Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 10
October 2023

Revolutionizing Oncology: Cutting-edge Classification Methods for Microarray Data

[1]Ankita Banerjee, [2]Ankan Bandyopadhyay, [3]Shreyasee Ghosh, [4]Dr. Abhishek Bandyopadhyay
[1]Computer Science & Engineering, Asansol Engineering College, [2]Computer Science & Engineering, Asansol Engineering College, [3]Danieli India Ltd., [4]Computer Science & Engineering (AI&ML), Asansol Engineering College

Abstract— Biologists grapple with the complexity of gene expression data, marked by a multitude of genes and limited samples. In the realm of Bioinformatics, pivotal challenges include gene subset selection, cancer profiling, and functional gene elucidation. Researchers are leveraging vast microarray gene expression datasets and employing machine learning on them for early cancer detection, prognosis, and biomarker identification, with a particular focus on survival analysis. The extensive gene dimensions inherent in microarray expression data, coupled with a limited number of patient samples, have ushered in a transformative era in cancer prediction and identification. Leveraging this technological advancement, precise cancer classification hinges on the meticulous selection of genes uniquely associated with each specific cancer subtype, marking a significant stride in the field of oncology research. This analysis delves into recent advancements in utilizing microarray gene expression data for disease diagnosis, particularly in cancer detection, through comprehensive coverage of data preprocessing, dimensionality reduction and machine learning algorithms, including supervised, unsupervised and semi-supervised approaches.

Index Terms— Biomarker, Classification, Machine Learning, Microarray, Oncology.

Introduction

An emerging technology called a microarray enables the simultaneous observation of the expression profiles of thousands of biomolecules under various experimental conditions or tissue samples. These days study of gene expression data enables the discovery of new biological system insights through clustering [1,2], classification [3,4], differential gene selection [5,6], cancer subtypes prediction [5-7] and so forth. For instance, the proper identification of many diseases is now possible due to the classification of DNA microarray data, leading to the discovery of previously unknown patterns in expression profiles.

Clinical decision support, encompassing tasks such as identifying diseases and predicting how patients will respond to treatment based on microarray expression profiles, is a widely recognized and continually advancing field within the realm of medical applications. Understanding the advantages and disadvantages of the various classification techniques is essential for the development of clinically effective microarray-based diagnostic models. Although earlier studies have demonstrated the viability of creating efficient models for cancer diagnosis, the

corresponding studies have restricted experiments in terms of the number of classifiers, gene selection algorithms, number of datasets and types of cancer involved [8-10]. Furthermore, since each study is built on a different experimental methodology and employs learning algorithms in a unique way, the findings of these studies cannot be compiled into a comparative meta-analysis. Therefore, it is not clear from the literature which classification method performs best among the numerous available alternatives. The dataset in gene expression analysis is frequently characterized by a large number of variables compared to a small number of records [11,12], making it difficult to pinpoint the features that are most crucial to solving a particular issue. By choosing relevant genes with feature selection methods, this issue can be solved. Gene selection is still an essential task to boost the effectiveness and speed of classification methods [13], making it

a crucial component of microarray data analysis [14].

When it comes to analyzing the microarray, the techniques can be broadly divided into three main groups:

1. Unsupervised methods, which involve exploratory approaches without predefined labels or guidance.
2. Supervised techniques, which rely on labeled data and provide clear guidance for the analysis process.
3. Semi-supervised techniques, which blend elements of both supervised and unsupervised methods to leverage available information while allowing for some degree of exploratory analysis.

Unbiased analysis of microarray data is known as unsupervised classification or class discovery. Unsupervised analysis involves organizing the data without the aid of outside classification knowledge. The samples are grouped based on the similarity/dissimilarity metric using clustering algorithms.

For supervised learning, there is a broad variety of classification techniques ranging from k-Nearest Neighbors Algorithms, Support Vector Machines, Decision Tree Classifiers, Naive Bayes Classifiers and others. Supervised learning has a downside of small data range due to the reduced availability of labelled data which forms a bottleneck in achieving a high degree of precision. There comes the need for supervised and semi-supervised learning.

For clustering, a wide range of statistical and computational techniques are available. These include self-organizing maps [15] and artificial neural networks [16] from the machine learning literature as well as hierarchical clustering [17,18], and k-means clustering [19] from the statistical literature. Gene clustering and gene marker identification are two methodological advancements for gene expression profiling studies that are documented in the literature [1,2,4,20-22]. In the fields of computational biology, bioinformatics, soft computing, and geoscience, a wide range of clustering techniques are presented in [23-25]. The use of prior knowledge is taken into account in supervised analysis. For example, patients with good and poor prognoses frequently provide tumor samples for microarray studies.

This survey addresses the identified knowledge gap by conducting an in-depth analysis of recent genomics studies. It delves into the formidable challenges posed by microarray datasets, particularly those pertaining to gene expression, and explores various proposed solutions. The study begins by presenting a comprehensive overview of data preprocessing techniques tailored for gene expression datasets. Moreover, the examination scrutinizes the utilization of machine learning algorithms in the diagnosis of cancer through DNA microarray analysis, encapsulating a thorough investigation of datasets employed in developing cancer classification models based on microarrays. The performance of these algorithms is systematically analyzed, offering insights into the strengths and limitations of each data reduction approach. The final goal is to find genes or create a model that can categorize patients into prognostic classes according to how well their corresponding tumors microarray data predicts their prognosis.

Literature Review

Learning Models	Strengths
Supervised Learning	Model generation relies on training data and involves various techniques. k-Nearest Neighbors (k-NN) handles multiple classes efficiently and is stable. Maximum Likelihood (ML) ensures accurate branch lengths and provides per-site likelihood estimates. Decision Trees (DT) are simple, interpretable, and non-parametric,

	avoiding outlier problems, with fast training and evaluation. Support Vector Machines (SVMs) guarantee accuracy and theoretical overfitting safeguards, while Artificial Neural Networks (ANN) excel with optimal parameter settings. These diverse methods enrich machine learning with tailored advantages.
Unsupervised Learning	Unsupervised learning, far more common than its supervised counterpart, operates without prior information and excels at discovering previously overlooked patterns. By employing cluster analysis, it reduces data while generating new hypotheses, making it an essential tool in data analysis.
Semi-supervised Learning	Semi-supervised learning, which combines labeled and unlabeled data for improved classification accuracy, encompasses several effective techniques. Self-training, a straightforward method applicable to all classifiers, offers a clear probabilistic framework. Generative approaches like the Correct model are highly effective. S3VM maximizes the margin using unlabeled data, while graph-based methods, with clear mathematical foundations, excel when the graph aligns with the task, even extending to directed graphs. These techniques diversify semi-supervised learning options for various scenarios and preferences.
Ensemble Learning	Enhancing generalization and minimizing classification errors are key objectives in machine learning. Various ensemble techniques contribute to achieving these goals. Bagging, for instance, improves unstable learning algorithms significantly by reducing variance while leaving bias unchanged, resulting in better generalization. Boosting, on the other hand, offers a powerful approach by easily implemented methods that effectively reduce both variance and bias, maximizing the likelihood. Random Forests (RF) provide efficient solutions for large databases and often rival or outperform Support Vector Machines (SVMs), making

	them a reliable choice for diverse learning tasks.

Table I. Table representing the advantages of each of the discussed learning techniques

Classification serves as a fundamental learning function, assigning data points to predefined classes [26][27]. Pattern classification endeavors encompass the automated recognition, description, classification, and categorization of similar patterns across diverse scientific and engineering disciplines. The identification of novel classes often relies on unsupervised techniques and clustering methodologies.

Supervised learning, conversely, involves establishing a mapping between input variables (X) and output variables (Y), subsequently applying this mapping to predict unseen data, serving as a pivotal methodology within machine learning, widely applicable in real-world scenarios.

Recent years have witnessed an escalating interest in leveraging unlabeled data in conjunction with labeled data in the realm of machine learning [28]. The rationale behind this lies in the cost-effectiveness and abundance of unlabeled data in many applications compared to labeled counterparts. Effective information extraction from unlabeled examples, enabling learning from a limited set of labeled examples, offers substantial advantages.

Several semi supervised learning techniques have emerged, demonstrating commendable performance across various learning paradigms. These techniques encompass label propagation for tasks like word-sense disambiguation [29], co-training for tasks such as web page classification [30] and enhanced visual detection [31], transductive SVM [7] for the identification of cancer subtypes, EM [32] for text classification, and graph-based methodologies, among others.

Beyond these categories, classifier combination [33][34] has garnered heightened attention within the machine learning community. Additionally, soft computing techniques like fuzzy and rough sets prove invaluable for pattern classification.

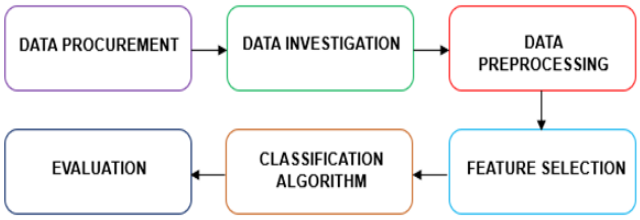

Fig.1 Diagrammatic representation of steps involved in microarray gene profile analysis and prediction

There are trillions of cells in every living thing, and each cell has a complete copy of the genome. The genome, which is stored as genetic code in DNA, is the blueprint for creating new organisms. Coding and non-coding segments exist in DNA. A functional RNA or protein product is encoded by a gene, which is a segment of DNA. Protein structures that carry out vital functions for the organism are indicated by genes. According to the Human Genome Project, 20,000–25,000 human protein-encoded genes exist. The process by which a gene's information is translated into an observable phenotype to produce protein and build the cell's structure is known as gene expression. Transcription and translation are the two primary phases of the gene expression process. The first stage of gene expression is transcription, in which a specific DNA segment is copied and converted into messenger RNA (mRNA) by the enzyme RNA polymerase. Translation is the

use of mRNA to control the synthesis of proteins. Gene expression, which is determined by the amount of mRNA in a tissue, is thus the level of a gene's activity in a specific body tissue.

An array of known DNA molecules is chemically bonded at specific locations on a glass slide to form a DNA microarray. This glass slide is put under a scanner to produce a picture of colored dots. Every data point symbolizes the level of gene expression within specific experimental contexts. At each location on the array, multiple identical copies of the same molecular entity serve as representatives for a probe. In essence, an array comprises a multitude of measurements spanning numerous genes, with each individual probe signifying the measurement pertaining to a single gene.

A microarray dataset consists of a s×t two-dimensional matrix M = m_{ij} containing s samples & t biomolecules. Within the dataset, each element holds information about the expression level corresponding to the i^{th} sample on the j^{th} microarray. Gene expression profiling, also known as microarray analysis, enables the simultaneous assessment of numerous genes within a single RNA sample, making it feasible to scrutinize thousands of genes at once. The identification and diagnostic prognosis of cancer have been successfully accomplished using this novel technique. Planning carefully and defining an objective are essential for successful microarray experiments [35]. The primary objective of most microarray investigations is to extract meaningful biological insights from the data and subsequently employ this newfound understanding in practical and valuable ways. Effective selection, class discovery, and classification are required steps to extract crucial information from microarray data. The procedure of selecting a compact subset of genes that exhibit the highest predictive capabilities for their respective class labels is referred to as gene identification or gene curation. The learning model benefits from increased classification accuracy as outcome. Data patterns can be thought of as the different types of information that can be gleaned from DNA microarray data. The goal of pattern analysis is to automatically identify and describe relationships in data. Data is typically assumed to be in vectorized form for statistical and machine learning purposes, and relations are expressed as classification rules, regression functions, or cluster structures. Survival models come in different flavours, falling into three main categories: parametric, non-parametric, and semi-parametric. Non-parametric models are most suitable when we seek a broad comparative perspective to discern which group exhibits a superior survival rate. In contrast, parametric models step into the picture when our aim is to anticipate specific time milestones, employing diverse probability distributions for this purpose. Semi-parametric models, on the other hand, adapt their strategy depending on the data characteristics, striking a balance between structure and flexibility in their survival analysis approach [93].

Novel genes, transcription factor binding sites, changes in DNA copy number, genes affected by treatment, time series (with and without a given treatment), patterns of gene activity (healthy vs. control), classification of tumors, identifying the target genes of tumor suppressors, identifying cancer biomarkers, antibiotic treatment, human heart failure, SNP linkage studies, and systems biology are all investigated by microarrays. Microarray analysis can provide a wealth of information on the pathology of diseases, how they progress, how they are resistant to treatment, and how they respond to cellular micro-environments. This information can help with early tumor diagnosis and new therapies. Using feature selection and classification methods, we prioritize our review on cancer studies among the many microarray study directions. The field of cancer research is both fascinating and important in the medical field. In the past, cancer diagnosis relied on morphological and clinical observations. However, thanks to the development of microarray technology, it is now possible to monitor the expression of thousands of genes simultaneously. This has greatly improved the accuracy of cancer classification based on gene expression data. To achieve this, specimens are grouped together based on the similarity or dissimilarity of their gene expression profiles. [36]. Statnikov et al. [37] conducted a comprehensive evaluation of classification methods for cancer

diagnosis based on microarray gene expression data.

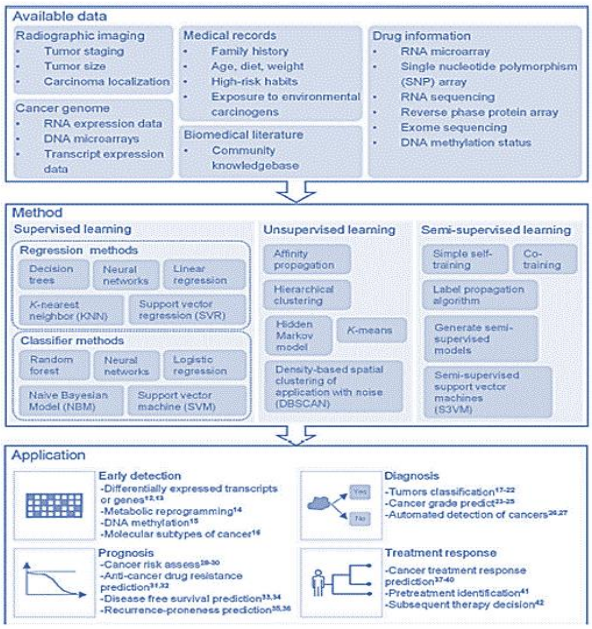

Fig.2. A Generalized preview of the entire process of prediction of oncogenes from microarray data

The selection of appropriate performance metrics for machine learning algorithms assumes a pivotal role in the evaluation and benchmarking of algorithmic performance. In this context, commonly employed metrics encompass classification accuracy, leave-one-out cross-validation (LOOCV), k-fold cross-validation, and receiver operating characteristics (ROC). Of notable significance is the classification accuracy, which serves as a fundamental criterion in numerous studies. Consequently, it becomes imperative to encompass a comprehensive array of performance metrics to ensure a holistic evaluation. This entails the inclusion of sensitivity, sensibility, and similarity metrics, which collectively offer a more nuanced perspective on algorithmic efficacy. In summary, the judicious choice and incorporation of diverse performance metrics are instrumental in providing a comprehensive assessment of machine learning algorithms in various applications [92].

Learning Models	Algorithms
Supervised Learning	Perception algorithm k-nearest neighbor (k-NN) Maximum likelihood (ML) Decision tree (DT) SVMs Artificial Neural networks (ANN)
Unsupervised Learning	k-means clustering Hierarchical Clustering Self-organizing map Principal components analysis Fuzzy-C-means clustering Genetic algorithm
Semi-supervised Learning	Self-training Generative Models

	Semisupervised SVMs Graph-based algorithms Multiview Learning
Ensemble Learning	Bagging Boosting Random Forests (RF)

Table II. Categorization of various algorithms based on the learning models

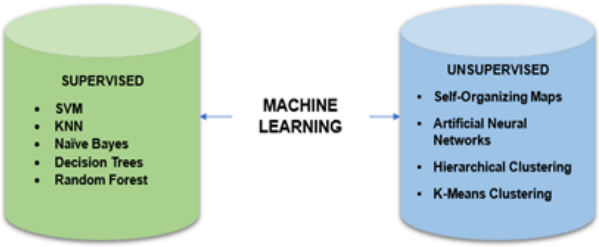

Fig.3. Categorization of the various algorithms into Supervised and Unsupervised classes

Molecular Classification of Tumors

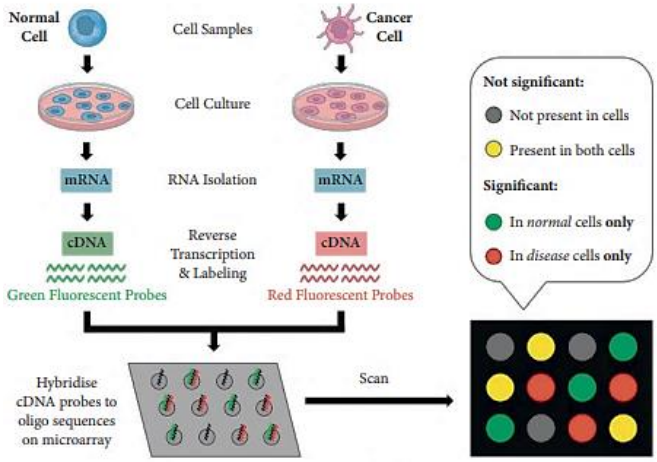

Fig.4. Visualization of gene extraction process

A. Unsupervised Cancer Classification Approaches

Clustering is an important tool in pattern recognition for identifying homogeneous gene clusters from microarray expression profiles. Gene expression profiling has given rise to a plethora of methodologies aimed at delineating distinct cancer subtypes. In a study by Alizadeh et al. [38], DLBCL subtypes were identified through hierarchical clustering. Meanwhile, Ben-Dor et al. [39] and von Heydebreck et al. [40] used a selected set of genes to cluster a microarray dataset in a more biologically meaningful way. They discovered that using a subset of genes produced better clusters than using all the genes. However, the method cannot provide the actual number of clusters since it did not use the clinical information to identify subgroups. This implies that the identified clusters may not be associated with the clinical outcome of interest. Van't Veer et al. [41] used agglomerative clustering for gene and tumor clustering using microarray expressions. They also developed a supervised method for classifying breast tumors. In [42], Pomeroy et al. separated different embryonal CNS tumors using PCA and hierarchical clustering. PCA was used to reduce

variables (genes). Tumor types were also identified in their work using the supervised methods: SNR and k-NN. PCA was also used in [43] to distinguish MLL from both ALL and AML. To demonstrate the effectiveness of gene expression profiles to identify

ALL, AML and ALL in leukemia dataset, they have employed kNN classifier.

In a scientific study conducted by researchers, fuzzy clustering was employed to enhance the quality of cluster formation for microarray data. Additionally, a fuzzy clustering approach rooted in real-coded Simulated Annealing (VSA) was devised, incorporating a variable-length configuration in conjunction with a classifier based on Artificial Neural Networks (ANN). The results showed that this method outperformed other clustering techniques for grouping expression profiles. Other works have also employed fuzzy clustering and its variants to group expression profiles, as seen in [44], [45]. Additionally, multi-objective genetic clustering has been combined with SVM to obtain compact clusters and efficient gene marker identification using three benchmark cancer datasets in [46]. A technique that simultaneously selects features and classifies samples was developed by Mitra et al. in their study [47]. The proposed method, called Statistical-Algorithmic Method for Bicluster Analysis (SAMBA), utilizes a bipartite graph-based model for biclustering. SAMBA is used for both feature (miRNA) and sample (tissue) selection (SFSS) to classify multiple classes of tumors and cancer cell lines. A bicluster is a subset of genes that are co-expressed over a subset of samples or experimental conditions. Empirical results demonstrated that the method outperformed other approaches on the cancer datasets.

Biologists are currently interested in analyzing 3D microarray datasets to discover co-expressed genes under certain experimental conditions and across a subset of time points. To tackle this challenge, Bhar et al. proposed a triclustering algorithm called δ-TRIMAX, which identifies genes that are co-expressed over a subset of samples across a subset of time points from a real-life time series dataset in estrogen-induced breast cancer cell line.

In a research paper titled [48], multi-objective genetic clustering was proposed for tissue samples. The method used a real-coded encoding of the cluster centers and optimized both cluster compactness and separation simultaneously. The resultant clustering information was then used to train the Support Vector Machines (SVMs) with different kernel functions. Finally, the agreement clustering results obtained from multiple SVMs were used to achieve the final clustering.

B. Supervised Classification Methods

Supervised machine learning is useful for medical problems such as disease diagnosis, prognostic prediction, drug discovery, and treatment selection. In case there are various subtypes of cancer, there are several supervised techniques available to identify which subtype is present in a patient. These techniques include methods such as [49], [50], [11], [51], [52], [53]. Major cancer studies using microarray data deals with class comparison or class prediction objectives [54], [55]. There are several supervised approaches that have been employed for the studies of microarray cancer datasets including SVM [56], [57], [8], k-NN [58], [59], NB [60], [61], [62], decision trees [63]. Lee et al. [10] proposed a Bayesian model to identify genes for cancer classification using microarray expression profiles. However, Support Vector Machines (SVM) have proven to be one of the most powerful supervised learning algorithms for analyzing biological data, including cancer gene expression data [64], [65]. Furey et al. [66] used SVM for classifying cancer tissue samples or cell types using microarrays. In a separate investigation, the study detailed in reference [67] employed Support Vector Machines (SVM) as a classification model for the analysis of cancer datasets, specifically focusing on leukemia, SRBCT, and lymphoma datasets. Effective genes were selected using Principal Components Analysis (PCA), class separability measure, Fisher ratio, and t-test. Mundra and Rajapakse exploited SVM and t-score in [68].

The aim of this study was to classify cancer. In the study, the researchers initially employed Support

Vector Machines (SVM) to discern pertinent samples. Subsequently, they utilized the t-score method to identify the most informative features from this subset of samples. Notably, a previously proposed SVM-based technique, as outlined in a prior study [69], was adopted for the purpose of cancer detection. Another entry in the realm of supervised machine learning algorithms is the Extra Trees Classifier. It requires labeled training data to learn and make predictions based on the patterns it identifies in the features and their corresponding target labels. In supervised learning, the algorithm is provided with both input data and the correct output (or labels), and its goal is to learn the mapping between the inputs and outputs to make predictions on new, unseen data [94].

In [70], a modular neural network-based approach was proposed for classifying breast cancer nuclei stained for steroid receptors in histopathological samples. Additionally, a neural network representation called simplified fuzzy ARTMAP was proposed in [71] to distinguish normal patients from those with diffuse large B- cell lymphoma (DLBCL), and to identify the differences between patients with molecularly distinct forms of DLBCL without using prior information on those subtypes.

In the realm of esophageal cancer (CA) and premalignancy research, the utilization of Artificial Neural Networks (ANNs) stands out as an efficacious and resilient approach for comprehensively analyzing cDNA microarray data, as demonstrated in [72]. ANNs are capable of recognizing patterns that are incomplete, biased, or extremely abundant. Yeung et al. [73] reported that traditional methods for gene selection and cancer classification do not consider model uncertainty and use a single set of selected genes for prediction. To mitigate the inherent ambiguity associated with the analysis of intersecting sets of pertinent genes, they introduced Bayesian Model Averaging (BMA) as a technique to elevate the accuracy of classification.

Murat et al. [74] conducted a study on early prostate cancer diagnosis using ANNs and SVMs. They found that the advanced level of cancer classification can be achieved by using the extreme learning machine (ELM) which is a novel learning scheme based on feed-forward neural networks. Another study [75] used ELM for multicategory cancer classification in cancer diagnosis with microarray data. The experimental results of this proposed technique on GCM, Lung, and Lymphoma data showed better performance than artificial neural networks (ANN) and SVM algorithms. In ELM, one can randomly choose and fix all the hidden node parameters and then analytically determine the output weights. Genetic programming integrated with mutual information-based feature selection was found to be the most effective alternative for predicting colon cancer compared to existing techniques.

In a previous study, the authors introduced two types of single gene classifiers, SGC-t and SGC-W, which were developed using t-test and Wilcoxon-Mann Whitney (WMW) test, respectively. The most effective genes were identified using their ability to discriminate between classes. It was demonstrated that these single gene classifiers achieved classification accuracy that is comparable to or better than those obtained using commonly used methods, such as diagonal linear discriminant analysis, k-NN, SVM, and RF.

C. Semi supervised Techniques

One major research area involves extending supervised techniques to handle semi-labeled data. This is achieved by training the base classifier using a set of labeled training samples while treating unlabeled samples as additional optimization variables. A study [76] unveils a methodology devised to identify diverse tumor subtypes through the amalgamation of both gene expression data and clinical information. The technique involves identifying a subset of genes that are correlated with a particular clinical variable of interest, and then clustering these genes using suitable techniques. The effectiveness of this method was demonstrated on various publicly available datasets. In a separate study, Xu et al. utilized a semi-supervised neural network (ssEAM) to predict multi-class cancer subtypes.

In their research paper, Rui et al. [77] introduced a semisupervised ellipsoid ARTMAP (SsEAM) that can be used for multiclass cancer diagnosis. They

used Particle Swarm Optimization (PSO) to select the genes that were most relevant to the cancer diagnosis. The proposed technique was tested on three different cancer datasets and was found to be effective. In another study [78], microarray data was used to develop a semisupervised method for class discovery in cancer research. In this method, the first step is to identify relevant classes based on clinical sample data. The second step involves identifying gene sets with strong differential expression based on known biological gene annotations. Chakraborty [79] proposed a semi-supervised Bayesian SVM classification model for binary classification (Semi-BSVM) using several benchmark microarrays, including two cancer datasets.

Koestler et al. [80] developed a method called semisupervised recursively partitioned mixture models (SS-RPMM) to identify cancer subtypes that are associated with patient survival, using both genetic and patient-level clinical data. The method identifies informative gene subsets that are linked to survival time, and uses this information to discover cancer subtypes. A method has been developed that is comparable to other semisupervised methods, such as semisupervised clustering and supervised principal components analysis. In [81], a low-density separation method was used to perform gene expression-based cancer subtype discovery. It has been shown that semisupervised learning can improve outcome prediction for cancer patients.

Pang et al. [82] introduced a new model called Personalized Transductive Learning (PTL) for classifying cancer datasets. PTL methods define the neighborhood using a predefined similarity/dissimilarity metric. The method uses the concept of an informative neighborhood and a transductive Support Vector Machine (TSVM) classification tree (t-SVMT) to achieve good performance in detecting class imbalance cancer datasets and solving overfitting problems.

In [7], the Transductive SVM (TSVM) method was introduced to enhance cancer classification, using four microarray cancer datasets. The authors applied a consistency-based feature selection [83] prior to classification, to identify superior gene markers. The experimental results confirm the effectiveness of the proposed technique compared to the inductive SVM and low-density separation technique in the area of semi-supervised cancer classification, as well as gene-marker identification.

In order to predict cancer recurrence using gene expression data, researchers proposed a semi supervised approach based on graph regularization in [84]. The method at hand revolves around the transformation of gene expression data into a graphical model, which in turn facilitates a semi-supervised learning process. This seamlessly integrates protein interaction information into the gene expression dataset, identifying pairs of genes with analogous functions, ultimately empowering the discovery of cancer recurrence. To construct the graph, both labeled and unlabeled nodes are used and a regularization technique is applied.

Learning Models	Weaknesses
Supervised Learning	In the realm of machine learning, various algorithms present unique challenges. Perceptrons lack solution uniqueness, while k-Nearest Neighbors (k-NN) can mislead in high-dimensional spaces, being computationally intensive and unclear in distance measure selection. Maximum Likelihood (ML) methods may falter when the model lacks sophistication, demanding

	significant computation. Decision Trees (DT) risk overfitting, Support Vector Machines (SVMs) are memory-intensive to tune, and Artificial Neural Networks (ANNs) grapple with slow convergence and complex parameter adjustment. Customizing algorithm selection becomes crucial in addressing these challenges in machine learning.
Unsupervised Learning	Cluster analysis encounters several challenges, including the ambiguity of defining the optimal number of clusters, a scarcity of statistical tests for assessing the strength of cluster membership, and subpar classification performance.
Semi-supervised Learning	Leveraging unlabeled data doesn't always yield performance improvements, as early mistakes can reinforce themselves, and it's challenging to predict convergence in techniques like Self-training. Generative models face difficulties in verifying correctness, encountering EM local optima, and potentially suffering from unlabeled data when the model assumption is incorrect. Optimization and avoiding bad local optima pose challenges in S3VM. Moreover, the performance of graph-based methods heavily depends on the quality of the graph structure and edge weights, rendering them sensitive to these factors.
Ensemble Learning	Overfitting is a common concern in various machine learning methods. Bagging can sometimes exacerbate this issue and diminish the performance of inherently stable procedures. Boosting, while effective, can be sensitive to noise and may require a substantial number of estimators for optimal results. Random Forests (RF) face the challenge of potentially biasing towards attributes with more levels, impacting their performance. These

	considerations underscore the need for careful handling and tuning of these techniques to mitigate their respective drawbacks.

Table III. Table representing the disadvantages of each of the discussed learning techniques

D. Ensemble Methods

An ensemble of classifiers was shown to perform well in classifying cancerous microarray data when using ensemble learning based on bagged and boosted decision trees. The results were reported in [85] where seven publicly available datasets were analyzed. In classification tasks, their experimental results frequently surpassed those of single decision trees.

In the context of tumor classification research, an enhanced classification performance was sought through the incorporation of boosting alongside decision-making. The researchers adapted the conventional boosting technique and incorporated a nonparametric scoring methodology developed by Park et al. for gene selection during the initial phase. For the second phase, they implemented the LogitBoost methodology, as introduced by Friedman et al., as a means to effectively manage noisy data. The experimental results showed that this method led to a small to significant increase in performance and yielded competitive results on several cancer datasets. In another experiment detailed in [86], the author used LogitBoost in their proposed BagBoosting algorithm to compare it with six other classifiers using microarray tumor datasets. They discovered that the BagBoosting algorithm provided the second-best classification performance, while SVM proved to be the most effective.

Liu et al. (2019) proposed a feature selection method that uses ensemble neural networks to improve the accuracy and robustness of sample classification for cancer datasets. The method is capable of extracting sufficient information from the datasets, resulting in more accurate classifications. Furthermore, it has the capability to identify specific genetic markers linked to various diseases, enabling more accurate diagnosis and treatment.

In their research documented in [87], the authors unveiled a multifaceted classifier tailored for the classification of cancer-related data. Conversely, in a separate investigation referenced as [88], an expert system emerged as a novel approach. This method serves as a valuable tool for mastering ensembles of process-oriented models, effectively addressing the challenges of the curse of dimensionality and overfitting in the realm of classification. This study shows that the proposed method performs well on numerous microarray cancer datasets. In [89], an ensemble of nonparallel plane proximal classifier (NPPC) was proposed for cancer classification, based on microarray gene expression profiles. Afterwards, several NPPC expert models are trained using a genetic algorithm-based feature and model selection scheme. The results of experiments on cancer data sets indicate that the NPPC ensemble has a testing accuracy similar to that of the SVM ensemble, but with shorter training times on average.

In another study [95], the researchers conducted a comprehensive analysis using two datasets, extracting varying numbers of top genes (ranging from 10 to 60) as training data. These gene subsets were subsequently evaluated using five machine learning classifiers, including SVM, kNN, DT, RF, and XGBoost. Notably, all classifiers exhibited commendable performance on the internal testing dataset, as detailed in Supplementary Tables S1 and S2. However, XGBoost emerged as the standout performer, achieving an impressive accuracy of 97% and demonstrating high auROC and PR-AUC scores of 0.996 on the top 40 feature subset from the complete dataset. Similarly, when dealing with the driver dataset, both XGBoost and

RF excelled, delivering accuracies of 93% and 92%, respectively, on the top 50 feature subsets. These findings underscore the efficacy of XGBoost in handling the selected feature subsets and highlight its potential in cancer-related applications.

Some more recent works

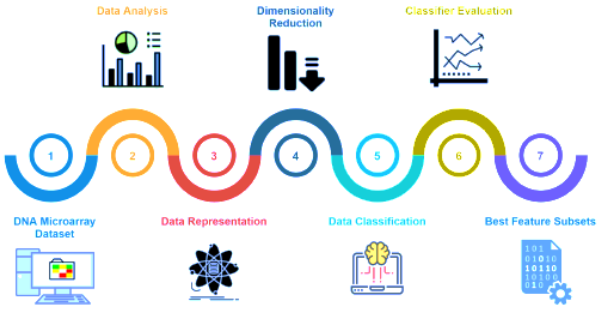

Fig.5. Diagrammatic representation of stages involved in pipelining of machine learning algorithms

One of the latest techniques is the use of pipeline on machine learning algorithms for cancer detection on micro array data. [90]. It presents a comprehensive approach that combines Feature Discretization (FD) and Feature Selection (FS) techniques as a prelude to classification tasks. The central aims encompass achieving minimal errors in classification, particularly in terms of reducing classification errors, false negatives, and false positives. Simultaneously, the goal is to pinpoint compact sets of essential features that hold high relevance for each unique classification undertaking. The methodology involves a multi-step process, commencing with the selection of evaluation techniques. Subsequently, a machine learning (ML) pipeline is constructed, encompassing data representation/discretization, dimensionality reduction, and data classification techniques. Comparative analysis is then performed using established performance metrics, culminating in the identification of the most effective technique and the optimal feature subsets for each dataset. These steps encompass label mapping for nominal class labels, imputing missing values, and the removal of constant features. Leave-one-out (LOO) cross-validation is employed to enhance generalization error estimation, particularly suited for small datasets. Finally, feature relevance is quantified by counting the number of times each feature is selected in the FS stage, facilitating the achievement of the study's objectives.

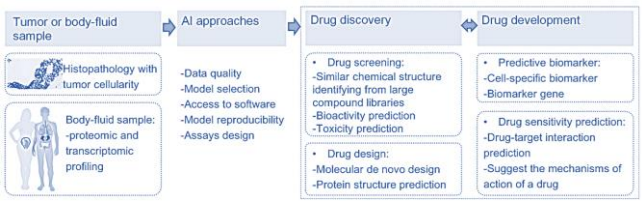

Fig. 6. Leveraging AI learning models in drug development to predict, diagnose, and effectively treat cancer.

Alternative approaches may encompass the application of deep learning methodologies. In a recent investigation cited as [91], the primary emphasis lies in harnessing biological networks, which encompass an array of organized knowledge repositories, including disease pathway networks, protein-protein interaction (PPI) networks, and disease similarity networks. These networks stand as invaluable assets, enabling the revelation of the complex interplay and attributes within biological

systems. They prove instrumental in driving noteworthy advancements in domains like the diagnosis of cancer, the prediction of genomic functions, and the exploration of new pharmaceuticals. Moreover, the adoption of network-based methodologies spans the spectrum from scrutinizing individual cells to examining entire populations, showcasing their adaptability in unearthing fresh perspectives from biological data. The study also highlights the significance of leveraging graph-based methodologies, particularly graph convolutional neural networks (GCNs), as a prevalent technique for incorporating graph-structured data into classification and regression tasks. Notably, the integration of GCNs with relation networks (RNs) has been proposed to classify breast cancer subtypes effectively. Additionally, a GCN-based approach has been employed to predict survival rates by optimizing graph representations of whole slide images (WSIs) in the context of lung and brain carcinoma. These advancements underscore the pivotal role of network-oriented strategies and AI techniques in advancing biological research and addressing complex challenges in the field.

In a recent investigation denoted as reference [96], a thorough understanding of the molecular underpinnings behind uterine leiomyomas (ULM) and leiomyosarcomas (ULMS) is unveiled. This is achieved through the fusion of multiomics data with the application of machine learning methodologies, offering holistic insights into these conditions.

Conclusion

This article inspects the significant contributions made in gene marker discovery and tumor prediction within the realm of bioinformatics. The primary challenges faced by researchers in this field are the high input dimensionality and limited sample sizes. To address these challenges, an array of techniques for feature selection has emerged, aimed at pinpointing genes of significance. Looking ahead, a promising avenue for future research in bioinformatics lies in the exploration of multivariate models for feature selection. Additionally, there is potential for advancing the field through the development of semi-supervised, ensemble, and integrated feature selection approaches to enhance the robustness of selected features. Combining these techniques with appropriate evaluation criteria offers an intriguing path forward to address the issue of small sample sizes in microarray data. The ultimate goal remains the precise identification of tumors, achievable through statistical or machine learning techniques. While supervised and unsupervised algorithms are commonly employed for class discovery, there is room for improvement by incorporating semi-supervised and multi-classifier techniques, an area currently underexplored in bioinformatics research. The second research focus revolves around enhancing detection efficiency by crafting appropriate semi-supervised and multi-classifier models.

References

[1] U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, “Combining pareto-optimal clusters using supervised learning for identifying coexpressed genes,” BMC Bioinformatics, vol. 10, no. 27, 2009.

[2] U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, “Multi-class clustering of cancer subtypes through SVM based ensemble of paretooptimal solutions for gene marker identification,” PLoS One, vol. 5, no.11, p. e13803, 2010.

[3] B. Wu, “Differential gene expression detection and sample classification using penalized linear regression models,” Bioinformatics, vol. 22, pp.472-476, 2006.

[4] U. Maulik and A. Mukhopadhyay, “Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzing microarray data,”’ Computers & Operations Research, vol. 37, no. 8, pp.1369-1380, 2010.

[5] U. Maulik, A. Mukhopadhyay and D. Chakraborty, “Gene-expression based cancer subtypes prediction through feature selection and transductive SVM,” IEEE Transactions on Biomedical Engineering, vol. 60, no.4, pp. 1111–1117, 2013.

[6] U. Maulik and D. Chakraborty, “Fuzzy preference-based feature selection /and semisupervised SVM for cancer classification,” IEEE Transactions on NanoBioscience, vol. 13, no. 2, pp. 152–160, 2014.

[7] D. Chakraborty and U. Maulik, "Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning,"IEEE Journal of Translational Engineering in Health and Medicine, vol2, 2014.

[8] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosisusing tumor gene expression signatures," Proceedings of the NationalAcademy of Sciences, USA, vol. 98, no. 26, pp. 15149–15154, 2001.

[9] C.H. Yeang et al., "Molecular classification of multiple tumor types," Bioinformatics, vol 17, suppl. 1, pp. S316–S322, 2001.

[10] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick, "Gene selection: A Bayesian variable selection approach," Bioinformatics, vol. 19, pp. 90-97, 2003.

[11] G. Piatetsky-shapiro and P. Tamayo, "Microarray data mining: facing the challenges," SIGKDD Explorations Newsletter, vol. 5, pp 1–5, December 2003.

[12] M.Rocha, R.Mendas, P.Maria, D.Glez-Pena, and F.Fdez-Reverola, "A platform for the selection of genes in DNA Microarray data using evolutionary algorithms," in Proceedings of 8th Annual Conference on Genetic and Evolutionary computation, London, England, 2007, pp 415–423.

[13] Q. Shen and C. Shang, "Aiding classification of gene expression data with feature selection: A comparative study," Journal of Computational Intelligence Research, vol. 1, pp. 68–76, 2006.

[14] J. C. Rajapakse and P. A. Mundra, "Multiclass gene selection using Pareto-fronts," IEEE Trans. Comput. Biol. Bioinformat., vol. 10, no. 1, pp. 87–97, Jan./Feb. 2013.

[15] T. Kohonen, "The self-organizing map", Proc. IEEE, 78, pp. 1464–1479, 1990.

[16] C.Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1996.

[17] B. Everitt and S. Rabe-Hesketh, The Analysis of Proximity Data, John Wiley, New York City, 1997.

[18] M. Eisen, P. Spellman, P. O. Brown et al., "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci., 95, pp. 14863–14868, 1998.

[19] J. Hartigan and M. Wong, "A k-means clustering algorithm," Applied Statistics, vol. 28, pp. 100–108, 1979.

[20] S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, "An improved algorithm for clustering gene expression data," Bioinformatics, vol. 23, no. 21, pp. 2859–2865, 2007.

[21] A. Mukhopadhyay and U. Maulik, "Towards improving fuzzy clustering using support vector machine: application to gene expression data," Pattern Recognition, vol. 42, no. 11, pp. 2744–2763, 2009.

[22] U. Maulik, "Analysis of gene microarray data in a soft computing framework," Applied Soft Computing, vol. 11, no. 6, pp. 4152–4160, 2011.

[23] U. Maulik, S. Bandyopadhyay and A. Mukhopadhyay, Multiobjective Genetic Algorithms for Clustering: applications in Data Mining and Bioinformatics, Springer-Verlag, New York Inc, 2011.

[24] S. Bandyopadhyay and U. Maulik, Analysis of Biological Data: A Soft Computing Approach, WLT Jason World Scientific Pub Co Inc, 2007.

[25] Luo et. al, "Improving the computational efficiency of recursive cluster elimination for gene selection," IEEE Transactions on Computational Biology and Bioinformatics, vol. 8, no. 1, pp. 122–129, 2011

[26] D. J. Hand, Discrimination and Classification, Chichester, U.K.: Wiley, 1981.

[27] S. M. Weiss and C. A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems, San Francisco, Calidf.: Morgan Kaufmann, 1991.

[28] O. Chapelle, B. Scholkopf, and A. Zien, (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, MA, USA, 2006.

[29] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," In Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp. 189–196, 1995.

[30] A. Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," COLT: In Proceedings of the Eleventh Annual Conference

on Computational Learning Theory, pp. 92–100, 1998.

[31] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," In Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 626–633, 2003.

[32] K. Nigam, A. K. McCallum, S. Thrun and T. M. Mitchell, "Text classification from labeled and unlabeled documents using EM," Machine Learning, vol. 39, no. 2-3, pp. 103-134, 2000.

[33] C. K. Chow, "Statistical independence and threshold functions," IEEE Transactions on Electronic Computers, vol. EC-14, no. 1, pp. 66–68, 1965.

[34] L. Xu, A. Krzyzak and C. Y. Suen, "Methods for combining multiple classifiers and their applications to handwriting recognition," IEEE trans- actions on System, Man, and Cybernetics, vol. 22, no. 3, pp. 418–435, 1992.

[35] B. Tjaden and J. Cohen, "A survey of computational methods used in microarray data interpretation," Applied Mycology and Biotechnology, Bioinformatics, vol. 6, pp. 7–18, 2006.

[36] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," American Journal of Obstetrics and Gynecology, vol. 195, no. 2, pp. 373–388, 2006.

[37] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," Bioinforamtics, vol. 21, pp. 631–643, 2005.

[38] A. Alizadeh and et. al., "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," Nature, vol. 403, pp. 503–511, 2000.

[39] A. Ben-Dor, N. Friedman and Z. Yakhini, "Class discovery in gene ex- pression data," Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal (Quebec): ACM Press. pp. 31–38, 2001.

[40] A. von Heydebreck, W. Huber, A. Poustka and M. Vingron, "Identify- ing splits with clear separation: A new class discovery method for gene expression data," Bioinformatics, vol. 17, Suppl 1, pp. S107–S114, 2001.

[41] L. J. Van't Veer et al., "Gene expression profiling predicts clinical outcome of breast cancer," Nature, vol. 415, pp. 530–536, 2002.

[42] S. L. Pomeroy et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, no. 6870, pp. 436–442, 2002.

[43] S. A. Armstrong et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol.30, no. 1, pp. 41-47, 2002.

[44] I. Saha, D. Plewczynski, U. Maulik and S. Bandyopadhyay, "Improved Differential Evolution for Microarray Analysis," International Journal of Data Mining and Bioinformatics, Vol. 6, No. 1, pp. 86–103, 2012.

[45] S. Saha, A. Ekbal, K. Gupta, and S. Bandyopadhyay, "Gene expression data clustering using a multiobjective symmetry-based clustering technique," Computers in Biology and Medicine, vol. 43, no. 11, pp. 1965–197, 2013.

[46] A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, Multi-class Clustering of Cancer Subtypes through SVM based Ensemble of Pareto- optimal Solutions for Gene Marker Identification," PLoS One, vol. 5, no.11, art. id. e13803, 2010.

[47] R. Mitra, S. Bandyopadhyay, U. Maulik and M. Q. Zhang, SFSSClass: An integrated approach for miRNA-based tumor classification," BMC Bioinformatics, vol. 11, Suppl. 1, p. S22, DOI 10.1186/1471-2105-11- S1-S22, 2010.

[48] A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, "Multi-class clustering of cancer subtypes through SVM based ensemble of pareto- optimal solutions for gene marker identification," PLoS One, vol. 5, no. 11, e13803, 2010.

[49] Golub et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531–537, 1999.

[50] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, "Supervised harvesting of expression trees," Genome Biol, vol. 2, pp. 1–12, 2001.

[51] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” Proc Natl Acad Sci, USA, vol. 99, pp. 6567–6572, 2002.

[52] M. J. van de Vijver et al., “A gene expression signature as a predictor of survival in breast cancer,” N Engl J Med, vol. 347, pp. 1999–2009, 2002.

[53] C. L. Nutt et al., “Gene expression-based classification of malignant gliomas correlates better with survival than histological classification,” Cancer Res., vol. 63, pp. 1602-1607, 2003.

[54] L. Wang, F. Chu and W. Xie, “Accurate cancer classification using ex- pressions of very few genes,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no. 1, pp. 40–53, 2007.

[55] R. Simon, “Analysis of dna microarray expression data,” Best Practice and Research Clinical Haematology, vol. 22, no. 2, pp. 271–282, 2009.

[56] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio, “Support vector machine classification of microarray data,” AI Memo 1677, Massachusetts Institute of Technology, 1999.

[57] N. Pochet, F. De Smet, J. Suykens and B. De Moor, “Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction,” Bioinformatics, vol. 20, no. 17, pp. 3185–3195, 2004.

[58] D. Berrar, I. Bradbury and W. Dubitzky, “Instance-based concept learning from multiclass DNA microarray data,” BMC Bioinformatics, vol. 7, no. 73, 2006.

[59] N. B. Prasad, H. Somervell, R. P. Tufano, et al., “Identification of genes differentially expressed in benign versus malignant thyroid tumors,” Clinical Cancer Research: An Official journal of the American Association for Cancer Research, vol. 14, no. 11, pp. 3327 –3337, 2008.

[60] A. Keller, M. Schummer, L. Hood and W. Ruzzo, “Bayesian classification of DNA array expression data,” Technical report, University of Washington, August, 2000.

[61] N. Friedman, M. Linial, M. Nachman and D. Peer, “Using Baysian networks to analyze expression data,” Journal of Computational Biology, vol. 7, pp. 601–620, 2000.

[62] A. Kelemen, Z. Hong, P. Lawhead and L. Yulan, “Naive Bayesian classifier for microarray data,” IEEE Proceedings of the International Conference on Neural Networks, vol. 3, pp. 1769–1773, 2003.

[63] H. Y. Chen, S. L. Yu, C. H. Chen, et al., “A five-gene signature and clinical outcome in non-small-cell lung cancer,” The New England Journal of Medicine, vol. 356, no. 1, pp. 11-20, 2007.

[64] V. Vapnik, Statistical Learning Theory, Wiley, 1998.

[65] B. Michael et al., “Knowledge-based analysis of microarray gene expression data by using support vector machines,” Proceedings of the National Academy of Sciences, vol. 97, no. 1, pp. 262–267, 2000.

[66] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” Bioinformatics, vol. 16, no. 10, pp. 906–914, 2000.

[67] F. Chu and L. Wang, “Applications of support vector machines to cancer classification with microarray data,” International Journal of Neural Systems, Vol. 15, No. 6, pp. 475–484, 2005.

[68] P. A. Mundra and J. C. Rajapakse, “Gene and sample selection for can- cer classification with support vectors-based t-statistic,” Neurocomputing, vol. 73, no. 1315, pp. 2353–2362, 2010.

[69] S. Mingjun and S. Rajasekaran, “A greedy algorithm for gene selection based on SVM and correlation,” Int J Bioinform Res Appl., vol. 6, no. 3, pp. 296–307, 2010.

[70] F. Schnorrenberget at al., “Improved detection of breast cancer nuclei using modular neural networks,” IEEE Engineering in Medicine and Biology Magazine, vol. 19, no. 1, pp. 48 63, 2000.

[71] F. Azuaje, “A computational neural approach to support the discovery of gene function and classes of cancer,” IEEE Transactions on

Biomedical Engineering, vol. 48, no. 3, pp. 332–339, 2001.

[72] Y. Xu et al., "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barretts esophagus and esophageal cancer," Cancer Research, vol. 62, pp. 3493–3497, 2002.

[73] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data," Bioinformatics, vol. 21, pp. 2394– 2402, 2005.

[74] M. Cinar, M. Engin, E.Z. Engin and Y.Z. Atesci, "Early prostate cancer diagnosis by using artificial neural networks and support vector machines," Expert Syst. Appl., vol. 36, no. 3, part 2, pp. 6357–6361, 2009.

[75] G. B. Zhang, N. Sundararajan and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene ex- pression cancer diagnosis," "IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 4, no.3, pp. 485–495, 2007.

[76] E. Bair and R. Tibshirani, "Semi-supervised methods to predict patient survival from gene expression data," Plos Biol, vol. 2, pp. 511–522, 2004.

[77] X. Rui, G. C. Anagnostopoulos and D.C.I.I. Wunsch, "Multiclass cancer classification using semi supervised ellipsoid ARTMAP and particle swarm optimization with gene expression data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.4, No.1, pp. 65–77, 2007.

[78] I. Steinfeld, R. Navon, D. Ardig'o, I. Zavaroni, and Z. Yakhini, "Clinically driven semisupervised class discovery in gene expression data," Bioinformatics, vol. 24, pp. 190–197, 2008.

[79] Sounak Chakraborty, "Bayesian semi-supervised learning with support vector machine," Statistical Methodology, vol. 8, no. 1, pp. 68–82, 2011.

[80] D. C. Koestler, C. J. Marsit, B. C. Christensen, M. R. Karagas, R. Bueno, D. J. Sugarbaker, K. T. Kelsey, and E. A. Houseman, "Semisu- pervised recursively partitioned mixture models for identifying cancer subtypes," Bioinformatics, vol. 26, pp. 2578–2585, 2010.

[81] M. Shi, B. Zhang, "Semisupervised learning improves gene expression- based prediction of cancer recurrence," Bioinformatics, vol. 27, no. 21, pp. 3017–3023, 2011.

[82] S. Pang, T. Ban, Y. Kadobayashi and N. Kasabov, "Personalized mode transductive spanning SVM classification tree," Information Sciences, vol. 181, pp. 2071–2085, 2011.

[83] M. Dash and H. Liu, "Consistency based search in feature selection," Artificial Intelligence, vol. 151, pp. 155-176, 2003.

[84] C. Park, J. Ahn, H. Kim, and S. Park, "Integrative gene network con- struction to analyze cancer recurrence using semi-supervised learning," PloS One, vol. 9, no. 1, e86309, 2014.

[85] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," Appl Bioinformatics, vol. 2, no. 3, pp. S75–S83, 2003.

[86] M. Dettling, "Bagboosting for tumor classification with gene expres- sion data," Bioinformatics, vol. 20, no. 18, pp. 3583–3593, 2004.

[87] H. Yu and S. Xu, "Simple rule-based ensemble classifiers for cancer DNA microarray data classification," International Conference on Com- puter Science and Service System (CSSS), pp. 2555–2558, 2011.

[88] H. Jorng-Tzong, W. Li-Cheng, L. Baw-Juine, K. Jun-Li, K. Wen- Horng, and Z. Jin-Jian, "An expert system to classify microarray gene expression data using gene selection by decision tree," Expert Syst. Appl., vol. 36, no. 5, pp. 9072–9081, Jul. 2009.

[89] S. Ghorai, A. Mukherjee, S. Sengupta and P. K. Dutta, "Cancer classification from gene expression data by NPPC ensemble," IEEE/Acm Transactions on Computational Biology and Bioinformatics, Vol. 8, No. 3, pp. 659–671, 2011.

[90] A. Nogueira, A. Ferreira and M. Figueiredo, "A Machine Learning Pipeline for Cancer Detection on Microarray Data: The Role of Feature Discretization and Feature Selection",2022.

[91] Shao, D, Dai, Y., Li, N., Cao, X., Zhao, W., Cheng, L., Rong, Z., Huang, L., Wang, Y., & Zhao, J. "Artificial intelligence in clinical research of

cancers. Briefings in Bioinformatics," 23(1), 2021.

[92] Osama, S., Shaban, H., & Ali, A. A., "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review.", Expert Systems with Applications, 213, 118946, 2023.

[93] Dhillon, A., Kaur, A., & Singh, A.," Application of Machine Learning for Prediction of Lung Cancer using Omics Data.", 2020.

[94] Mirza, Z., Ansari, M. S., Iqbal, M. S., Ahmad, N., Alganmi, N., Banjar, H., Al-Qahtani, M., & Karim, S.," Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis." Cancers, 15(12), 3237,2023.

[95] Joon, H. K., Thalor, A., & Gupta, D., "Machine Learning Analysis of Lung Squamous Cell Carcinoma Gene Expression Datasets Reveals Novel Prognostic Signatures.",2023.

[96] Upadhyay, S., Bhushan, R., Tripathi, A., Chaubey, L., Diwakar, A., & Dubey, P. K., "Differential gene expression profile evaluation between the human uterine leiomyoma and leiomyosarcoma using a machine learning approach." Gynecology and Obstetrics Clinical Medicine, 2023.