B. Shyamala Gowri[1],*,
S. Anu H. Nair[2],
K. P. Sanal Kumar[3],
S. Kamalakkannan[4]

Machine Learning in DNA Microarray Analysis for Cancer Classification

Abstract: - Successful patient treatment depends on the early detection of cancer utilizing gene expression data. Since incorrect detection can lead to greater complexity and higher fatality rates, accurate data identification is crucial to preventing it. Many features, each representing a different gene, are commonly found in gene expression data. High dimensionality brought about by the number of characteristics raises computing complexity and resource requirements. Moreover, multicollinearity problems might arise from the presence of duplicated, strongly linked chosen characteristics. Overall classification accuracy may be jeopardized by some constraints in the current works, such as poor performance brought on by deteriorated data quality, overly storage space needs, overfitting problems, and lack of resilience. This study uses an effective framework based on a Machine learning (ML) method to overcome these issues and improve classification results. The data is first collected using five gene cancer databases, data transformation techniques are then used to enhance the data. We employ Min-Max Adjusting to pre-processed data. The most important genes are chosen while superfluous or undesirable ones are removed using the LDA, or linear discriminant analysis technique. The XGBoost is used to classify various malignant and non-cancerous classes based on the chosen gene collection. The resolution of the dimensionality and overfitting issues significantly enhances the performance of the suggested model. Python is used for the implementation, and it is shown that the XGBoost model's overall accuracy across all datasets is 99.37%. The model's overall performance is also assessed using measures including accuracy, remember as well as F1 score. The suggested model has a notably improved performance in terms of effectiveness than existing approaches.

Keywords: DNA Microarray, Data transformation, Min-Max Normalization, Linear Discriminant Analysis, XGBoost, Accuracy, recall, F1-sscore.

INTRODUCTION

The ability to measure thousands of genes' expression levels at once in a single experiment thanks to DNA microarray technology has completely changed the study of gene expression. This powerful tool has proven invaluable in cancer research, providing insights into the molecular differences between normal and cancerous cells, as well as among different types of cancers. DNA microarrays generate vast amounts of data, capturing the complex patterns of gene expression that define various cancer subtypes and their progression. However, the large-scale, high-dimensional nature of this data presents significant challenges in terms of analysis and interpretation.Machine learning (ML) techniques have become integral in addressing these challenges. The creation of algorithms that can automatically learn from data and generate predictions or reveal hidden patterns is known as machine learning. In the context of DNA microarray analysis, ML algorithms can be applied to classify cancer types, identify biomarkers, and predict patient outcomes based on gene expression profiles. By analyzing the complex data generated from microarray experiments, machine learning models can uncover subtle, yet important, patterns that traditional statistical methods might miss.

[1],*Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, (Deputed To WPT, Chennai), Chennai, India.
[3]Assistant Professor, Department of CS, RV Government Arts College, Chengalpattu, India
[4]Associate Professor, Department of Information Technology, School of Computing Sciences, (VISTAS), Chennai, Tamil Nadu, India
1,*shyamalagowribalaraman@gmail.com, 2anu_jul@yahoo.co.in, 3sanalprabha@yahoo.co.in , 4kannan.scs@velsuniv.ac.in

Numerous machine learning algorithm types have demonstrated potential in the categorization of cancer. including supervised learning approaches such as support vector machines, random forests, and neural networks, which are particularly useful for training models to distinguish between different cancer subtypes based on labeled datasets. Unsupervised Learning techniques like dimensionality reduction and grouping are also employed to explore the underlying structure of the data and identify novel patterns of gene expression linked to cancer.The integration of machine learning with DNA microarray technology holds the potential to significantly improve cancer diagnosis, prognosis, and treatment. By enabling more accurate and efficient analysis of gene expression data, ML models can assist in the development of personalized medicine, where treatments are customized to each patient's unique genetic cancer features. Ultimately, the combination of DNA microarray analysis and machine learning is paving the way for more precise, individualized cancer care.A novel automated approach is presented for the effective classification of cancer data. Some of the potential objectives for promoting the complete tailored to the particular hereditary cancer characteristics of each patientTo introduce an accurate classification system based on novel feature selection and classification process for enhancing overall accuracy with minimized error.

- To undergo data pre-processing using the Min-Max normalization process for normalizing the data within a particular range.
- To develop a Linear Discriminant Analysis (LDA) for selecting suitable genes from the large dataset, minimizing the feature dimensionality and improving the training ability.
- To develop a Machine Learning model, XGboost model for effectively classifying cancerous and non-cancerous cells from the selected genes in various classes.
- To perform classification analysis using different parameters and comparison with different machine learnin models for estimating the performance superiority.

The remainder of the paper is organized into important sections that are explained as follows: Section II lists the current research projects in gene expression data categorization that have been completed by different authors. Section III outlines the workflow of the proposed method, which consists of feature selection, classification, and data pre-processing. The gene categorization findings and performance analysis are shown in Section IV. In Section V, together with references, is the conclusion of the suggested work that will be undertaken in a future scope.

1. RELATED WORK

Mahmood khalasan et al., (2022) Machine learning techniques are useful approaches commonly utilized to develop cancer prediction models using relevant gene expression and mutation data. Numerous facets of machine learning-related cancer research are covered by the survey, such as microarray, RNA-Seq, biomarker gene discovery, cancer categorization, and cancer prediction. We also look into how determining potential biomarker gene expression patterns might help forecast future cancer risk and guide the delivery of individualized care.

Khosro Rezaee et al., (2022) Microarray data can be used to diagnose and categorize diseases and cancers with a variety of genetic causes employing a special deep neural network for classification and gently ensembling to identify the most effective genes stacked deep neural network was used to classify all three datasets, with average accuracy rates of 96.34%, 99.6%, and 97.51%, in that order. Two previously unreleased datasets from brain tissue lesions and small, round blue cell tumors (SRBCTs) associated with multiple sclerosis were examined in order to further illustrate the generalizability of the model approach.

Bingsheng He et al., (2020) In the United States, 3–5 out of every 100 cancer patients have carcinoma of unknown primary, which is a term used to describe metastatic malignancies with an unknown malignant origin. The treatment and follow-up diagnosis of CUP are further complicated by malignant heterogeneity and metastasis. In this study, we offer a machine learning technique that tracks the origin of tumor tissue utilizing somatic data from mutation sequencing that has been adjusted for gene length.

Tarek Khorshed et al., (2020) Our algorithm achieves on human samples, the categorization accuracy was 98.9%. that include 33 different types of malignant tumors distributed across26 organ locations. We show how our

transfer learning approach may be used to create classifiers for tumors with too few examples for self-learning. We present visualization techniques to shed biological light on our model's classification performance across a variety of cancers.

Alejandro Lopez-Rincon, et al., (2020). The suggested method is evaluated on a classification of cancer subtypes in order to differentiate triple negative breast cancer from another breast cancer subtypes task. Using samples collected from ten distinct clinical studies, it is first evaluated on a tumor classification issue to distinguish between ten distinct forms of cancer. All things considered, the methodology that has been described is successful and performs well when compared to other cutting-edge feature selection techniques.

Abdu Gumae et al., (2021) One of the primary areas of research in computer-based medical science is the use of machine learning algorithms for cancer diagnosis. Applying methods from soft computing and machine learning to analyze gene expression data from microarrays is a useful method for diagnosing prostate cancer. The experimental results show that, with a precision of 95.098%, the proposed method performs better than comparable methods using the same dataset.

E. H. Houssein, D et al., (2021) These days, distinguishing between malignant and healthy tissues as well as between other cancer kinds is a crucial problem. The primary difficulty in diagnosing cancer is thought to be choosing the few informative genes. According to the experimental results, the suggested BMO-SVM approach outperformed a number of popular Optimization algorithms that use meta-heuristics, including Particle Swarm Optimization (PSO), Tunicate Swarm Algorithm (TSA), Genetic Algorithm and Artificial Bee Colony (ABC). Notably, in terms of informational superiority percentage, our proposed method performs better than existing algorithms.

Hajieskand, etal., (2020) The accuracy of the suggested approach was around 100. Additionally, the suggested approach was contrasted with decision tree algorithms, one vs. RBF, the nearest neighbor, Naive Bayes, linear regression, and linear support vector machine classification. The suggested approach improved the LUAD dataset by 0.57, optimized the STAD dataset by 1.11, and developed the BRCA dataset by 0.78.

Hussam Jasim Ali, et al., (2024). Machine learning algorithms have demonstrated the ability to classify cancers using gene expression data. hybrid model achieved classification accuracies of 91.46% on data related to breast cancer, 91.54% on data related to brain cancer, and 95.16% on data related to colon cancer. The results demonstrate that cancer gene expression profiles can be accurately classified using an evolutionary hybrid method that combines PSO and PNN. The suggested method performs better than traditional machine learning techniques. These results can be confirmed by additional study using additional cancer datasets.

Muhammed Abd-Elnaby, et al., (**2021**) The World Health Organization lists cancer, especially breast cancer, as one of the world's primary causes of death. Much study has been done in the area of accurate and fast cancer diagnosis in an effort to increase the likelihood of a cure. To enhance the microarray-based classification, the paper reviews the primary feature selection and classification methods that have been discussed in cancer literature, particularly with relation to breast cancer.

Haiyan Liu, et al., (2021) One kind of metastatic cancer where the primary tumor site is unknown is called carcinoma of unknown primary. The 10-fold cross-validation procedure was used to compare the effectiveness of each biomarker and combination. Our findings demonstrated that the most accurate TOO tracing profiles were gene expression profiles, followed by DNA methylation, whereas somatic mutation had the lowest accuracy. In the meanwhile, we discovered that merely adding several biomarkers has no impact on increasing prediction accuracy.

Kyle Swanson, et al., (2023) Clinical oncology is using machine learning (ML) more and more to detect malignancies, forecast patient outcomes, and guide therapy decisions. Here, we examine current ML uses in clinical oncology process. Lastly, we look at ML models that regulatory bodies have approved for use with cancer patients and talk about ways to make ML more clinically helpful.

Rajpal, et al., (2021) The suggested work for the molecular classification of breast cancer subtypes identifies a small number of biomarker genes. We introduced the state-of-the-art deep learning framework Triphasic DeepBRCA for identifying biomarkers and detecting breast cancer subtypes. Additionally, the analysis of 54 biomarkers for prognosis showed that more than 30 of them are substantially associated with the prognostic outcome.

Priya Ravindrana, et al., (2021) The astounding forecast in cancers based on levels of genomic expression. Utilizing neurotic techniques, Significant progress has been made in both conclusion and explanation, and the applicability in malignancy has been proved.DNA microarray information regarding genes often has several characteristics, most of which are found to be excessive and uninformative. Meanwhile, the accuracy of factual models' determination is compromised by the small size of microarray information tests.

Das, N. Neelima, K, et al., (2024) The DSCNN model's total accuracy across all datasets is found to be 99.18% after the implementation is completed in Python. Metrics including the model's total performance is also assessed using precision, recall, and F1 score. The proposed model performs noticeably better in terms of effectiveness than existing approaches.

2. PROPOSED METHODOLOGY

Building a classification system that links gene expression data has emerged as a lively area of bioinformatics study in recent years. Generally speaking, it was created to distinguish between healthy and malignant samples in order to identify cancer in patients. The four primary stages of the suggested paradigm are data acquisition, data pre-processing, gene classification, and optimal gene selection. Five datasets are used to assess the suggested model. To increase classification accuracy, pre-processing is a crucial step. The normalizing process in this study is carried out using the Min-Max normalization approach, which maintains the relationship between the baseline data values.

Samples are presented in the microarray's dataset, leading to dimensionality problems in large datasets. As the data quantity is huge, all the features cannot be considered, resulting in dimensionality issues and degraded training ability So, a gene selection method is utilized to overcome this problem by selecting the independent genes and removing repeated or noisy genes from the dataset. Better training of valuable genes can enhance accuracy, reduce time complexity, and improve convergence in the proposed approach. The third stage is gene selection and, in this stage, the LDA optimization strategy is developed with enhanced exploration and exploitation capability. By implementing a novel optimization strategy, the most relevant features can be selected, and thus, the trainability of the classifier model can be improved. The final stage is categorized, and at this point, the XGBoost model is developed for an effective Classification of cancer. The key innovation of the proposed model lies in integrating the LDA algorithm with XGBoost model. The novelty of the proposed approach is underscored by its ability to maximize classification performance while minimizing error rates. The cooperative utilization of LDA and XGBoost yields superior results compared to existing methods like Visual Geometry Group (VGG), dense network, residual network and so on. Figure 1 depicts the general framework of the suggested work.

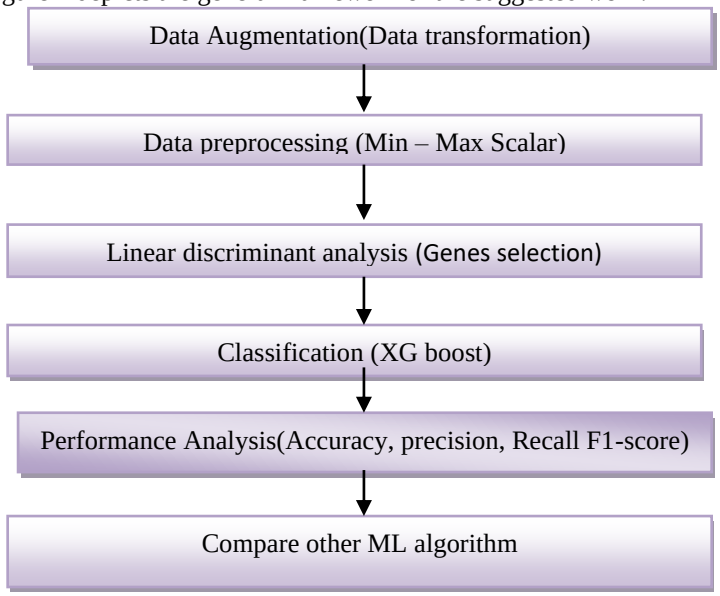

Figure 1. Proposed Block diagram

3.1 Data Pre-Processing

Data scaling is one stage in getting data ready for numerical characteristics. Many machines learning methods, including the Extreme Gradient Boosting approach, require data scaling, KNN algorithm, logistic and linear regression, etc., to yield satisfactory results.

- A feature's mean and standard deviation are calculated using the fit(data) approach so that it can be utilized for scaling.
- Scaling is done using the transform (data) method with mean and standard deviation determined using the.fit () method.
- Both fitting and transforming are accomplished using the fit transform () method.

The standard scalar is used to create a standardized distribution with a zero mean and a one standard deviation (unit variance). By deducting It standardizes features by taking the mean value of the feature and dividing the result by the feature standard deviation.
The standard scaling is calculated

$$z = (x - u)/s \quad \ldots(1)$$

Where,

- Scaled data is represented by z.
- Scaled data is x.
- U is the training samples mean.
- The training samples standard deviation is denoted by s.

3.2 Standard Scaler

This can be accomplished directly in just two to three stages with Sklearn preprocessing's support for the Standard Scaler () technique.
Class sklearn. preprocessing is the syntax. (*, copy=True, with mean=True, with_std=True); Standard Scaler
Parameters:

- copy: In-place scaling is carried out if False. Instead of inplace scaling, a copy is made.
- with_mean: Data is centered before scaling if this is true.
- with_std: Data is scaled to unit variance if true.

3.3 Feature selection

In supervised machine learning, linear discriminant analysis is one method for simplifying classification problems. Set differences are modeled using it since it necessitates differentiating between two or more groups. With this tool, a characteristic can be moved from one dimension to a space with fewer dimensions. The technique aims to maximize the ratio of within-class variance to between-class variation by converting attributes from a high-dimensional space into a lower-dimensional one, hence providing optimal class separability [31, 32]. This work primarily uses linear discriminant analysis (LDA), a well-liked preprocessing technique for machine-learning classification applications. Through the reduction of the ratio between the within-class variation and the between-class variance, it can convert characteristics into a lower dimensional space. In this article, LDA (Algorithm 1) is improved. It simply enters the independent variables of the sample to determine the prior probability and class means. We find the average vector dimensions for each group of datasets. The associated eigenvalues (1, 2, d) and the eigenvectors (e1, e2..., ed) must be used to calculate the scattering matrix. By sorting the Eigenvectors by decreasing Eigen values, you may determine which k Eigenvectors have the highest Eigen values. Consequently, a dk matrix W is created. In a new subspace, the data is then recreated using the eigenvector matrix W that was previously constructed. It can be represented in matrices as simply as Y = XW. The covariance matrices for every group are calculated, and the pooled covariance matrix is estimated. Next, it calculates

The LDA discriminant function is then calculated, and the classes are given names. This article presents an improvement on LDA (Algorithm 1) that just requires the input of the sample's independent variables to determine the class means and prior probability. We find the average vector dimensions for each group of datasets. The associated eigenvalues (1, 2..., d) and the eigenvectors (e1, e2, ed) must be used to determine the scattering matrix. To do this, sort the Eigenvectors by decreasing Eigen values to determine which k Eigenvectors have the highest Eigen values. Consequently, a dk matrix W is created. Then, using the previously produced eigenvector matrix W, the data is recreated in a new subspace. It can be represented in matrices as simply as Y = XW. The covariance matrices for every group are calculated, and the pooled covariance matrix is estimated. The LDA discriminant function is then calculated, and the classes are given names.

Algorithm 1: Analysis Linear Discriminant Algorithm

	Improved Linear Discriminant Analysis
1	$En = xU\ j\|yj = cn, j = 1, \dots, m-1, n = 1,2//$ explicit subcategories in /class
2	$\mu i = mean(Ei), i = 1,2//$class means
3	$C = (\mu 1 - \mu 2)(\mu 1 - \mu 2)U//$ among class scatter mediums
4	$Zn = En - 1mn\mu Ui, n = 1,2//$ criteria for the midway class
5	$Tn = ZUnZn, n = 1,2//$class scatter conditions
6	$T = T1 + T2//$ in−class scatter medium
7	$\lambda 1, x = eigen(T - 1C)//$calculate central eigenvector m

3.4 Classification (XG boost)

Extreme Gradient Boosting (XGBoost), a ground-breaking tree-based algorithm, has recently gained traction for data classification and has shown itself to be a highly successful technique. The XGBoost end-to-end tree boosting technique is very scalable for machine learning applications that need to classify and do regression. DenseNet201's Fully Connected Layer (FCL) has been swapped out for the XGBoost classifier. This is due to the fact that the initial FCL was based on the Image Net dataset, which included images that have nothing to do with medicine. Chen and Guestrin, who developed the method, have provided a detailed description of their approach. Because this approach is new, the next section summarizes the definitions and calculations.

First, a tree ensemble approach using regression and classification trees

When K^(i) E|i∈1…K nodes are present, CARTs The ultimate prediction output of class label y ̂i is calculated using the cumulative prediction scores at a leaf node f_k or each tree k^th, as indicated in Equation (2).

$$\hat{y}\ i\ = \varphi x_1 = \sum_{k-1}^{K} f_k\ (x_i), f_k \in \quad (2)$$

The training set is denoted by x_i, and the set containing all K scores for all CARTs is denoted by F. The findings are then improved by a regularization step, as shown by Eq. (3).

$$L(\varphi) = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega\,(f_k) \dots . (3)$$

Using l as the differentiable loss function, this is the error difference between the expected class labels (y_i)ˆ and the goal yi. The second part prevents over-fitting issues by penalizing the model complexity by Ω. Using equation (4), the function for the penalty Ω is computed.

$$\Omega(f) = \gamma^T + \frac{1}{2}\,\lambda \sum_{j=1}^{T} w^2\, j, \quad \dots . (4)$$

where γ and λ, programmable parameters, regulate the degree of regularization. Each leaf's weight values are maintained in w, even though the tree's leaves are represented by T. The classification issue and loss function are subsequently effectively solved using a second Taylor expansion and Gradient Boosting (GB). At step t, the constant term will be removed in line with Eq (5) to produce a more straightforward aim.

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f t^2(x_i)\right] + \Omega(f_t)$$
$$= \sum_{i=1}^{n}\left[g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$
$$= \sum_{j=1}^{T}\left[(\sum_{iI_j} g_i) w_j + \frac{1}{2}(\sum_{i\epsilon I_j} h_i + \lambda w_j^2\right] + \gamma T \ldots . (5)$$

The instance of leaft is represented by Ij= {i|q (x_ (i)) =j}, and equations (6)– (7) define the equation for the first gi and second hi order gradient statistics of the loss function.

$$g_i = \frac{\partial \ell(\hat{y}\, i^{(t-1)}, y_i)}{\partial\, \hat{y_i}\, i^{(t-1)2}} \ldots (6)$$

$$h_i = \frac{\partial \ell(\hat{y}\, i^{(t-1)}, y_i)}{\partial\, \hat{y_i}\, i^{(t-1)2}} \ldots (7)$$

Therefore, the optimal weight w*j of leaf j can be found using Eq. (8).

$$w *_j = \frac{\sum i\epsilon Ij\ g_i}{\sum i\epsilon Ij h_i + \lambda} \ldots . (8)$$

The function to be utilized as a scoring function to gauge the quality of a tree structure q can be computed using Eq. (9), given a tree structure q(x_(i)).

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1} T \frac{(\sum i\epsilon Ij\ g_i\ )2}{\sum i\epsilon Ij h_i + \lambda} + \gamma T \ldots . (9)$$

The split nodes are quantified by scoring the instanceset of left I_L and right I_R nodes when splitting is complete, and the loss reduction is typically computed in Eq (10).

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum i\epsilon I_L\ g_i\ )2}{\sum i\epsilon I_L h_i + \lambda} + \frac{(\sum i\epsilon I_R\ g_i\ )2}{\sum i\epsilon I_R h_i + \lambda} + \frac{(\sum i\epsilon I\ g_i\ )2}{\sum i\epsilon I h_i + \lambda}\right] - \gamma \ldots . (10)$$

Where $I = I_R \cup I_L$

3. Result and discussion

4.1 Dataset details

Five distinct datasets including those related to breast, brain, colon, ALL-AML, and prostate cancer are used to gather binary and multi-class microarray gene expression data. Details on the dataset, including the number of supplemented and considered samples, are provided in this section. Five datasets are shown in detail in Table 1, including the number of samples, classes, and enhanced data samples.

Table 1: Total Number of Genes and Selected Genes

S. No	Dataset name	Classes	Samples	Augmented data samples
1	ALL-AML	5	64	1280
2	Colon	2	62	1116
3	Brain	5	130	1950
4	Breast	6	151	1208
5	Prostate	2	102	1530

This section identifies the experimental outcomes of a proposed XGBost classification model. Through Python implementation, the anticipated model's performance is evaluated. The suggested gene data classification performance is estimated by comparing several current approaches. A batch size of 32 and an epoch size of 10 are among the hyperparameters; six hidden layers, one output layer, and a 0.001 learning rate are used with the ReLu activation function. The parts that follow give information about the dataset, performance measurements and their

mathematical formulation, analysis, and performance comparison. Details of the device configuration used to implement the suggested model are shown in Table 2.

Table 2: Hyper-parameters and Values

Hyper-parameters	Values
Scaling factor	2.5
Iteration	1500
Population size	20
Mutation rate	0.05
Alpha	0.3
Kernal size	3
Batch size	32
Maximum iteration	100
Learning rate	0.01
Epoch size	100
Loss function	Categorical cross entropy
Activation function	Softmax
Learning algorithm	Adam

A scaling factor of 2.5 is applied, and the algorithm performs 1000 iterations with a population size of 20, ensuring diversity in optimization. The selective pressure is set to 2, balancing exploration and exploitation in the evolutionary process, while a mutation rate of 0.05 introduces variability. The alpha parameter is set to 0.3, likely influencing weight adjustments or regularization. A kernel size of 3 is used for local feature extraction in the convolutional layers, and the batch size is 32, optimizing memory usage and computation. The maximum iteration count is limited to 100 for specific processes, with a learning rate of 0.01 ensuring gradual model updates. The model employs 32 convolution filters for feature extraction and is trained over 100 epochs for sufficient dataset exposure. Categorical cross-entropy is used as the loss function, which is suitable for multi-class classification, while the Softmax activation function is applied for output probability generation. The Adam optimizer is selected for efficient and adaptive training. These hyperparameters are tuned to enhance model performance and accuracy while maintaining computational efficiency.

4.2 Performance Evaluations

Performance Evaluations In the nearly equilibrated target variable classes of the data, accuracy the number of accurate forecasts the model generates across all prediction types in the categorization tasks is a valuable indicator.

$$Accuracy = \frac{TN+TN}{TP+FP+FN+TN} \ldots.(11)$$

One indicator of how accurate the predictions are accurate.

$$Precision = \frac{TP}{TP+FP} \ldots.(12)$$

A test's accuracy is assessed using the F1 score, which is the harmonic mean of precision and recall.

$$F1 = \frac{2TP}{(2TP+FP+FN)} \ldots.(13)$$

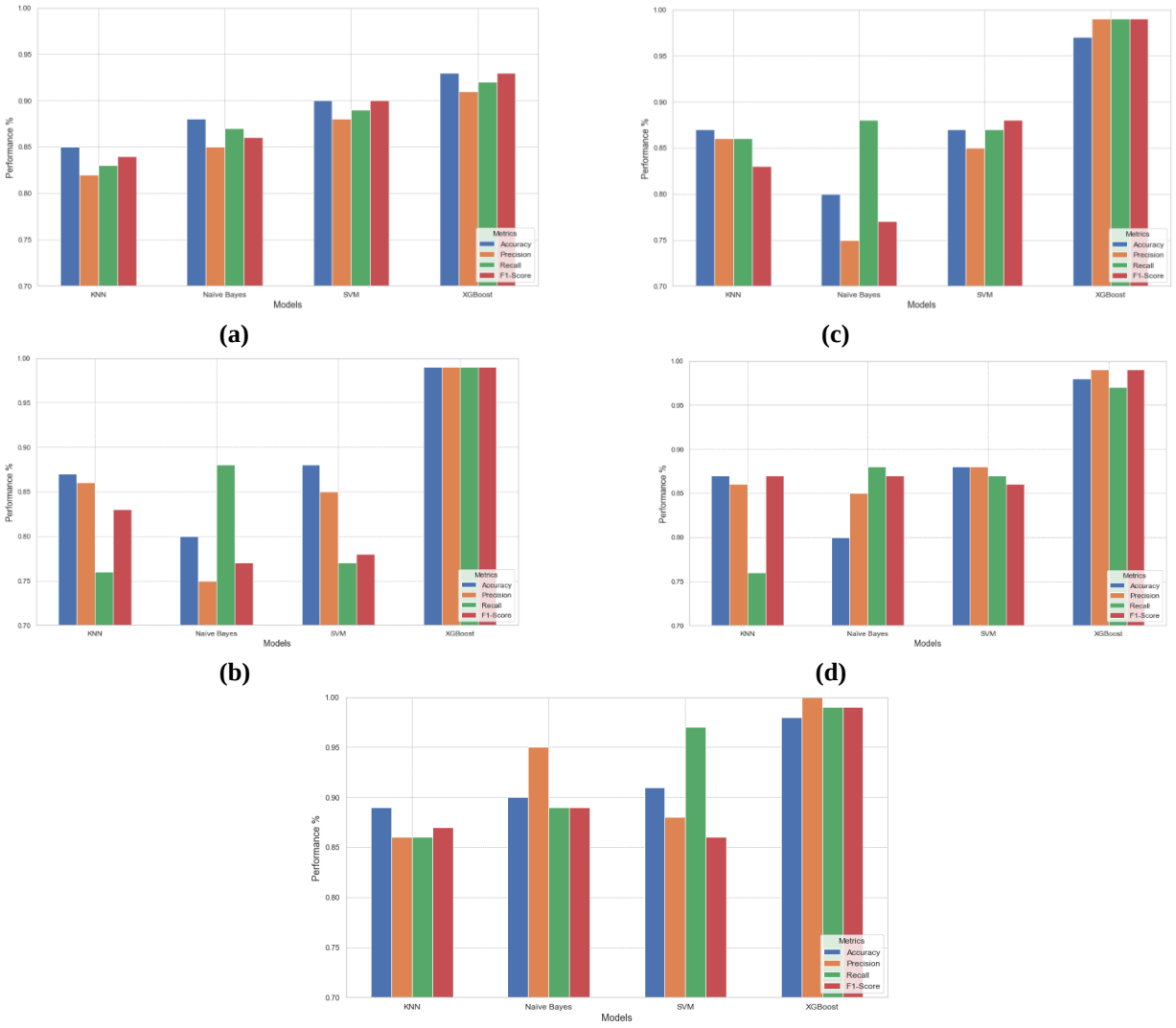

Figure 2. Performance comparison analysis (a) ALL-AML (b) Brain (c) Breast (d) Colon (e) Prostate.

The performance of four different models SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Naïve Bayes, and XGBoost was compared across key performance metrics: Classification performance is assessed by accuracy, precision, recall, and F1-score on five gene expression datasets: ALL-AML, Colon, Brain, Breast, and Prostate, which is shown in Figure 2. These outcomes clearly hint that XGBoost yields better results than the other models for the majority of the datasets in terms of classification efficiency. More precisely, it is noticed that the XGBoost model has the highest values of accuracy, precision, recall, and F1-score than the other models. For instance, when we use the Breast dataset, XGBoost has a high accuracy of 92% and an excellent performance of precision, recall, and F1-Score scores. Likewise, in the Prostate dataset, the results for XG Boost remain fair with approximately 99.37 per cent accuracy and do the best job in the identification of the positive and negative instances in case of class imbalance. SVM and KNN are still powerful overall in performance, albeit slightly lower, especially in the ALL-AML and Brain datasets, where we identified that these classifiers had an aspect of precision vs. recall imbalance. Though Naïve Bayes turned out to be the worst performer across the board, particularly with lower values of precision and recall, it suggests that it might fail to generalize positive instances in these datasets. Therefore, it identifies XGBoost as the most efficient model when it comes to classification, offering the highest accuracy, precision, recall, and F1 score. Hence, it is advantageous for gene expression data classification, in

particular for cancer detection tasks, as well as to solve other problems arising in machine learning, such as class imbalance and overfitting.

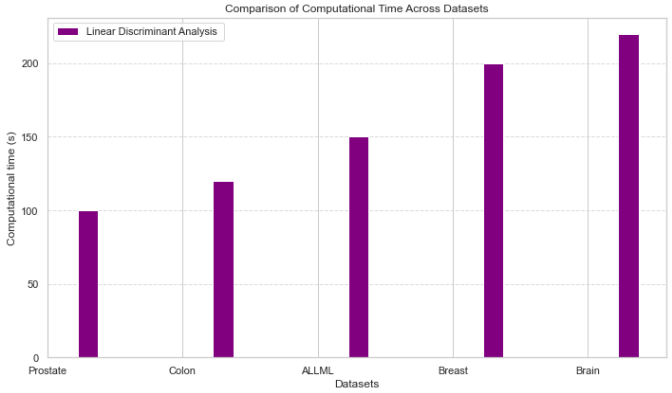

Figure 3. Computational time analysis.

The existing approaches displayed less convergence rates because they only had a poor ability of training and a higher risk of overfitting as well as having greater time complexities. It also demonstrates improved results in selecting features appropriately across 5 data sets for the proposed model's convergence performance evaluation. Figure 3 depicts the time taken by neighbouring set for the feature selection ability, high overfitting issues, and increased time complexities. The convergence performance is evaluated across five datasets, with the proposed model showing improved outcomes in selecting optimal features. Figure 3 illustrates the computational time involved in the feature selection process. The proposed model also decreases the number of used features, which also means a decrease of the dimensionality of the gene data which in turn decreases the computational load. Within this regard, algorithms for feature selection are used in order that main information could be preserved. In detail, customizing XGBoost for feature selection allows for maximum reduction of computational time. The model utilizes these algorithms to filter a number of relevant features in order to reduce the computational cost since only the informative set of features is utilized for the classification process, thus improving the pace of classification without reducing the efficacy of results. Cost of classification is the number two concern when it comes to machine learning and that is why every second counts. This paper also features the comparison of the feature selection process of the proposed model into algorithms such as XGBoost with other conventional methods KNN, SVM, and Naïve Bayes. In terms of computational time, the model performs as follows ALLML dataset require 300 seconds, Breast dataset 229 second, the Brain dataset 287 second, the Colon dataset 198 second and the Prostate dataset 201 seconds. The following results undeniably indicate that the proposed model takes far lesser time than the existing methods.

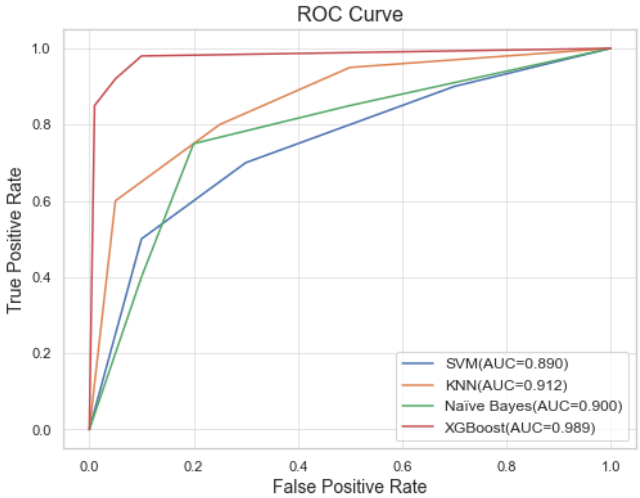

Figure 4. ROC curve analysis (a) ALL-AML (b) Brain (c) Breast (d) Colon (e) Prostate.

The ROC curve analysis for the prostate dataset compares the performance of four classification algorithms: SVM, KNN, Naïve Bayes and XGBoost. The Figure 4 shows the ROC cure for Comparing them all, XGBoost has shown the highest accuracy with an AUC of 0.989, making it the most effective algorithm for correctly classifying the positives from the negatives or, in this case, true positives with false positives. This makes XGBoost the best algorithm to use for the classification of this dataset. KNN, with an accuracy of 78% (AUC= 0.912), performs significantly similar to this model, thereby indicating that it offers a good measure of reliability as a substitute model to XGBoost. Naïve Bayes has an AUC of 0.900, which may be considered reasonable classification performance but seems a bit poorer than KNN and XGBoost. SVM performs poorly, with the lowest AUC score of 0.890 in this regard. All in all, XGBoost is the best solution in terms of classification accuracy, with KNN as an equivalent. Naïve Bayes and SVM may have to be tweaked in order to be more effective for this dataset.

4. CONCLUSION

Death and advanced stages of cancer might result from a failure to diagnose the disease in time. Despite the fact that many techniques have been used to classify gene cancer data, there are still certain disadvantages, such as poor feature representation, reduced robustness, decreased feature learning capacity, increased time consumption, decreased accuracy, and elevated error rates. In the suggested research project, precise gene expression data categorization is achievable with effective techniques. Five gene cancer datasets are collected, and the amount of data is then increased through augmentation. Utilizing min-max normalization, the data is arranged in a pertinent manner. By choosing key characteristics responsible for accurate classification, LDA significantly reduces the dimensionality of the data. Ultimately, the XGBoost classification model successfully classifies the chosen genes. Python is used to evaluate the performances, and the suggested XGBoost model's total accuracy across five datasets is 99.37%. Drug exposures and the identification of illness molecular signatures are two examples of applications for the suggested paradigm. The suggested results are somewhat better than the current models, and further datasets will be considered for performance analysis in the future. In order to further increase the categorization accuracy, more valuable features will also be considered.

REFERENCES

[1] Mahmood khalsan, Lee r. Machado, Eman salih al – shamrey, suraj ajit, Karen anthony mu, and michael opoku agyeman, “A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction”, IEEE ACCESS, Vol.10,2022.

[2] Khosro Rezaee, Gwanggil Jeon, Mohammad R. Khosravi, Hani H. Attar, Alireza Sabzevari, “Deep learning-based microarray cancer classification and ensemble gene selection approach”, IET Syst. Biol, vol.16, 2022.

[3] Bingsheng He, Chan Dai, Jidong Lang, Pingping Bing, Geng Tian, Bo Wang, “Jialiang Yang, A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation”, Elsevier, Vol no. 1866 ,2020.

[4] Tarek Khorshed, Mohamed N. Mosutafa, and Ahmed rafea**, “**Deep Learning for Multi-Tissue Cancer Classification of Gene Expressions (GeneXNet)”, IEEE Access, VOLUME 8, 2020.

[5] Alejandro Lopez-Rincon, Lucero Mendoza-Maldonado, Marlet Martinez-Archundia, Alexander Schonhuth, Aletta D. Kraneveld, Johan Garssen, and Alberto Tond, “Machine Learning-Based Ensemble Recursive Feature Selection of Circulating miRNAs for Cancer Tumor Classification”, Cancers, vol. 12, 2020.

[6] Gumaei A, Sammouda R, Al-Rakhami M, AlSalman H, El-Zaart, “A. Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression”, Health Informatics Journal; vol.27, no.1,2021.

[7] E. H. Houssein, D. S. Abdelminaam, H. N. Hassan, M. M. Al-Sayed and E. Nabil, “A Hybrid Barnacles Mating Optimizer Algorithm with Support Vector Machines for Gene Selection of Microarray Cancer Classification”, IEEEAccess, vol. 9, 2021.

[8] Hajieskandar, A., Mohammadzadeh, J, Khalilian, M. et al, “Molecular cancer classification method on microarrays gene expression data using hybrid deep neural network and grey wolf algorithm”, J Ambient Intell Human Comput, vol. 14 ,2023.

[9] Hussam Jasim Ali, Sameer Alani, Riyadh Jameel Toama, Tabarak Ali Abdulhussein, “Evolutionary Hybrid Machine Learning Techniques for DNA Cancer Data Classification”, IIETA, Vol. 38, No. 2, April, 2024.

[10] Muhammed Abd-Elnaby , Marco Alfonse , Mohamed Roushdy, “Classification of breast cancer using microarray gene expression data: A survey”,Journal of Biomedica Informatics,Volume 117, 2021.

[11] Haiyan Liu, Chun Qiu, Bo Wang, Pingping Bing, Geng Tian, Xueliang Zhang, Jun Ma, Bingsheng He1 and Jialiang Yang, “Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-of-Origin”, Frontiers in Cell and Developmental Biology, Volume 9,2021.

[12] Kyle Swanson ,Eric Wu , Angela Zhang, “From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment”,Volume 186, Issue 8, 2023.

[13] S. Rajpal, M. Agarwal, V. Kumar, A. Gupta and N. Kumar, "Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification," in IEEEAccess, vol. 9, 2021.

[14] Priya Ravindrana,1, S. Jayanthia, Arun Kumar Sivaramanb, Dhanalakshm, “Proficient Mining of Informative Gene from Microarray Gene Expression Dataset Using Machine Intelligence”, Smart Intelligent Computing and Communication Technology,2021.

[15] A. Das, N. Neelima, K. Deepa and T. ozer, "Gene Selection Based Cancer Classification with Adaptive Optimization Using Deep Learning Architecture," in IEEEAccess, vol. 12, 2024.