Identification of Cancer Types from Gene Expressions using Learning Techniques

Swati B Bhonde[1*], Sharmila K Wagh[2], Jayashree R Prasad[3]
[1]Department of Bioinformatics, Kashibai Navale College of Engineering, Pune, India
[2]Department of Bioinformatics, MES College of Engineering, Pune, India
[3]Department of Bioinformatics, MIT ADT University, Pune, India

ABSTRACT

Around the globe, the tumor is the leading cause of death. Early detection and prediction of a cancer type are important for a patient's wellbeing. Functional genomic data has recently been used in the effective and early detection of cancer. According to previous research, the use of microarray data in cancer prediction has evidenced two main problems as high dimensionality and limited sample size. Several researchers have used numerous statistical and machine learning - based methods to classify cancer types but still, limitations are there which makes cancer classification a difficult job. Deep Learning (DL) and Convolutional Neural Network (CNN) have proven effective in analyzing a wide range of unstructured data including gene expression data. In the proposed method gene expression data of five types of cancer is collected from The Cancer Genome Atlas (TCGA). Prominent features are selected using a hybrid Particle Swarm Optimization (PSO) and Random Forest (RF) algorithm followed by the use of Principal Component Analysis (PCA) for dimensionality reduction. Finally, for classification blend of Convolutional Neural Network (CNN) and Bi - Directional Long Short Term Memory (Bi - LSTM) is used to predict the target type of cancer. We demonstrate that accuracy of the proposed method is 96.89 %. As compared to existing work, our method outperformed and showed better results.

KEYWORDS
Cancer prediction, Gene Expression (GE), Random Forest (RF), Particle Swarm Optimization (PSO), PCA (Principal Component Analysis), Bi-LSTM (Bidirectional - Long Short Term Memory)

*Corresponding Author:
Swati B Bhonde, Department of Bioinformatics, Kashibai Navale College of Engineering, Pune, India;
E-mail: scholar.swatibbhonde@gmail.com

How to Cite This Article:
Bhonde SB, Wagh SK, Prasad JR. Identification of Cancer Types from Gene Expressions using Learning Techniques. J Evid Based Med Healthc 2022;9(11):60.

Received: 04-May-2022,
Manuscript No: JEBMH-22-59446;
Editor assigned: 06-May-2022,
PreQC No. JEBMH-22-59446(PQ);
Reviewed: 20-May-2022,
QC No. JEBMH-22-59446;
Revised: 04-Jul-2022,
Manuscript No. JEBMH-22-59446(R);
Published: 14-Jul-2022,
DOI: 10.18410/jebmh/2022/09/11/60.

Copyright © 2022 Bhonde SB, et al. This is an open access article distributed under Creative Commons Attribution License [Attribution 4.0 International (CC BY 4.0)]

INTRODUCTION

Cancer is the second most common cause of mortality worldwide, accounting for roughly one of every six fatalities. To reduce the effect of cancer on human health, significant research efforts have been dedicated to cancer detection and treatment techniques.[1] the goal of cancer detection is to classify tumor categories and establish indicators for each malignancy so that we can develop a learning technique that can automatically recognize certain metastatic tumors or diagnose cancer in an early phase. Cancer prediction focuses on cancer susceptibility, recurrence, and prognosis by offering precise cancer treatment depending on unique genetic biomarkers.[2] The last decade has witnessed abundant use of DL algorithms, which has the exciting potential to uncover complicated interactions buried in large - scale information including bioinformatics.[3] Although it is often considered synonymous with computational biology, bioinformatics is a discipline of science that is related to but distinct from the biological computation.[4] Bioinformatics uses computing to better understand biology, while biological computation uses bioengineering and biology to construct biological computers.[5] The use of DL algorithms has grown rapidly in bioinformatics, demonstrating exciting abilities to mine the intricate relationships concealed in extensive biological and biomedical evidence. DL is a class of multi - layer Neural Network models (NN) that progressively succeeds at learning from the enormous amount of data.[6] It also comprises a training phase wherein the network characteristics are predicted from a training dataset and a testing phase in which the learned network is used to estimate subsequent outputs.[7] The development of the DL model for improved accuracy and interpretability for cancer type prediction is now made possible by the accumulation of whole transcriptomic profiling of tumor samples.[8] Cancer is caused by variations or changes in gene regulators that regulate cell division and development, resulting in highly expressed genes.[9] In such cases, a group of genes known as oncogenes plays an important role in the transformation of normal cells into cancerous cells.[10] Somatic mutations, Copy Numbers (CNs), profiles, and various epigenetic changes are distinct in each kind of tumor.[11] As a consequence, cell differentiation, environmental factors, and genetic inheritance by parents may interrupt Gene Expression (GE). Changes in GE can affect the development of proteins, which can affect normal cell behavior.[12] The damaged cells begin to reproduce at a faster rate than normal, eventually forming a tumor in the affected region. Such tumors may sometimes develop into cancer.[13] This is one of the reasons why cancer cases are steadily growing year after year, eventually becoming the world's second leading cause of death.[14] Despite its importance in directing patient care, histologic - based cancer diagnosis remains difficult in many patients, especially in those who first present with metastatic, poorly differentiated neoplasms, where unclear or inaccurate classification may have a negative impact on treatment options and outcomes.[15] Tumors are comprised of an array of cancer cells of varying properties and anticancer drug susceptibilities. Tumor heterogeneity also made it impossible to match patients with the right drug at the right time. Furthermore, the wide range of health - related characteristics described by non - comics data, such as clinical and epidemiological variables, may account for some of genomic data's low predictive ability.[16] As a consequence, it's important to combine Omics and non - omics (On) data in a single model. This opens the door to gain a better understanding of biological mechanisms of health and illness.[17] This endeavor, without a doubt, poses a host of challenges in terms of data creation, capture, curation, dissemination, analysis, emulation, and data security and storage.[18] Identifying candidate genes that could explain major reaction variations is also important.[19] The patient's top priority is to get a prompt diagnosis of essential illnesses, such as tumors, with the least amount of error possible.[20] Furthermore, a genomics profile involves a large amount of multidimensional data that must be analyzed using the required statistical approach to obtain precise information.[21] Because of biomedical analysis on genome data, we can analyze cancer omics data in the form of raw sequencing data, Single Nucleotide Polymorphism (SNP) data, Copy Number Variation (CNV) data, DNA methylation data, and miRNA gene expression data.[23] A massive amount of gene expression data is publicly available in databases like The Cancer Genome Atlas (TCGA), Catalogue of Somatic Mutations (COSMIC), Genebank, University of California Irvine (UCI), National Center for Biotechnology Information (NCBI), etc.[24] The sensitivity of a clinical drug in terms of predicting patient response to various diseases is a major concern. There is still scope to deal with algorithmic inequity, outcome noise, process bias, and model variance.[25] DL algorithms, on the other hand, have set new benchmarks in image processing, natural language processing, voice recognition, and, most recently, bioinformatics.[26] DL is a theoretically useful method for large - scale and deep Artificial Neural networks that have received a lot of attention in recent years.[27] DL models can reliably estimate the complex nonlinear relationship between environmental parameters, thanks to multi - layer learning, which helps to capture the possible interaction between environmental variables for remote sensing retrieval, fusion, and downscaling.[28] To solve the current challenges and difficulties of cancer prediction, a new approach using DL is insistently needed to make cancer prediction easier and more reliable.[29] In this context, this research work proposes a novel approach to predict cancer type using a hybrid algorithm in the feature selection and classification phase. We have used the miRNA PAN cancer dataset which addresses five types of cancers consisting of 20,531 gene columns and 801 patient records as rows.[30] Attributes of each sample are RNA - Seq gene expression levels measured by the Illumine -HiSeq platform. As an extension to the preliminary study carried out in section - II, this research work has the following contributions:

- Our system can investigate five types of cancers (Breast Carcinoma – BRCA, Colon Adenocarcinoma (COAD), Kidney Renal Clear-cell carcinoma (KIRC), Lung Adenocarcinoma (LUAD), Prostate Adenocarcinoma (PRAD).
- To select prominent features, we have introduced a novel feature selection algorithm that combines

Particle Swarm Optimization (PSO) with Random Forest (RF) algorithm.
- For classification blend of Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) algorithm is used to attain high accuracy.
- The precision achieved by our model is high as compared to existing systems and is of key importance in highlighting the effectiveness of our model.

This paper is divided into five main sections. The previous section described the theoretical background of cancer and current trends in cancer prediction with pros and cons. Section - 2 reviews related work on cancer prediction. Section - 3 emphasizes the architecture of the proposed system including experimental results obtained for novel hybrid algorithms in respective phases. Section - 4 gives insights on validation and performance comparison of our system with baseline algorithms followed by concluding remark in section - 5 and future scope in section - 6.

LITERATURE REVIEW

The prediction accuracy of a Weighted - Particle Swarm Optimization (WPSO) using a Smooth Support Vector Machine (SSVM) was 98.42 percent. The suggested system in used the voting classifier technique to combine SVM, Naive Bayes, and J 48 to obtain an accuracy of 97.13 percent, which is better than each of the separate classifiers. For NB, RepTree, and K - NNs, the method established in achieved 70 percent, 76.3 percent, and 66.3 percent accuracy, respectively. They discovered four characteristics that are best for this classification assignment after implementing PSO. The accuracy values for NB, RepTree and K - NNs with PSO were 81.3 percent, 80 percent, and 75 percent, respectively. According to the findings in, 91.7 percent, 91.7 percent, and 94.11 percent accuracy were attained for BBN, BAN, and TAN, respectively, using gradient boosting. The acquired findings in showed that the Naive Bayes algorithm worked well with a 97.36 percent accuracy, the RBF network performed well with a 96.77 percent accuracy, and the J 48 came in third with a 93.41 percent accuracy uses data from TCGA to compile Copy Number Variations (CNVs) for 8000 cancer patients with 14 distinct cancer types. Then, using 578 oncogenes and 20,308 protein - coding genes, two alternative sparse representations of Copy Number Variations (CNVs), encompassing genomic deletions and duplication across samples, are created. The researchers then used both representations to train Convolutional - Long Short Term Memory (Conv - LSTM) and Convolutional Auto Encoder (CAE) networks and produce snapshot models. To distinguish various five types of cancer, based on tumor Ribonucleic acid - sequence (RNA - Seq) data, proposes a novel integrated DL approach based on Binary Particle Swarm Optimization (PSO) with a Decision Tree (BPSO - DT) and CNN. The architecture of two main convolutional layers for featured extraction and two fully connected layers is introduced to classify the 5 different types of cancer according to the availability of images on the dataset. Used GE profiles to classify cancer have revealed new information about the origins of cancer and how to cure it.[37] addresses 21 types of cancer. Here, 300 most important genes expressed in each cancer were used to train 7,398 cancer samples and 640 normal samples from 21 tumors and normal tissues in the DL.[38] collected microarray gene expression data for 238 samples from Colo Rectal Cancer (CRC) and regular samples from the Gene Expression Omnibus (GEO) database, which contains 13,487 genes. On 173 samples, Weighted Gene Co - Expression Network Analysis (WGCNA) yielded 12 gene modules. For the classifier, the authors have used Variation Auto Encoder (VAE) for 1159 genes.[39] Incorporated distinct molecular modifications and inferred features such as mutational signatures, a machine learning method was developed to predict tumor type from targeted panel DNA sequence data collected at the point of treatment. This algorithm was trained on 7791 tumors from a prospectively sequenced population of patients with advanced cancer, spanning 22 cancer types.[39] Targeted tumor sequencing estimated likely tissues of origin in 95 of 141 patients (67.4 %) with cancers of uncertain primary location.

CANCER PREDICTION SYSTEM

Variations or modifications in gene regulators that control cell division and growth result in the highly expressed gene, which causes cancer. Generally, building a stable history mutation model is challenging due to the heterogeneity of tumors. So, there is a crucial need to introduce a novel solution that can overcome the complexities of the gene expression data and enhance the accuracy of prediction.[40] this section demonstrates the architecture of the proposed system. Figure 1 shows the architecture of the proposed system which is carried out in five phases:

- Data collection
- Feature extraction (PSO + RF)
- Dimensional Reduction using PCA
- Classification (RNN + LSTM)
- Validation and performance comparison

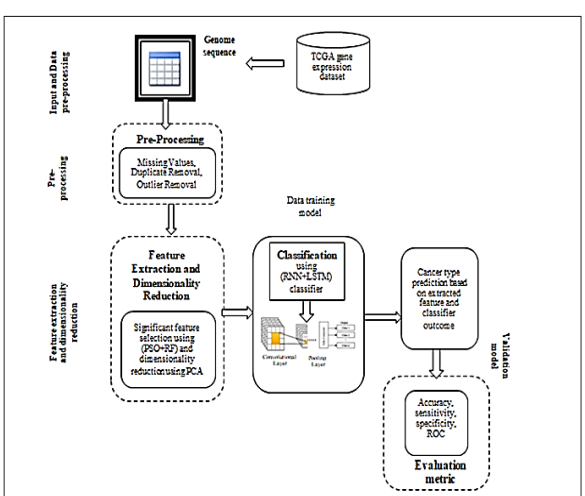

Figure 1. Architecture of the proposed system.

Details of Each Phase are Presented below

Data collection: Publicly accessible mi - RNA TCGA PANCAN dataset is used in this work where attributes of each sample are RNA - Seq gene expression levels

measured by the Illumina HiSeq platform. It consists of 801 rows and 20531 columns which addresses five types of cancer profiles as, BRCA - Breast Carcinoma, COAD - Colon Adenocarcinoma, KIRC - Kidney Renal Clear – Cell Carcinoma, LUAD - Lung Adenocarcinoma, PRAD - Prostate Adenocarcinoma (Table 1).

Cancer class	Count
BRCA	300
COAD	78
KIRC	146
LUAD	141
PRAD	136

Table 1. Frequency for Each Cancer Type.

The pie chart below shows the class frequency for each type of cancer in the dataset. It is also observed that BRCA has the highest number of samples (Figure 2).

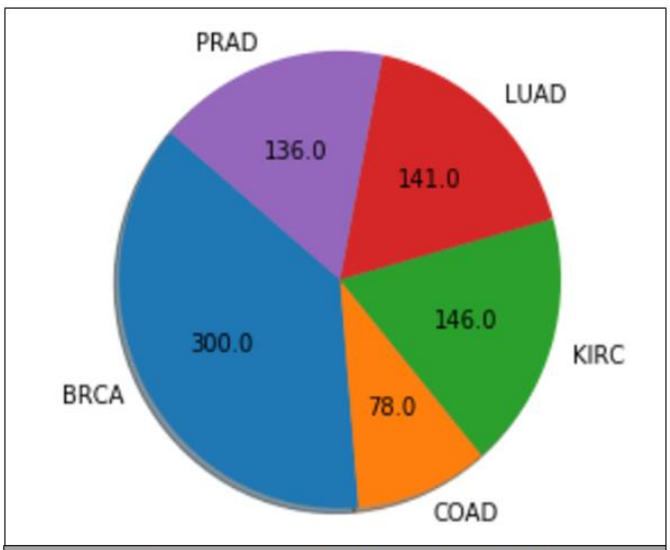

Figure 2. Pie Chart Visualization for Cancer Class Frequency.

For each particle
Initialize particle
End
Do
For each particle
Calculate fitness value
If the fitness value is better than the best fitness value (pBest) in history
set current value as the new pBest
End
Choose the particle with the best fitness value of all the particles as the gBest
For each particle
Calculate particle velocity according to equation (a)
Update particle position according to equation (b)
End

The proposed method selected 12556 genes out of 20531 genes based on a hybrid feature selection algorithm. After applying hybrid RF and PSO these features are passed to the PCA for capturing the variance and make the data linearly separable within 500 PCA components.

Dimensionality Reduction using Principal Component Analysis (PCA):

PCA is used to find the eigenvectors of a covariance matrix with the highest eigenvalues and then the same is used to protect the data into a new subspace of equal or fewer dimensions. Eigenvector and Eigenvalues capture more variance and correlation. It drops the variables with the low variance within a compressed dimension of components. This is a transformation that entails linear algebra to compress a dataset. The feature extracted by PCA generally has high variance and it will be linearly separable which makes our LSTM model fit with less complexity.

Following is Pseudo code for PCA:

X - > input samples
Components -> 500
Compute dot product matrix: XT X = PN
$i=1(x_i - \mu)\ T\ (x_i - \mu)$
Eigenanalysis: XT X = VΛVT
Compute eigenvectors: U = XVΛ− 1/2
Keep feature vector specific number of first components: Ud = $[u_1, u_2 . . , u_{500}]$
Return Computed features: Y = Ud T X

From the below Figure 3 scree plot, we can interpret that there are 500 optimal components that capture most of the variance. Genes get compressed within 500 features components of PCA.

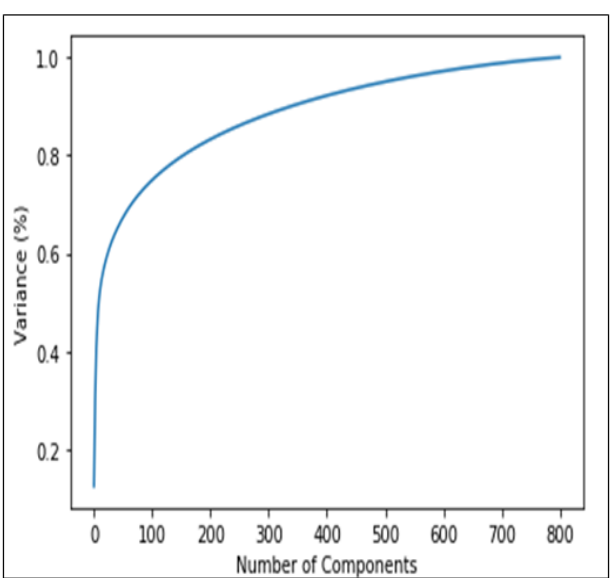

Figure 3. Variance graph.

Bi - directional LSTM

RNN - LSTM model is trained for the RNA gene sequence dataset. RNN model is trained in a sequence that is by looking at the previous state it predicts the current state. For example, if we take the sentence "Ram is playing Kabaddi and his knee got injured". From this sentence, we can see each word is dependent on the previous one, so if we use the RNN model on the half of the sentence like "Ron is playing soccer and" the model will predict the next word based on the all previous words and predict "his" if it is girl name it will predict as "her". Likewise, RNA gene sequence is also trained using the LSTM - RNN model where it learns some patterns from the previous genes and predicts the current gene. LSTM is the same as RNN but RNN forgets the long back previous state if the sequence is long, for example, paragraph, gene sequence. To tackle this problem we used the LSTM model to train the long gene sequence RNA model due to its memory cell tries to store the long back previous information in memory. The processed gene sequence is passed to the LSTM model to do sequence prediction Example first inputs to the model will be Model Sample 1 inputs: gene 1, gene 2 …….gene N Model sample 1 target: "BRCA" The LSTM models first calculate linear transformation Wx + C and apply activation function for first gene 1. Further, it computes the loss or error of the model by comparing model output and actual output we already know. It tries to reduce the loss by updating the weights W of the input gene X where c is the bias term. Finally, that output of gene 1 is carried forward to the next gene 2 by adding the output to the current gene which is going to predict current outputs by looking into the previous gene by applying linear transformation and activation function. Likewise, it repeats the same procedure for all cancer types and finally, it learns the exact gene pattern for each cancer type (Figure 4).

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_3 (Bidirection (None, 400)               1121600
_________________________________________________________________
dense_3 (Dense)              (None, 5)                 2005
=================================================================
Total params: 1,123,605
Trainable params: 1,123,605
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: RNN - LSTM Output .

We used a Bi - directional LSTM layer with one lag of period and the parameters are one input layer, one bidirectional LSTM hidden layer with 400 neurons, and an output layer. The following snapshot shows the sample output of the model training phase (Figure 5).

```
Train on 640 samples, validate on 161 samples
Epoch 1/20
640/640 [==============================] - 2s 3ms/step - loss: 1.4161 - acc: 0.3469 - val_loss: 1.3034 - val_acc: 0.4
907
Epoch 2/20
640/640 [==============================] - 0s 316us/step - loss: 1.2100 - acc: 0.5391 - val_loss: 1.1989 - val_acc:
0.5590
Epoch 3/20
640/640 [==============================] - 0s 330us/step - loss: 1.1025 - acc: 0.6766 - val_loss: 1.1254 - val_acc:
0.6335
Epoch 4/20
640/640 [==============================] - 0s 323us/step - loss: 1.0214 - acc: 0.7781 - val_loss: 1.0654 - val_acc:
0.6770
Epoch 5/20
640/640 [==============================] - 0s 324us/step - loss: 0.9548 - acc: 0.8687 - val_loss: 1.0148 - val_acc:
0.7267
Epoch 6/20
640/640 [==============================] - 0s 311us/step - loss: 0.8971 - acc: 0.9156 - val_loss: 0.9708 - val_acc:
0.7640
Epoch 7/20
640/640 [==============================] - 0s 318us/step - loss: 0.8463 - acc: 0.9422 - val_loss: 0.9306 - val_acc:
0.8075
Epoch 8/20
640/640 [==============================] - 0s 323us/step - loss: 0.8014 - acc: 0.9641 - val_loss: 0.8948 - val_acc:
0.8634
```

Figure 5. Sample Model Training Process Output.

VALIDATION and PERFORMANCE

After conducting trials, we evaluated the proposed model's efficiency using classification accuracy and the Confusion matrix. The accuracy of the classifier is measured by how accurately it predicted the cancer type. The confusion matrix is a common metric for evaluating the efficiency of a classification model. It calculates true accuracy against a classifier. We have used two more evaluation criteria for the best outcomes. Time and number of iterations elapsed. The number of loops refers to the number of rounds taken by the network to train and test data, while the elapsed time refers to the amount of time taken by the network to train and test data.

Model Evaluation

The model on the test dataset was evaluated and elapsed time was measured as shown below Table 2.

Model	Training time	Testing time
Bi-directional LSTM	2 Min 3 Sec	83 Milli sec

Table 2. Elapsed Timing.

The proposed RNN-LSTM has taken a training time of 2 min 3 sec and a testing time of just 83 milli sec. The following shows the confusion matrix obtained (Figure 6).

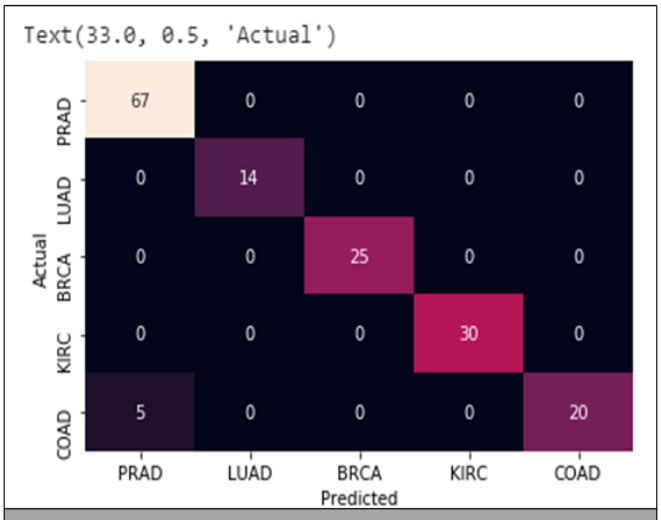

Figure 6. Confusion Matrix.

Performance Measurement

Performance of the proposed system is evaluated using precision, recall and F_1 measure, as indicated in the following Figure 7.

	result_NN
acc	0.968944
FP_rate	0.053191
precision	0.986111
recall	0.960000
F1	0.970584

Figure 7. Confusion Matrix.

A bar chart is plotted for the same as below Figure 8.

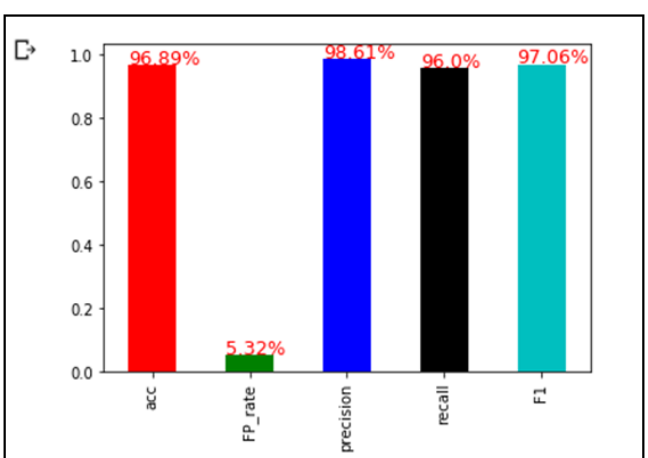

Figure 8. Confusion Matrix.

We also plotted ROC for each type of cancer & the results are illustrated in (Figure 9).

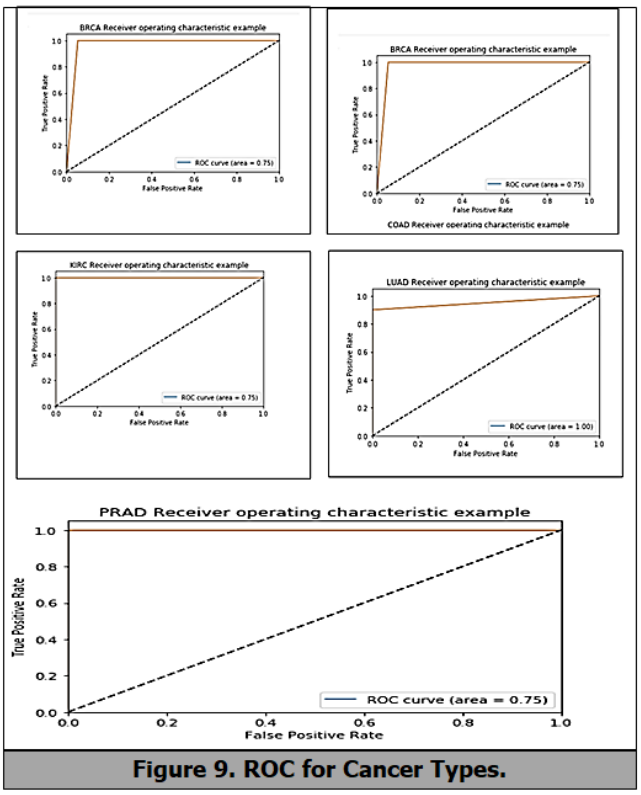

Figure 9. ROC for Cancer Types.

The proposed methodology achieved 96.89 % accuracy, 5.32 % FP rate, 98.61 % precision, 96 % recall, 97.06 % f_1. By the metrics, the novel solutions account to be the better method to predict cancer eliminating mentioned complexities in section - 1.

Comparison Metrics

We also compared the performance of our system with existing systems. Following table shows the comparative analysis (Table 3).

Author	Method	Accuracy
	Genome deep learning	
Sun (2019)		94.70 %
	Logistic Regression	92.10 %,
Shravya (2019)	Support Vector Machine (SVM)	92.23 %,
	K Nearest Neighbor (KNN)	92.78 %
	XG Boost	
Maurizio Polano (2019)		87.80 %
Chang S (2019)	SVM	75 %
Proposed Method	Bi - LSTM	96.89 %

Table 3. Comparative Analysis.

Results of our proposed methodology are compared with existing methodologies proposed by previous researchers who used genome DL, Logistic Regression, Support Vector Machine (SVM), and K Nearest Neighbor (KNN), XG Boost,

SVM (Figure 10).

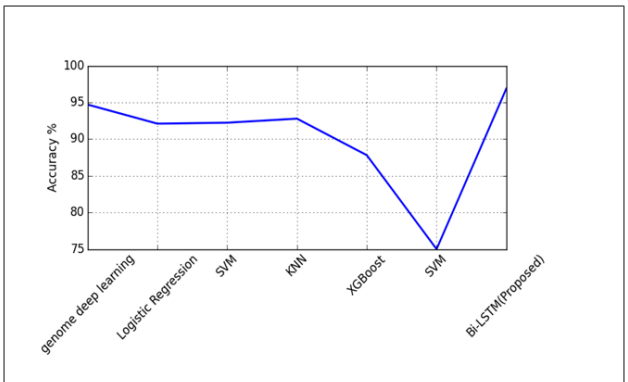

Figure 10. Accuracy Comparisons with Existing Systems.

Table 4 below shows comparative analysis concerning sensitivity and precision.

Methods	Sensitivity	Precision
Logistic Regression	91 %	95.31 %
K Nearest Neighbor	90.32 %	96.55 %
SVM[44]	91.07 %	95.94 %
Gaussian Method	96.80 %	70.09 %
Random Forest	80 %	72.95 %
ANN	59.93 %	84.58 %
SVM - RBF	87.56 %	65.42 %
SVC - W	95.56 %	89.47 %
Proposed methodology	97.06 %	98.61 %

Table 4. Comparison of Sensitivity and Precision.

As compared to above works, our methodology outperformed and generated better results (Figure 11).

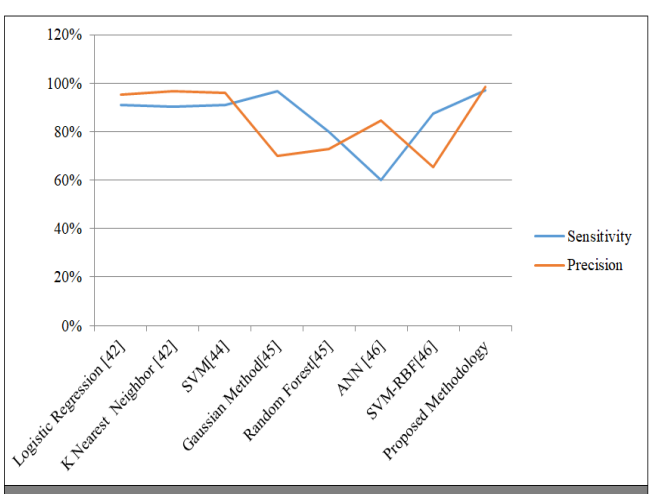

Figure 11. Sensitivity and Precision Comparison Metric.

DISCUSSION

The sensitivity and precision of the Logistic regression, K nearest neighbor, SVM, Gaussian, Random Forest, ANN, SVM + RBF is compared with our proposed method with 97.06 % sensitivity and 98.61 % precision the proposed technique has been proven the best comparatively. The proposed method resulted in the highest precision with Performance metric shows that the proposed methodology outperformed in the evaluations and outcome as compared to existing systems for a cancer type prediction.

CONCLUSION

In this paper, we propose a method to enhance cancer diagnosis and improve classification accuracy from gene expression data. Proposed hybrid of RF and PSO extract desirable features and optimizes parameters which are further passed to PCA to emphasize variation and extract the most significant patterns from a dataset. Furthermore, the Bi - LSTM algorithm learns from the extracted features and predicts the target type of cancer. Experimental results show 96.89 % accuracy and 98.61 % precision. Applying this method to gene expression data and comparing it with baseline algorithms our method not only shows that it can be used to enhance the accuracy but also shows that it can be scaled further to deal with different types of cancer genomic profiles.

FUTURE DIRECTION

Though we achieved 96.89 % accuracy, we would like to address the following issues to extend this work as a part of future direction.

- A generalized cancer prediction system covering all types of cancers can be built.
- Omics and non - omics data can be integrated which will reveal few more clues to ease classification performance. Still, the scope is there to enhance the accuracy of the system.

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers for constructive suggestions and valuable comments which have reshaped the contents of our paper. I express my gratitude towards my research guide Dr. Sharmila K Wagh. She has shaped my view for how research process should go, and inspired me to know the importance of quality research with promising impact. Also I extend my heartfelt thanks to Jayashree R Prasad whose kind advice & few words of encouragement used to enlighten me to work with even more zeal & enthusiasm. Last but not least, thanks to my family members and friends for keeping trust in me and elevating me whenever I was in need.

REFERENCES

1. Bahri Y, Kadmon J, Pennington J, et al. Statistical Mechanics of Deep Learning. Annu Rev Condens Matter Phys 2020;11:501–528.
2. Li H. Modern deep learning in bioinformatics. J Mol Cell Biol 2020;12(11):823–827.
3. Luo P, Ding Y, Lei X, et al. Deep Driver: Predicting cancer driver genes based on somatic mutations using

deep convolutional neural networks. Front Genet 2019;101-113.
4. Chiu YC. Deep learning of pharmacogenomics resources: Moving towards precision oncology. Brief Bio Inform 2020;21(6):2066–2083.
5. Karim MR, Rahman A, Jares JB, et al. A snapshot neural ensemble method for cancer-type prediction based on copy number variations. Neural Comput Appl 2020;32(19):15281–15299.
6. Mallik S, Seth S, Bhadra T, et al. A linear regression and deep learning approach for detecting reliable genetic alterations in cancer using dna methylation and gene expression data. Genes (Basel) 2020;11(8):1–15.
7. Shanthi S, Rajkumar N. Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods. Neural Process Lett 2021;53(4):2617–2630.
8. Tabares - Soto R, Orozco - Arias S, Romero - Cano V, 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;4:22.
9. Daniel E Shumer NJNNPS. HHS Public Access. Physiol Behav 2017;176(12):139–148.
10. Basavegowda HS, Dagnew G. Deep learning approach for microarray cancer data classification. CAAI Trans Intell Technol 2020;5(1):22–33.
11. Greene CS, Costello JC. Biologically Informed Neural Networks Predict Drug Responses. Cancer Cell 2020;38(5):613–615.
12. Hajieskandar AR, 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 Humaniz Comput 2020;1-11.
13. Jang HJ, Lee A, Kang J, et al. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. World J Gastroenterol 2020;26(40):6207–6223.
14. Echle A, Rindtorff NT, Brinker TJ, et al. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 2021;124(4):686–696.
15. Carrasco Pro S, Bulekova K, Gregor B, et al. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. Sci Rep 2020;10(1): 17632.
16. Choi J, Park S, Ahn J. Redon: a reference drug based neural network for more accurate prediction of anticancer drug resistance. Sci Rep 2020;10(1):1–11.
17. Ramirez R. Classification of Cancer Types Using Graph Convolutional Neural Networks. Front Phys 2020;8:1–14.
18. Adam G, Rampasek L, Safikhani Z, et al. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020;4(1):1–10.
19. Wang Y. Identification of a putative competitive endogenous RNA network for lung adenocarcinoma using tcga datasets. Peer J 2019;4:2019.
20. Lathwal A, Kumar R, Raghava GPS. Computer-aided designing of oncolytic viruses for overcoming translational challenges of cancer immunotherapy. Drug Discov Today 2020;25(7):1198–1205.
21. Sanavia T, Birolo G, Montanucci L, et al. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. Comput Struct Biotechnol J 2020;18:1968–1979.
22. Cai J, Yang F, Chen X, et al. Signature panel of 11 methylated mrnas and 3 methylated lncrnas for prediction of recurrence-free survival in prostate cancer patients, Pharmgenomics. Pers Med 2021;14:797–811.
23. Pare L. Association between PD_1 mRNA and response to anti-PD_1 monotherapy across multiple cancer types. Ann Oncol 2018;29(10):2121–2128.
24. Way Gt. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep 2018;23(1):172-180.
25. Chetan MR, Gleeson FV. Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives. Eur Radiol 2021;31(2):1049–1058.
26. Liu J. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 2018;173(2):400-416.
27. Huang S, Yang J, Fong S, et al. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett 2020;471:61–71.
28. Gan TQ. Clinical value and prospective pathway signaling of microRNA-375 in lung adenocarcinoma: A study based on the Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) and bioinformatics analysis. Med Sci Monit 2017;23:2453–2464.
29. Zhang Y, Li Y, Jiang W, et al. The clinical significance of microRNA-122 in predicting the prognosis of patients with hepatocellular carcinoma a meta-analysis validated by the Cancer Genome Atlas dataset. Med 2019;98(13):14810.
30. Samuelefiorini (2016) Gene expression cancer RNA-Seq Data Set.
31. Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl 2014;24(5):1163–1177.
32. Kumar UK, Nikhil MBS, Sumangali K, et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN Comput Sci 2020;1(5):1–14.
33. Mostavi M, Chiu YC, Huang Y, et al. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 2020;13(5):1–13.
34. Tsai MJ, Tao YH. Deep learning techniques for the classification of colorectal cancer tissue. Electron. 2021;10(14).
35. Yu K, Kim BH, Lee PCW. Abstract 5104: Pan-cancer classification on gene expression data by neural network. Cancer Res 2019;5104.
36. Penson A. Development of Genome Derived Tumor Type Prediction to Inform Clinical Cancer Care. JAMA Oncol 2020;6(1):84–91
37. Liang HW. Utility of miR-133 a - 3 p as a diagnostic indicator for hepatocellular carcinoma: An investigation combined with GEO, TCGA, meta-analysis and bioinformatics. Mol Med Rep 2018;17(1):1469–1484.

38. Sun Y. Identification of 12 cancer types through genome deep learning. Sci Rep 2019;9(1):1–9.
39. Shravya CH, Pravalika K, Subhani S. Prediction of breast cancer using supervised machine learning techniques. Int J Innov Technol Explor Eng 2019;8(6):1106–1110.
40. Polano M. A pan-cancer approach to predict responsiveness to immune checkpoint inhibitors by machine learning. Cancers (Basel) 2019;11(10):1–16.
41. Williams JK, Carlson GW, Cohen C, et al. Tumor angiogenesis as a prognostic factor in oral cavity tumors. Am J Surg 1994;168(5):373–380.
42. Aytekin S, Yarman BS and Gokbay IZ. Microarray Gene Expression Data Classification with Random Forest. Int J Eng Sci 2016;3898.
43. Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. Pacific Symp Biocomput 2017;22:219–229.
44. Bhalla S, Kaur H, Dhall A, et al. Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. Sci Rep 2019;9(1):1–16.