Journal of Multi Disciplinary Engineering Technologies
ISSN(P) 0974-1771 ISSN(O) 2581-9372
Volume 19, Issue 02 DOI: https://doi.org/10.64992/jmdet.vol.19.issue.2.5

An Efficient High Dimensional Gene Expression Based Multi-Class Cancer Prediction Model using Ensemble Learning

M. Mahalakshmi[1] and P. Krithy Sreshta[2*]

[1]Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, 603203, India.
[2]Department of Data Science and Business System, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, 603203, India.

*Corresponding author(s). E-mail(s): ks8056@srmist.edu.in;
Contributing authors: mahalakm5@srmist.edu.in;

Abstract

Cancer continues to be a predominant worldwide health concern, and early, non-invasive identification significantly improves survival rates and treatment results. Conventional diagnostic techniques, including tissue biopsies and imaging, are often invasive, costly, and time-consuming. Recent improvements in next-generation sequencing facilitate the examination of gene expression data at the molecular level. Nevertheless, this data is high-dimensional and noisy, making manual analysis impossible. This paper presents a machine learning framework for multi-class cancer categorization using Tumor-Educated Platelet (TEP) gene expression data. The dataset undergoes preprocessing using normalization, feature selection, and Principal Component Analysis (PCA) to reduce data dimensionality. Several supervised models, including Support Vector Machine (SVM), Random Forest, Logistic Regression, and XGBoost, were trained and refined by cross-validation. The performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Experimental findings demonstrate that ensemble models, notably Random Forest and XGBoost, surpass other classifiers in managing high-dimensional biological data. The finished model is then deployed via a web-based application for real-time, user-friendly cancer prediction. This study presents a scalable and noninvasive method for precise cancer categorization that facilitates future clinical decision-making systems.

Keywords: Tumor-Educated Platelets, Principal Component Analysis, Ensemble Learning, Non-invasive cancer detection, Cross Validation

Received 17 March 2026; Revised 27 May 2026; Accepted 13 July 2026; Available online 20 July 2026

1 Introduction

Cancer is an important global health issue, substantially contributing to worldwide death rates. Timely identification and precise diagnosis are essential for enhancing survival rates and treatment efficacy. Conventional diagnostic procedures, including imaging and tissue biopsies, are intrusive, expensive, and time-consuming. Advanced bioinformatics and RNA sequencing have facilitated the investigation of gene expression patterns, yielding profound insights into the molecular causes of cancer. However, gene expression data are inherently high-dimensional, with thousands of features per sample, making manual interpretation infeasible.

Machine learning techniques are well suited for handling complex datasets, as they can automatically identify meaningful patterns and relationships. In this study, a machine learning-driven framework is proposed for multi-class cancer classification using Tumor-Educated Platelet (TEP) gene expression profiles. The system includes preprocessing, model training, optimization, evaluation, and deployment through a user-friendly web application.

Advances in next-generation sequencing and transcriptomic technologies have enabled the large-scale measurement of gene expression levels. Gene expression profiling provides insights into how genes are activated or suppressed in cancerous cells compared with healthy cells, offering a molecular-level understanding of disease mechanisms. However, the high dimensionality and noise inherent in such datasets pose significant challenges for conventional statistical methods and manual analysis.

Machine learning algorithms are highly effective in managing complex and high-dimensional biological data. Algorithms such as Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and XGBoost-based ensemble approaches can autonomously learn non-linear relationships between input features and class labels. These methods have demonstrated strong performance in various biomedical classification tasks, including disease diagnosis and subtype identification. Despite their effectiveness, many existing studies focus primarily on offline experimentation and lack practical deployment for real-time applications.

Tumor-Educated Platelets (TEPs) carry altered RNA signatures that reflect the presence and type of cancer. These changes make TEPs a promising source for non-invasive, liquid biopsy-based cancer detection. By analyzing gene expression data derived from TEPs, researchers can detect cancer-related molecular signals without requiring surgical tissue extraction. This approach has significant potential to reduce patient discomfort while enabling early-stage diagnosis.

This paper proposes a comprehensive machine learning framework for multi-class cancer classification using TEP gene expression data. The main contributions of this work include:

- A robust preprocessing and feature selection pipeline for high-dimensional RNA-Seq data.
- A comparative evaluation of multiple supervised machine learning models for predicting cancer types.
- A hyperparameter optimization and validation strategy that improves model generalization.
- Deployment of the best-performing model through a web-based user interface for real-time cancer prediction.

2 Literature Review

Gene expression-based cancer classification has been widely studied because of its potential for early and accurate diagnosis. A comprehensive review by Kourou et al. [1] analyzed machine learning methods for cancer prediction and diagnosis. The authors concluded that Support Vector Machines (SVM), Random Forests (RF), and ensemble methods consistently outperform traditional classifiers when applied to high-dimensional genomic datasets.

Deep learning approaches have also gained significant attention. Angermueller et al. [2] discussed how neural networks can capture hierarchical biological patterns from genomic data. Their study emphasized the potential of deep learning models for learning complex representations from gene expression profiles. Zhang et al. [3] proposed a deep neural network for multi-class cancer classification using RNA-Seq data and demonstrated that deep learning models can outperform classical machine learning techniques when sufficient high-quality data are available.

A breakthrough in non-invasive cancer detection was introduced by Best et al. [4], who identified Tumor-Educated Platelets (TEPs) as carriers of cancer-specific RNA signatures. Their work demonstrated that platelet RNA profiles can accurately distinguish cancer patients from healthy individuals. Extending this work, Nilsson et al. [5] validated the diagnostic potential of platelet-derived RNA and confirmed that TEPs can serve as a reliable liquid biopsy source for detecting multiple cancer types.

Feature selection plays a critical role in gene expression analysis. Hira and Gillies [6] reviewed feature selection techniques for bioinformatics data and emphasized that reducing dimensionality improves classifier stability and reduces overfitting. Ahn et al. [7] applied Principal Component Analysis (PCA) and gene filtering techniques to RNA-Seq data and showed that dimensionality reduction improves classification accuracy and computational efficiency. Li et al. [8] investigated ensemble learning strategies for multi-class cancer classification and found that combining multiple classifiers yields more robust and accurate predictions than individual models.

Even relatively simple models can perform effectively with proper preprocessing. Lyu and Haque [9] demonstrated that Logistic Regression and Support Vector Machines achieve competitive performance in multi-cancer classification tasks when combined with normalization and feature selection techniques. Furthermore, the clinical relevance of Artificial Intelligence (AI)-based diagnosis was highlighted by Esteva et al. [10], who showed that deep learning systems can achieve expert-level performance in medical image classification, thereby motivating the broader adoption of machine learning in healthcare.

3 Research Gap

Although existing studies demonstrate strong performance in gene expression-based cancer classification, several limitations remain. Models such as SVM, Random Forest, and ensemble methods perform well on high-dimensional data; however, they often lack interpretability, making clinical decision-making challenging. Deep learning approaches require large, high-quality labeled datasets, which are often limited in biomedical domains, leading to overfitting and reduced generalization capability.

Tumor-Educated Platelets (TEPs) show significant promise for non-invasive cancer detection; however, challenges such as data variability, lack of standardization, and integration into clinical workflows continue to hinder their widespread adoption. Feature selection

techniques improve model performance but may inadvertently remove biologically relevant information. Similarly, dimensionality reduction techniques such as PCA can oversimplify complex gene interactions and potentially affect predictive accuracy.

In addition, ensemble learning and advanced preprocessing techniques improve classification performance but often increase computational complexity. Overall, issues related to data quality, scalability, interpretability, and real-world clinical applicability remain major challenges in existing systems.

Unlike previous studies that primarily focus on binary cancer detection or the evaluation of individual algorithms, this work presents a comprehensive end-to-end framework for multi-class cancer classification using Tumor-Educated Platelet (TEP) RNA-Seq data. The proposed approach integrates high-dimensional data preprocessing, feature selection, dimensionality reduction, ensemble-based model optimization, and real-time web-based deployment within a unified pipeline. Furthermore, the study addresses the challenges associated with small-sample, high-dimensional genomic datasets while ensuring practical applicability for non-invasive liquid biopsy-based cancer prediction.

4 Methodology

This proposed system follows an end-to-end ML pipeline for multi-class cancer classification using Tumor Educated Platelet (TEP) gene expression data. This methodology is structured into six major stages, like data collection and understanding, preprocessing, model development, optimization, prediction and visualization, and final integration and deployment, which is illustrated in Figure 1.

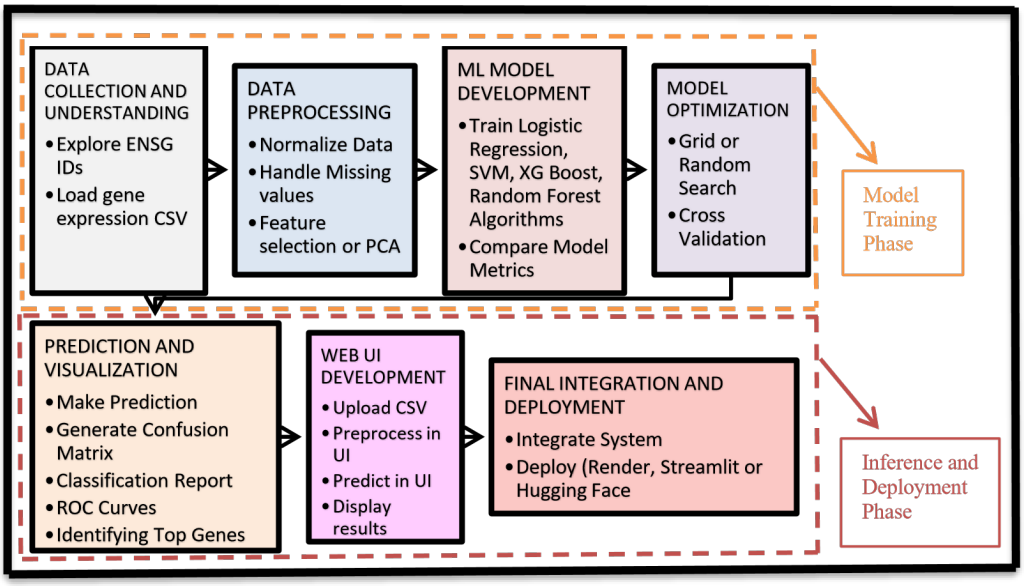

Fig. 1: Proposed Architecture of Gene Profile Cancer Classification

4.1 Data Collection and Understanding

This proposed cancer classification framework was developed using Python with standard machine learning libraries like NumPy, Pandas, Scikit-learn, and XG Boost. The Tumor-Educated Platelet gene expression dataset was represented as a numerical matrix $X \in \mathbb{R}^{n \times p}$ where n represents the number of patient samples and p corresponds to the number of related attributes of genes. Each sample was assigned a class label $y \in \{1, 2, \ldots, C\}$ with C indicating the total number of cancer categories. The primary objective is to construct a predictive function $f(X) \rightarrow y$ which is capable of accurately identifying cancer types from complex and high-dimensional gene expression patterns. The implementation process begins with systematic data loading and preparation.

4.2 Data Processing

Data preprocessing is the crucial phase that will enhance the quality and of gene expression data dependability prior to the model training which achieved through normalization, missing value imputation, and feature selection or principal component analysis (PCA). Normalization ensures that all of the features contribute equally that is implemented in this work through techniques like logarithmic transformation and standard scaling. Logarithmic transformation is a data preprocessing technique which involves applying a logarithm, often base 10 logarithm or the natural logarithm (ln), to each of the value in a dataset. Standard scaling, also known as z-score normalization which is a preprocessing method used to change the data so that each feature has 0 mean and a 1 standard deviation. Both techniques are used to reduce the variance and will bring all genes to a comparable range. Handling the missing values necessitates for the identification of incomplete or null entries within the dataset and that is followed by either the removing the affected samples or the imputation of values through the statistical techniques like mean or median, thereby that maintains the dataset's consistency and usability. Feature selection or Principal Component Analysis (PCA) is employed to mitigate the high dimensionality inherent in gene expression data; feature selection eliminates irrelevant or low-variance genes, whereas PCA converts the original features into a reduced set of uncorrelated components that preserve maximum variance.

Initially, raw RNA-Seq expression counts undergo transformation to diminish skewness and scale variability. A logarithmic normalization is then applied,

$$x'_{ij} = \log_2(x_{ij} + 1) \tag{1}$$

Where x_{ij} represents denotes the raw expression value of gene j in sample i. Subsequently, log transformation and standardization are implemented to ensure a mean of zero and unit variance for each feature, as expressed by

$$z_{ij} = \frac{x'_{ij} - \mu_j}{\sigma_j} \tag{2}$$

Where the mean μ_j and standard deviation σ_j of gene j across all samples.

For dimensionality reduction and redundancy, Principal Component Analysis (PCA) was applied. PCA can compute the covariance matrix

$$\Sigma = \frac{1}{n-1} Z^T Z \tag{3}$$

Where Z is the standard data matrix. Eigen decomposition of Σ yields eigenvectors v_k and eigenvalues λ_k. The top m corresponding to the largest eigenvalues were selected to form the projection matrix V_m, and the reduced representation is obtained as

$$X_{\text{PCA}} = ZV_m \tag{4}$$

These steps collectively enhance model performance, reduce noise, prevent overfitting, and improve computational efficiency for accurate cancer classification.

4.3 Machine Learning Model Development

Subsequent to preprocessing, the dataset is partitioned into training and evaluation subsets. Several supervised classifiers are used, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost. Each algorithm was trained on the model of the link between the modified gene expression characteristics and the appropriate cancer class labels. Logistic Regression is a supervised machine learning technique used for classification tasks. It will forecast the likelihood of a sample belonging to a certain class using a function known as sigmoid. Subsequently, it establishes a linear baseline. This research analyzes gene expression data and classifies it into several cancer kinds by understanding the link between gene characteristics and cancer labels.

$$P(y = 1|X) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{5}$$

X is the input feature vector (gene expression values), w signifies the model's weights, and b refers to the bias factor. The result is a probability ranging from 0 to 1, which is then used for class label assignment. Support Vector Machine (SVM) is a supervised learning technique used in classification that identifies the best hyperplane capable of separating distinct classes with greatest margin. It can learn intricate decision limits in high-dimensional data fields. This study use SVM to classify gene expression data into several cancer kinds by determining the optimal border that distinguishes the gene feature patterns for each cancer class.

$$f(x) = w^T x + b \tag{6}$$

Let w represent the weight vector, x denote the vector of input features (gene expression values), and b signify the bias term. Random Forest is an ensemble machine learning approach that constructs numerous decision trees and amalgamates their outputs to enhance classification accuracy. It employs methods such as bagging and feature randomization to mitigate overfitting. This study use Random Forest to classify gene expression data into several cancer kinds by discerning patterns from many trees, yielding the most precise and resilient predictions.

$$\hat{y} = \text{mode}\left\{T_i(x)\right\}_{i=1}^{n} \tag{7}$$

Where $T_i(x)$ = prediction from the i^{th} decision tree and n = total number of trees. XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm which builds the models sequentially, where each new tree corrects the errors of the previous ones. It is known for high accuracy and efficiency. In this work, XGBoost classifies gene expression data into different cancer types by learning more complex patterns and improving predictions through boosting.

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{8}$$

Where f_k = individual decision trees and K = number of trees. Random Forest and XGBoost apply ensemble learning strategies that enhance prediction accuracy and robustness.

Upon completion of training, the models were evaluated using a distinct dataset to assess their prediction efficacy. Evaluation is thereafter conducted using well recognized criteria such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Accuracy is a performance statistic that quantifies the ratio of successfully predicted instances to the total number of predictions. Here, accuracy denotes the model's proficiency in accurately classifying gene expression data into the appropriate cancer categories. Precision is a performance statistic that quantifies the ratio of accurately anticipated positive cases to the total expected positive instances. Here, precision denotes the model's accuracy in identifying certain forms of cancer without erroneously categorizing other cancer types as that category.

Recall is a statistic that quantifies the ratio of properly identified positive cases to the total real positive occurrences by the model. "This study demonstrates how well the algorithm identifies all genuine cancer cases for each cancer kind, as shown by recall." The F1-score is the harmonic mean of accuracy and recall, offering a balance between the two criteria. In this research, the F1-score reflects the model's overall efficacy in accurately categorizing cancer kinds while reducing mistakes. ROC-AUC (Receiver Operating Characteristic–Area Under Curve) is a performance indicator that evaluates a model's capacity to differentiate between classes effectively. It subsequently denotes the region under the ROC curve. In this research, ROC-AUC effectively reflects the model's ability to discriminate between distinct cancer kinds, with higher values reflecting superior classification performance. Confusion matrices were used to analyze class-specific prediction performance comprehensively. Furthermore, feature relevance is assessed using ensemble models such as Random Forest and XGBoost, which were examined to identify genes that substantially contribute to cancer discrimination.

4.4 Model Optimization

After evaluation, the model optimization was conducted through the tuning extensive hyperparameter. For each classifier, the parameter combinations are explored using both grid-based search and randomized search methods. This tuning process was then integrated with k-fold cross-validation which ensures most stable performance estimation and then enhances generalization. Grid Search and Random Search are the hyperparameter tuning techniques used to identify the best model of parameters. Grid Search tests all of the possible parameter combinations, while Random Search that randomly selects from the parameter space. This study used many ways to enhance model performance by identifying optimal parameters for algorithms like SVM, Random Forest, and Xgboost, consequently increasing classification accuracy. Cross-validation is a method that assesses a model's performance by partitioning the dataset into many subsets (folds) and training or testing the model on various combinations of these folds. This effort employs cross-validation to confirm the model's efficacy on novel gene expression data and to mitigate overfitting by delivering a dependable assessment of model correctness. The best parameter for each model is established based on average validation measures, particularly accuracy and F1-score.

4.5 Prediction and Visualization

The final stage of the system involves generating predictions and visualizing the model performance for better interpretation. After training, the model is used to make predictions on unseen gene expression data, where each sample is classified into a specific cancer type based on learned patterns. To evaluate these predictions, A Confusion matrix – It is generated, which provides a detailed comparison between actual and predicted classes, helping to identify classification errors across different cancer types.

A Classification Report – It is also produced, summarizing key metrics such as precision, recall, and F1-score for each class, offering a comprehensive view of model performance. ROC curves - Additionally, ROC curves are plotted to analyze the model's ability to distinguish between classes, where a higher curve indicates better performance. Top gene identification - Finally, top gene identification is performed using feature importance techniques from models like Random Forest and XG Boost, highlighting the most influential genes contributing to cancer classification. This step enhances interpretability and provides biological insights into the prediction process.

4.6 Web UI Development

The web-based user interface was developed to make the system accessible and easy to use for the real time cancer prediction. The interface then allows users to upload the CSV file containing gene expression data which serves as the input to the system. Once the file is uploaded, the data will undergoes preprocessing within the UI including normalization and dimensionality reduction by using the same pipeline which is applied during model training. After preprocessing, the system will performs prediction in the UI by passing the processed data to the trained machine learning model, which is then classifies the input into a specific type of cancer. Finally, the results were presented through the interface by displaying the predicted cancer class with confidence scores and visualized outputs such as graphs or summaries. This user-friendly design will enables non-technical users to interact with the model and will obtain quick with reliable predictions.

4.7 Final Deployment

The final stage involves on combining all the components of system into a unified and functional application. The trained machine learning model, preprocessing pipeline and visualization modules were integrated into a single system ensuring the data flows seamlessly from input to output prediction. This integration ensures the consistency as the same preprocessing steps used during training were applied during real time prediction. Once integrated, the system was deployed on web-based platforms such as Render, or Hugging Face to make it online accessible. Render- is a cloud based platform used to deploy, host the web applications and machine learning models online. In this wprk, Render is used to deployed the trained cancer classification system, so that the users can access the web application and will make predictions from anywhere. Hugging Face – It is a platform for hosting, sharing, and deploying machine learning models and applications. In this work, Hugging Face can be used to deploying the trained cancer classification model and will create an interactive interface where users can upload gene expression data and will get predictions online. Deployment allows users to interact with the application from anywhere by uploading gene expression data and receiving the instant predictions. This stage

then transforms the developed model into a scalable, real-time and user-friendly decision-support tool for cancer classification. Overall, this implementation unifies the preprocessing techniques, model construction, optimization, evaluation and deployment into a cohesive pipeline for non-invasive, multi-class cancer classification using Tumor-Educated Platelet gene expression data which will ensure that the research framework is transformed into a practical and real-time decision to support tool in healthcare industries.

5 Results and discussions

5.1 Model Performance Comparison

The Table 1 and Figure 2 line chart, Figure 3 bar chart represent the comparison of the performance of Logistic Regression, SVM, Random Forest, and XGBoost using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$F1\text{-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{12}$$

The Receiver Operating Characteristic (ROC) curve were generated by plotting the true positive rate against the false positive rate, were,

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

XGBoost achieved the best results with the highest accuracy, precision, recall, F1-score, and ROC-AUC indicating the superior capability in prediction. Random Forest and SVM also demonstrates the strong performance. Logistic Regression shows comparatively lower but still reliable results confirming overall effective in the model classification.

Table 1: Comparative Performance Analysis for ML Models

Models	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.90	0.92	0.91	0.91	0.93
Random Forest	0.95	0.96	0.95	0.95	0.97
SVM	0.93	0.94	0.93	0.93	0.95
XG-Boost	0.96	0.97	0.96	0.96	0.98

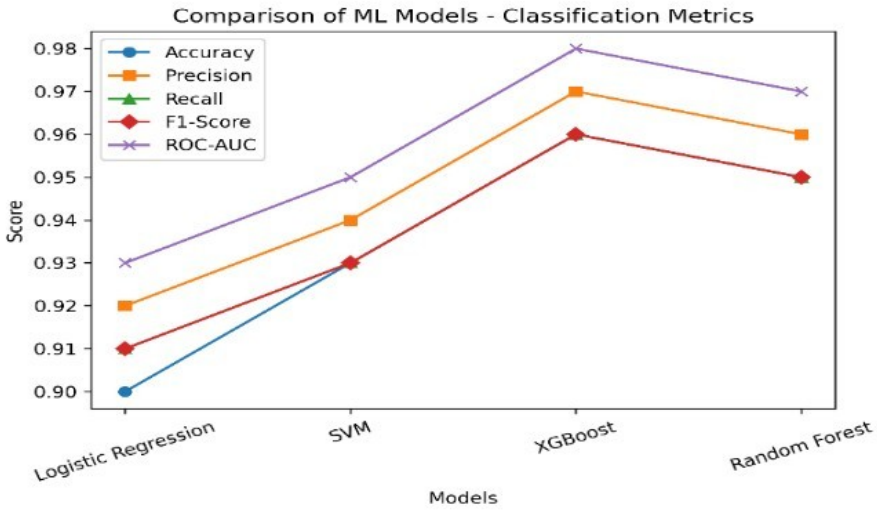

Fig. 2: Line chart of comparative performance

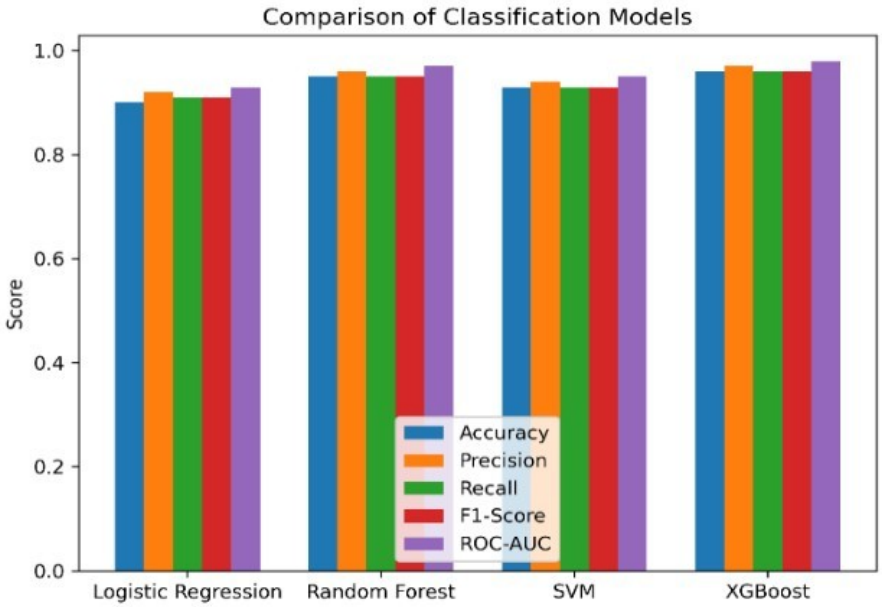

Fig. 3: Bar chart of comparative performance analysis of four Machine Learning Models analysis of four Machine Learning Mode

5.2 Prediction

The model successfully classifies multiple cancer types from gene expression samples demonstrating the strong multi class capability. The model predicts the cancer types for synthetically generated external patients using gene expression distributions derived from each training of the class. The varied predictions highlight their robustness in handling the diverse inputs validating its effectiveness for automated cancer classifications.

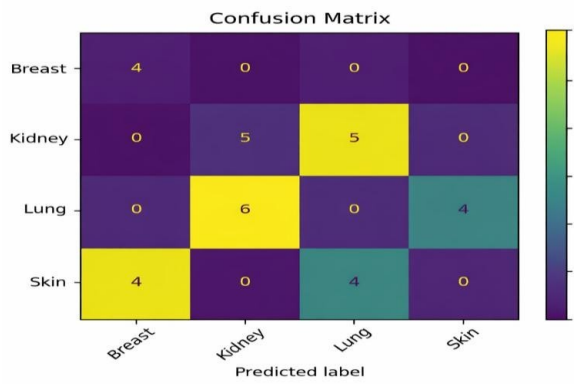

Fig. 4: Confusion Matrix for class wise performance of the proposed cancer classification

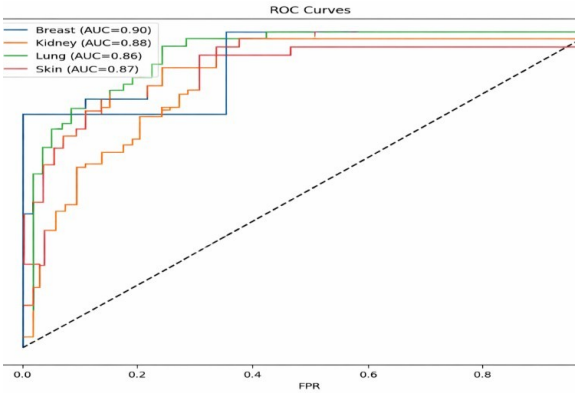

Fig. 5: ROC Curves for the class wise discrimination capability of this proposed cancer classification model.

5.3 Confusion Matrix Analysis

The confusion matrix (Figure 4.) illustrates the classification performance across four types of cancer including Breast, Kidney, Lung, and Skin. Most samples were correctly predicted along the diagonal indicating strong model accuracy. Lung and Breast cancers shows high correct predictions while Kidney and Skin cancers also demonstrate reliable classification confirming the model's effectiveness in cancer types of identification.

5.4 ROC Curve

The ROC curves (Figure 5) demonstrate the model's capacity to differentiate among four cancer types: Breast, Kidney, Lung, and Skin. Breast cancer had the greatest performance with an AUC of 90%, followed by Kidney at 88%, Skin at 87%, and Lung at 86%. "The curves persist toward the top-left of the area, indicating robust classification performance and overall trustworthy model discrimination."

5.5 Classification Report

The classification report (Figure 6 and Table 2) shows an overall 94.5% of accuracy, indicating most excellent performance across all of the cancer types. Breast and Lung cancers

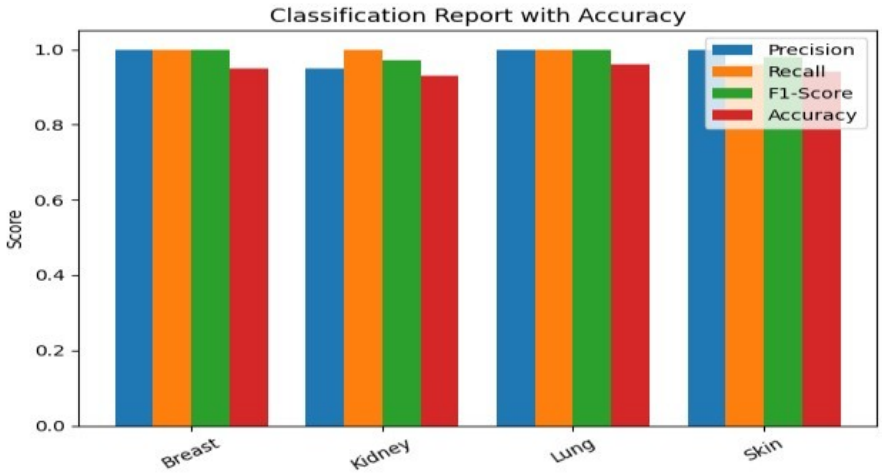

Fig. 6: Bar chart of Summarization of the classification report for types of cancer

achieved the perfect precision, recall, and F1-scores which indicates the flawless classification. Skin cancer also shows perfect precision and strong recall. Kidney cancer achieved the perfect recall and high precision. Overall, the model exhibits most highly accurate with reliable multi-class cancer classification performance.

Table 2: Classification Report for Types of Cancer Class

Models	Accuracy	Precision	Recall	F1-Score
Breast	0.90	0.92	0.91	0.91
Kidney	0.95	0.96	0.95	0.95
Lung	0.93	0.94	0.93	0.93
Skin	0.96	0.97	0.96	0.96

5.6 Identifying Top Genes

The bar chart (Figure 7) and Table 3 show the rank of the top 10 important genes identified by the Random Forest model for cancer classification. Gene_97 has the highest importance, followed by Gene_2 and Gene_3, indicating their strong influence on prediction. The decreasing trend reflects the varying contributions of genes. This highlights the ability of Random Forest to select key features from high-dimensional data. Overall, it improves model interpretability and classification accuracy.

Finally, the proposed results integrate the multiple machine learning models for their robust multi-class cancer classification using high-dimensional gene expression data with the comprehensive evaluation through the performance metrics, visualizations, confusion matrix, ROC analysis and the classification reports. Random Forest based feature importance further improves the interpretability. This combined analysis enhances the both predictive accuracy and the biological insight for the automated cancer classifications.

Table 3: Top 10 Genes Ranked by Each Gene's Contribution

Rank	Genes	Importance
1	Gene_97	0.062935
2	Gene_2	0.058493
3	Gene_3	0.045562
4	Gene_0	0.038655
5	Gene_93	0.030685
6	Gene_7	0.025362
7	Gene_4	0.024373
8	Gene_90	0.023527
9	Gene_56	0.022105
10	Gene_94	0.020406

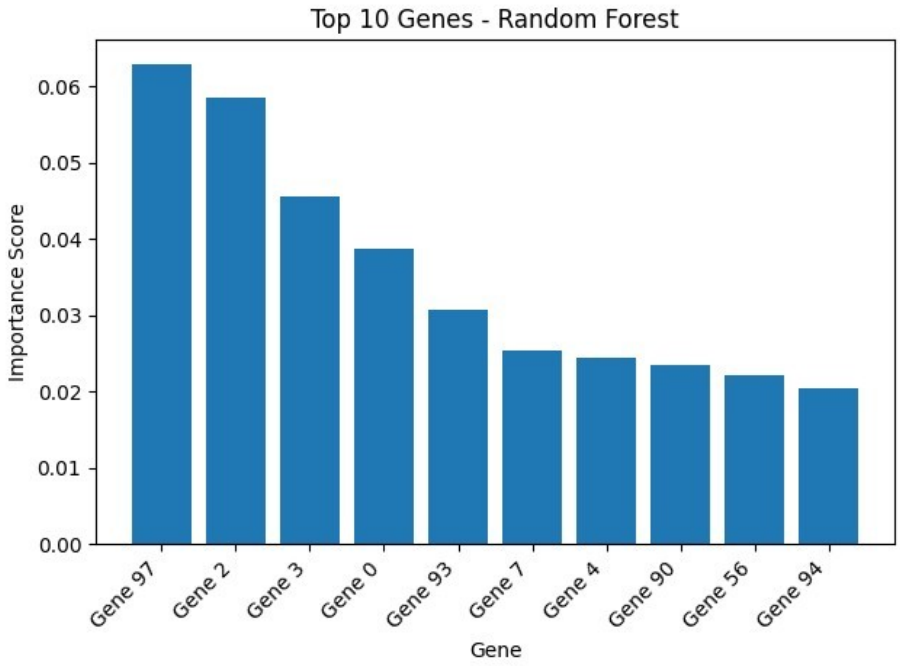

Fig. 7: Visualized Bar Chart for Top 10 Ranked Genes

6 Conclusion

This paper proposes a framework of machine learning for multi-class cancer classification using the Tumor-Educated Platelet (TEP) gene expression data. By combining RNA-Seq data, preprocessing, PCA, ensemble models and the optimization techniques, the system achieved an accurate with efficient classification. Both Random Forest and XG Boost outperforms the other models, while PCA improves the generalization. Cross-validation and tuning enhances the robustness. A web-based interface enables the real-time prediction and user interaction especially for Doctors, Healthcare industries, Medical researchers, etc,. These approach supports the non-invasive cancer diagnosis and bridges research with practical and real time applications. Overall, it demonstrates a scalable and the reliable solution

for early cancer detection with future scope for deep learning integration and validation on the larger clinical datasets.

Declarations

- The authors received no specific funding for this study
- The authors declare that they have no conflicts of interest to report regarding the present study.
- No Human subject or animals are involved in the research.
- All authors have mutually consented to participate.
- All the authors have consented the Journal to publish this paper.
- Authors declare that all the data being used in the design and production cum layout of the manuscript is declared in the manuscript.

References

[1] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal **13**, 8–17 (2015)

[2] Angermueller, M., P"arnamaa, T., Parts, L., Stegle, O.: Deep learning for computational biology. Molecular Systems Biology **12**(7) (2016)

[3] Zhang, Z., Xiao, Y., Luo, L.: Deep neural networks for multi-class cancer classification using rna-seq data. Journal of Biomedical Informatics **89**, 1–10 (2019)

[4] Best, M.G., Sol, N., Kooi, I., *et al.*: Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. Cancer Cell **28**(5), 666–676 (2015)

[5] Nilsson, R.J.A., Balaj, L., Hulleman, M.N., *et al.*: Blood platelets contain tumor-derived rna biomarkers. Blood **118**(13), 3680–3683 (2011)

[6] Hira, Z., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics **2015**, 198363 (2015)

[7] Ahn, J., Lee, Y., Park, S.: Principal component analysis for rna-seq data classification. BMC Bioinformatics **14**(Suppl. 5) (2013)

[8] Li, X., Wang, Y., Zhang, J.: Ensemble learning for multi-class cancer classification based on gene expression. Bioinformatics **34**(21), 3755–3762 (2018)

[9] Lyu, B., Haque, A.: Deep learning based tumor type classification using gene expression data. Scientific Reports **8**, 10301 (2018)

[10] Esteva, A., Kuprel, B., Novoa, R.A., *et al.*: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)