Li Z (2019) Validation of miRNAs as breast cancer biomarkers with a machine learning approach. Cancer 11(3):431
Ren L, Li J, Wang C, Lou Z, Gao S, Zhao L et al (2021) Single cell RNA sequencing for breast cancer: present and future. Cell Death Dis 7(1):1–11
Rish I (2001) An empirical study of the Naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. SAGE, Montreal
Sakthikumar S, Roy A, Haseeb L, Pettersson ME, Sundström E, Marinescu VD et al (2020) Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes. Genome Biol 21:1–22
Saurabh R, Nandi S, Sinha N, Shukla M, Sarkar RR (2020) Prediction of survival rate and effect of drugs on cancer patients with somatic mutations of genes—an AI based approach. Chem Biol Drug Des 96:1005–1019
Shui L, Ren H, Yang X, Li J, Chen Z, Yi C et al (2020) Era of radiogenomics in precision medicine: an emerging approach for prediction of the diagnosis, treatment and prognosis of tumors. Front Oncol 10:3195
Sood A, Miller AM, Brogi E, Sui Y, Armenia J, McDonough E et al (2016) Multiplexed immunofluorescence delineates proteomic cancer cell states associated with metabolism. JCI Insight 1(6):e87030
Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A (2019) A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Front Genet 10:256
Tomczak K, Czerwińska P, Wiznerowicz M (2015) The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 19(1A):A68
Torgerson W (1952) The first major MDS breakthrough. Psychometrika 17:401–419
Tuo Y, An N, Zhang M (2018) Feature genes in metastatic breast cancer identified by MetaDE and SVM classifier methods. Mol Med Rep 17(3):4281–4290
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63
Wang L-H, Wu C-F, Rajasekaran N, Shin YK (2018) Loss of tumor suppressor gene function in human cancer: an overview. Cell Physiol Biochem 51(6):2647–2693
Wang N, Zheng J, Chen Z, Liu Y, Dura B, Kwak M et al (2019) Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nat Commun 10(1):1–12
Weigelt B, Bi R, Kumar R, Blecua P, Mandelker DL, Geyer FC et al (2018) The landscape of somatic genetic alterations in breast cancers from ATM germline mutation carriers. J Natl Cancer Inst 110(9):1030–1034
Welch DR, Hurst DR (2019) Defining the hallmarks of metastasis. Cancer Res 79(12):3011–3027
Witten IH, Frank E (2000) Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, Burlington
Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ (1999) Weka: Practical machine learning tools and techniques with Java implementations. ACM Sigmod Rec 31(1): 76–77
Wu J, Hicks C (2021) Breast cancer type classification using machine learning. J Pers Med 11(2):61
Wu H, Xu T, Feng H, Chen L, Li B, Yao B et al (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res 43(21):e141

Xi Y, Shi J, Li W, Tanaka K, Allton KL, Richardson D et al (2018) Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. BMC Genomics 19(1):1–11

Yates LR, Knappskog S, Wedge D, Farmery JH, Gonzalez S, Martincorena I et al (2017) Genomic evolution of breast cancer metastasis and relapse. Cancer Cell 32(2):169–184

Zhang X, Zhang M, Hou Y, Xu L, Li W, Zou Z et al (2016) Single-cell analyses of transcriptional heterogeneity in squamous cell carcinoma of urinary bladder. Oncotarget 7(40):66069

Zhang M, Yang D, Gold B (2019a) Origin of mutations in genes associated with human glioblastoma multiform cancer: random polymerase errors versus deamination. Heliyon 5(3):e01265

Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H et al (2019b) The international cancer genome consortium data portal. Nat Biotechnol 37(4):367–369

Zhang Y-H, Li Z, Zeng T, Pan X, Chen L, Liu D et al (2020) Distinguishing glioblastoma subtypes by methylation signatures. Front Genet 11:1482

Zhao EY, Jones M, Jones SJ (2019) Whole-genome sequencing in cancer. Cold Spring Harb Perspect Med 9(3):a034579

Zhao Z, Zhang K-N, Wang Q, Li G, Zeng F, Zhang Y et al (2021) Chinese Glioma Genome Atlas (CGGA): a comprehensive resource with functional genomic data from Chinese gliomas. Genomics Proteomics Bioinformatics 19(1):1–2