Analyzing RNA-Seq Gene Expression Data for Cancer Classification Through ML Approach

Abdul Wahid, M Tariq Banday
Department of Electronics & Inst. Technology
University of Kashmir, Srinagar, India

***Abstract*—*Purpose*: Ribonucleic Acid Sequencing (RNA-Seq) is a technique that allows an efficient genome-wide analysis of gene expressions. Such analysis is a strategy for identifying hidden patterns in data, and those related to cancer-specific biomarkers. Prior analyses without samples of different cancer kinds used RNA-Seq data from the same type of cancer as the positive and negative samples. Therefore, different cancer types must be evaluated to uncover differentially expressed genes and perform multiple cancer classifications. *Problem*: Since gene expression reflects both the genetic make-up of an organism and the biochemical activities occurring in tissue and cells, it can be crucial in the early identification of cancer. The aim of this study is to classify the RNA-Sequence data into five different cancer forms, such as LUAD, BRCA, KIRC, LUSC, and UCEC, through an ensemble approach of machine learning algorithms. RNA-Seq data for five different cancer types from the UCI Machine Learning Repository are examined in this research. *Methods*: As a first step, the relevant features of RNA-Seq are extricated using Principal Component Analysis (PCA). Then, the extricated features are given to the ensemble of machine learning classifiers to classify the type of cancer. The ensemble of classifiers is built using Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN). *Results*: The results demonstrated that the proposed ensemble classifier outperformed the existing machine-learning approaches with an accuracy of 99.59%.**

Keywords*—*RNA-Sequence; gene expression; feature extraction; voting classifier; ensemble approach

I. Introduction

Cancer is a complex disease characterized by the uncontrolled division and growth of abnormal cells in the body, often forming tumors and potentially spreading to other tissues. When cells behave abnormally and divide abnormally, they can damage neighboring cells and form tumors that can be lethal depending on the circumstances. Early detection and appropriate therapy can reduce the chances of harming other cells. Researchers are working to evolve new systems for preliminary cancer detection and categorization in response to the high cancer mortality rate. However, it is challenging to diagnose cancer early due to the disorganized nature of cancer cells. As a result, RNA-Seq analysis can be instrumental in this case [1]. RNA (Ribonucleic acid) is a molecule that plays a critical role in protein synthesis in cells. RNA is made up of a sequence of four different nucleotide bases: adenine (A), guanine (G), cytosine (C), and uracil (U). RNA sequencing (RNA-Seq) is a powerful technique used to study gene expression by determining the sequence of RNA molecules in a sample. In RNA sequencing, RNA is first isolated from the sample and then converted into complementary DNA (cDNA) using reverse transcription. Next, the cDNA is sequenced using high-throughput sequencing technologies to generate extensive RNA sequence data. RNA sequential datasets can be used for various purposes, such as studying gene expression, identifying genetic mutations, and developing new disease therapies. These datasets can be generated through various techniques, such as RNA sequencing, microarrays, and hybridization. RNA-Seq is a recent and well-liked method for discovering new transcripts and isoforms by delivering more normalized and less noisy data for prediction and classification purposes. The most crucial role of transcriptome profiling is identifying the differentially expressed genes in the body or finding gene variances at various levels. Using RNA-sequencing, identification and quantification may be done all at one spot. To categorize diseases like breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), renal chromophobe, etc. RNA-Seq data are freely accessible from many databases [2]. However, many dimensions, complexity, and duplication of features make studying RNA gene expression data particularly challenging. Thus, Machine Learning (ML) and deep learning algorithms can be used to extract features [3], [4] automatically.

Machine Language is a subset of Artificial Intelligence (AI) which is accustomed to identifying underlying patterns in data to identify associations between them [5], [6]. In the age of big data, ML is becoming crucial since it is becoming increasingly difficult for humans to recognize trends and patterns in data to make predictions [7], [8]. ML is thus taking over from humans when identifying and forecasting unseen data to enable informed decision-making. By retrieving features from a database without human input, ML generates predictions. There is a growing use of ML almost everywhere [9]. Its common uses include natural language processing, forecasting, aviation management, and biology to identify protein and RNA sequences [10], [11].

The most crucial aspect of RNA-Seq analyses is differential analysis [12]. Traditional differential analysis techniques often match tumor samples to standard samples of the same tumor kind [13], [14]. However, due to its ignorance of additional tumor forms, such a technology could not distinguish between distinct tumor types [15]. Therefore, conducting an in-depth analysis using RNA-Seq data is necessarily better for understanding the causes of different cancers [16]. Furthermore, most studies attempt to locate genes with differential expression to extract the most pertinent properties. Therefore, developing a strategy that incorporates an understanding of various tumors kinds in the study is essential.

Although RNA-Seq data help detect changes at the gene level, working with RNA-Seq data can be difficult due to its spatial properties [17]. Feature engineering, a technique used to address the challenges of high dimensionality and the relatively small number of samples in gene expression data, is a crucial part of computer approaches for gene expression research. In the current study, gene expression features are extracted in order to overcome the curse of dimensionality and an ensemble of three ML methods for cancer classification using gene expression data have been applied with hard voting strategy. Five tumors of RNA-Seq data are used in this investigation. The current study has applied an ensemble of three ML methods for cancer classification using gene expression data. Five tumors of RNA-Seq data is used in this investigation.

The key contributions of this study are following:

- The proposed framework applies multiple ML models to produce a final ensemble model that is rich in diversity.
- Relevant features extricated from the RNA sequence dataset for cancer prediction.
- RNA Sequence data has been analyzed and visualized to infer knowledge.
- Receiver operating characteristics analysis and state-of-the-art analysis has been done to prove the superiority of the proposed approach.

The remaining paper is organized as follows: The literature relating to the current investigation is discussed in Section II. In Section III, the proposed method is covered. The experimental findings are covered in Section IV, and the article is wrapped up in Section V.

II. Literature Review

First, to categorise cancer, Sterling Ramroach et al. used various machine learning techniques [18]. A dataset for several cancer kinds was downloaded for their study from the online data portal COSMIC. The machine learning models that were used were support vector machine (SVM), neural networks, K closest neighbour (KNN), and random forest (RF). For various cancer types and primary sites, the authors conducted numerous tests. In contrast to other algorithms, RF distinguished itself by achieving significant classification accuracy and being simple to tune.

The boosting deep cascade forest (BCDForest) deep learning algorithm was presented by Yang Guo et al. as the preference for deep neural networks for categorising the cancer RNA. This strategy was used to publicly available microarray data sets encompassing adenocarcinoma, brain, and colon cancer and RNA-Seq data sets containing BRCA, GBM, pan cancers, and LUNG. Each deep forest in this ensemble methodology worked well in predicting the classification outcomes. First, Cascade forests are built using decision tree-based random forests trained to find relevant characteristics in raw data. Next, this result was placed against state-of-the-art classifiers like SVM, KNN, LR, RF, and the original gcforest [18]. The authors claimed that their suggested approach produced more precise results.

Yawen Xiao et al. suggested that the multimodal ensemble technique includes KNN, SVM, DTs, RFs, and Gradient Boosting Decision Trees (GBDTs) [19]. Three different cancers were treated using their suggested approach: LUAD, stomach adenocarcinoma (STAD), and BRCA. This tactic was used to train each classifier individually using the supplied data to produce predictions, which were then used to inform a multimodal ensemble approach using deep learning. This technique predicts cancer more accurately than data produced by a single classifier.

Using Voom, Dincer Goksuluk et al. developed a new range of classifiers termed “voomNSC”, “voomNBLDA”, and “voomPLDA” to classify and assess RNA-Sequencing data. VoomNSC uses the NSC approach in conjunction with voom transformation to create classifiers that are more reliable and accurate [3]. Because VoomDLDA and voomDQDA are not sparse bases, they take advantage of all the model’s properties. The sparse base classifier voomNSC uses only the subset of features in the model. The results showed that voomNSC produced the best outcomes compared to PLDA, NBLDA, and NSC.

Paul Ryvkin et al. provided a brand-new numerical method for CoRAL (classification of RNA by analysis of length) [20]. For this reason, the authors sequenced databases of short RNA sequences. Three trimmed adapter sequences were then applied to the dataset, and a FASTA file was generated after completing numerous pre-processing steps. Next, aligned reads were recorded in SAM files by comparing them to a reference file. A SAM file was then created based on the mismatch rate of the readings. Finally, a BAM file containing the aligned and matched genes was created and delivered to CoRAL. CoRAL categorises various RNA sequence types and draws out salient traits from them. This technique categorises short RNA sequences and gives the user a more significant direction.

Hamid Reza Hassanzadeh et al. suggested a cutting-edge pipeline technique to predict the prognosis of cancer patients [21]. The proposed method used Laplacian Support Vector Machines for semi-supervised learning. This technique predicted the survival of patients with neuroblastoma (NB) and kidney cancer (KIRC). It involved four steps where pre-processing is the first step which includes feature metric storage and data analysis. The second step is feature extraction and then next step removes overfitting problems. Using a generalisation strategy as the final step will enable to assess the precision of each model and determine the weights accordingly. In terms of accuracy, this pipeline method performed better than supervised SVM.

Jiande Wu et al. have suggested using several machine-learning algorithms to detect triple-negative breast cancers [5]. In this study, TCGA data were used to evaluate the gene expression levels of 110 breast cancer samples that were triple-negative with 992 non-triple-negative samples. SVM, KNN, Naive Bayes (NB), and DT were the machine learning classification models that were employed. Due to the enormous dimensions of the data, a further step known as feature selection was carried out before classification to obtain the essential features. The categorisation job had accuracy rates of

90%, 87%, 85%, and 87%, respectively. The results demonstrate that SVM outperformed the other techniques.

GeneQC (gene expression quality control), a machine learning-based technique, was proposed by Adam McDermaid et al. to determine the reliability of expression levels precisely from RNA sequencing datasets [15]. The authors used data from seven plant and animal taxa's RNA sequencing. Three different types of information were entered into GeneQC. A SAM file is read by the first mapping, a reference genome FASTA file by the second, and a species-specific annotation file by the third. GeneQC uses two processes: a Perl script to extract features and an R programme to model the mathematical relationships between those features. GeneQC then categorises the reading alignment category for each Genome.

Yawen Xiao et al. presented a stacked sparse auto-encoder, utilising a semi-supervised deep learning methodology [19]. LUAD, STAD, and BRCA were just a few of the cancer types that this approach predicted. This model integrated supervised classification methods with semi-supervised feature extraction techniques to handle labelled and unlabelled data and extract more precise information for cancer prediction. The results demonstrated that the suggested method gave more accurate prediction results when compared to several cutting-edge machine learning classifiers, including SVM, RF, NN, and auto-encoders. In addition, several studies have considered using technologies, including wireless sensor networks, networks, software-defined networking, and the Internet of Things (IoT) [22].

To find biomarkers in high throughput sequencing, Brian Aevermann et al. suggested combining feature selection and the binary manifestation method of a random forest [23]. The authors' analysis supports this by using the NS-Forest version 2.0. Identifying active cell types and under investigation are two goals for which the most recent iteration of NS-Forest is effective. Their study sent a cell with a clustered gene expression assignment to the RF, from which significant features were gleaned using the Gini index. To overcome unfavourable indicators, genes were further prioritised. The top-ranked genes were then determined using a binary expression score. To adjudicate the least number of features, a criterion based on a decision tree and F-Beta score was employed to investigate various combinations of biomarkers. Finally, the human middle temporal gyrus (MTG) was used in tests to gauge the technique's efficiency [24].

Barbara Pes used the homogenous ensemble approach and applied the selection algorithm to several diversified datasets derived from the original set of records. The author worked on high-dimensional benchmarks from various domains, and this ensemble approach led to a significant gain without any degradation of the predictive performance [25].

Table I tabulates the existing literature on cancer classification with advantages and disadvantages, which pave the way to propose a novel ensemble machine learning technique in this study. Compared to the current cancer classification approaches, the proposed method is different in the way that the RNA features are extricated using PCA and the type of cancer is classified using the proposed ensemble classifier that reduces the computation complexity as the model is constructed using the extricated features alone.

III. Proposed Approach

Most traditional cancer classification systems use a single classification method, relying heavily on a specific classification algorithm for accuracy. The performance of a particular classifier may differ depending on the dataset. Therefore, to increase prediction accuracy, a framework must be developed for combining complementary information from different classifiers. The proposed approach is the hybridization of feature extraction and an ensemble of machine learning classifiers that classify cancer using RNA sequence data. Fig. 1 illustrates the block diagram of the proposed approach. This approach consists of the following modules: feature extraction, data splitting, model selection, and voting ensemble classification.

A. Feature Extraction

Feature extraction is an essential step in machine learning. It involves selecting and transforming the most relevant information from the input data to create a set of new, more informative features. This can help improve machine learning algorithms' performance by reducing the data's dimensionality and removing noise or irrelevant information. There are many techniques for feature extraction, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA). PCA identifies the directions of maximum variance in the data and projects the data onto a new coordinate system defined by these directions, called principal components. The first principal component is the direction of maximum variance in the data. Each subsequent principal component is orthogonal to the previous components and captures the maximum remaining variance. The data is first standardised by performing PCA to have zero mean and unit variance. Then, the covariance matrix is computed, and its eigenvectors and eigenvalues are calculated. The eigenvectors represent the directions of the principal components, and the eigenvalues represent the amount of variance explained by each component. The data can then be projected onto the principal components by multiplying the original data matrix by eigenvectors [34], [35].

Let the dataset be D consisting of x+1 dimensions. Ignore the labels such that new dataset become x dimensional.

The mean for every dimension of the whole dataset is computed as follows:

$$D_{\mu} = \frac{x}{D_{size}} \tag{1}$$

The covariance matrix of the whole dataset is computed as follows:

$$\mathrm{Cov_{mat}}(\mathrm{D_A},\mathrm{D_B}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{A}-\overline{\mathrm{A}}\right)\left(\mathrm{B}-\overline{\mathrm{B}}\right) \tag{2}$$

TABLE I. REVIEW OF EXISTING CANCER CLASSIFICATION SYSTEMS

Ref.	Methodology	Used Dataset	Metrics	Advantages	Disadvantages
Goksuluk et al., 2019 [3]	Microarray-based classifiers	Synthetic dataset	Accuracy, sparsity, sensitivity, specificity	User-friendly and simple	Prior knowledge of packages is required
Khalifa et al., 2020 [4]	Optimised deep learning	Tumour gene expression dataset	Precision, recall, F1-score, accuracy	Less complex and requires less time to train	Performance is low
Wu et al., 2021 [5]	SVM, KNN, NB, and DT	Cancer Genome Atlas dataset	Accuracy, recall, specificity, precision, F1-score	Efficient	Complexity is high
Ramroach et al., 2020 [9]	RF and Gradient boosting machine	Cancer Genome Atlas dataset	Accuracy	High performance	Complexity is high
Arowolo et al., 2020 [26]	Ensemble classifier	RNA sequence dataset	Accuracy, sensitivity, specificity, precision, recall, F1-score	Less complex	Low accuracy
Yu et al., 2020 [27]	NB, RF, SVM	RNA sequence dataset	Sensitivity, specificity, accuracy, F1-score, AUC	Complexity is low	Interpretation is low
Garcıa-Dıaz et al., 2020 [28]	Grouping genetic algorithm	RNA sequence dataset	Standard deviation, accuracy	Computation speed is fast	Incomplete exploration of solution space
Mohammed et al., 2023 [29]	Reinforcement learning	Omics dataset	Accuracy	High processing speed	The optimisation is done partially
Arowolo et al., 2021 [30]	KNN and Decision tree	Western Kenya RNA sequence dataset	Accuracy, sensitivity, specificity, precision, recall, F-score	Less complex	Low accuracy
Arowolo et al., 2021 [31]	Genetic algorithm and Ensemble classification	Anopheles Gambiae dataset	Accuracy, sensitivity, specificity, precision, recall, F-score	High specificity	Works for only small datasets
Ramamurthy et al., 2020 [32]	Deep learning	Synthetic dataset	Recall Jaccard index, dice index, correlation coefficient, specificity, F1-score, computational time	High accuracy	More complex
Mohammed et al., 2021 [33]	Stacking ensemble	Cancer Genome Atlas dataset	Accuracy, F1-score, precision, sensitivity, AUC	High accuracy	Less inference

The eigenvectors and the corresponding eigenvalues are computed as follows:

$$\det(D-\Lambda I)=0 \quad (3)$$

A $d \times k$ dimensional matrix is created by selecting the k eigenvectors with the most significant eigenvalues after the eigenvectors are sorted in decreasing order. Next, the samples are transformed into the new subspace using the eigenvector matrix, yielding the principal components.

B. Data Splitting

Training and test sets are created from the primary component data. The test set assesses how well each classification model performed, and the training set is used to create classification models. The total sample size determines the ratio for dividing the data into two portions. For example, 70% of the training set is typically used in research, and the remaining 30% is used as the test set. However, the split ratio can be lowered to 50% when there are fewer samples [36] [37]. Like the last example, this ratio might be raised to 80% or 90% if the total number of samples is high enough. The fundamental idea behind determining the ideal splitting ratio is to select a splitting ratio with a sufficient number of samples in both the training and test sets to generate a trustworthy fitted model and test predictions. The test accuracy is sensitive to unit misclassifications even though the fitted model is ultimately reliable. In our proposed approach, data has been split on the ratio of 70:30.

C. Model Selection

Selecting the right classifier for a particular machine-learning task is essential to the modelling process. There are a variety of classifiers to choose from, each with its strengths and weaknesses. The factors to consider when selecting a classifier include the type of problem, dataset size, data complexity, and interpretability and performance metrics. Some commonly used classifiers in machine learning are Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forests, and Naive-Bayes. It is often a good idea to try multiple classifiers and compare their performance on the given task to determine the best option. The ensembles of multiple classifiers can often perform better than a single classifier. After building these machine learning models, only the top-performing models are considered for proposed ensemble model building.

D. Ensemble of Classifiers

The ensemble of classifiers is built by combining the advantages of three classifiers such as SVM, NB, and KNN.

- Support Vector Machine

The SVM is mainly used for categorisation due to its excellent accuracy and capacity for managing enormous amounts of data. It is a supervised ML algorithm. The goal of the SVM method is to find a hyper-plane that divides the data set into distinct groups in a suitable way for training sets [38]. Linearly separable data can be divided into two groups by a straight line. A line can separate data that are linearly separable in two dimensions. The function of the line can be represented as follows:

$$y=ax+b \tag{4}$$

The above equation can be re-written as follows by replacing x with x_1 and y with x_2:

$$ax_1-x_2+b=0 \tag{5}$$

If x and w are defined as $x = (x_1, x_2)$ and $w = (a, -1)$, then (4) is defined as follows:

$$wx+b=0 \tag{6}$$

It is the equation of the hyperplane, which is derived from two-dimensional vectors. This hyperplane is used to make predictions. For example, cancer is defined as having a point above or on the hyperplane and not having a threshold below the hyperplane.

- Naive Bayes

Naive Bayes (NB) classifiers are scalable because the number of parameters required is linear in the learning process's number of variables (features/predictors). A closed-form expression, which takes linear time, can be evaluated to perform maximum-likelihood training [39]. The classifier is a function that is computed as follows:

$$NB_{cl}=\mathrm{argmax}_{k\epsilon\{1,\cdots,K\}}P\left(C_k \pi_{i=1}^{n} p(x_i|C_k)\right) \tag{7}$$

- K Nearest Neighbor

K-Nearest Neighbor (KNN) is a supervised algorithm based on the distance function. The distance function, which assesses the degree of similarity or difference between two samples, is the basis of this classifier. The Minkowski distance metric is computed as follows:

$$MD(x,z)=\left(\sum_{r=1}^{d} \|x_r-z_r\|^p\right)^{\frac{1}{p}} \tag{8}$$

With KNN, the function is locally approximated, and all computation is delayed until the function is assessed. Normalising the training data can significantly improve accuracy if the features represent different physical units or sizes because this technique relies on distance for classification. In addition, applying weights to neighbour contributions can help classification and regression because it encourages neighbours closer to one another to contribute more to the average than neighbours farther away. When utilising KNN classification or KNN regression, the neighbours are selected from a group of objects for which the class or object property value is known [40].

IV. Results and Discussion

The experiments were evaluated on an Intel(R) Core(TM) i7-6700 processor with 8 GB of RAM under Windows 10. The proposed approach was implemented in Python using the available machine learning packages.The UCI Machine Learning Repository hosts an RNA sequencing dataset containing gene expression data obtained from RNA sequencing of cancer cells and healthy cells. The gene expression levels are measured for over 20,000 genes, and more than 5,000 samples are in the dataset. The dataset used for experimentation is the RNA sequence dataset. This dataset is from the UCI Machine Learning Repository. The dataset contains information on the gene expression levels of five different cancer forms [41]. They are listed as follows:

a) LUng ADenocarcinoma (LUAD)

b) BReast invasive CArcinoma (BRCA)

c) KIdney Renal Clear cell Carcinoma (KIRC)

d) LUng Squamous Cell Carcinoma (LUSC)

e) Uterine Corpus Endometrial Carcinoma (UCEC)

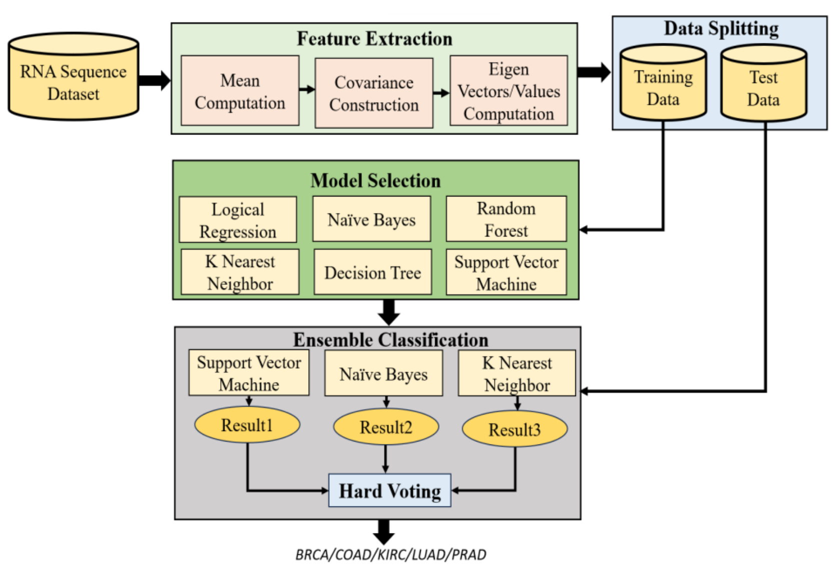


Fig. 1. Block schematic of proposed cancer classification approach.

Algorithm I: Cancer RNA Sequence Classification Algorithm

Input*:* RNA Sequence Dataset, R_D

Output*:* RNA Cancer type, BRCA/COAD/KIRC/LUAD/PRAD

Process:

for all records in R_D
    Compute mean of RNA sequence data, D_μ
    Construct covariance matrix, Cov_{mat}
    Calculate Eigen vectors/Eigen values of Cov_{mat}
    Return top K principal components
end for
Traindata,Testdata=split(CancerRNASequencefeatures,label)
Return Traindata, Testdata
voting=”hard”
M1=SVM(Traindata, Trainlabel, Testdata)
M2=NB(Traindata, Trainlabel, Testdata)
M3=KNN(Traindata, Trainlabel, Testdata)
VotingEnsembleModel(Traindata, Trainlabel, Testdata)
hardvotingclassifier=concatenate(M1, M2, M3)
hardvotingclassifier.fit(Traindata, Trainlabel)
classification=hardvotingclassifier.predict(Testdata)
Return RNAcancerclass

There are 20531 attributes over 801 occurrences. The most dangerous type of cancer for women is BRCA. The most common type of kidney carcinoma, known as KIRC, accounts for 70–80% of instances of the disease and has a high mortality rate globally. LUAD is a common type of cancer. Around 40% of all lung cancer diagnoses are due to it. It primarily attacks non-smokers. LUAD is typically discovered by accident and spreads more slowly than other forms of lung cancer. Smokers are likelier to get LUSC, the second most prevalent lung cancer. Airborne smoke particles often reside in the middle of the lung and transmit LUSC cancer. Undiagnosed in its early stages, UCEC is a recurrent prenatal malignancy. It affects more women than any other type of cancer. Due to the lack of information on its biomarkers for early detection and treatment, it has a high mortality rate. Fig. 2 depicts the distribution of cancer classes.

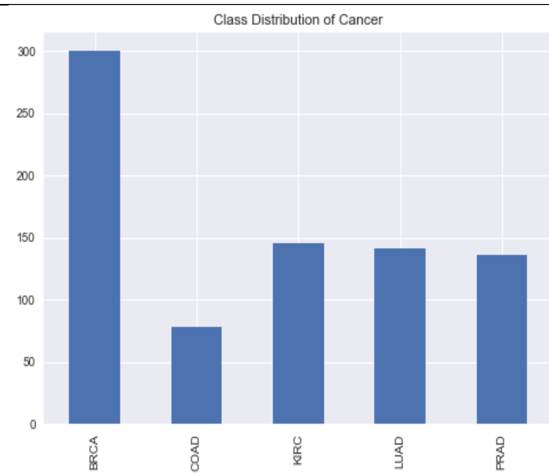


Fig. 2. Distribution of cancer classes.

E. Principal Component Analysis

Fig. 3 depicts the scatter plot of principal components. The dimension of the RNA sequence data is high and in order to improve the performance of the classification task, the dimension of the dataset has been reduced and features are extricated using PCA. Experimentation has been done with with varying number of principal components and using trial and error approach the number of principal components used in the proposed approach is five. The reason behind the achievement of significant results using five principal components is that the dataset consists of five cancer classes. It is observed from the scatter plot that there are similarities in LUAD, BRCA, and COAD cancer classes. The KIRC and PRAD are scattered separately as there are dissimilarities exist in these classes compared to LUAD, BRCA, and COAD.

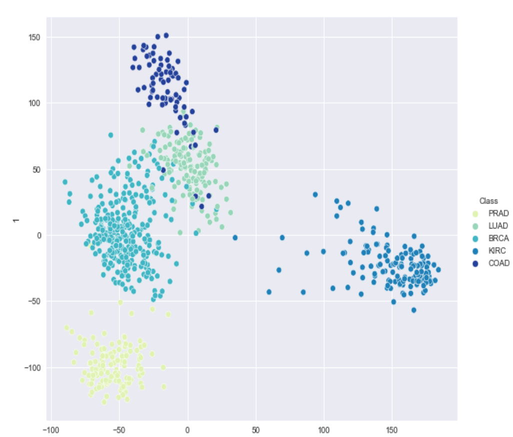


Fig. 3. Scatter plot of principal components.

F. Performance Evaluation

Performance evaluation in machine learning is assessing the accuracy and effectiveness of a trained model. It is essential to evaluate the performance of a machine learning model to determine its effectiveness in solving a specific problem [42]. The model's performance can be improved by tuning the hyper-parameters. Various metrics for evaluating a model's performance include Accuracy, Confusion Matrix, Precision, Recall, F1-Score, AUC (Area- Under-the-Curve)-ROC.

Fig. 4 depicts the confusion matrix for the classification of cancer RNA sequences. There are n columns and n rows in a confusion matrix, where each column represents a predicted classification, and each row represents the true classification [43]. To determine the model's accuracy, it is possible to examine the values along the diagonal - a good model will have a high diagonal value and low values off it. Furthermore, one can determine where the model is having difficulty by examining the highest values, not on the diagonal. These analyses help identify cases where the model's accuracy is high but consistently misclassifies the same data.

A classification report is a technique used to evaluate the performance of machine learning models in multiclass classification problems. It comprehensively summarises the model's performance on various evaluation metrics such as precision, recall, F1-score, and support. Fig. 5 depicts the classification report with the considered performance metrics. The precision, recall, f1-score, and support are computed for all the cancer classes. Furthermore, the macro average and weighted average are also computed to know the performance of the studied cancer ensemble classifier. The accuracy obtained is approximately 100% using the proposed ensemble approach for classifying the cancer RNA sequences.

Table II compares training and testing scores of the existing and proposed cancer classifications. It is seen that the proposed approach performed significantly well in training, but the performance is not significant in terms of testing compared to the proposed hybrid ensemble approach.

TABLE II. TRAINING AND TESTING SCORE ANALYSIS

Model	**Training Score (%)**	**Testing Score (%)**
LR	99.46	98.59
NB	98.75	99.17
RF	99.46	98.76
KNN	99.46	98.75
DT	98.75	97.51
Proposed	***99.64***	***99.59***

G. ROC Analysis

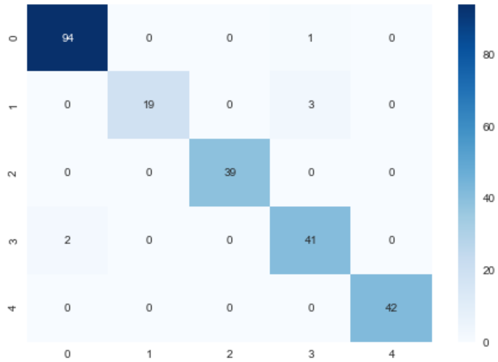
The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of the classifier, showing the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at different classification thresholds [44], [45]. To create a ROC curve, the classifier is applied to a dataset with known outcomes (i.e., a labelled dataset), and the true positive rate (TPR) and false positive rate (FPR) are calculated for different classification thresholds. The TPR is the proportion of true positive predictions among all positive cases in the dataset, and the FPR is the proportion of false positive predictions among all negative cases in the dataset. These rates are plotted on the y-axis and x-axis for different thresholds, resulting in a curve that starts at the origin (TPR=0, FPR=0) and ends at (TPR=1, FPR=1). The area under the ROC curve (AUC) is a standard metric summarising the classifier's overall performance. For example, an AUC of 0.5 indicates random performance, while an AUC of 1 indicates perfect performance. A higher AUC value indicates better classifier performance distinguishing between the positive and negative classes.

The One-vs-Rest (OvR) classifier and the One-vs-One (OvO) classifier are two common approaches for multiclass classification problems [46]. In the OvR approach, a separate binary classifier is trained for each class, which distinguishes that class from all the other classes. In contrast, the OvO approach trains a binary classifier for each pair of classes. Both approaches can be used to generate ROC curves for multiclass classification problems. In the case of OvR, the ROC curve is generated by computing the false positive rate (FPR) and true positive rate (TPR) for each class's binary classifier. The overall ROC curve is then obtained by combining the individual curves for each class. In the case of OvO, the ROC curve is generated by comparing the predicted class probabilities for each pair of classes and computing the FPR and TPR based on the number of correct and incorrect predictions for each pair.

Finally, the overall ROC curve is obtained by combining the FPR and TPR values for all the pairs of classes. When applied to gene selection methods, OvR and OvO can help to improve the results by reducing the number of false positives and false negatives in the classification process. By treating each class as a separate binary classification problem or training separate models for each pair of classes, OvR and OvO can help to better capture the subtle differences between the different classes, leading to more accurate classification results.

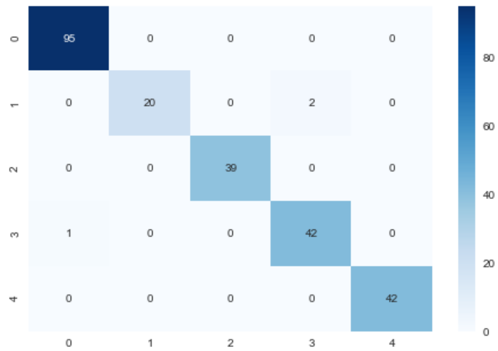
Fig. 6 depicts the ROC analysis for One-vs-Rest (OvR) classifier. The AUC is high for the proposed approach compared to the existing approaches such as LR, NB, RF, and KNN. The reason behind the high performance of the proposed approach is that the extracted features are used for classification. Furthermore, the advantages of the existing classifiers are combined to build the ensemble classifier.

Fig. 7 depicts the ROC analysis for One-vs-One (OvO) classifier. All the plots depict high AUC except COAD versus LUAD, as these cancer classes have high similarity by which classifier cannot differentiate these two classes efficiently.



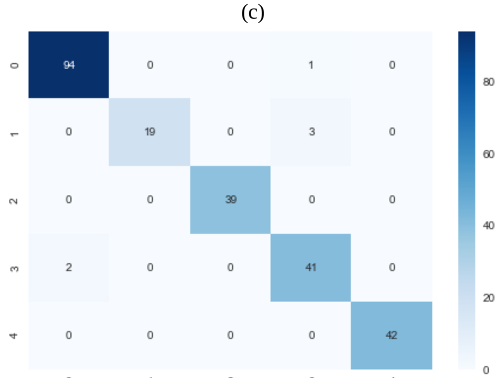
(a)

(b)

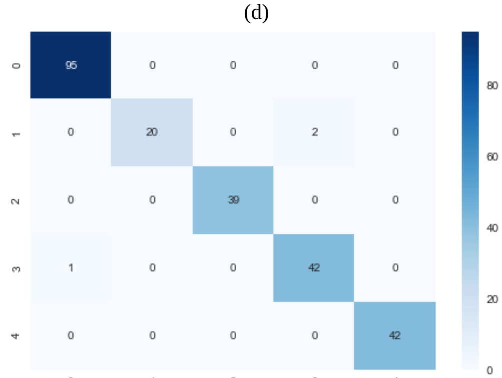


(c)

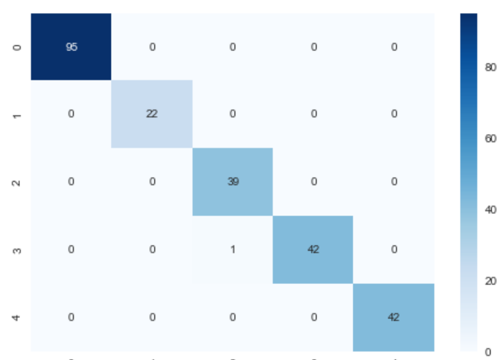
(d)



(e)



(f)



(g)

Fig. 4. Confusion matrix (a) Logistic regression (b) Naive Bayes (c) Random forest (d) K nearest neighbor (e) Decision tree (f) Support vector machine (g) Proposed approach.

	precision	recall	f1-score	support
BRCA	0.98	0.99	0.98	95
COAD	1.00	0.86	0.93	22
KIRC	1.00	1.00	1.00	39
LUAD	0.91	0.95	0.93	43
PRAD	1.00	1.00	1.00	42
accuracy			0.98	241
macro avg	0.98	0.96	0.97	241
weighted avg	0.98	0.98	0.97	241

(a)

	precision	recall	f1-score	support
BRCA	1.00	1.00	1.00	95
COAD	1.00	0.91	0.95	22
KIRC	1.00	1.00	1.00	39
LUAD	0.96	1.00	0.98	43
PRAD	1.00	1.00	1.00	42
accuracy			0.99	241
macro avg	0.99	0.98	0.99	241
weighted avg	0.99	0.99	0.99	241

(b)

	precision	recall	f1-score	support
BRCA	0.99	1.00	0.99	95
COAD	1.00	0.91	0.95	22
KIRC	1.00	1.00	1.00	39
LUAD	0.95	0.98	0.97	43
PRAD	1.00	1.00	1.00	42
accuracy			0.99	241
macro avg	0.99	0.98	0.98	241
weighted avg	0.99	0.99	0.99	241

(c)

	precision	recall	f1-score	support
BRCA	0.99	1.00	0.99	95
COAD	1.00	0.91	0.95	22
KIRC	1.00	1.00	1.00	39
LUAD	0.95	0.98	0.97	43
PRAD	1.00	1.00	1.00	42
accuracy			0.99	241
macro avg	0.99	0.98	0.98	241
weighted avg	0.99	0.99	0.99	241

(d)

	precision	recall	f1-score	support
BRCA	0.98	0.99	0.98	95
COAD	1.00	0.86	0.93	22
KIRC	1.00	1.00	1.00	39
LUAD	0.91	0.95	0.93	43
PRAD	1.00	1.00	1.00	42
accuracy			0.98	241
macro avg	0.98	0.96	0.97	241
weighted avg	0.98	0.98	0.97	241

(e)

	precision	recall	f1-score	support
BRCA	0.99	1.00	0.99	95
COAD	1.00	0.91	0.95	22
KIRC	1.00	1.00	1.00	39
LUAD	0.95	0.98	0.97	43
PRAD	1.00	1.00	1.00	42
accuracy			0.99	241
macro avg	0.99	0.98	0.98	241
weighted avg	0.99	0.99	0.99	241

(f)

	precision	recall	f1-score	support
BRCA	1.00	1.00	1.00	95
COAD	1.00	1.00	1.00	22
KIRC	0.97	1.00	0.99	39
LUAD	1.00	0.98	0.99	43
PRAD	1.00	1.00	1.00	42
accuracy			1.00	241
macro avg	0.99	1.00	1.00	241
weighted avg	1.00	1.00	1.00	241

(g)

Fig. 5. Classification report (a) Logistic regression (b) Naive Bayes (c) Random forest (d) K nearest neighbor (e) Decision tree (f) Support vector machine (g) Proposed approach.

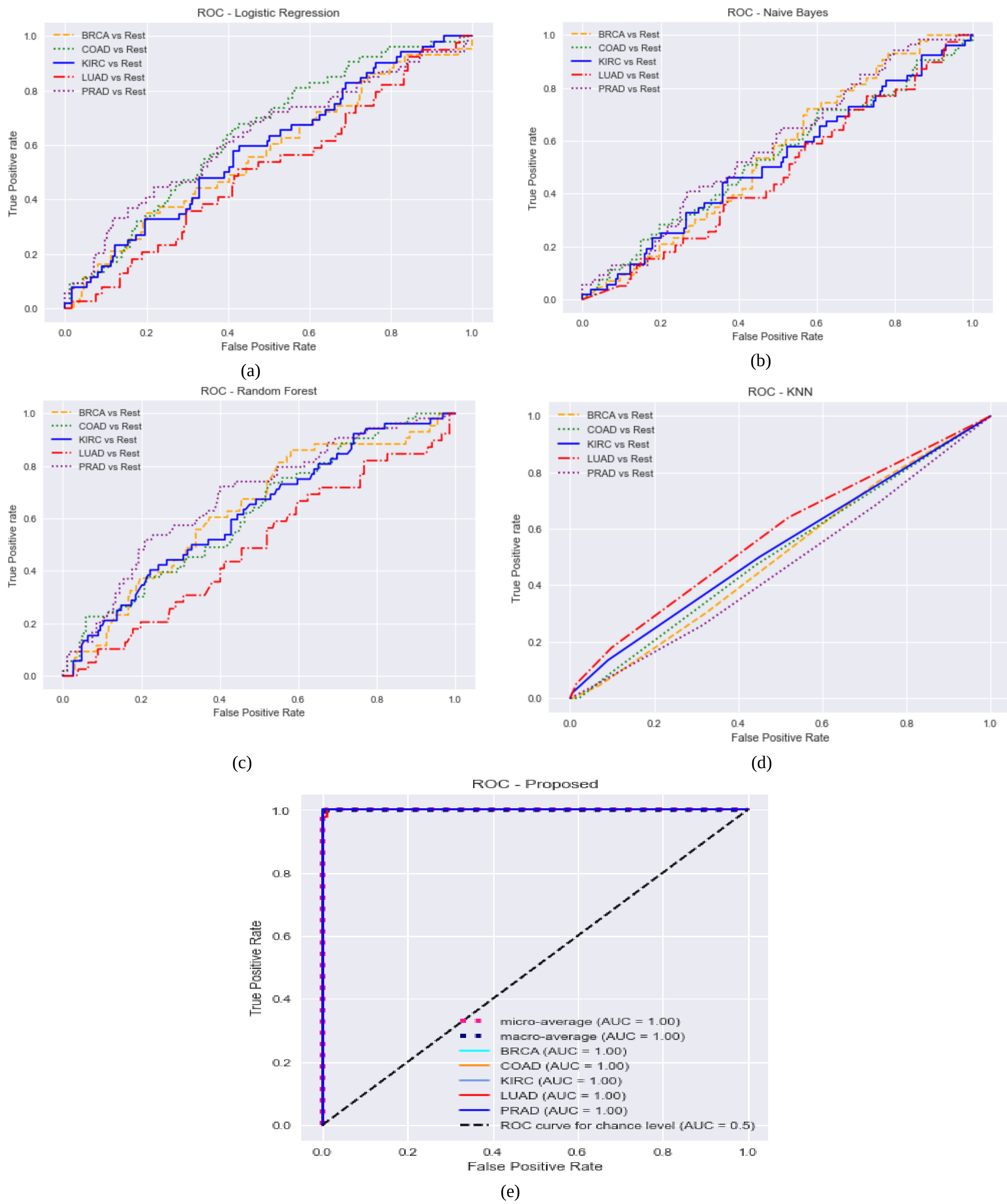


Fig. 6. Receiver operating characteristics curve – One-vs-rest (OvR) (a) Logistic regression (b) Naive Bayes (c) Random forest (d) K nearest neighbor (e) Proposed approach.

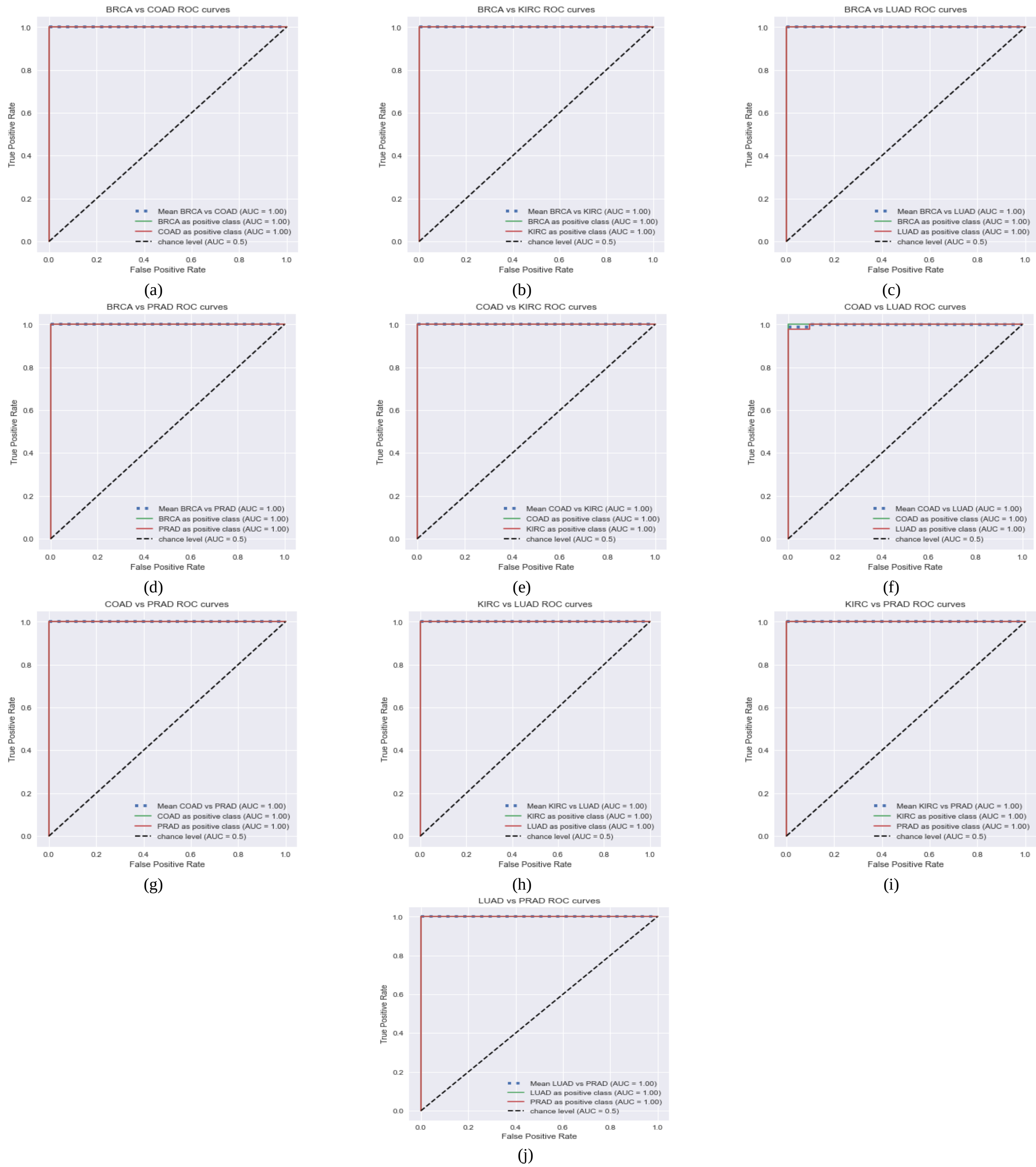


Fig. 7. Receiver Operating Characteristics Curve of Proposed Approach – One-vs-One (OvO) (a)BRCA vs. COAD (b) BRCA vs. KIRC (c) BRCA vs. LUAD (d) BRCA vs. PRAD (e) COAD vs. KIRC (f) COAD vs. LUAD (g) COAD vs. PRAD (h) KIRC vs. LUAD (i) KIRC vs. PRAD (j) LUAD vs. PRAD.

H. State-of-the-Art Analysis

The state-of-the-art analysis with respect to the reported results of existing cancer RNA classification systems is tabulated in Table III. The proposed approach is compared with optimized deep learning, ensemble classifier, SVM, grouping genetic algorithm, marker gene selection, dimensionality reduction with neural network, and dimensionality reduction with SVM. It is evident that the proposed approach surpasses the existing cancer classification systems. The reason behind the significant performance is that the curse of dimensionality problem existing in gene sequence data has been overcome using the feature extraction process and extracted features are utilized for the ensemble classification task. Furthermore, a hard voting classifier has been built using the combination of best-performing classifiers that are chosen based on the trial-and-error process. Thus, the superiority of the proposed approach has been proved.

TABLE III. STATE-OF-THE-ART ANALYSIS

Method	Year	Accuracy (%)
Optimised deep learning [4]	2020	96.9
Ensemble classifier [26]	2020	93.3
Support vector machine [27]	2020	97.37
Grouping genetic algorithm [28]	2020	98.81
Marker gene selection [23]	2021	97.0
PCA-NN [30]	2023	96.6
PCA-SVM [30]	2023	96.5
Proposed	-	***99.59***

V. CONCLUSION

The study successfully classified the RNA cancer types from a huge database using the proposed voting ensemble classifier approach. The RNA cancer sequence features were extracted using feature extraction process of PCA to reduce the dimension of the sequence data. The extracted features were used for ensemble classification model building and a hard voting ensemble classifier was effectively applied. In this work a dataset from the UCI Repository was used that includes 801 samples and 20,531 attributes representing five forms of cancer (Breast, Kidney, Colon, Lung, and Prostate). The proposed system used to find an ideal response for the classification of cancer RNA sequences. The accuracy percentage for ensemble categorization is 99.59%. The ROC analysis had been performed with respect to one versus one class and one versus rest of the classes. It is evident that the AUC for the proposed approach is high. Furthermore, the state-of-the-art analysis proved that the proposed ensemble approach outperforms the existing RNA cancer classification systems. In future, the work can be improved by employing a wider variety of exhaustive and thorough techniques, which might be used with other kinds of high-dimensional datasets.

ACKNOWLEDGMENT

All data were collected and handled in accordance with ethical standards, including anonymization and secure storage, to ensure the protection of participants' privacy and confidentiality.

The dataset used and analyzed during the current study are available in the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq.The data is available publicly and can be used by the machine learning community for the empirical analysis of machine learning algorithms.

Moreover, the first author receives the grant under FDP scheme of UGC India. The second Author has no conflict of Interest.

REFERENCES

[1] S. Wesolowski, M. R. Birtwistle, G. A. Rempala, A comparison of methods for RNA-seq differential expression analysis and a new empirical Bayes approach, Biosensors 3 (3) (2013) 238–258.

[2] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcze´sniak, D. J. Gaffney, L. L. Elo, X. Zhang, et al., A survey of best practices for RNA-seq data analysis, Genome Biology 17 (1) (2016) 1–19.

[3] D. Goksuluk, G. Zararsiz, S. Korkmaz, V. Eldem, G. E. Zararsiz, E. Ozcetin, A. Ozturk, A. E. Karaagaoglu, Mlseq: Machine learning interface for RNA-sequencing data, Computer methods and programs in biomedicine 175 (2019) 223–231.

[4] N. E. M. Khalifa, M. H. N. Taha, D. E. Ali, A. Slowik, A. E. Hassanien, Artificial intelligence technique for gene expression by tumor RNA-seq data: a novel optimised deep learning approach, IEEE Access 8 (2020) 22874–22883.

[5] J. Wu, C. Hicks, Breast cancer type classification using machine learning, Journal of personalised medicine 11 (2) (2021) 61.

[6] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, H. Alinejad-Rokny, A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues, Journal of Biomedical Informatics 113 (2021) 103627.

[7] D. Sachin et al., Dimensionality reduction and classification through PCA and LDA, International Journal of Computer Applications 122 (17) (2015).

[8] R. Zhang, T. Du, S. Qu, A principal component analysis algorithm based on dimension reduction window, IEEE Access 6 (2018) 63737–63747.

[9] S. Ramroach, A. Joshi, M. John, Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers, Molecular omics 16 (2) (2020) 113–125.

[10] B. Gunasundari, S. Arun, Ensemble classifier with hybrid feature transformation for high dimensional data in healthcare, in 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, 2022, pp. 886–892.

[11] Y. Xu, Z. Yu, W. Cao, C. P. Chen, A novel classifier ensemble method based on subspace enhancement for high-dimensional data classification, IEEE Transactions on Knowledge and Data Engineering 35 (1) (2021) 16–30.

[12] G. Zararsiz, D. Goksuluk, B. Klaus, S. Korkmaz, V. Eldem, E. Karabulut, A. Ozturk, voomdda: discovery of diagnostic biomarkers and classification of RNA-seq data, PeerJ 5 (2017) e3890.

[13] A. Ishii, K. Yata, M. Aoshima, Geometric classifiers for high-dimensional noisy data, Journal of Multivariate Analysis 188 (2022) 104850.

[14] N. Song, K. Wang, M. Xu, X. Xie, G. Chen, Y. Wang, Design and analysis of ensemble classifier for gene expression data of cancer, Adv. Genet. Eng 5 (2015).

[15] A. McDermaid, X. Chen, Y. Zhang, C. Wang, S. Gu, J. Xie, Q. Ma, A new machine learning-based framework for mapping uncertainty analysis in RNA-seq read alignment and gene expression estimation, Frontiers in genetics 9 (2018) 313.

[16] G. Zararsız, D. Goksuluk, S. Korkmaz, V. Eldem, G. E. Zararsiz, I. P. Duru, A. Ozturk, A comprehensive simulation study on classification of RNA-Seq data, PloS one 12 (8) (2017) e0182507.

[17] Y. Guo, S. Liu, Z. Li, X. Shang, Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data, in 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), IEEE, 2017, pp. 1664–1669.

[18] S. Ramroach, M. John, A. Joshi, The efficacy of various machine learning models for multiclass classification of RNA-seq expression data, in Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 1, Springer, 2019, pp. 918–928.

[19] Y. Xiao, J. Wu, Z. Lin, X. Zhao, A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-Seq data, Computer methods and programs in biomedicine 166 (2018) 99–105.

[20] P. Ryvkin, Y. Y. Leung, L. H. Ungar, B. D. Gregory, L.-S. Wang, Using machine learning and high-throughput RNA sequencing to classify the precursors of small non-coding RNAs, Methods 67 (1) (2014) 28–35.

[21] H. R. Hassanzadeh, J. H. Phan, M. D. Wang, A multimodal graph-based semi-supervised pipeline for predicting cancer survival, in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016, pp. 184–189.

[22] A. M. McCarthy, Y. Liu, S. Ehsan, Z. Guan, J. Liang, T. Huang, K. Hughes, A. Semine, D. Kontos, E. Conant, et al., Validation of breast cancer risk models by race/ethnicity, family history and molecular subtypes, Cancers 14 (1) (2021) 45.

[23] B. Aevermann, Y. Zhang, M. Novotny, M. Keshk, T. Bakken, J. Miller, R. Hodge, B. Lelieveldt, E. Lein, R. H. Scheuermann, A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing, Genome Research 31 (10) (2021) 1767–1780.

[24] R. Zhu, Z. Wang, N. Sogi, K. Fukui, J.-H. Xue, A novel separating hyperplane classification framework to unify nearest-class-model methods for high-dimensional data, IEEE transactions on neural networks and learning systems 31 (10) (2019) 3866–3876.

[25] B. Pes, Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains, Neural Computing and Applications 32 (10) (2020) 5951–5973.

[26] M. O. Arowolo, M. Adebiyi, A. Adebiyi, O. Okesola, Pca model for RNA-seq malaria vector data classification using KNN and decision tree algorithm, in 2020 international conference in mathematics, computer engineering and computer science (ICMCECS), IEEE, 2020, pp. 1–8.

[27] Z. Yu, Z. Wang, X. Yu, Z. Zhang, et al., Rna-seq-based breast cancer subtypes classification using machine learning approaches, Computational intelligence and neuroscience 2020 (2020).

[28] P. Garcıa-Dıaz, I. S´anchez-Berriel, J. A. Mart´ınez-Rojas, A. M. Diez-Pascual, Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-seq data, Genomics 112 (2) (2020) 1916–1925.

[29] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, B. Garcia-Zapirain, A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (sarsa), Computers in Biology and Medicine 154 (2023) 106617.

[30] M. O. Arowolo, M. O. Adebiyi, A. A. Adebiyi, A genetic algorithm approach for predicting ribonucleic acid sequencing data classification using knn and decision tree, TELKOMNIKA (Telecommunication Computing Electronics and Control) 19 (1) (2021) 310–316.

[31] M. O. Arowolo, M. Adebiyi, A. A. Adebiyi, J. OKesola, Predicting RNA-seq data using genetic algorithm and ensemble classification algorithms, Indonesian Journal of Electrical Engineering and Computer Science 21 (2) (2021) 1073–1081.

[32] M. Ramamurthy, I. Krishnamurthi, S. Vimal, Y. H. Robinson, Deep learning-based genome analysis and NGS-RNA ll identification with a novel hybrid model, Biosystems 197 (2020) 104211.

[33] M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir, B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on tcga data, Scientific reports 11 (1) (2021) 1-22.

[34] M. O. Arowolo, M. Adebiyi, A. A. Adebiyi, An efficient PCA ensemble learning approach for prediction of RNA-seq malaria vector gene expression data classification, International Journal of Engineering Research and Technology 13 (1) (2020) 163–169.

[35] M. F. Kabir, T. Chen, S. A. Ludwig, A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction, Healthcare Analytics 3 (2023) 100125.

[36] K. Pradhan, P. Chawla, Medical internet of things using machine learning algorithms for lung cancer detection, Journal of Management Analytics 7 (4) (2020) 591–623.

[37] A. A. Osuwa, H. Oztoprak, Importance of continuous improvement of machine learning algorithms from a health care management and management information systems perspective, in 2021 International Conference on Engineering and Emerging Technologies (ICEET), IEEE, 2021, pp. 1–5.

[38] F. Alharbi, A. Vakanski, Machine learning methods for cancer classification using gene expression data: A review, Bioengineering 10 (2) (2023) 173.

[39] W. M. Ead, M. A. Abdelazim, M. M. Nasr, Feedforward deep learning optimiser-based RNA-Seq women's cancers detection with a hybrid classification models for biomarker discovery, International Journal of Advanced Computer Science and Applications 13 (12) (2022).

[40] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, M. A. Moni, Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning, Expert Systems with Applications 205 (2022) 117695.

[41] K. Ferles, Y. Papanikolaou, 'cancer types: RNA sequencing values from tumour samples/tissues, Distributed by Mendeley (2018).

[42] Japkowicz, N., Shah, M. (2015). Performance Evaluation in Machine Learning. In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham.

[43] Visa, Sofia & Ramsay, Brian & Ralescu, Anca & Knaap, Esther. (2011). Confusion Matrix-based Feature Selection. CEUR Workshop Proceedings. 710. 120-127.

[44] M. H. Zweig, G. Campbell, Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine, Clinical chemistry 39 (4) (1993) 561–577.

[45] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology. 1982;143(1):29–36.

[46] Student S, Fujarewicz K. Stable feature selection and classification algorithms for multiclass microarray data. Biol Direct. 2012 Oct 2;7:33. doi: 10.1186/1745-6150-7-33. PMID: 23031190; PMCID: PMC3599581.