HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Integrative analysis of RNA expression data unveils distinct cancer types through machine learning techniques

Saad Awadh Alanazi [a,*], Nasser Alshammari [a], Maddalah Alruwaili [b], Kashaf Junaid [c], Muhammad Rizwan Abid [d], Fahad Ahmad [e]

[a] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia
[b] Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72341, Saudi Arabia
[c] School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4NS, United Kingdom
[d] Department of Computer Science, Florida Polytechnic University, Lakeland, FL 33805, United States
[e] Department of Basic Sciences, Common First Year, Jouf University, Sakaka 72341, Saudi Arabia

ARTICLE INFO

Keywords:
Gene
RNA expression, cancer
Tumour, diagnosis
Machine learning
Clustering
Classification

ABSTRACT

Cancer is a highly complex and heterogeneous disease. Traditional methods of cancer classification based on histopathology have limitations in guiding personalized prognosis and therapy. Gene expression profiling provides a powerful approach to unraveling molecular intricacies and better-stratifying cancer subtypes. In this study, we performed an integrative analysis of RNA sequencing data from five cancer types - BRCA, KIRC, COAD, LUAD, and PRAD. A machine learning workflow consisting of dataset identification, normalization, feature selection, dimensionality reduction, clustering, and classification was implemented. The k-means algorithm was applied to categorize samples into distinct clusters based solely on gene expression patterns. Five unique clusters emerged from the unsupervised machine learning based analysis, significantly correlating with the known cancer types. BRCA aligned predominantly with one cluster, while COAD spanned three clusters. KIRC was represented within two main clusters. LUAD is associated strongly with a single cluster and PRAD with another cluster. This demonstrates the ability of machine learning approaches to unravel complex signatures within transcriptomic profiles that can delineate cancer subtypes. The proposed study highlights the potential of integrative analytics to derive meaningful biological insights from high-dimensional omics datasets. Molecular subtyping through machine learning clustering enhances our understanding of the intrinsic heterogeneities and pathways dysregulated in different cancers. Overall, this study exemplifies a powerful computational framework to classify gene expressions of patients having different types of cancers and guide personalized therapeutic decisions. Finally, Wide Neural Network demonstrates a significantly higher accuracy, achieving 99.834% on the validation set and an even more impressive 99.995% on the test set.

1. Introduction

Cancer is a highly prevalent and profoundly impactful global disease that affects people across the world. According to the World Health Organization Global Cancer Report, it is projected that the global incidence of cancer will increase by a significant 57 % over the next two decades. This disease, characterized by pathological disruptions in the natural process of cellular division, is responsible for a substantial global mortality rate. In 2020 alone, there were more than 19.3 million newly diagnosed cancer cases, resulting in an estimated 10 million fatalities, as reported by the Global Cancer Report (Arslan et al., 2022). Cancer imposes a significant healthcare burden, affecting not only individuals diagnosed with the disease but also their families and the healthcare systems that serve them (Malebari et al., 2020).

According to (Sung et al., 2021), breast, lung, colorectal, prostate, stomach, and liver cancers are among the most frequently diagnosed kinds of cancer worldwide. Breast cancer is widely recognized as the predominant form of cancer affecting women, ahead of lung cancer on a global level (Mei and Wu, 2022). On the other hand, prostate cancer ranks highest among cancer types in men, closely followed by colorectal

* Corresponding author.
E-mail addresses: sanazi@ju.edu.sa (S.A. Alanazi), nashamri@ju.edu.sa (N. Alshammari), madallah@ju.edu.sa (M. Alruwaili), kashaf.junaid@qmul.ac.uk (K. Junaid), mabid@floridapoly.edu (M.R. Abid), fahad.ahmad@ieee.org (F. Ahmad).

https://doi.org/10.1016/j.sjbs.2023.103918
Received 14 November 2023; Received in revised form 15 December 2023; Accepted 22 December 2023
Available online 30 December 2023

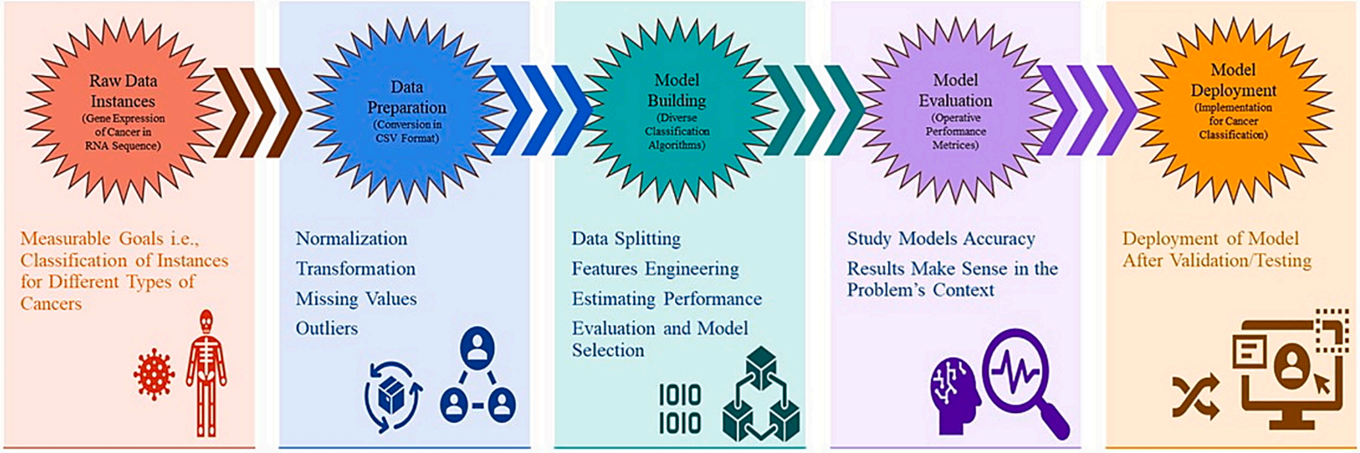

Raw Data Instances
(Gene Expression of Cancer in RNA Sequence)
Measurable Goals i.e., Classification of Instances for Different Types of Cancers
Data Preparation
(Conversion in CSV Format)
Normalization
Transformation
Missing Values
Outliers
Model Building
(Diverse Classification Algorithms)
Data Splitting
Features Engineering
Estimating Performance
Evaluation and Model Selection
1010
1010
Model Evaluation
(Operative Performance Metrices)
Study Models Accuracy
Results Make Sense in the Problem's Context
Model Deployment
(Implementation for Cancer Classification)
Deployment of Model After Validation/Testing

Fig. 1. Process gene expression classification in various tumour types.

and lung cancer (Sanko and Kuralay, 2023). Moreover, thyroid, colorectal, and cervical cancers are among the other malignancies that are commonly identified in females.

The effects of cancer on general health are profound and diverse. Each year, cancer takes the lives of millions of people and puts a huge strain on individuals, families, and healthcare systems. This burden consists of financial pressure brought on by the high expense of cancer treatment and care, in addition to physical and emotional difficulties (Alsayari et al., 2021). In addition, the growing prevalence of cancer worldwide emphasizes the critical need for enhanced preventative measures, early detection techniques, and efficient therapeutic alternatives to lessen the impact of cancer (Rigel and Carucci, 2000).

The molecular basis of cancer is a complex and multifaceted topic that involves various genetic, molecular, and cellular alterations that drive the development and progression of cancer (Gil-Hernández et al., 2021). Hanahan and Weinberg's key article highlighted fundamental characteristics of cancers that have been widely acknowledged (Gyamfi et al., 2022). The model they presented provided robust evidence in favour of the genetic basis of cancer, which suggests that the disease arises from the accumulation of mutations, epigenetic modifications, and genetic alterations in critical genes responsible for governing cell growth, cell division, cell metabolism, and cell replication. Based on their effect after mutation, these genes are classed as oncogenes or tumor suppressors (Martínez-Jiménez et al., 2020). Although incomplete, the genomic model for cancer formation has provided important insights into the genetic events driving cancer origin, progression, metastasis, response to therapy, and drug resistance development (Hanahan and Weinberg, 2000, Dancey et al., 2012, Hawkes, 2019). The theory is broadly supported by the identification of mutations in specific genes across a diverse array of tumor types. An extensive catalogue of the numerous genes mutated in malignancies has been generated since the advent of sequencing. Over 1000 genes have been linked to cancer and are classed as cancer-associated genes (250 oncogenes, 700 tumor suppressors) (Wishart, 2015).

Additionally, the variation in treatment response, disease progression, and prognosis among cancer patients is influenced by the heterogeneity of cancer, which is the manifestation of unique genetic, molecular, and cellular attributes in various types and subtypes of cancer. To deal with this issue over the last two decades, scientists have profiled the gene expression of human malignancies in detail, with data from thousands of studies released into the public domain. Many different -omics levels can be used to study cancer, but the level of transcriptomics has the most data so far because most of the academic labs started using DNA microarrays in the late 1990s. Afterward, the next-generation sequencing has made RNA sequencing (RNA-seq) a mainstream transcriptomics platform. Gene expression encompasses both messenger RNA (mRNA) and protein, and the correlation between the two may not always be robust. RNA-Seq is a relatively new and widely used technology for detecting novel isoforms and transcripts by providing more normalized and less noisy data for prediction and classification (Mehmood et al., 2022; Munawar et al., 2022). The primary goal of transcriptome profiling is to identify differentially expressed genes in the body or to discover alterations in genes at different levels (Wang et al., 2023). RNA sequencing allows for both identification and quantification in a single step (Wesolowski et al., 2013).

RNA-Seq data from various databases are widely available and are being used to classify many cancers (Urda et al., 2017). However, due to their high dimensions, complexity, and the presence of feature value duplications, assessments of RNA gene expression data are highly complex (Danaee et al., 2017). Hence, an algorithmic approach utilizing machine learning (ML) and deep learning (DL) may be employed to execute automatic feature extraction (Mehmood et al., 2022; Alharbi and Vakanski, 2023).

When analyzing RNA expression data to differentiate between different forms of cancer, ML approaches are essential. These techniques use sophisticated algorithms to find intricate relationships and patterns in huge datasets that human analysis could overlook (Liu et al., 2023). This ability is especially important in oncology because various cancer types can have quite varied RNA expression profiles (Xiao et al., 2023). Personalized medicine and focused therapy development depend on ML's capacity to process and interpret these changes. Moreover, it helps in the early identification and categorization of malignancies, which may result in better patient outcomes and more efficient interventions. As science progresses, ML remains a vital weapon in the battle against several intricate forms of cancer.

Our understanding of genetic abnormalities and diseases is changing because of the application of ML in the study of RNA expression data. ML algorithms can detect minute patterns in RNA sequences, providing insights into gene regulation and expression, by effectively analyzing large datasets (Li et al., 2022). This capacity is essential for comprehending intricate biological processes and the molecular causes of numerous disorders. Furthermore, through the analysis of RNA expression patterns, ML helps to anticipate the course of a disease and its response to treatment, enabling more accurate and customized therapeutic approaches. Its application is having a substantial impact on the fields of genomics and personalized medicine by speeding up the identification of novel biomarkers and improving diagnostic accuracy (Steyaert et al., 2023).

The continuous emerging incidences of cancer worldwide that causes millions of deaths annually have generated the need and demand for developing advanced and sophisticated tools for ML based classification tools that can take RNA-seq as input and identify different type of cancer

(Liñares Blanco et al., 2019). Although RNA-Seq data are useful for detecting alterations at the gene level, working with them is difficult due to their spatial characteristics. For classifying cancer based on gene expression data, thirty-four different ML techniques have been implemented in this work. In addition, we analyzed RNA-Seq data from five distinct tumors. The process of classification is represented in Fig. 1.

1.1. Problem statement

In the recent research endeavor, we focused on distinguishing between various tumor types: BRCA, KIRC, COAD, LUAD, and PRAD. The crux of the challenge is to employ ML techniques to accurately categorize these tumor types based on specific gene expressions. Thus, the problem at hand is to design and optimize an ML model that can precisely classify the tumor into one of the aforementioned categories, thereby aiding in more specific and targeted medical interventions.

1.2. Aims and objectives

The primary aim of this study is to enhance the precision and realism of cancer tumor classifications derived from RNA gene expression. Leveraging advanced ML techniques, we aim to surpass the efficacy of current classification systems. Our objective is not merely to refine the analytical approach, but also to provide medical practitioners with a more reliable tool, facilitating more informed decision-making in the treatment of various cancer tumors.

The principal contribution of this study lies in its systematic evaluation of various ML algorithms to determine their efficacy in classifying diverse cancer types. Not only does this research pinpoint the most optimal ML algorithm for general cancer classification, but it also provides a nuanced understanding, highlighting which specific algorithms excel in identifying cancer types. This granularity in algorithmic performance for individual cancer categories sets our study apart and offers valuable insights for targeted oncological research.

The structure of this paper is delineated as follows: Section 1 provides a comprehensive introduction to the topic. Section 2 outlines the methodologies employed in the research. Section 3 delves into the experiments conducted and their respective results. A detailed discussion of these findings is presented in Section 4. The paper concludes in Section 5, summarizing the experimental procedures, outcomes, and directions for future research.

2. Materials and methods

2.1. System specifications

On the RNA-seq dataset, the anticipated ML based classification algorithms were evaluated for gene expression classification in various tumor types, including BRCA, KIRC, COAD, LUAD, and PRAD. Experiments were carried out on a Lenovo Mobile Workstation equipped with a Processor: 12th Generation Intel Core i9, Operating System: Windows 11 Pro 64, Memory: 128 GB DDR4, Hard Drive: 4 TB SSD, Graphics: NVIDIA RTX A4000. We have used MATLAB R2023a and Orange-v3.36 tools for the explanation and results.

2.2. Dataset collection and preprocessing

The RNA-seq data in question has been sourced from the UCI ML Repository, specifically the gene expression cancer RNA-Seq dataset. This dataset is a component of the RNA-Seq (HiSeq) PANCAN collection. It presents a random extraction of gene expressions from patients diagnosed with various tumor types, including - BReast CAncer (BRCA), KIdney Renal cell Carcinoma (KIRC), COlon ADenocarcinoma (COAD), LUng ADenocarcinoma (LUAD), and PRostate ADenocarcinoma (PRAD). Organized row-wise, each sample or instance holds RNA-Seq gene expression levels, as captured by the Illumina HiSeq platform. Each attribute within the dataset is given a placeholder name in the format "gene_XX". Accompanying the dataset is a CSV manifest which offers file annotations and supplementary details for every file. With 801 instances and 16,383 features, the dataset encompasses a total of 801 samples.

2.3. Comparison between machine learning models

In our study, we sourced gene expression cancer RNA-Seq data from the UCI ML Repository, specifically a subset of the RNA-Seq (HiSeq) PANCAN dataset. This subset contained random extractions of gene expressions from patients presenting various tumor types: BRCA, KIRC, COAD, LUAD, and PRAD. Utilizing this dataset with 801 instances and 16,383 features, we embarked on a comparative analysis of 34 ML models. Our goal was to assess their classification accuracy in distinguishing these tumor types. Ensuring rigor, we adopted a cross-validation scheme with five folds to mitigate the risk of overfitting. The dataset, represented row-wise, exhibited RNA-Seq gene expression levels gauged by the illumina HiSeq platform, with each attribute labelled with a placeholder name, such as "gene_XX". Thirty-four models were tested, and the dataset's primary characteristic was its multivariate nature in the realm of life sciences, suitable for classification tasks.

2.4. Train test split

The data was divided into two main subsets, with 75 % allocated for training and the remaining 25 % reserved for testing. This division was done randomly to eliminate any potential biases. Within the training data, further subdivisions were made into training and validation sets using the K-Fold cross-validation method, specifically with a K value of 5. The rationale behind using K-Fold cross-validation was to optimize hyperparameters effectively. This method, along with others such as leave-one-out, leave-p-out, and Monte-Carlo sampling, facilitates the division of data for training the model and validating its performance. The core objective of such a process is to gauge the ML model's capacity to generalize its learning to fresh, unseen data and to pinpoint the best hyperparameters.

2.5. Hyperparameter optimization

Hyperparameter optimization plays a pivotal role in determining the performance of an ML model. The choice of hyperparameters can profoundly impact how efficiently an algorithm learns from the data and, ultimately, its predictive power. Various optimization strategies exist that aim to refine these parameters to enhance the performance of ML models. This importance is underscored when looking at different ML models and their respective hyperparameters.

A diverse range of ML models, from decision trees and support vector machines to neural networks and ensemble methods have been adopted. For instance, decision trees have parameters such as the maximum number of splits and split criterion, while support vector machines have kernel functions, kernel scales, and box constraint levels. Neural networks introduce another layer of complexity with parameters indicating the number of layers, sizes of each layer, and activation functions. Across all these models, specific hyperparameters were chosen and fine-tuned to optimize their performance.

Additionally, the study provides insights into other crucial metrics, like prediction speed, training time, and model size. A consistent trend in feature selection, based on the Analysis of Variance (ANOVA) feature ranking algorithm, can also be seen, with models utilizing nearly the full feature set available (16,345 out of 16,383 features). The choice of hyperparameters and their optimization, combined with the right feature selection, ensures that these models capture the underlying patterns in the data efficiently while minimizing overfitting. The emphasis on maintaining a high explained variance (95 % in this case) further solidifies the importance of selecting and tuning the correct hyperparameters.

2.6. Performance metrics

In the training of ML models, various performance metrics were employed to evaluate the efficacy of the models. These metrics included Sensitivity (True Positive Rate), False Negative Rate, Precision (Positive Predictive Value), False Discovery Rate, and the Area Under the Curve (AUC). Each of these metrics offers a different perspective on the model's ability to correctly predict and classify data points. For instance, while Sensitivity gauges the proportion of positives that are correctly identified, Precision provides insight into the ratio of correctly predicted positive observations to the total predicted positives. Moreover, AUC provides a comprehensive overview of the model's classification ability across all possible thresholds.

During the optimization of hyperparameters, performance metrics represented by Eq. (1), Eq. (2), Eq. (3), and Eq. (4) served as vital benchmarks. The exact formulas to compute them were provided. It's essential to note that selecting the right metric depends on the specific problem and the costs associated with different types of errors. For instance, in scenarios where false negatives could have grave consequences, one might prioritize Sensitivity over other metrics. On the other hand, in situations where false positives are more detrimental, Precision might be given more importance.

$$\text{Sensitivity}: \quad \text{True Positive Rate} = TP/(TP + FN) \tag{1}$$

$$\text{Miss Rate}: \text{False Negative Rate}: \quad FNR = FN/(FN + TP) \tag{2}$$

$$\text{Precision}: \quad \text{Positive Predictive Value} = TP/(TP + FP) \tag{3}$$

$$\text{False Discovery Rate}: \quad FDR = FP/(FP + TP) \tag{4}$$

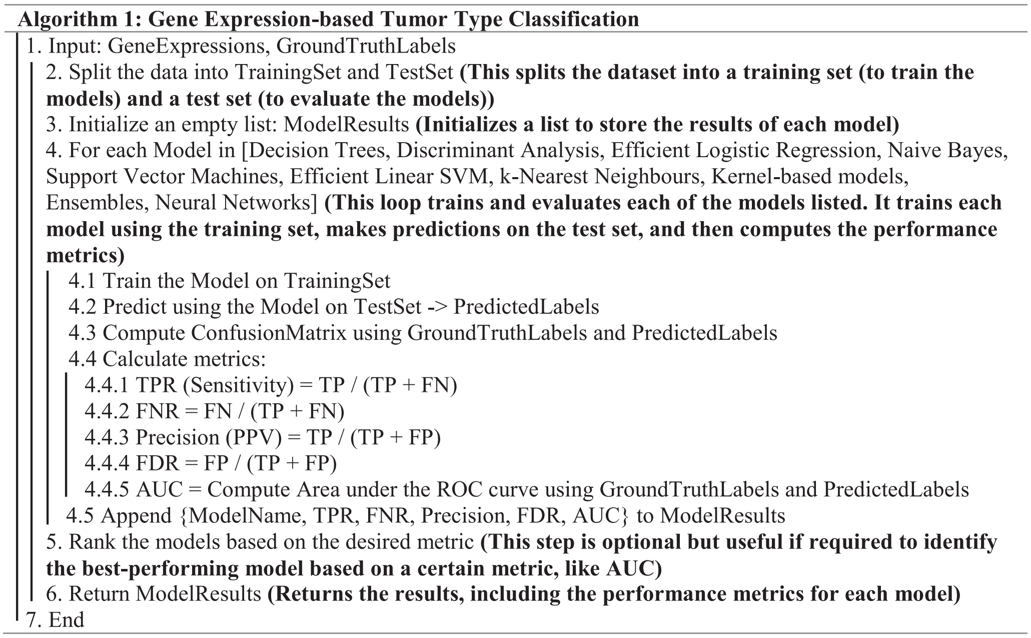
Here,

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Area Under the Curve (AUC): The AUC refers to the area under the Receiver Operating Characteristic (ROC) curve. It is a measure of a model's ability to distinguish between the positive and negative classes across all possible thresholds. AUC is usually computed using numerical methods to assess the integral of the ROC curve.

The comprehensive procedure, spanning from the collection of RNA gene expression datasets to the classification of various cancer types, is detailed in Algorithm 1 provided below.

Algorithm 1: Gene Expression-based Tumor Type Classification



```
Algorithm 1: Gene Expression-based Tumor Type Classification
1. Input: GeneExpressions, GroundTruthLabels
    2. Split the data into TrainingSet and TestSet (This splits the dataset into a training set (to train the
    models) and a test set (to evaluate the models))
    3. Initialize an empty list: ModelResults (Initializes a list to store the results of each model)
    4. For each Model in [Decision Trees, Discriminant Analysis, Efficient Logistic Regression, Naive Bayes,
    Support Vector Machines, Efficient Linear SVM, k-Nearest Neighbours, Kernel-based models,
    Ensembles, Neural Networks] (This loop trains and evaluates each of the models listed. It trains each
    model using the training set, makes predictions on the test set, and then computes the performance
    metrics)
        4.1 Train the Model on TrainingSet
        4.2 Predict using the Model on TestSet -> PredictedLabels
        4.3 Compute ConfusionMatrix using GroundTruthLabels and PredictedLabels
        4.4 Calculate metrics:
            4.4.1 TPR (Sensitivity) = TP / (TP + FN)
            4.4.2 FNR = FN / (TP + FN)
            4.4.3 Precision (PPV) = TP / (TP + FP)
            4.4.4 FDR = FP / (TP + FP)
            4.4.5 AUC = Compute Area under the ROC curve using GroundTruthLabels and PredictedLabels
        4.5 Append {ModelName, TPR, FNR, Precision, FDR, AUC} to ModelResults
    5. Rank the models based on the desired metric (This step is optional but useful if required to identify
    the best-performing model based on a certain metric, like AUC)
    6. Return ModelResults (Returns the results, including the performance metrics for each model)
7. End
```

3. Experimental results

3.1. Clustering of datasets

In our study, we employed a clustering method to categorize datasets based on shared characteristics. The aim was to elucidate significant groupings and gain a deeper understanding of the data's intrinsic structure. This approach revealed discerning RNA-seq patterns and classifications pertaining to various cancers. The outcomes for the RNA-seq clustering are depicted in Fig. 2, showcasing five distinct clusters. The Silhouette plot provides a visual representation of the cohesion within each data cluster, allowing for an intuitive evaluation of cluster integrity. The Silhouette score quantifies the similarity of an object to its respective cluster relative to others. A score nearing 1 suggests that the data point is proximate to the cluster's center, while scores approaching 0 indicate a position near the boundary of two adjacent clusters.

The relationship between cancer types and clusters is illustrated in Fig. 3. The BRCA category is primarily represented by C3. COAD consists of three clusters: C4, C5, and C2, with C2 being the most dominant. KIRC predominantly has C2 and a smaller portion of cluster 4. LUAD is represented by C4, while PRAD is mainly represented by C1. Statistical analysis indicates a significant association between the type of cancer and clustering.

Results of mean distribution of dataset among each cancer type are represented in Fig. 4. In a comparative analysis of various datasets relating to different cancer types, significant variations were observed as indicated by a p-value of 0.001. COAD presented the highest mean value of (0.57414 ± 0.0273) followed by KIRC (0.567084 ± 0.0203), PRAD

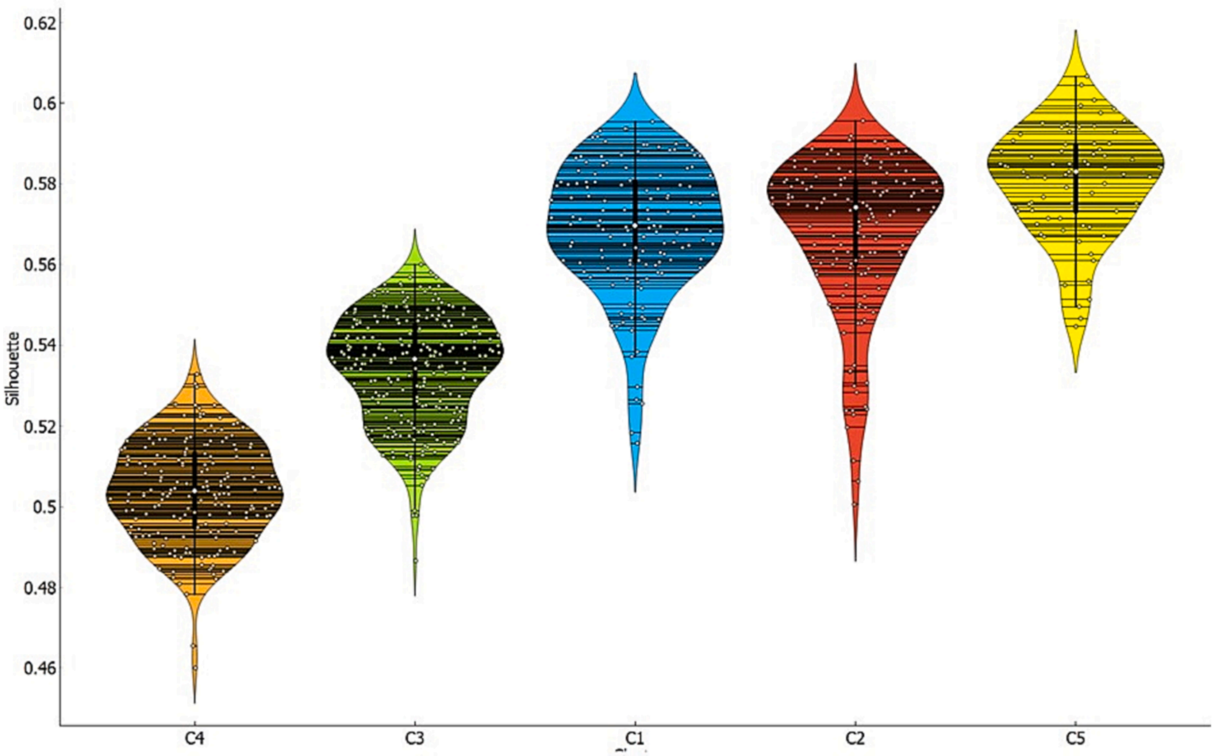


Fig. 2. This visualization showcases five violin plots representing data distributions for five different clusters labelled C4, C3, C1, C2, and C5. Each plot is colour-coded for differentiation: C4 is in brown, C3 is in green, C1 is in blue, C2 is in red, and C5 is in yellow. The y-axis labelled "Silhouette" has values ranging approximately from 0.46 to 0.62. The width of each violin plot at various y-values indicates the density of data points, with wider sections symbolizing higher data density and vice versa.

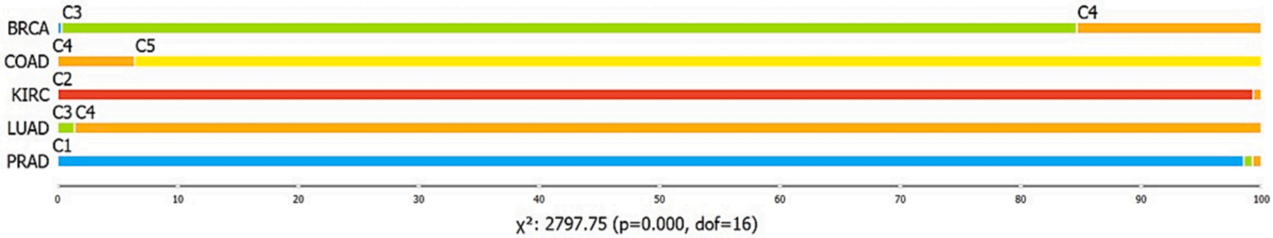


Fig. 3. The provided visualization represents a horizontal bar chart that depicts the distributions of various categories, possibly referring to medical or research datasets. The chart consists of seven primary labels on the left: BRCA, COAD, KIRC, LUAD, and PRAD. Next to each primary label, there are coloured horizontal bars with labels such as C1, C2, C3, C4, and C5, each corresponding to different lengths on the x-axis. The x-axis is numerical, starting from 0 and extending up to 100, marked at intervals of 10.

(0.568109 ± 0.019), BRAC (0.528186 ± 0.0193), and LUAD (0.508626 ± 0.0098). Below the graphical representation, there's a reference to an ANOVA statistical test, indicating a value of 325.269 and an associated p-value of 0.000. The sample size for the test, as denoted by "N" is 801. The very low p-value suggests that the differences between the categories are statistically significant. The x-axis of the graph ranges from 0.5 to 0.65, which defines the scale for the represented data points.

3.2. Classification accuracy during training and testing

The experimental setup for a classification task involved preprocessing 600 samples for training, and validation and 201 samples for testing. After feature selection, 38 zero-score features were removed. Principal Component Analysis (PCA) was then applied, resulting in 16,340 numeric components that explained 95 % of the data variance. Hyperparameters for the ML models were optimized using Bayesian Optimization, with a limit of 500 s and 10 grid divisions. This process aimed to enhance data preparation and model accuracy.

Table 1 presents classification accuracy and total cost results for various ML models during both the training (Validation) and testing (Test) phases. The models include Decision Trees, Discriminant Analysis, Efficient Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Efficient Linear SVM, k-Nearest Neighbours (KNN), Kernel-based models, Ensembles, and Neural Networks.

Decision Trees have repeatedly exhibited excellent levels of accuracy in both training and testing phases, with accuracy percentages ranging from 86.31 % to 99.01 %. Discriminant Analysis demonstrates impeccable accuracy in both training and testing phases, whereas Efficient Logistic Regression exhibits commendable performance with accuracy ratings beyond 99 %. The Naive Bayes classifier demonstrates fluctuating levels of accuracy, with training accuracy reaching a maximum of 95.83 % and testing accuracy reaching a maximum of 91.58 %. Support Vector Machines (SVM) often exhibit favourable performance; nonetheless, there exists variability in accuracy across different instances,

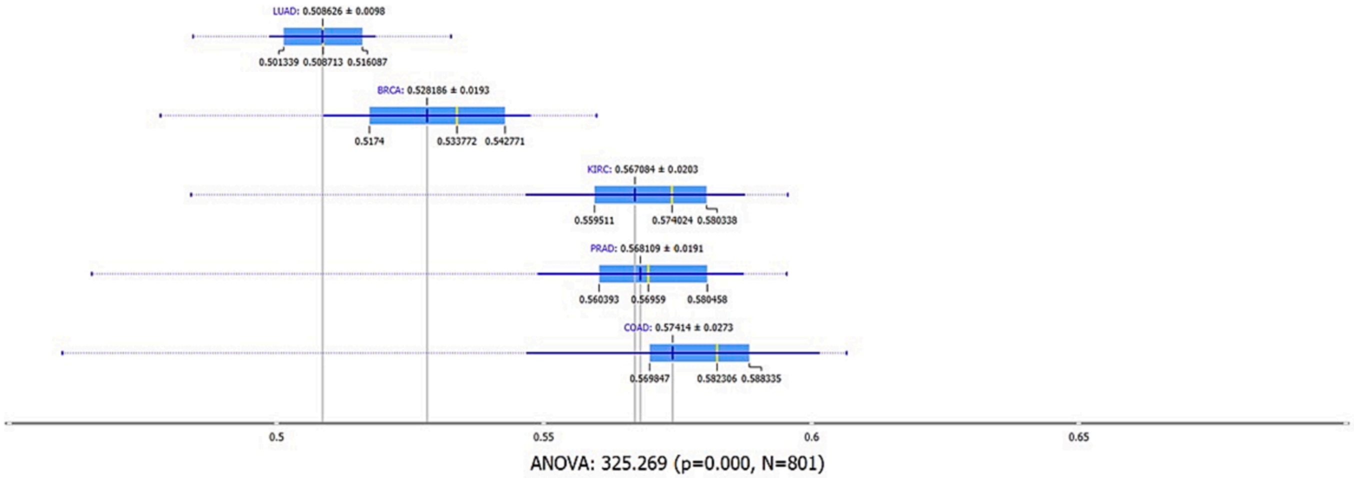


Fig 4. The visualization presents a graphical representation of data related to various categories, possibly referring to specific types of conditions or samples. These categories are labelled as “LUAD”, “BRCA”, “KIRC”, “PRAD”, and “COAD”. Each category has a central data point, denoted by a vertical line, with accompanying horizontal error bars that likely represent variability or uncertainty in the data (standard deviation or confidence intervals).

Table 1
Classification accuracy during training and testing.

Model Type	Accuracy % (Validation)	Total Cost (Validation)	Accuracy % (Test)	Total Cost (Test)
Fine Tree	97.496	15	99.009	2
Medium Tree	97.496	15	99.009	2
Course Tree	86.311	82	88.119	24
Linear Discriminant	99.967	0	99.978	0
Quadratic Discriminant	NaN	NaN	NaN	NaN
Efficient Logistic Regression	99.833	1	99.505	1
Gaussian Naïve Bayes	89.6493	62	79.208	42
Kernel Naïve Bayes	95.8263	25	91.585	17
Linear SVM	99.6663	2	99.505	1
Quadratic SVM	99.8333	1	99.505	1
Cubic SVM	99.499	3	99.505	1
Fine Gaussian SVM	80.968	114	74.257	52
Medium Gaussian SVM	36.394	381	40.594	120
Coarse SVM	36.394	381	40.594	120
Efficient Logistic Regression	99.666	2	99.505	1
Efficient Logistic SVM	99.833	1	99.856	0
Fine KNN	41.235	352	38.119	125
Medium KNN	26.711	439	25.743	150
Coarse KNN	16.861	498	17.327	167
Cosine KNN	99.666	2	99.505	1
Cubic KNN	27.713	433	26.238	149
Weighted KNN	26.210	442	25.248	151
SVM Kernel	99.833	1	99.009	2
Logistic Regression Kernel	98.998	6	98.515	3
Boosted Tree	36.394	381	40.595	120
Bagged Trees	98.165	11	98.515	3
Subspace Discriminant	99.891	0	99.900	0
Subspace KNN	99.834	1	99.163	0
RUSBoosted Tree	97.997	12	97.525	5
Narrow Neural Network	97.829	13	99.836	0
Medium Neural Network	99.812	0	99.505	1
Wide Neural Network	99.834	1	99.995	0
Bilayered Neural Network	93.489	39	95.049	10
Trilayered Neural Network	88.481	69	83.664	33

with training accuracy ranging from 80.97 % to 99.83 % and testing accuracy ranging from 74.26 % to 99.50 %. The accuracy of k-Nearest Neighbours (KNN) and Kernel-based models is found to be moderate. During the training phase, KNN achieves accuracy ranging from 16.86 % to 99.67 %, while during the testing phase, the accuracy ranges from 17.33 % to 99.50 %. Ensembles consistently demonstrate high performance, achieving accuracy rates that surpass 97 %. Finally, it is seen that Neural Networks exhibit a range of accuracy levels, ranging from 88.48 % to 99.99 % during the training phase and from 83.66 % to 99.50 % during the testing phase.

Table 2 provides a detailed summary of various classification models applied to a dataset, focusing on several key aspects of each model’s performance and characteristics. These models aim to classify cancer types based on gene expression cancer RNA-Seq data.

3.3. Model performance and characteristics

Diverse sets of classification models were employed, each described

Table 2
Characteristics of deployed classification techniques.

Model Type	Preset	Prediction Speed (obs/sec)	Training Time (sec)	Model Size (bytes)	Hyperparameters
Tree	Fine Tree	33.27	95.05	60812.00	Max Splits Allowed: 100; Criterion for Splitting: Gini Diversity Index; Use of Surrogate Splits: Disabled
Tree	Medium Tree	32.14	97.43	60812.00	Max Splits Allowed: 20; Criterion for Splitting: Gini Diversity Index; Use of Surrogate Splits: Disabled
Tree	Coarse Tree	31.25	99.26	58556.00	Max Splits Allowed: 4; Criterion for Splitting: Gini Diversity Index; Use of Surrogate Splits: Disabled
Discriminant	Linear Discriminant	30.60	101.72	2267717.00	Covariance Structure: Full
Discriminant	Quadratic Discriminant	NaN	11.10	NaN	Covariance Structure: Full
Efficient Logistic Regression	Efficient Logistic Regression	30.17	103.79	761299.00	Model: Logistic Regression; Solver Selection: Automatic; Regularization Parameter (Lambda): Automatic; Coefficient Convergence Threshold (Beta Tolerance): 0.000
Naive Bayes	Gaussian Naive Bayes	8.54	316.34	322549.00	Numeric Predictor Distribution Type: Gaussian; Categorical Predictor Distribution: Not Applicable
Naive Bayes	Kernel Naive Bayes	26.67	124.57	11161160.00	Numeric Predictor Distribution: Kernel; Categorical Predictor Distribution: Not Applicable; Selected Kernel: Gaussian; Range of Kernel: Unbounded
Support vector Machine	Linear Support Vector Machine	30.50	102.55	712505.00	Kernel Choice: Linear; Kernel Scaling: Set Automatically; Box Constraint Level: 1; Multiclass Strategy: One-vs-One; Data Normalization: Enabled
Support vector Machine	Quadratic Support Vector Machine	26.76	214.46	7664705.00	Kernel Type: Quadratic; Kernel Scaling: Auto-Determined; Box Constraint Setting: 1; Multiclass Classification Technique: One-vs-One; Data Standardization: Enabled
Support vector Machine	Cubic Support Vector Machine	30.44	102.19	7738505.00	Kernel Selection: Cubic; Kernel Scaling: Set Automatically; Box Constraint Intensity: 1; Approach for Multiclass Classification: One-vs-One; Normalize Input Data: True
Support vector Machine	Fine Gaussian Support Vector Machine	20.47	121.27	7354705.00	Kernel Type: Gaussian; Kernel Scaling Value: 32; Box Constraint Setting: 1; Multiclass Classification Technique: One-vs-One; Data Standardization: Enabled
Support vector Machine	Medium Gaussian Support Vector Machine	30.96	112.14	7579057.00	Kernel Choice: Gaussian; Kernel Scale Parameter: 130; Box Constraint Intensity: 1; Approach for Multiclass Classification: One-vs-One; Normalize Input Data: True
Support vector Machine	Coarse Gaussian Support Vector Machine	30.62	101.38	7053601.00	Kernel Type: Gaussian; Kernel Scaling Factor: 510; Box Constraint Value: 1; Multiclass Strategy: One-vs-One; Data Normalization: Enabled
Efficient Logistic Regression	Efficient Logistic Regression	30.99	100.63	761299.00	Model: Logistic Regression; Solver Method: Automatic; Regularization Level (Lambda): Automatic; Coefficient Convergence Threshold (Beta Tolerance): 0.0001
Efficient Linear SVM	Efficient Linear Support Vector Machine	18.99	138.38	778313.00	Model: Support Vector Machine (SVM); Solver Selection: Automatic; Regularization Parameter (Lambda): Automatic; Coefficient Convergence Tolerance (Beta Tolerance): 0.0001
K Nearest Neighbour	Fine K Nearest Neighbour	31.05	100.09	1832266.00	Neighbour Count: 1; Distance Measurement: Euclidean; Distance Weighting Method: Uniform; Data Normalization: Enabled
K Nearest Neighbour	Medium K Nearest Neighbour	30.96	100.04	1832266.00	Neighbours Quantity: 10; Distance Formula: Euclidean; Weight Assignment for Distance: Uniform; Data Standardization: Active
K Nearest Neighbour	Coarse K Nearest Neighbour	21.66	148.39	1832266.00	Neighbour Count: 100; Distance Calculation Method: Euclidean; Distance Weighting: Uniform; Data Normalization: Enabled
K Nearest Neighbour	Cosine K Nearest Neighbour	30.82	101.17	1832254.00	Total Neighbours: 10; Distance Measure: Cosine; Weighting Method: Uniform; Data Standardization: Enabled
K Nearest Neighbour	Cubic K Nearest Neighbour	29.20	105.21	1832282.00	K-Neighbours: 10; Distance Measurement: Minkowski (Power: 3); Weighting Scheme: Uniform; Normalize Input Data: True
K Nearest Neighbour	Weighted K Nearest Neighbour	25.48	120.95	1832284.00	Neighbour Count: 10; Distance Calculation: Euclidean; Weighting by Distance: Inverse Square; Data Normalization: Enabled
Kernel	Support Vector Machine Kernel	30.62	120.72	1978295.00	Model: Support Vector Machine (SVM); Dimension Expansion: Automatic; Regularization Level (Lambda): Automatic; Kernel Scaling: Automatic; Multiclass Approach: One-vs-One; Iteration Cap: 1000
Kernel	Logistic Regression Kernel	31.20	105.60	1978707.00	Classifier: Logistic Regression; Expansion Dimension Count: Automatic; Regularization Parameter (Lambda): Automatic; Kernel Scale: Automatic; Multiclass Strategy: One-vs-One; Maximum Iterations: 1000
Ensemble	Boosted Trees	31.60	99.01	56833.00	Ensemble Technique: AdaBoost; Base Classifier: Decision Tree; Maximum Splits per Tree: 20; Total Number of Classifiers: 30; AdaBoost Learning Rate: 0.1; Predictor Sampling Strategy: Use All Predictors
Ensemble	Bagged Trees	11.64	277.23	1850887.00	Ensemble Strategy: Bagging; Base Estimator: Decision Tree; Max Splits Allowed per Tree: 598; Total Base Estimators: 30; Predictor Sampling Method: Use All Predictors
Ensemble	Subspace Discriminant	5.16	623.94	67742447.00	Ensemble Technique: Subspace; Base Model: Discriminant Analysis; Total Classifiers in Ensemble: 30; Dimensionality of Each Subspace: 8173
Ensemble	Subspace K Nearest Neighbour	29.26	104.56	54715033.00	Ensemble Technique: Subspace Method; Base Classifier: K-Nearest Neighbours; Total Number of Base Classifiers: 30; Dimension of Each Subspace: 8173
Ensemble	RUSBoosted Trees	32.13	97.53	1856503.00	Ensemble Technique: RUSBoost; Base Estimator: Decision Tree; Max Splits per Tree: 20; Total Estimators in Ensemble: 30; Learning Rate: 0.1; Predictor Sampling Strategy: Sample All Features

(*continued on next page*)

Table 2 (*continued*)

Model Type	Preset	Prediction Speed (obs/sec)	Training Time (sec)	Model Size (bytes)	Hyperparameters
Neural Network	Narrow Neural Network	24.07	114.57	94544.00	Number of Dense Layers: 1; Size of Initial Dense Layer: 10; Activation Function: ReLU; Maximum Iterations: 1000; Regularization Parameter (Lambda): 0; Normalize Input Data: True
Neural Network	Medium Neural Network	12.63	239.94	139304.00	Number of Dense Layers: 1; Size of Initial Dense Layer: 25; Activation Function: ReLU; Maximum Iterations: 1000; Regularization Parameter (Lambda): 0; Normalize Input Data: True
Neural Network	Wide Neural Network	23.80	150.88	363104.00	Number of Dense Layers: 1; Size of Initial Dense Layer: 100; Activation Function: ReLU; Maximum Iterations: 1000; Regularization Parameter (Lambda): 0; Normalize Input Data: True
Neural Network	Bilayered Neural Network	30.48	102.71	96344.00	Number of Dense Layers: 2; Size of Initial Dense Layer: 10; Second Dense Layer: 10; Activation Function: ReLU; Maximum Iterations: 1000; Regularization Parameter (Lambda): 0; Normalize Input Data: True
Neural Network	Trilayered Neural Network	24.91	117.65	98144.00	Number of Dense Layers: 3; Size of Initial Dense Layer: 10; Second Dense Layer: 10; Third Dense Layer: 10; Activation Function: ReLU; Maximum Iterations: 1000; Regularization Parameter (Lambda): 0; Normalize Input Data: True

with its preset configuration, prediction speed in observations per second, training time in seconds, and model size in bytes. Notably, Decision Trees of varying complexity (Fine, Medium, Coarse) exhibit differences in prediction speed and training time, despite having the same model size. Discriminant Analysis models, Linear and Quadratic, present varying training times, while Efficient Logistic Regression emphasizes its efficiency in terms of both prediction speed and model size. Naive Bayes models showcase the trade-off between prediction speed and training time, with Gaussian Naive Bayes being notably faster than Kernel Naive Bayes. Support Vector Machines exhibit differences in prediction speed, training time, and model size, with various kernel functions and scale settings. K Nearest Neighbours models show varying prediction speeds based on the number of Neighbours and distance metrics, and Ensemble methods demonstrate their capabilities with different learning techniques and model sizes. Finally, Neural Networks with different architectures offer insights into the impact of layer sizes on prediction speed and training time.

3.4. Hyperparameters and feature selection

Each model is characterized by a set of hyperparameters that govern its behavior. For example, Decision Trees specify the maximum number of splits and the split criterion (Gini's diversity index) among others. Naive Bayes models define distribution names and kernel types, while SVMs allow users to specify the kernel function and box constraint level. Ensemble methods detail the number of learners, learning rates, and other ensemble-specific parameters. The optimization of model performance heavily relies on the manipulation of hyperparameters, which are essential in achieving optimal outcomes. Furthermore, Table 2 demonstrates that all models employ feature selection techniques, resulting in the retention of 16,345 out of a total of 16,383 characteristics. This observation underscores the significance of reducing features and managing dimensionality when dealing with gene expression data of high dimensionality. The feature ranking algorithm utilized in this study is ANOVA, and PCA is employed to get a 95 % explained variance, hence facilitating dimensionality reduction.

3.5. Comparison and implications

The influence of trade-offs between prediction speed, training time, and model size is apparent in the process of model selection, as certain models demonstrate greater accuracy in specific domains. The models under consideration cover a diverse array of methods, with each technique being assessed based on several factors such as prediction speed, training time, model size, hyperparameters, selected features, feature ranking algorithm, and the use of PCA.

The models exhibit varying levels of prediction speed, as indicated by the metric "Prediction Speed (obs/sec)". The range of values spans from a minimum of 5.16 observations per second for the "Subspace Discriminant" model to a maximum of 33.27 observations per second for the "Fine Tree" model.

The duration of the training, expressed in seconds, is recorded in the column labelled "Training Time (sec)". There is considerable variation in training times among different models, with the fastest model, namely "Quadratic Discriminant", requiring 11.10 s for training, while the slowest model, "Subspace Discriminant", takes 623.94 s to complete the training process.

The size of a model, measured in bytes, indicates the amount of memory required to store the model. The sizes exhibited variation, ranging from 56,833 bytes for the "Boosted Trees" model to 67,742,447 bytes for the "Subspace Discriminant" model.

The column labelled "Hyperparameters" presents details regarding the hyperparameters employed in each model. The behavior of the algorithms is governed by these hyperparameters, which can be adjusted to achieve optimal performance.

The column labelled "Selected Features" denotes the quantity of features that have been chosen from a pool of 16,383 in total. This emphasizes the significance of feature selection and dimensionality reduction in the management of gene expression data with a high number of dimensions.

The column titled "Feature Ranking Algorithm1" denotes the algorithm utilized to rank features. In this scenario, the statistical technique known as "ANOVA" is employed to assess the variability observed across different groups as well as within each group, to identify significant factors that exert an influencing effect.

The application of PCA is described in the "Principal Component Analysis" section, highlighting its usage in achieving a 95 % explained variance. PCA is employed to minimize dimensionality and improve model performance.

In brief, the table offers significant insights into the effectiveness and attributes of diverse classification models when employed in the analysis of gene expression cancer RNA-Seq data. This information can assist individuals in identifying the most appropriate classification algorithm for their specific requirements in the field of cancer research and diagnosis.

3.6. Performance metrics of machine learning models across different cancer types

Results of various ML models across different cancer types, with performance metrics such as True Positive Rates (TPR), False Negative Rates (FNR), Positive Predictive Values (PPV), False Discovery Rate (FDR), and Area Under the Curve (AUC) for both training and testing datasets are presented in Table 3.

Table 3
Classification accuracy during training and testing.

Model Type	Cancer Type	Training Results					Testing Results				
		TRP	FNR	PPV	FDR	AUC	TRP	FNR	PPV	FDR	AUC
Fine Tree	BRCA	98.2	1.8	98.2	1.8	0.986	99.2	0.8	98.8	1.2	0.996
	COAD	95	5	98.3	1.7	0.974	94.4	5.6	100	0	0.972
	KIRC	99.1	0.9	99.1	0.9	0.995	99.3	0	100	0	0.997
	LUAD	96.3	3.7	92.9	7.1	0.974	96.9	3.1	96.9	3.1	0.981
	PRAD	97	3	99	1	0.984	100	0	100	0	1
Medium Tree	BRCA	98.2	1.8	98.2	1.8	0.986	100	0	98.8	1.2	0.996
	COAD	95	5	98.3	1.7	0.974	94.4	5.6	100	0	0.9722
	KIRC	99.1	0.9	99.1	0.9	0.995	100	0	100	0	1
	LUAD	96.3	3.7	92.9	7.1	0.974	96.9	3.1	96.9	3.1	0.981
	PRAD	97	3	99	1	0.984	100	0	100	0	1
Course Tree	BRCA	91.7	8.3	99.5	0.5	0.976	92.7	7.3	100	0	0.976
	COAD	0	100	0	0	0.888	0	100	0	0	0.902
	KIRC	99.1	0.9	99.1	0.9	0.997	100	0	100	0	1
	LUAD	98.2	1.8	57.8	42.2	0.912	100	0	59.3	40.7	0.935
	PRAD	99	1	98	2	0.996	100	0	94.6	5.4	0.994
Linear Discriminant	BRCA	100	0	100	0	1	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	100	0	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Quadratic Discriminant	BRCA	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
	COAD	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
	KIRC	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
	LUAD	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
	PRAD	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
Efficient Logistic Regression	BRCA	100	0	99.5	0.5	1	100	0	98.8	1.2	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	100	0	1	96.9	3.1	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Gaussian Naïve Bayes	BRCA	79.4	20.6	99.4	0.6	0.98	70.7	29.3	98.3	1.7	0.935
	COAD	96.7	3.3	100	0	0.995	88.9	11.1	100	0	0.984
	KIRC	100	0	90.2	9.8	1	97.1	2.9	87.2	12.8	0.991
	LUAD	86.2	13.8	100	0	0.988	53.1	46.9	89.5	10.5	0.956
	PRAD	100	0	67.3	32.7	0.99	100	0	50.7	49.3	1
Kernel Naïve Bayes	BRCA	95	5	97.6	2.4	0.996	93.9	6.1	95.1	4.9	0.989
	COAD	98.3	1.7	100	0	0.998	88.9	11.1	100	0	0.986
	KIRC	100	0	94.9	5.1	1	97.1	2.9	94.4	5.6	0.999
	LUAD	88.1	11.9	100	0	0.996	71.9	28.1	95.8	4.2	0.992
	PRAD	100	0	87.8	12.2	1	100	0	77.8	22.2	1
Linear SVM	BRCA	100	0	99.5	0.5	1	100	0	98.8	1.2	1
	COAD	98.3	1.7	100	0	1	94.4	5.6	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	99.1	0.9	0.999	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Quadratic SVM	BRCA	100	0	99.5	0.5	1	100	0	100	0	1
	COAD	100	0	100	0	1	94.4	5.6	100	0	1
	KIRC	100	0	100	0	1	100	0	97.2	2.8	1
	LUAD	99.1	0.9	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Cubic SVM	BRCA	100	0	99.1	0.9	1	100	0	100	0	1
	COAD	100	0	100	0	1	94.4	5.6	100	0	1
	KIRC	100	0	99.1	0.9	1	100	0	97.2	2.8	1
	LUAD	98.2	1.8	100	0	1	100	0	100	0	1
	PRAD	99	1	100	0	1	100	0	100	0	1

(*continued on next page*)

Table 3 (*continued*)

Model Type	Cancer Type	Training Results					Testing Results				
		TRP	FNR	PPV	FDR	AUC	TRP	FNR	PPV	FDR	AUC
Fine Gaussian SVM	BRCA	98.2	1.8	97.7	2.3	0.999	89	11	98.6	1.4	0.999
	COAD	1.7	98.3	100	0	1	0	100	0	0	1
	KIRC	100	0	50.5	49.5	0.996	97.1	2.9	40.5	59.5	0.988
	LUAD	53.2	46.8	100	0	0.999	25	75	88.9	11.1	0.995
	PRAD	100	0	100	0	1	100	0	100	0	1
Medium Gaussian SVM	BRCA	100	0	36.4	63.6	0.838	100	0	40.6	59.4	1
	COAD	0	100	0	0	0.996	0	100	0	0	0.988
	KIRC	0	100	0	0	0.834	0	100	0	0	1
	LUAD	0	100	0	0	0.71	0	100	0	0	0.844
	PRAD	0	100	0	0	1	0	100	0	0	1
Coarse SVM	BRCA	100	0	36.4	63.6	0.837	100	0	40.6	59.4	1
	COAD	0	100	0	0	0.996	0	100	0	0	0.986
	KIRC	0	100	0	0	0.834	0	100	0	0	1
	LUAD	0	100	0	0	0.713	0	100	0	0	0.845
	PRAD	0	100	0	0	1	0	100	0	0	1
Efficient Logistic Regression	BRCA	99.5	0.5	100	0	1	100	0	98.8	1.2	1
	COAD	100	0	98.4	1.6	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	99.1	0.9	0.997	96.9	3.1	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Efficient Logistic SVM	BRCA	100	0	99.5	0.5	1	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Fine KNN	BRCA	17	83	100	0	0.585	13.4	86.6	100	0	1
	COAD	18.3	81.7	100	0	0.592	33.3	66.7	100	0	1
	KIRC	78.4	21.6	79.8	20.2	0.869	68.6	31.4	85.7	14.3	1
	LUAD	10.1	89.9	100	0	0.551	3.1	96.9	100	0	1
	PRAD	100	0	23.4	76.6	0.669	100	0	22.4	77.6	1
Medium KNN	BRCA	0	100	0	0	0.63	1.2	98.8	100	0	0.725
	COAD	0	100	0	0	0.589	0	100	0	0	0.555
	KIRC	53.2	46.8	98.3	1.7	0.924	45.7	54.3	100	0	0.905
	LUAD	0	100	0	0	0.491	0	100	0	0	0.443
	PRAD	100	0	18.7	81.3	0.901	100	0	18.9	81.1	0.939
Coarse KNN	BRCA	0	100	0	0	0.724	0	100	0	0	0.848
	COAD	0	100	0	0	0.868	0	100	0	0	0.886
	KIRC	0	100	0	0	0.914	0	100	0	0	0.975
	LUAD	0	100	0	0	0.7538	0	100	0	0	0.833
	PRAD	100	0	16.9	83.1	0.7728	100	0	17.3	82.7	0.785
Cosine KNN	BRCA	99.5	0.5	99.5	0.5	1	98.8	1.2	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	99.1	0.9	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	97.2	2.8	1
Cubic KNN	BRCA	0.5	99.5	100	0	0.613	1.2	98.8	100	0	0.729
	COAD	0	100	0	0	0.763	0	100	0	0	0.674
	KIRC	57.7	42.3	86.5	13.5	0.897	48.6	51.4	100	0	0.864
	LUAD	0	100	0	0	0.491	100	0	0	0	0.439
	PRAD	100	0	19.3	80.7	0.918	100	0	19	81	0.945
Weighted KNN	BRCA	0	100	0	0	0.642	1.2	98.8	100	0	0.735
	COAD	0	100	0	0	0.605	0	100	0	0	0.597
	KIRC	50.5	49.5	98.2	1.8	0.935	42.9	57.1	100	0	0.919
	LUAD	0	100	0	0	0.494	0	100	0	0	0.444
	PRAD	100	0	18.6	81.4	0.924	100	0	18.8	81.2	0.962
SVM Kernel	BRCA	100	0	99.5	0.5	1	100	0	98.8	1.2	1
	COAD	100	0	100	0	1	94.4	5.6	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	0.999
	LUAD	99.1	0.9	100	0	1	96.9	3.1	96.9	3.1	1
	PRAD	100	0	100	0	1	100	0	100	0	1

(*continued on next page*)

Table 3 (*continued*)

Model Type	Cancer Type	Training Results					Testing Results				
		TRP	FNR	PPV	FDR	AUC	TRP	FNR	PPV	FDR	AUC
Logistic Regression Kernel	BRCA	100	0	97.3	2.7	1	100	0	97.6	2.4	1
	COAD	98.3	1.7	100	0	1	88.9	11.1	100	0	1
	KIRC	99.1	0.9	100	0	1	100	0	100	0	1
	LUAD	98.2	1.8	100	0	0.999	96.9	3.1	96.9	3.1	0.999
	PRAD	98	2	100	0	1	100	0	100	0	1
Boosted Tree	BRCA	100	0	36.4	63.6	0	100	0	40.6	59.4	0
	COAD	0	100	0	0	0	0	100	0	0	0
	KIRC	0	100	0	0	0	0	100	0	0	0
	LUAD	0	100	0	0	0	0	100	0	0	0
	PRAD	0	100	0	0	0	0	100	0	0	0
Bagged Trees	BRCA	99.5	0.5	97.3	2.7	0.999	100	0	98.8	1.2	0.999
	COAD	95	5	98.3	1.7	0.999	88.9	11.1	100	0	0.999
	KIRC	99.1	0.9	99.1	0.9	0.999	100	0	100	0	1
	LUAD	95.4	4.6	97.2	2.8	0.998	96.9	3.1	93.9	6.1	0.998
	PRAD	99	1	100	0	1	100	0	100	0	1
Subspace Discriminant	BRCA	100	0	100	0	1	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	100	0	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Subspace KNN	BRCA	100	0	99.5	0.5	0.999	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	99.1	0.9	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
RUSBoosted Tree	BRCA	98.6	1.4	98.6	1.4	0.998	98.8	1.2	98.8	1.2	0.998
	COAD	98.3	1.7	95.2	4.8	0.998	94.4	5.6	100	0	1
	KIRC	99.1	0.9	99.1	0.9	0.994	100	0	100	0	1
	LUAD	94.5	5.5	95.4	4.6	0.998	96.9	3.1	91.2	8.8	0.998
	PRAD	99	1	100	0	1	94.3	5.7	97.1	2.9	0.999
Narrow Neural Network	BRCA	98.2	1.8	96.8	3.2	0.93	100	0	100	0	1
	COAD	100	0	98.4	1.6	1	100	0	100	0	1
	KIRC	97.3	2.7	99.1	0.9	0.995	100	0	100	0	1
	LUAD	94.5	5.5	100	0	0.999	100	0	100	0	1
	PRAD	100	0	96.2	3.8	0.999	100	0	100	0	1
Medium Neural Network	BRCA	100	0	100	0	1	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	100	0	100	0	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Wide Neural Network	BRCA	99.5	0.5	100	0	1	100	0	100	0	1
	COAD	100	0	100	0	1	100	0	100	0	1
	KIRC	100	0	100	0	1	100	0	100	0	1
	LUAD	100	0	99.1	0.9	1	100	0	100	0	1
	PRAD	100	0	100	0	1	100	0	100	0	1
Bilayered Neural Network	BRCA	96.3	3.7	96.8	3.2	0.989	91.5	8.5	98.7	1.3	0.985
	COAD	85	5	87.9	12.1	0.992	94.4	5.6	100	0	0.998
	KIRC	94.6	5.4	95.5	4.5	0.997	100	0	85.4	14.6	0.994
	LUAD	91.7	8.3	92.6	7.4	0.979	96.9	3.1	91.2	8.8	0.999
	PRAD	93.1	6.9	88.7	11.3	0.977	97.1	2.9	100	0	0.999
Trilayered Neural Network	BRCA	89	11	94.6	5.4	0.961	91.5	8.5	87.2	12.8	0.95
	COAD	91.7	8.3	88.7	11.3	0.987	94.4	5.6	100	0	0.998
	KIRC	88.3	11.7	83.8	16.2	0.978	94.3	5.7	80.5	19.5	0.984
	LUAD	84.4	15.6	78.6	21.4	0.948	40.6	59.4	81.2	18.8	0.778
	PRAD	90.1	9.9	92.9	7.1	0.972	88.6	11.4	73.8	26.2	0.953

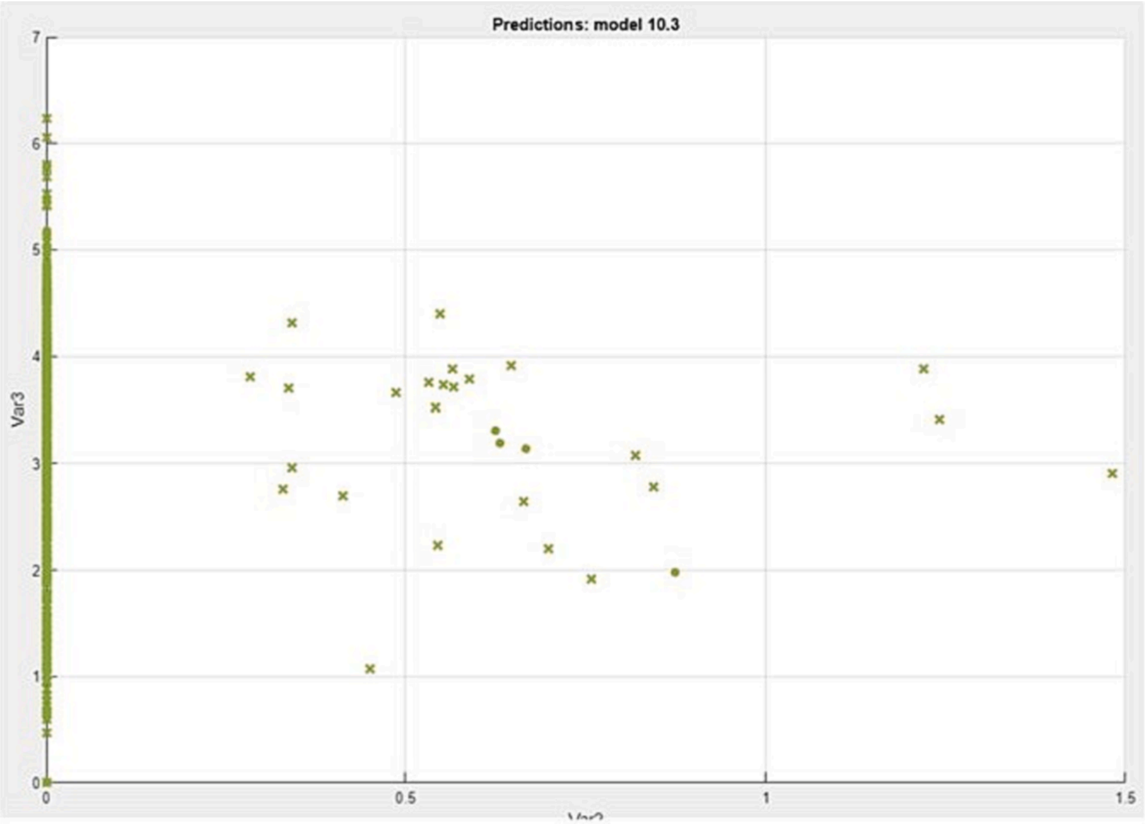


Fig. 5a. Scatter plot.

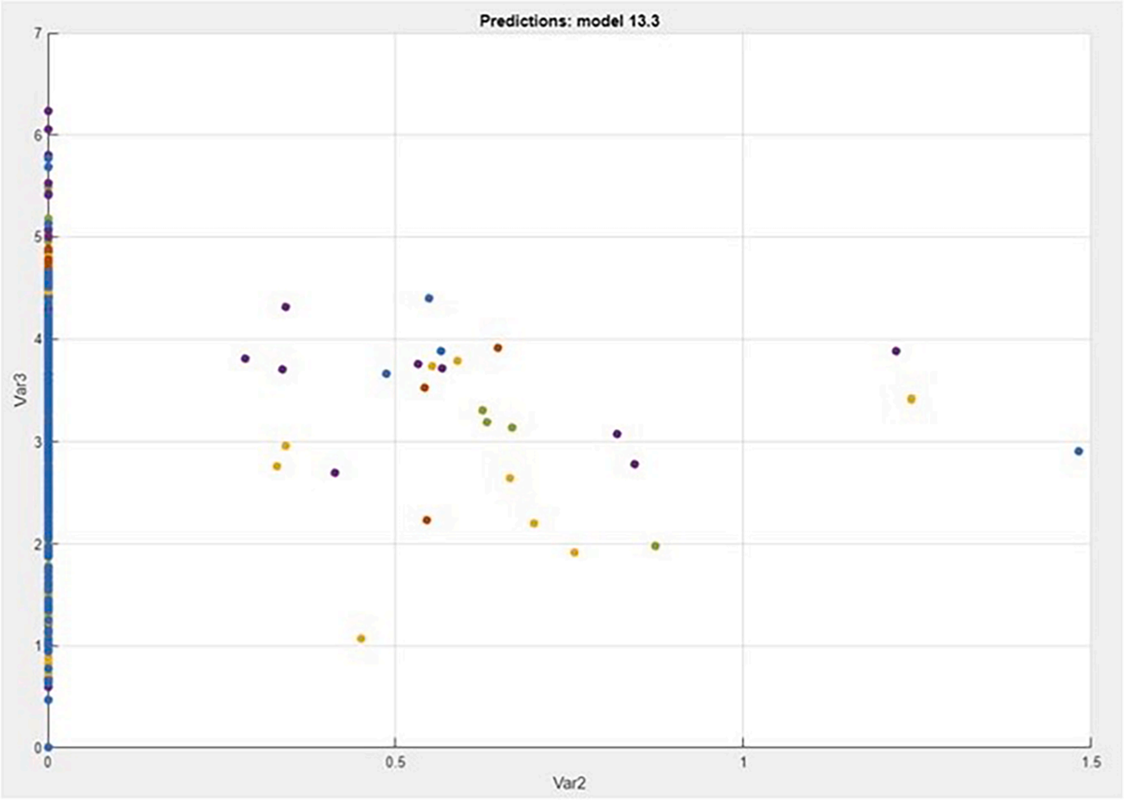


Fig. 5b. Scatter plot.

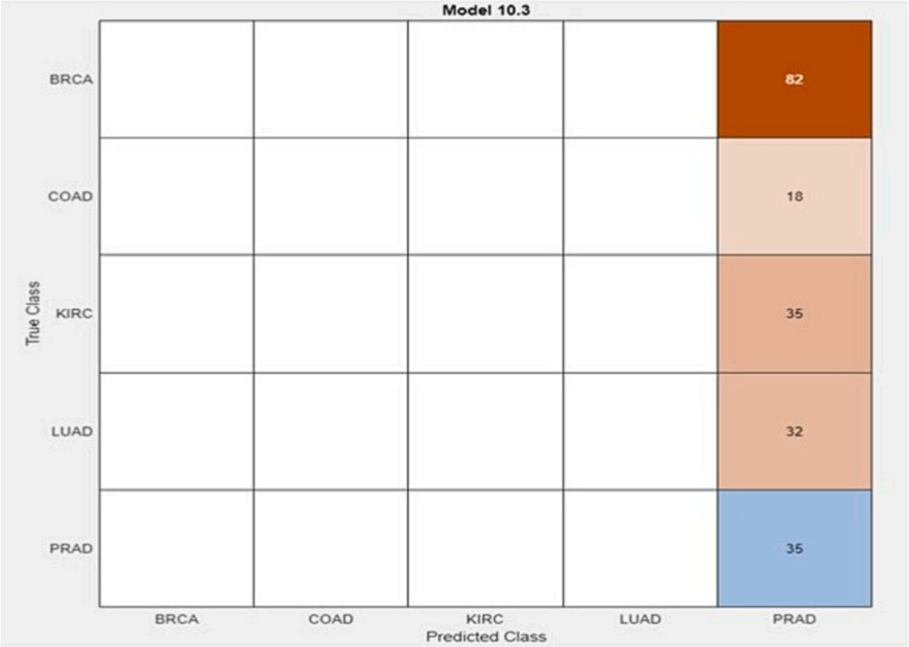


Fig. 5c. Validation confusion matrix.

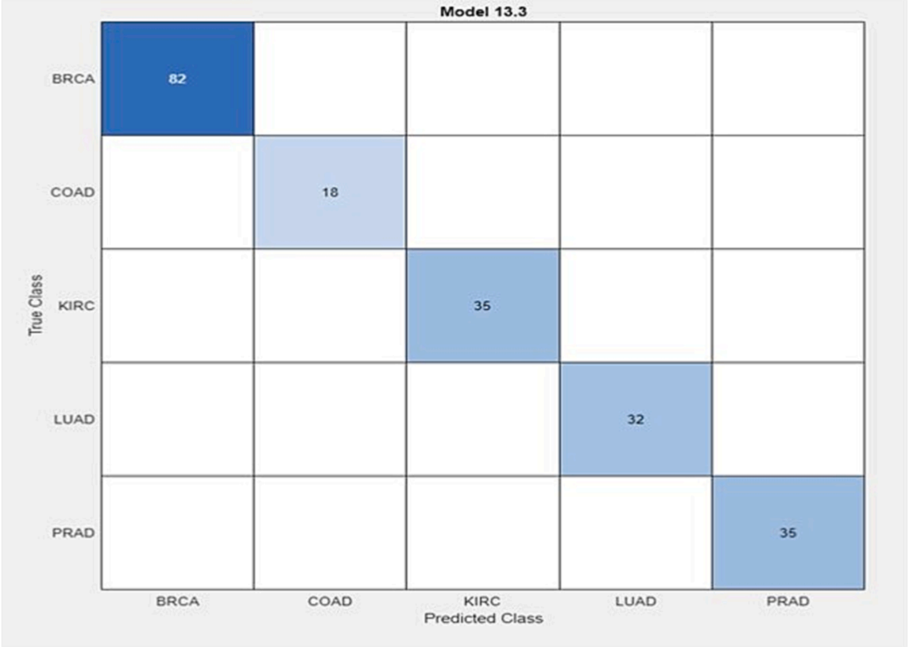


Fig. 5d. Validation confusion matrix.

3.6.1. *True positive rates (TPR)*

Results of the TRP analysis indicate that the Fine KNN model consistently underperforms with the lowest rates for BRCA at 17 %, LUAD at 10.1 %, and PRAD at 0 %. In contrast, the Linear Discriminant, Efficient Logistic Regression, and Efficient Logistic SVM models consistently achieved 100 % TPR for these cancers. Both the Fine and Medium Gaussian SVM models have a 0 % TPR for COAD, failing to identify any true positives. However, many models report a 100 % TPR for COAD. For KIRC, every model except the Medium KNN reaches a 100 % TPR.

3.6.2. *False negative rates (FNR)*

For BRCA, the Fine KNN model revealed the highest FNR at 83 %. Conversely, the Linear Discriminant, Efficient Logistic Regression, and Efficient Logistic SVM models featured 0 % FNR. In the case of COAD, both the Fine and Medium Gaussian SVM models misclassify every true positive, resulting in a 100 % FNR, while the rest attain 0 % FNR. For KIRC, the Medium KNN model has a notable FNR of 46.8 %, with others at 0 %. The Fine KNN model again underperforms with LUAD and PRAD, having FNRs of 89.9 % and 76.6 % respectively. Other models for these cancers record near or exactly 0 % FNRs.

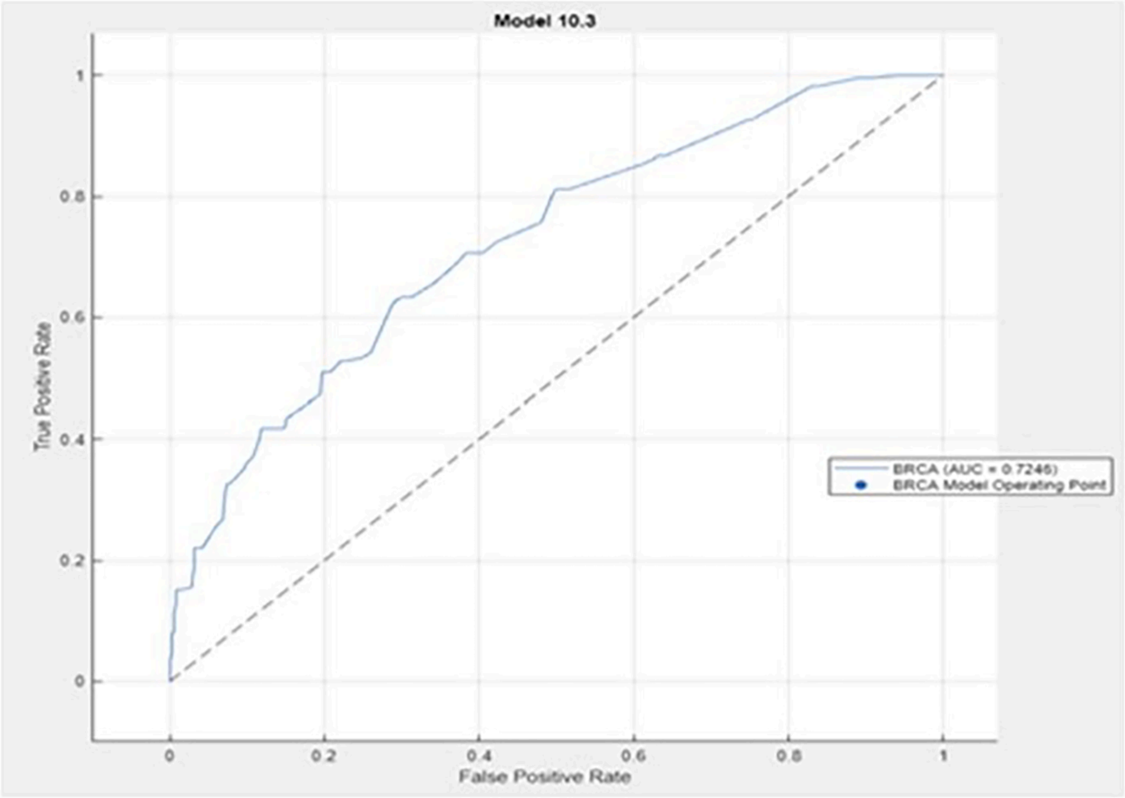


Fig. 5e. Validation receiver operating characteristic curve.

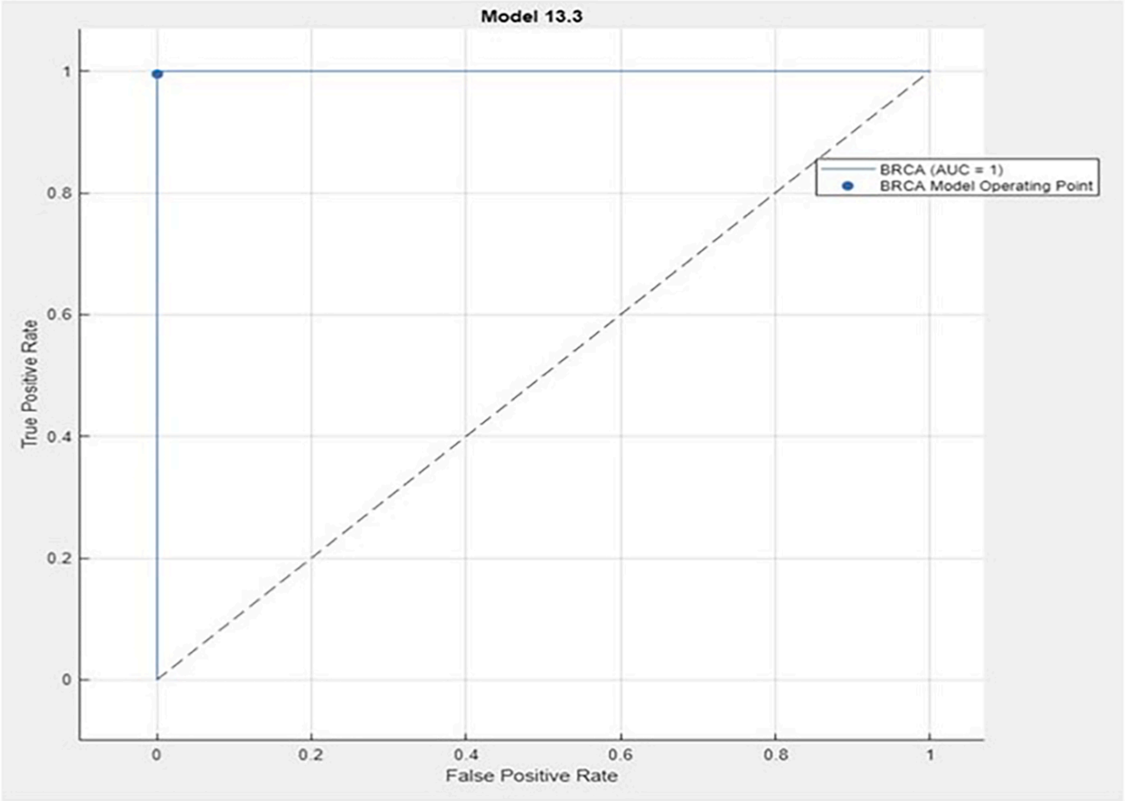


Fig. 5f. Validation receiver operating characteristic curve.

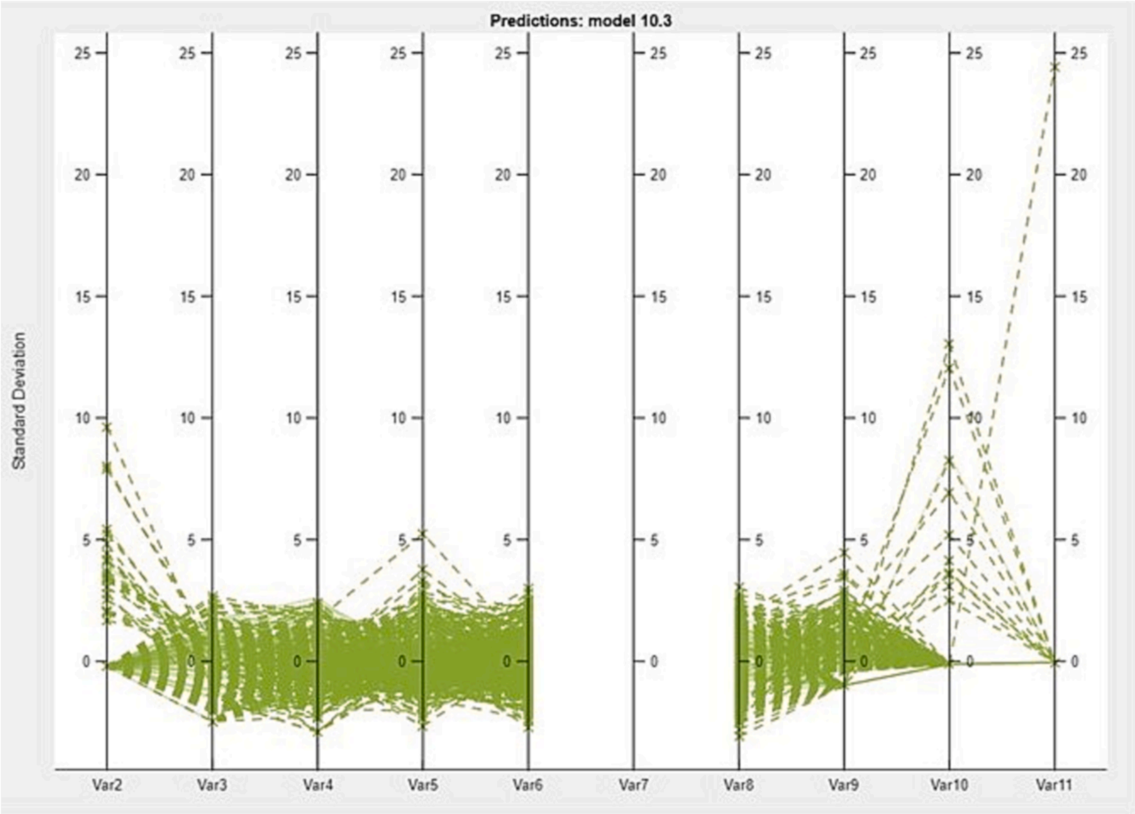


Fig. 5g. Parallel coordinates plot.

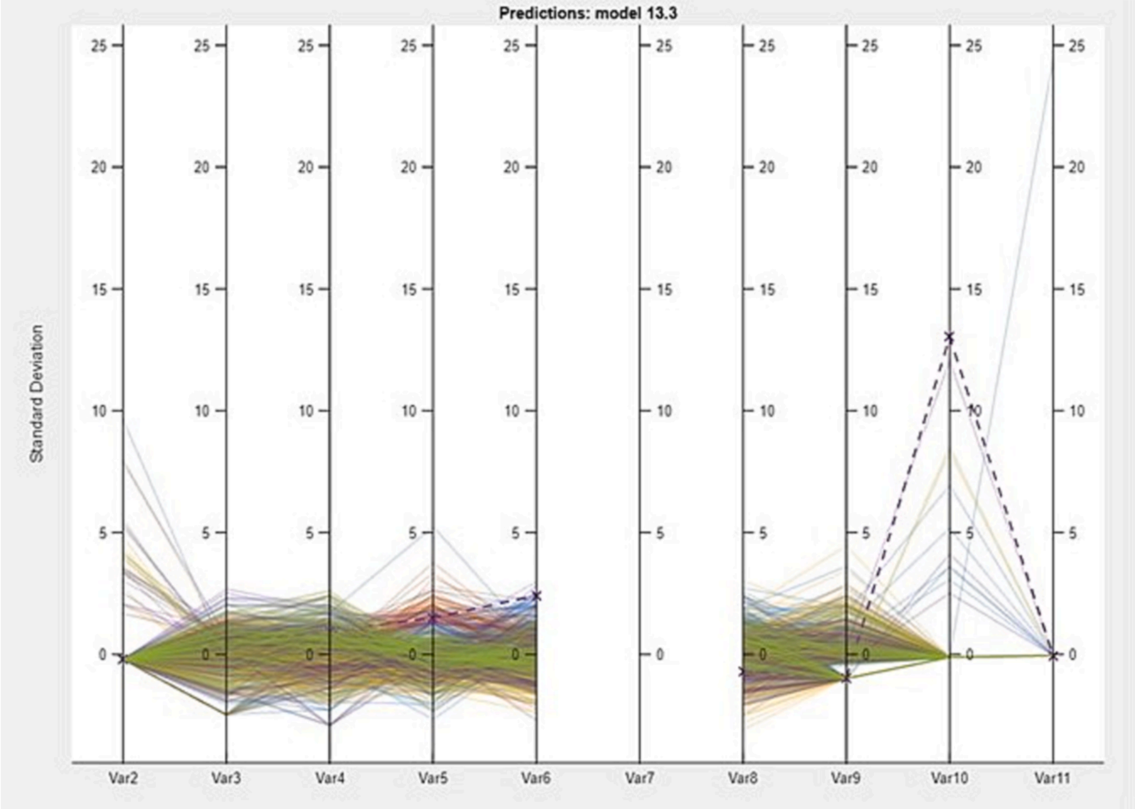


Fig. 5h. Parallel coordinates plot.

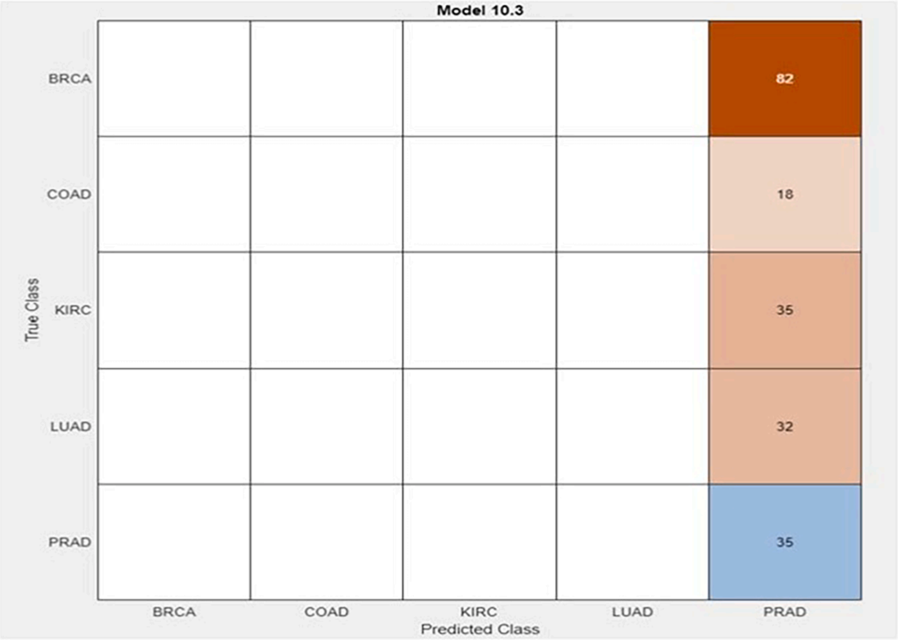


Fig. 5i. Test Confusion Matrix.

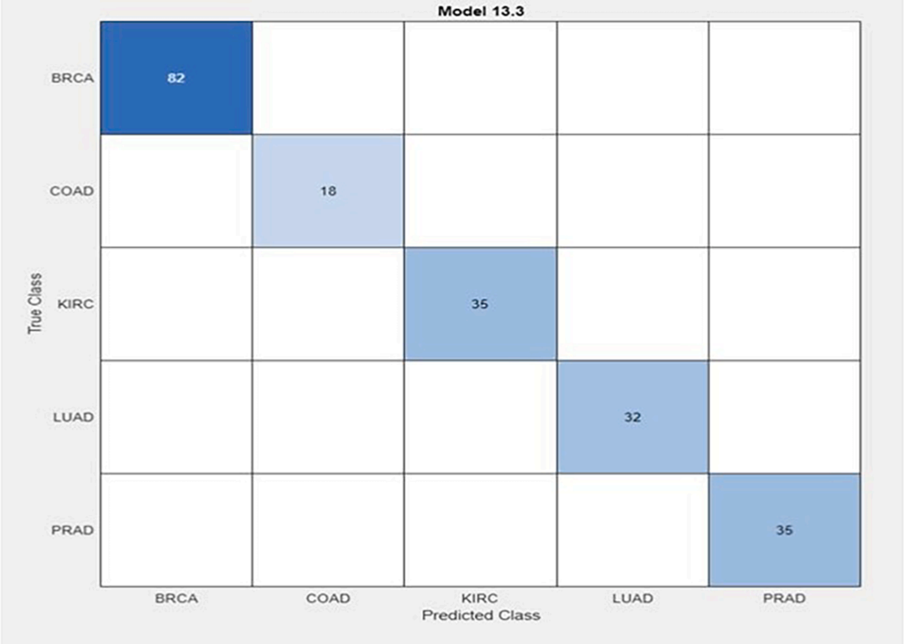


Fig. 5j. Test Confusion Matrix.

3.6.3. *Positive predictive values (PPV)*

For BRCA, PPVs across all models are commendably high, ranging from 96.8 % to 100 %. In the COAD category, the Fine Gaussian SVM model is the outlier with a 0 % PPV, whereas most models indicate 100 % PPV. In the case of KIRC, all models showcase a 0 % FDR, translating to impeccable PPVs. Regarding LUAD, the Fine KNN model lags with a 0 % PPV, but other models score between 88.9 % and 100 %. Finally, for PRAD, the Fine KNN model reports the lowest PPV at 50.7 %, with most other models clinching a 100 % PPV.

3.6.4. *False discovery rate (FDR)*

In the evaluation of several models across different cancer types, the Fine KNN model consistently underperformed in terms of FDR. For BRCA, FDR was 100 %, meaning all its positive predictions were inaccurate. Other models, however, performed better with FDRs between 0 % and 3.2 %. Similarly, for COAD and KIRC, the Fine Gaussian SVM and Medium KNN models, respectively, also showed a 100 % FDR. For PRAD, the Fine KNN model trailed with a 77.6 % FDR, while most other models flawlessly stood at 0 %.

3.6.5. *Area under the curve (AUC)*

Across different cancer types, the AUC values provide insights into model performances. For BRCA, while most models display commendable results, the Fine KNN model underperforms with an AUC of 0.585.

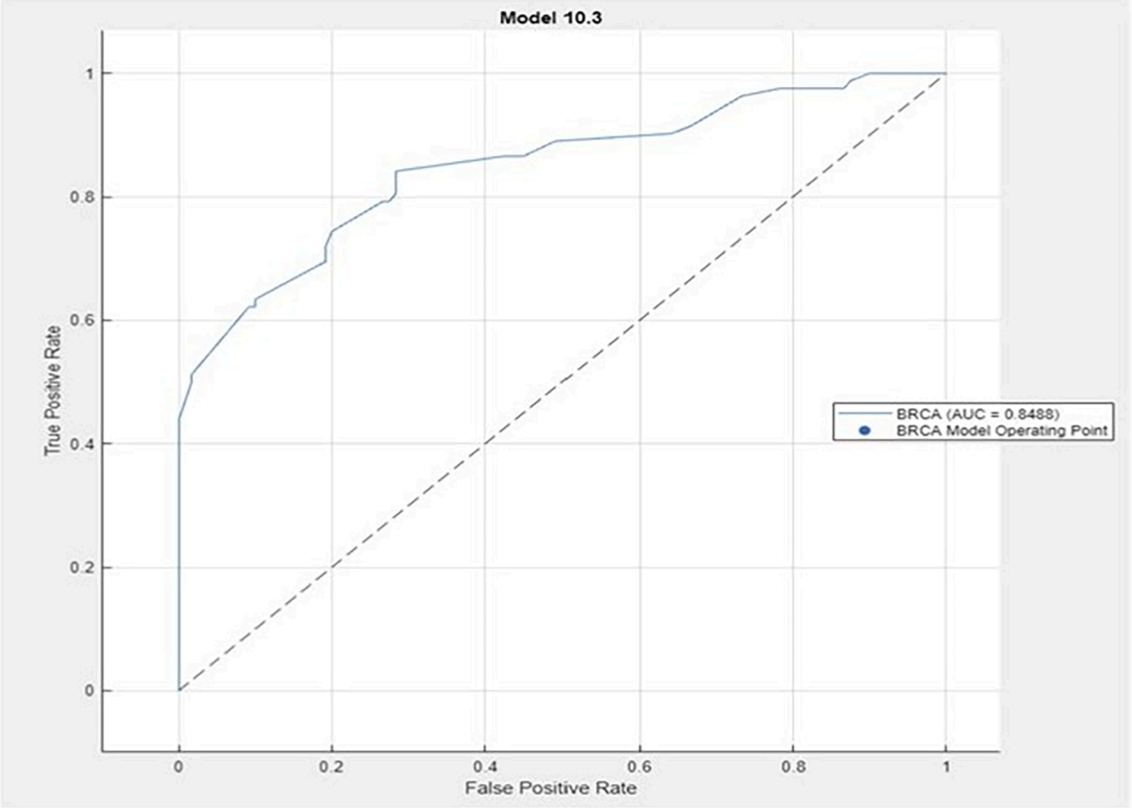


Fig. 5k. Test rreceiver operating characteristic curve.

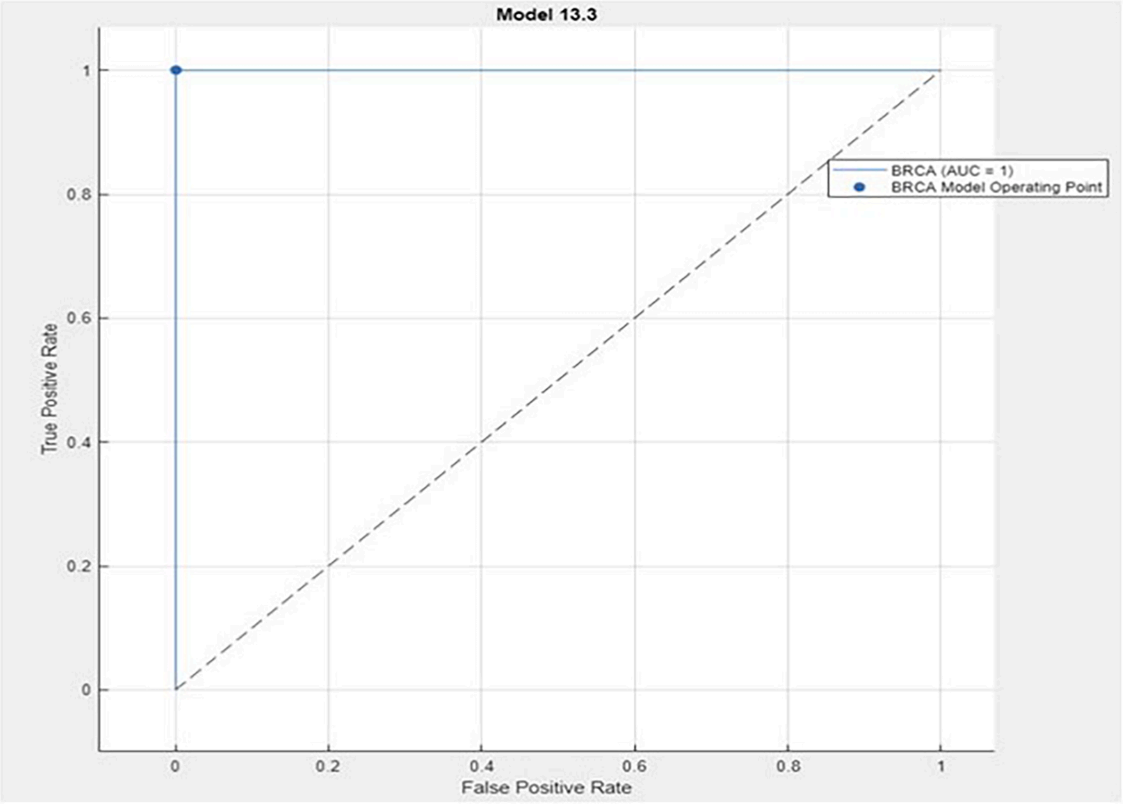


Fig. 5l. Test receiver operating characteristic curve.

Table 4
Performance comparison of the anticipated classification models with benchmark studies in the identified domain.

Reference	Dataset	Algorithm	Dataset Type	Performance: Accuracy
(Hijazi and Chan, 2013)	Mixed-Lineage Leukemia	SVM Linear Gene	Gene Expression	99.89 %
(Zhang et al., 2018)	Breast Cancer	SVM-RFE-PSO	Gene Expression	81.54 %
(Alshamlan, 2018)	Binary and Multi-Class Cancer Datasets	DBQ	Gene Expression	close to 100 %
(Yuan et al., 2020)	Tumor-Educated Platelets	Evolutionary Programming-trained SVM	Gene Expression	95.93 %
(Alshareef et al., 2022)	Medical Databases (PubMed, CENTRAL, EMBASE, OASIS, and CNKI)	DNN	Gene Expression	96.21 %
(Yuan et al., 2020	Lung Adenocarcinoma and Lung Squamous Cell Cancer	RF	Gene Expression	94.9 %
		RF	Gene Expression	93.3 %
		SVM	Gene Expression	94.7 %
(Gao et al., 2019)	Breast Cancer	DeepCC	Gene Expression	89 %
(Saheed, 2023)	Lymphoma	LR	Microarray	99 %
Proposed Method	Gene Expression Cancer RNA-Seq	Coarse KNN (Lowest)	Gene Expression	16.861 %
		Wide Neural Network (Highest)	Gene Expression	99.995 %

Similarly, in the COAD category, while most models displayed AUCs from 0.972 to 0.999, the Medium KNN model only manages an AUC of 0.555. The KIRC evaluations show near-perfect results for all but the Medium KNN model, which scores an AUC of 0.905, with the rest achieving an AUC of 1. In LUAD, the Fine KNN model once again falls behind with an AUC of 0.571, yet other models score between 0.973 and 1. Lastly, for PRAD, despite most models hovering between AUCs of 0.964 and 1, the Fine KNN model stands out with a lower AUC of 0.546.

4. Discussion

Understanding the subtype of cancer is essential for classifying the disease, determining prognosis, and devising treatment strategies. Breast carcinoma can be divided into five transcriptome-based subtypes, each of which results in a different therapeutic response and outcome (Dai et al., 2015). Conventional approaches to cancer classification rely on the visual examination of histological staining or immunohistochemically stained cancer sections by trained pathologists. Bavafaye Haghighi et al. (2019) report that the integration of multi-omics data into ML tools has improved the accuracy of patient diagnoses.

The gene expression cancer RNA-Seq dataset from the UCI ML Repository is a subset of the larger RNA-Seq (HiSeq) PANCAN collection, encompassing gene expression profiles from patients diagnosed with tumor types including BRCA, KIRC, COAD, LUAD, and PRAD. Consisting of 801 instances, each record is represented by 16,383 real-valued features, corresponding to a gene's expression level.

BRCA refers to breast invasive carcinoma. KIRC denotes kidney renal clear cell carcinoma. COAD stands for colon adenocarcinoma. LUAD is lung adenocarcinoma, and PRAD signifies prostate adenocarcinoma. Each of these cancer types possesses unique molecular and genetic profiles, offering opportunities for in-depth biological and computational analysis.

The k-means clustering algorithm is designed to segregate datasets into distinct clusters based on feature similarities. Applied to the dataset, k-means sought to unravel hidden patterns within the gene expressions of diverse cancer types, potentially unveiling specific RNA signatures characteristic of each type.

Upon analyzing the gene expression data, distinctive clustering patterns emerged. The BRCA tumor type was predominantly clustered within C3, COAD spanned across C4, C5, and C2, with C2 emerging as dominant. KIRC data mostly settled within C2, LUAD samples with C4, and PRAD samples with C1. These patterns underscore a significant statistical correlation between the specific cancer type and its respective clustering.

In the analysis of gene expression data, clusters illuminated notable groupings based on inherent similarities. The significant differences observed across the datasets indicate distinct genetic signatures. COAD displayed the highest mean gene expression value, followed by KIRC and PRAD, while BRCA and LUAD had lower values. Understanding these clusters aids in deciphering the molecular intricacies of each tumor and highlights the importance of personalized treatment approaches in oncology.

The results provide insight into the performance of different ML models on the classification task. A combination of data preprocessing, dimensionality reduction through PCA, and hyperparameter optimization using Bayesian Optimization has allowed for an evaluation of multiple ML models. The performance metrics include TRP, FNR, PPV, FDR, and AUC.

Linear Discriminant and Efficient Logistic Regression models consistently perform the best across all datasets. The Quadratic Discriminant model does not seem to be working as indicated by NaN values. KNN models have the most variability, with significant differences in performance. The Linear Discriminant or Efficient Logistic Regression would be recommended due to their consistently high performance.

The importance of model selection in life science is crucial, especially when dealing with specific cancer datasets like BRCA, KIRC, COAD, LUAD, and PRAD. Such datasets consist of 16,383 features and 801 instances, and the choice of ML model significantly impacts the results.

The oncology and cancer research domains are transforming with the incorporation of ML, which holds the potential for enhanced patient outcomes, precision medicine, and early detection. ML can assist in the development of prediction models, predict tumor responses to treatments, and more. Its implementation must be approached cautiously, with ethical concerns in mind.

A comparison of accuracy percentages for distinct model types was performed during this study. The Coarse KNN model yields an accuracy of 16.861 % on the validation set and 17.327 % on the test set, while the Wide Neural Network achieves 99.834 % on the validation set and 99.995 % on the test set. The performance of the Coarse KNN and Wide

Neural Network models is comprehensively illustrated through various graphical representations, including Scatter plots, Validation Confusion Matrices, Validation Receiver Operating Characteristic (ROC) curves, Parallel Coordinates Plots, Test Confusion Matrices, and Test ROC curves. These visual comparisons are depicted in Figs. 5a-5l.

Table 4 shows the performance comparison of anticipated models Coarse KNN, and Wide Neural Network with the existing studies. The Wide Neural Network achieves much better performance in terms of quantitative measures than the existing studies.

5. Conclusion

In this research, we employed an RNA sequencing (RNA-seq) based dataset to categorize distinct cancer types, emphasizing the importance of classification accuracy during both training and testing phases. Our findings underscore the power of RNA-seq as a tool in precision oncology, enabling us to differentiate cancer subtypes with notable accuracy. The RNA-Seq data used in this study is part of the RNA-Seq (HiSeq) PANCAN dataset, it is a random extraction of gene expressions of patients having different types of tumors: BRCA, KIRC, COAD, LUAD, and PRAD. The clustering patterns observed provide invaluable insights into the molecular signatures and heterogeneities inherent to different cancer types. Such granularity not only enhances our understanding of cancer biology but also lays the groundwork for tailored therapeutic interventions. As we move towards an era of personalized medicine, the emphasis on tools like RNA-seq and rigorous validation methodologies, such as ensuring classification accuracy, become paramount. This study marks a significant step forward in leveraging high-throughput genomics for precision oncology, and we anticipate that further refinements in this approach will catalyze advancements in cancer diagnostics, prognostics, and therapeutics.

The consistent classification performance observed during the training and testing phases attests to the robustness of our analytical approach. As the last step, a classification task is performed through ML based algorithms (Decision Trees, Discriminant Analysis, Efficient Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Efficient Linear SVM, k-Nearest Neighbours (KNN), Kernel-based models, Ensembles, and Neural Networks). Finally, Wide Neural Network demonstrates a significantly higher accuracy, achieving 99.834 % on the validation set and an even more impressive 99.995 % on the test set.

Although ML has several benefits when it comes to RNA expression data analysis, one of its drawbacks is that it needs huge, high-quality datasets to function at its best. These datasets may contain biases or inaccuracies that produce false findings and jeopardize the validity of the prognosis and diagnosis of diseases gathered from this analysis.

The identification of gene expressions associated with cancer types is a significant area of future research, where scholars might explore various approaches for this purpose.

Funding

This work was funded by the Deanship of Scientific Research at Jouf University under Grant Number (DSR2022-RG-0102).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We extend our gratitude to our institutions and colleagues for their unwavering moral support.

References

Alharbi, F., Vakanski, A., 2023. Machine learning methods for cancer classification using gene expression data: a review. Bioengineering. https://doi.org/10.3390/bioengineering10020173.

Alsayari, A., Asiri, Y.I., Muhsinah, A.B., Hassan, M., 2021. Anticolon cancer properties of pyrazole derivatives acting through xanthine oxidase inhibition. J. Oncol. https://doi.org/10.1155/2021/5691982.

Alshamlan, H.M., 2018. Dqb: A novel dynamic quantitive classification model using artificial bee colony algorithm with application on gene expression profiles. Saudi J. Biol. Sci. https://doi.org/10.1016/j.sjbs.2018.01.017.

Alshareef, A.M., Alsini, R., Alsieni, M., Alrowais, F., Marzouk, R., Abunadi, I., Nemri, N., 2022. Optimal deep learning enabled prostate cancer detection using microarray gene expression. J. Healthcare Eng. https://doi.org/10.1155/2022/7364704.

Arslan, A.K.K., Uzunhisarcıklı, E., Yerer, M.B., Bishayee, A., 2022. The golden spice curcumin in cancer: A perspective on finalized clinical trials during the last 10 years. J. Cancer Res. Ther. https://doi.org/10.4103/jcrt.JCRT_1017_20.

Danaee, P., Ghaeini, R., Hendrix, D.A., 2017. A deep learning approach for cancer detection and relevant gene identification. Pacific symposium on biocomputing 2017. World Scientific.

Dancey, J.E., Bedard, P.L., Onetto, N., Hudson, T.J., 2012. The genetic basis for cancer treatment decisions. Cell. https://doi.org/10.1016/j.cell.2012.01.014.

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., Wang, X., 2019. Deepcc: A novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis. https://doi.org/10.1038/s41389-019-0157-8.

Gil-Hernández, A., Arroyo-Campuzano, M., Simoni-Nieves, A., Zazueta, C., Gomez-Quiroz, L.E., Silva-Palacios, A., 2021. Relevance of membrane contact sites in cancer progression. Front. Cell Dev. Biol. https://doi.org/10.3389/fcell.2020.622215.

Gyamfi, J., Kim, J., Choi, J., 2022. Cancer as a metabolic disorder. Int. J. Mol. Sci. https://doi.org/10.3390/ijms23031155.

Hanahan, D., Weinberg, R.A., 2000. The Hallmarks of Cancer. Cell. https://doi.org/10.1016/S0092-8674(00)81683-9.

Hawkes, N., 2019. Cancer survival data emphasise importance of early diagnosis, British Medical Journal Publishing Group.

Hijazi, H., Chan, C., 2013. A classification framework applied to cancer gene expression profiles. J. Healthcare Eng. https://doi.org/10.1260/2040-2295.4.2.255.

Li, R., Li, L., Xu, Y., Yang, J., 2022. Machine learning meets omics: Applications and perspectives. Brief. Bioinform. https://doi.org/10.1093/bib/bbab460.

Liñares Blanco, J., Gestal, M., Dorado, J., Fernandez-Lozano, C., 2019. Differential gene expression analysis of RNA-seq data using machine learning for cancer research. Mach. Learn. Paradigms: Appl. Learn. Anal. Intell. Syst. https://doi.org/10.1007/978-3-030-15628-2_3.

Liu, Y., Chi, H., Chen, H., Wang, R., Jiang, L., Zhang, S., Jiang, C., Huang, J., Zhang, Q., Yang, G., 2023. Proposing new early detection indicators for pancreatic cancer: Combining machine learning and neural networks for serum miRNA-based diagnostic model. Front. Oncol. https://doi.org/10.3389/fonc.2023.1244578.

Malebari, A.M., Ibrahim, T.S., Salem, I.M., Salama, I., Khayyat, A.N., Mostafa, S.M., El-Sabbagh, O.I., Darwish, K.M., 2020. The anticancer activity for the bumetanide-based analogues via targeting the tumour-associated membrane-bound human carbonic anhydrase-ix enzyme. Pharmaceuticals. https://doi.org/10.3390/ph13090252.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., 2020. A compendium of mutational cancer driver genes. Nat. Rev. Cancer. https://doi.org/10.1038/s41568-020-0290-x.

Mehmood, M., Alshammari, N., Alanazi, S.A., Ahmad, F., 2022a. Systematic framework to predict early-stage liver carcinoma using a hybrid of feature selection techniques and regression techniques. Complexity. https://doi.org/10.1155/2022/7816200.

Mehmood, M., Alshammari, N., Alanazi, S.A., Basharat, A., Ahmad, F., Sajjad, M., Junaid, K., 2022b. Improved colourization and classification of intracranial tumour expanse in MRI images via a hybrid scheme of Pix2Pix-CGANS and NasNet-Large. J. King Saud Univ. - Comput. Inf. https://doi.org/10.1016/j.jksuci.2022.05.015.

Mei, Y., Wu, K., 2022. Application of multi-objective optimization in the study of anti-breast cancer candidate drugs. Sci. Rep. https://doi.org/10.1038/s41598-022-23851-0.

Munawar, Z., Ahmad, F., Alanazi, S.A., Nisar, K.S., Khalid, M., Anwar, M., Murtaza, K., 2022. Predicting the prevalence of lung cancer using feature transformation techniques. Egypt. Inform. J. https://doi.org/10.1016/j.eij.2022.08.002.

Rigel, D.S., Carucci, J.A., 2000. Malignant melanoma: Prevention, early detection, and treatment in the 21st century. CA: a cancer journal for clinicians. https://doi.org/10.3322/canjclin.50.4.215.

Saheed, Y.K., 2023. Effective dimensionality reduction model with machine learning classification for microarray gene expression data. Data science for genomics, Elsevier, pp. 153-164.

Sanko, V., Kuralay, F., 2023. Label-free electrochemical biosensor platforms for cancer diagnosis: Recent achievements and challenges. Biosensors. https://doi.org/10.3390/bios13030333.

Steyaert, S., Qiu, Y.L., Zheng, Y., Mukherjee, P., Vogel, H., Gevaert, O., 2023. Multimodal deep learning to predict prognosis in adult and pediatric brain tumours. Communications Medicine. https://doi.org/10.1038/s43856-023-00276-y.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. https://doi.org/10.3322/caac.21660.

Urda, D., Montes-Torres, J., Moreno, F., Franco, L., Jerez, J.M., 2017. Deep learning to analyze RNA-seq gene expression data. Advances in Computational Intelligence:

14th International Work-Conference on Artificial Neural Networks, IWANN 2017, Cadiz, Spain, June 14-16, 2017, Proceedings, Part II 14, Springer.

Wang, J., Wang, L., Liu, Y., Du, C., Hou, C., Xie, Q., Tang, D., Liu, F., Lou, B., Zhu, J., 2023. Change to the transcriptomic profile, oxidative stress, apoptotic and immunity in the liver of small yellow croaker (larimichthys polyactis) under hypoxic stress. Aquaculture. https://doi.org/10.1016/j.aquaculture.2023.739854.

Wesolowski, S., Birtwistle, M.R., Rempala, G.A., 2013. A comparison of methods for RNA-seq differential expression analysis and a new empirical Bayes approach. Biosensors. https://doi.org/10.3390/bios3030238.

Wishart, D.S., 2015. Is cancer a genetic disease or a metabolic disease? EBioMedicine. https://doi.org/10.1016/j.ebiom.2015.05.022.

Xiao, H., Hu, L., Tan, Q., Jia, J., Xie, P., Li, J., Wang, M., 2023. Transcriptional profiles reveal histologic origin and prognosis across 33 of the cancer genome atlas tumour types. Transl. Cancer Res. https://doi.org/10.21037/tcr-23-234.

Yuan, F., Lu, L., Zou, Q., 2020. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. https://doi.org/10.1016/j.bbadis.2020.165822.

Yuan, L.-M., Sun, Y., Huang, G., 2020b. Using class-specific feature selection for cancer detection with gene expression profile data of platelets. Sensors. https://doi.org/10.3390/s20051528.

Zhang, Y., Deng, Q., Liang, W., Zou, X., 2018. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. BioMed Res. Int. https://doi.org/10.1155/2018/7538204.