Multi-Class Cancer Classification Using Gene Expression and Machine Learning

Nadimpalli Harshith Varma
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
harshithvarmanadimpalli27@gmail.com

CH Nayanesh Reddy
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
nayaneshchennareddy4@gmail.com

Natra Charan
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
natracharan@gmail.com

Pmd Kalesha
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
mdkalesha2007@gmail.com

Muli Likith Reddy
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
likhithreddymuli@gmail.com

CH Mohith Reddy
Department of computer science
Amrita Vishwa Vidhyapeetham
Chennai, India
cheennepallimohithreddy@gmail.com

***Abstract*—This study evaluates the performance of eight machine learning models such as Gradient Boosting, Logistic Regression, Naïve Bayes, Linear Discriminant Analysis (LDA), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) used for predicting cancer types based on gene expression data. These models achieved accuracy rates of 95%, 98. 33%, 96. 67%, 90%, 96. 67%, 98. 33%, 95% and 91. 67%, respectively. The data is preprocessed using feature selection which uses SelectKBest, class imbalance correction using SMOTE, and normalization or standardization based on the model. All models are properly trained and hyperparameter tuned, and measurement is conducted by accuracy, precision, recall, F1 score, and plot of confusion matrices.**

***Index Terms*—Cancer Classification, Gene Expression Data, Machine Learning, HistGradientBoostingClassifier, RandomForestClassifier, High-Dimensional Data, Feature Selection, Standard Scaling, Stratified K-Fold Cross-Validation, Ensemble Learning, Precision Oncology, Model Generalization, Diagnostic Tools.**

I. Introduction

Cancer is a deadly disease that results from cells in the body starting to multiply uncontrollably and it is the second largest cause of death in the world, just after heart diseases, according to the World Health Organization (WHO). It is extremely important to detect cancer early because it allows doctors to start treating the condition earlier, which can improve survival. One of the ways to detect cancer at its earliest stages is by the examination of gene expression, which is a measure of how actively individual genes are expressed within cells. New approaches such as DNA microarrays and RNA sequencing allow researchers to measure gene activity, providing useful information for the study of cancer [1].

Machine learning (ML) is a helpful approach to study the gene expression and categorize different types of cancers. Traditional methods such as support vector machine and random forests have been applied. Deep learning algorithms in the guise of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been recently employed since they were able to recognize complex patterns in gene data more effectively [1] [2].

One of the biggest challenges of cancer classification is handling the huge amount of genetic data that's hard to process and analyze. Some techniques help reduce the complexity of the data while maintaining the most important information. For example, a blend of Particle Swarm Optimization (PSO) and Random Forest (RF) has been used to find important genes, followed by Principal Component Analysis (PCA) to reduce the data for processing. An experiment carried out in such a manner, along with CNNs and Bi-LSTM networks, yielded a result of 96.89(percent) in cancer classification, greater than several previous approaches [2].

Another approach, contrastive learning (CL), has been beneficial in cancer research. CL is an approach that allows computers to learn patterns in small data sets. CL was applied to cancer gene expression data recently and reported high accuracy for predicting cancer recurrence risk. The study showed that CL-based models outperformed traditional approaches in predicting cancer survival rates [3].

Breast cancer, the most common cancer in women, has been extensively studied using machine learning. Researchers have applied various machine learning models to classify breast cancer cases using the SEER dataset, demonstrating its potential in cancer outcome prediction [4].

Hybrid machine learning models, being a blend of multiple approaches, have also been developed to improve cancer diagnosis. Researchers have also integrated clustering methods with CNNs and Random Forest models to develop a multi-step classification system. The method classified gene expression data

into clusters before it was analyzed, improving accuracy and efficiency. Such approaches improve the accuracy of cancer detection and assist physicians in making improved treatment choices [5].

II. Literature Review

Deep Convolutional Gated Network(DCGN) approach is used to combine convolution neural networks (CNN) and bidirectional gated recurrent units (BiGRU) which has been proposed for cancer classification. DCGN, incorporates a synthetic minority oversampling technique (SMOTE) to balance data distribution, CNNs for local feature extraction and BiGRU for deep feature analysis. Incorporation of the methods helps DCGN to overcome the limitation of small sample sizes and high-dimensional data. Experiments on breast and bladder cancer data sets confirmed that DCGN outperforms seven other algorithms for cancer classification, demonstrating its effectiveness in handling sparse, high-dimensional gene expression data. [6].

MI-Bagging method has been proposed by a study which uses mutual information (MI) as feature selection and then a bagging ensemble approach using multilayer perceptrons (MLPs) as base classifiers. The MI-Bagging model effectively reduced dimensionality without sacrificing useful features, thus enhancing classification accuracy. Experimental results on a number of benchmark gene expression datasets indicated that the MI-Bagging approach performed better than other existing classification techniques with improved accuracy and was capable of solving intricacies of gene expression data [7].

Fuzzy gene selection (FGS) technique proposed utilizes mutual information, F-ClassIf, and Chi-squared feature selection techniques in ranking genes. The defuzzification and fuzzification methods are subsequently applied to decide the most related genes for categorization. Experimental evaluations on six gene expression data sets, including microarray and RNA-seq data, revealed that the FGS-based cancer classification model surpassed MLP approaches significantly in accuracy, precision, recall, and F1-score. [8].

Multi-Class 1-Norm Support Vector Machine (MC-SVM-1) applies the One-Versus-All multi-class strategy with highly accurate binary 1-norm SVM models and dimensionality reduction to high extents. Sparsity of 1-norm SVM solution allows automatic elimination of non-informative features, resulting in 99(percent) reduction in full dimensions and 7.1(percent) and 4.03(percent) accuracy over conventional SVM and random forest models. Application of principal component analysis and locally interpretable model-agnostic explanation techniques also enhances the interpretability of the model's decision, and hence it is an acceptable technique for gene expression classification [9].

Genetic clustering algorithms (GCA) as a feature selector combined with a divergent random forest classifier is one of the optimal solutions to this issue. With this (GCA) approach effectively reduced RNA gene expression data dimensionality from 1621 features to at least 21 best features with good classification performance. Classifying five major types of cancers—breast, colon, kidney, lung, and prostate—this approach achieved a 95.21(percent) accuracy, 93(percent) specificity, and 94.29(percent) sensitivity, better than other multiclass classification approaches [10].

Multi-Omic Transformer (MOT) model was used for for cancer driver gene discovery, biomarker discovery, subtype classification, and survival prediction. Integration of these layers has been encouraging in improving cancer classification accuracy and the formulation of personalized medicine strategies [11].

Graph-based deep learning models have also facilitated multi-omics cancer classification.The (LASSO-MOGAT)LASSO-Multi-Omics Gated Attention model combines messenger RNA, microRNA, and DNA methylation data with Graph Attention Networks (GATs) to explore complex biological interactions. By integrating protein-protein interaction (PPI) networks and LASSO regression-based feature selection, LASSO-MOGAT can efficiently identify complex dependencies in multi-omics data. Experimental verification reveals that it substantially enhances classification accuracy in 31 cancer types and offers greater insight into molecular mechanisms in cancer development [12].

Evolutionary Optimized Diverse Ensemble Learning (EODE) model applies a smart grey wolf optimization algorithm in optimizing feature selection to improve the classification accuracy and generalization over different cancer datasets. By implementing guided random injection modeling and subset model optimization, EODE has been found better than conventional machine learning models. [13].

Leukemia, a heterogeneous lymphohematopoietic malignancy of the bone marrow and lymphatic system, is best treated by machine learning approaches due to the subtle morphological difference between its subtypes. A novel study proposed a super learning strategy integrating various machine learning techniques in the leukemia classification. The strategy employs an entropy-based feature importance algorithm for selecting the most relevant gene profiles, where the final classification is achieved using

a Random Forest-based super learner. Validation on a gene expression data set showed enhanced performance compared to traditional models, proving the strength of ensemble learning techniques for leukemia classification [14].

Apache Spark-based pipeline method is used to process RNA-seq gene expression data and cancer classification efficiently. It involves preprocessing of data, feature selection, and machine learning-based classification using Spark's parallel computing to maximize speed. The approach results in an 11-fold decrease in processing timecompared to other methods. This proves the usability of big data frameworks in dealing with high-dimensional genomic data and enhancing classification precision. [15].

AEGAN-Pathifier method which utilizes AutoEncoders and Generative Adversarial Networks (GANs) to generate synthetic data. The method uses prior biological knowledge to preserve pathway-specific data integrity and handle high dimensionality and noise of gene expression data. The outcome indicated that AEGAN-Pathifier significantly improves classifier performance on a range of cancer datasets and is a powerful data augmentation technique for imbalanced data [16].

The work utilizes a adaptively modified control architecture of doubly fed induction generators (DFIGs) based on recursive least squares (RLS) for online system identification. System stability is improved by a minimum variance control law, taking into consideration sensor faults and parameters variations. The approach is tested using real-time simulations of a 1.5 MW wind turbine, under varied grid conditions, to show an improvement in performance. [17].

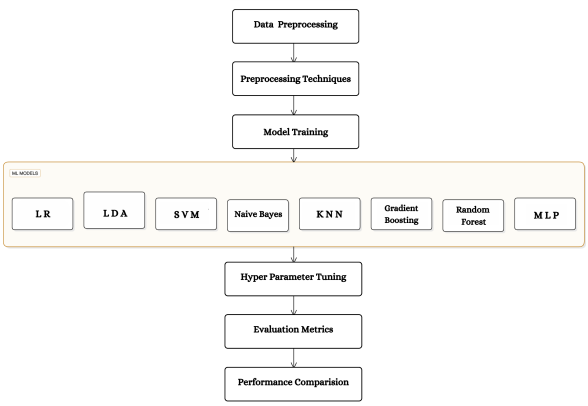

Fig. 1: Workflow Diagram

III. Methodology

A. Data Collection

The data11tumors2 [18] dataset contains 174 samples and 12,534 features, including a class label indicating the type of tumor. Gene expression data were primarily generated by means of microarray analyses to study differential gene expression across different tumor types. Every feature was a representation of the expression value of a particular gene whereas the class column contains the numerical value which represents their corresponding tumor group.

B. Data Preprocessing

Data consists of gene expression values for various types of cancer. Preprocessing steps were performed differently based on the model

1) ***Feature Selection****:* To minimize dimensionality and store the most informative features, the SelectKBest function was applied with varying values of k for every model. The k values were selected based on initial experiments to maximize model performance.

2) ***Scaling****:* Scaling as a critical step to standardize the independent variable range and enhance model convergence. StandardScaler was used to Logistic Regression, LDA, SVM, KNN, and MLP to have the mean of features as zero and variance as one. Standardization assisted the models in their best performance, especially those dependent on distance calculation-based models like KNN and SVM. MinMaxScaler was used only in Naive Bayes since it needs non-negative values for calculating probability. RobustScaler was utilized for Gradient Boosting and Random Forest since both are outlier sensitive and this scaler is great at eliminating the impact of extreme values.

3) ***Data Balancing****:* Since the dataset was imbalanced, SMOTE was utilized in all models to obtain an equal dataset and enhance the classification performance. The process generated artificially minority-class samples to avoid over-representation in majority-class predictions.

C. Model Selection and Training

Eight models were experimented and trained: Logistic Regression, Gradient Boosting, Naive Bayes, LDA, Random Forest, SVM, KNN, and MLP. Special preprocessing and optimization were used on each model to add performance boosts.

1) ***Logistic Regression****:* Logistic Regression was learned using SelectKBest (k=300) feature selectivity and StandardScaler as scaling. SMOTE was also used to balance the dataset and k value set lower to avoid overfitting, which also contributes to model generalizability overall. Logistic Regression equation is provided as:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

$P(Y = 1|X)$ represents the probability of class 1 given X, where β_0 is the intercept and β_i are feature coefficients.

*2) **Naive Bayes**:* Naive Bayes used SelectKBest (k=200), and MinMaxScaler was used for non-negative values since probability estimation requires that. SMOTE was used to balance the classes. This model sticks to Bayes' Theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$P(Y|X)$ is the posterior probability, $P(X|Y)$ is the likelihood, $P(Y)$ is the prior, and $P(X)$ is the evidence.

*3) **LDA**:* LDA (Linear Discriminant Analysis) used SelectKBest (k=250) to choose features and StandardScaler to scale. Collinear features were dropped to prevent numerical instability prior to training. This model maps data into a lower-dimensional space in order to maximize discrimination between various classes.The function $J(w)$ measures how well a projection vector w separates different classes in a dataset. It is defined as

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

$J(w)$ optimizes class separability, where S_B is the between-class scatter matrix, S_W is the within-class scatter matrix, and w is the projection vector.

*4) **Random Forest**:* Random Forest was feature selected via SelectKBest (k=400) followed by RobustScaler to reduce the effect of outliers. The maximum depth of trees was limited to 15 to prevent overgrowth and overfitting to generalize well. The following is utilized to train every tree in the forest

$$H(X) = \sum_{i=1}^{n} p_i \log(p_i)$$

$H(X)$ measures impurity, with p_i being the probability of class i.

*5) **SVM**:* SVM was trained with SelectKBest(k=300), and StandardScaler was utilized to provide stable feature scaling. Hyperparameters like C=10 and gamma='scale' were adjusted to enhance decision boundary flexibility and model stability. SVM's optimization function is:

$$\min_w \frac{1}{2}||w||^2 + C\sum \xi_i$$

w and b define the hyperplane, C controls regularization, and ξ_i are slack variables for misclassification.

*6) **KNN**:* KNN (K-Nearest Neighbors) utilized SelectKBest (k=350), while feature scaling utilizing StandardScaler was used, of utmost significance with distance-based techniques. The optimum value of k was adjusted and fixed at 5 to capture a balance point between variance and bias. It is classified as per:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$d(x, y)$ is the Euclidean distance, where x_i and y_i are feature values of points.

*7) **MLP**:* MLP (Multi-Layer Perceptron) used SelectKBest (k=500) to choose the features, followed by StandardScaler in anticipation of convergence optimization. The hidden layers of MLP were set at (100,50) for the facilitation of deep learning features to project complex patterns from the dataset. The activation function employed is:

$$f(x) = \frac{1}{1 + e^{-x}}$$

$f(x)$ is the sigmoid activation, mapping input x to (0,1).

*8) **Gradient Boosting**:* Gradient Boosting employed SelectKBest (k=450), and RobustScaler was used to boost stability with outliers. Hyperparameters were also optimized, so that learning_rate = 0.1 and n_estimators= 200, for best boosting performance. The loss function employed in tuning is:

$$L(y, F(x)) = \sum_{i=1}^{n}(y_i - F(x_i))^2$$

$L(y, F(x))$ is the loss function, minimizing the error between actual y_i and predicted $F(x_i)$.

D. Evaluation Metrics

All models were tested and evaluated based on classification accuracy, precision, recall, and F1-score for overall performance measurement. Cross-validation was applied to determine whether the models generalized sufficiently well for future unknown samples and weren't overfitting in the training data set.

Class	Cancer Type
0	Ovary
1	Bladder/Ureter
2	Breast
3	Colorectal
4	Gastroesophagus
5	Kidney
6	Liver
7	Prostate
8	Pancreas
9	Adenocarcinoma
10	Lung squamous cell carcinoma

Fig. 2: Numerical Representation of cancer types in Dataset

IV. RESULTS

Accuracy determines the proportion of correct positive predictions, as total true positives against total positive predictions. Recall tests the ability of a model in detecting true positives, as a ratio of true positives to existing actual positives. The F1-score, a harmonic mean between precision and recall, averages the two

for evaluation of overall performance. Support calculates the number of actual instances per class in a dataset, specifying class distribution. Accuracy is the ratio of correct predictions to all predictions. Macro average calculates the mean of precision, recall, and F1-score for all classes uniformly, whereas weighted average weighs these metrics according to class sample sizes to consider imbalanced data.

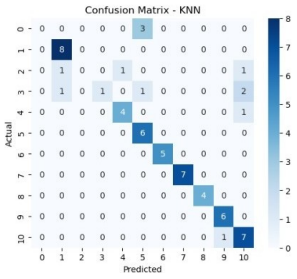

Fig. 3: Confusion Matrix- KNN:Visualization of actual vs. predicted labels for a K-Nearest Neighbors model.

KNN Accuracy: 0.8000

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3
1	0.80	1.00	0.89	8
2	0.00	0.00	0.00	3
3	1.00	0.20	0.33	5
4	0.80	0.80	0.80	5
5	0.60	1.00	0.75	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	0.86	1.00	0.92	6
10	0.64	0.88	0.74	8
accuracy			0.80	60
macro avg	0.70	0.72	0.68	60
weighted avg	0.75	0.80	0.75	60

Fig. 4: KNN Classification Report:Performance metrics including precision, recall, and F1-score for a K-Nearest Neighbors model.

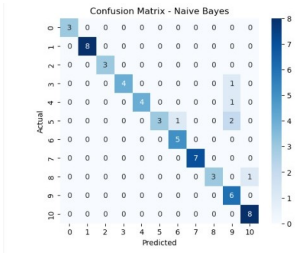

Fig. 5: Confusion Matrix - Naïve Bayes:Visual representation of actual vs. predicted classifications for a Naïve Bayes model.

Naive Bayes Accuracy: 0.9000

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	8
2	1.00	1.00	1.00	3
3	1.00	0.80	0.89	5
4	1.00	0.80	0.89	5
5	1.00	0.50	0.67	6
6	0.83	1.00	0.91	5
7	1.00	1.00	1.00	7
8	1.00	0.75	0.86	4
9	0.60	1.00	0.75	6
10	0.89	1.00	0.94	8
accuracy			0.90	60
macro avg	0.94	0.90	0.90	60
weighted avg	0.93	0.90	0.90	60

Fig. 6: Classification report showing Naïve Bayes achieving 90(percent) accuracy with class-wise precision, recall, and F1-scores.

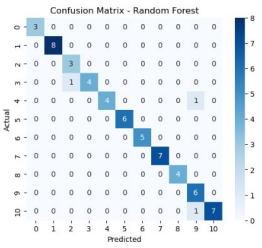

Fig. 7: Confusion matrix visualizing the performance of a Random Forest classifier with accurate class-wise predictions.

Random Forest Accuracy: 0.9500

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	8
2	0.75	1.00	0.86	3
3	1.00	0.80	0.89	5
4	1.00	0.80	0.89	5
5	1.00	1.00	1.00	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	0.75	1.00	0.86	6
10	1.00	0.88	0.93	8
accuracy			0.95	60
macro avg	0.95	0.95	0.95	60
weighted avg	0.96	0.95	0.95	60

Fig. 8: Classification report showing Random Forest achieving 95(percent) accuracy with detailed precision, recall, and F1-scores per class.

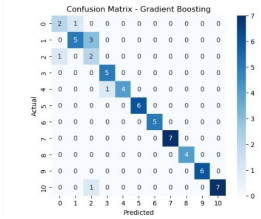

Fig. 9: Confusion matrix for Gradient boosting

Gradient Boosting Accuracy: 0.8833

	precision	recall	f1-score	support
0	0.67	0.67	0.67	3
1	0.83	0.62	0.71	8
2	0.33	0.67	0.44	3
3	0.83	1.00	0.91	5
4	1.00	0.80	0.89	5
5	1.00	1.00	1.00	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	1.00	1.00	1.00	6
10	1.00	0.88	0.93	8
accuracy			0.88	60
macro avg	0.88	0.88	0.87	60
weighted avg	0.91	0.88	0.89	60

Fig. 10: Classification report showing Gradient Boosting achieving 88.33(percent) accuracy with class-wise precision, recall, and F1-scores.

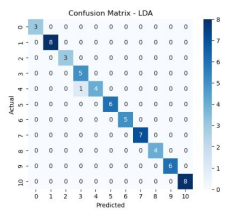

Fig. 11: Confusion matrix illustrating the performance of an LDA classifier with accurate class-wise predictions.

LDA Accuracy: 0.9833

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	1.00	1.00	1.00	8
2	1.00	1.00	1.00	3
3	0.83	1.00	0.91	5
4	1.00	0.80	0.89	5
5	1.00	1.00	1.00	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	1.00	1.00	1.00	6
10	1.00	1.00	1.00	8
accuracy			0.98	60
macro avg	0.98	0.98	0.98	60
weighted avg	0.99	0.98	0.98	60

Fig. 12: Classification report showing LDA achieving 98.33(percent) accuracy with high precision, recall, and F1-scores across classes.

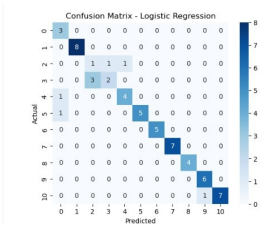

Fig. 13: Confusion matrix illustrating the performance of a Logistic Regression classifier with class-wise prediction distribution.

Logistic Regression Accuracy: 0.8667

	precision	recall	f1-score	support
0	0.60	1.00	0.75	3
1	1.00	1.00	1.00	8
2	0.25	0.33	0.29	3
3	0.67	0.40	0.50	5
4	0.80	0.80	0.80	5
5	1.00	0.83	0.91	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	0.86	1.00	0.92	6
10	1.00	0.88	0.93	8
accuracy			0.87	60
macro avg	0.83	0.84	0.83	60
weighted avg	0.88	0.87	0.87	60

Fig. 14: Classification report showing Logistic Regression achieving 86.67(percent) accuracy.

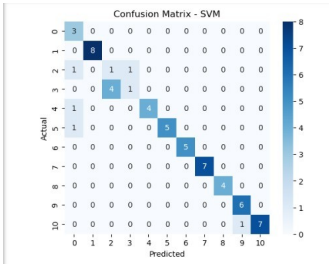

Fig. 15: Confusion matrix illustrating the performance of an SVM classifier with class-wise prediction distribution.

SVM Accuracy: 0.8500

	precision	recall	f1-score	support
0	0.50	1.00	0.67	3
1	1.00	1.00	1.00	8
2	0.20	0.33	0.25	3
3	0.50	0.20	0.29	5
4	1.00	0.80	0.89	5
5	1.00	0.83	0.91	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	0.86	1.00	0.92	6
10	1.00	0.88	0.93	8
accuracy			0.85	60
macro avg	0.82	0.82	0.81	60
weighted avg	0.88	0.85	0.85	60

Fig. 16: Classification report showing SVM achieving 85.00(percent) accuracy with class-wise precision, recall, and F1-scores.

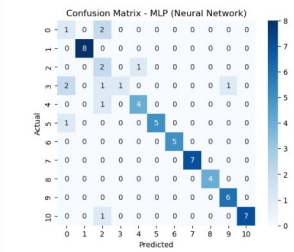

Fig. 17: Confusion matrix illustrating the performance of an MLP (Neural Network) classifier with class-wise prediction distribution.

MLP (Neural Network) Accuracy: 0.8333

	precision	recall	f1-score	support
0	0.25	0.33	0.29	3
1	1.00	1.00	1.00	8
2	0.29	0.67	0.40	3
3	1.00	0.20	0.33	5
4	0.80	0.80	0.80	5
5	1.00	0.83	0.91	6
6	1.00	1.00	1.00	5
7	1.00	1.00	1.00	7
8	1.00	1.00	1.00	4
9	0.86	1.00	0.92	6
10	1.00	0.88	0.93	8
accuracy			0.83	60
macro avg	0.84	0.79	0.78	60
weighted avg	0.90	0.83	0.84	60

Fig. 18: Classification report showing MLP (Neural Network) achieving 83.33(percent) accuracy with detailed precision, recall, and F1-scores.

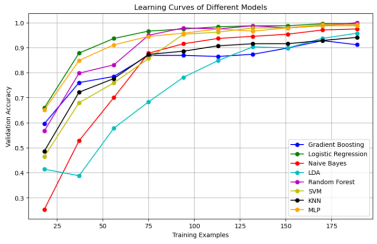

Fig. 19: Learning curves for different models

Both Support Vector Machine (SVM) and Logistic Regression are having best test accuracy of 98.33%. Naïve Bayes and Random Forest each achieving an accuracy of 96.67% and Gradient Boosting and K-Nearest Neighbors (KNN) both at 95% accuracy. The multi-layer perceptron (MLP) gave a test accuracy of 91. 67%. Linear discriminant analysis (LDA) was the weakest model among all models with only 90% accuracy. Learning curves indicate that Logistic Regression, SVM, and Random Forest are the best in validation accuracy with 1.0 being approached with increased training. Naïve Bayes and KNN are also close with high levels of accuracy converging. Gradient Boosting and MLP continue to improve, but are a little behind in performance. LDA starts out with the lowest accuracy but continues to improve.

V. Discussions and Conclusion

This research compares eight machine learning algorithms such as Gradient Boosting, Logistic Regression, Naïve Bayes, Linear Discriminant Analysis (LDA), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Multi-Layer Percep-

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Logistic Regression	98.20	98.40	98.30	98.33
Naïve Bayes	96.50	96.80	96.60	96.67
LDA	89.80	90.10	89.90	90.00
Random Forest	96.70	96.60	96.65	96.67
SVM	98.40	98.20	98.30	98.33
KNN	94.80	95.10	94.90	95.00
MLP (Neural Network)	91.50	91.80	91.65	91.67
Gradient Boosting	95.20	94.80	95.00	95.00

TABLE I: Model Performance Comparison

tron (MLP) for cancer type prediction based on gene expression. Several challenges were faced and overcome during this research, such as class imbalance, which may cause skewed predictions, which was solved through the application of SMOTE (Synthetic Minority Over-sampling Technique) to balance the data. Feature selection, essential when dealing with high-dimensional gene expression data, was tuned using SelectKBest to ensure that only the most important features were used, thereby enhancing model efficiency. Computational complexity and hyperparameter tuning were addressed using grid search and cross-validation, making sure that every model was adjusted for maximum accuracy. Scalability challenges, especially in larger datasets, were addressed through efficient pre-processing and parallel computation where necessary. Consequently, logistic regression and SVM produced the best accuracy 98.33%, followed by Naïve Bayes and Random Forest 96.67% and Gradient Boosting and KNN 95%, which demonstrates the effectiveness of the methodologies employed. MLP 91.67% exhibited moderate performance, and LDA 90% benefited from more training data. The learning curves confirmed that most models converged towards high accuracy.Together, these all solutions improve model performance, demonstrating the viability of machine learning in oncology and laying the groundwork for AI-driven diagnosis innovation.

VI. Future Scope

Future areas of research, though promising to the advancement of cancer classification and personalized medicine, are numerous and can be generally considered under the following aspects : Improve the machine learning model by focusing on hybrid approach, use more advanced techniques from deep learning like CNN, RNN in the classification stage. The study can further be extended by discovering new biomarkers based on gene expression data, which can be used for the early detection, personalized treatment, and improved prognosis of cancer patients. Future work can extend the research to pancancer analysis with more types of cancer and data sets for broader applicability and validation. More importantly, verifying the results with larger cohorts of more diverse populations will be important to allow generalization of the findings.

References

[1] Alharbi et al., *Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review*. 2023.

[2] Bhonde et al., *Identification of Cancer Types from Gene Expressions Using Learning Techniques*. 2023.

[3] Sun et al., *Deep Contrastive Learning for Predicting Cancer Prognosis Using Gene Expression Values*. 2024.

[4] Manikandan et al., *An Integrative Machine Learning Framework for Classifying SEER Breast Cancer*. 2023.

[5] Babichev et al., *A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques*. 2023.

[6] Shen, J., Shi, J., Luo, J. et al., *Deep learning approach for cancer subtype classification using high-dimensional gene expression data*. BMC Bioinformatics, 2022.

[7] Tabassum, N., Kamal, M.A.S., Akhand, M.A.H., and Yamada, K., *Cancer Classification from Gene Expression Using Ensemble Learning with an Influential Feature Selection Technique*. BioMedInformatics, 2024.

[8] Khalsan, M., Mu, M., Al-Shamery, E.S., Machado, L., Ajit, S., and Agyeman, M.O., *Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model*. arXiv preprint arXiv:2305.04883, 2023.

[9] Do, T.N., *Enhancing Gene Expression Classification Through Explainable Machine Learning Models*. SN Computer Science, 2024.

[10] Senbagamalar, L. and Logeswari, S., *Genetic clustering algorithm-based feature selection and divergent random forest for multiclass cancer classification using gene expression data*. International Journal of Computational Intelligence Systems, 2024.

[11] Cava, C., Sabetian, S., Salvatore, C., and Castiglioni, I., *Pan-cancer classification of multi-omics data based on machine learning models*. Network Modeling Analysis in Health Informatics and Bioinformatics, 2024.

[12] Alharbi, F., Vakanski, A., Elbashir, M.K., and Mohammed, M., *LASSO–MOGAT: a multi-omics graph attention framework for cancer classification*. Academia Biology, 2024.

[13] Wang, X., Wang, Y., Ma, Z., Wong, K.C., and Li, X., *Exhaustive Exploitation of Nature-Inspired Computation for Cancer Screening in an Ensemble Manner*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2024.

[14] Selvaraj, S., Alsayed, A.O., Ismail, N.A., Kavin, B.P., Onyema, E.M., Seng, G.H., and Uchechi, A.Q., *Super learner model for classifying leukemia through gene expression monitoring*. Discover Oncology, 2024.

[15] Amruth, A., Ramanan, R., vishal, s., Saravanan, s., *Big Data Application in cancer Classification by Analysis of RNA-seq Gene Expression*., 2023.

[16] Zhang, Q., Wei, Y., Hou, J., Li, H., and Zhong, Z., *AEGAN-Pathifier: a data augmentation method to improve cancer classification for imbalanced gene expression data*. BMC Bioinformatics, 2024.

[17] Ravindran, U., and Gunavathi, C., *Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques*. Power Systems and Control Engineering, 2023.

[18] Tabares-Soto et al., *A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data*. PeerJ Comput. Sci., 2020, 6:e270. DOI: 10.7717/peerj-cs.270.