Optimized Gene Selection And Classification Of Disease From Genes Data Using Machine Learning

Anjali P. Honmute [1], Shailaja. C. Patil[2]
Department of E&TC Engineering
[1, 2] JSPM'S Rajarshi Shahu College Of Engineering, Tathawade Pune

Abstract- *Cancer is the major leading reason of death around the world. However, the early identification and prediction of a cancer type is very critical for patient's health. The classification of cancer's genes data from different types of disease genes data is very importance in concern with cancer diagnosis and its treatment. Earlier studies on cancer classification have limited diagnostic ability. The recent development of DNA microarray technology has made monitoring of thousands of metagenomes gene expression simultaneously. By using this abundance of gene expression data researchers are exploring the possibilities of cancer classification. There are number of methods proposed with good results, but still many issues are need to be checked. This project presents an overview of RF, XGboost and CNN cancer classification methods and evaluate these proposed methods based on their classification accuracy, computational time and ability to reveal gene information. We have also evaluated the performance of ML algorithm with Deep learning algorithm. The experimental results conclude that XGboost algorithm outperformed compared to Machine learning and recently used deep learning approaches.*

Keywords- Microgenonims genes data • Machine learning, XGBoost , Random forest, Deep learning, Convolution neural network (CNN)

I. INTRODUCTION

Modern deep learning-based techniques has proven to very successful in dealing variety of structure and unstructured data comprising of image, audio, video, text and disease related data. In cancer disease, cells in some tissues undergo uncontrolled division in the body. Because of this condition, malignant growth occurs in the body and cancer effected cells destroy neighbor's healthy tissues and organs. According to National Cancer Institute (NCI) report currently there are more than 200 different cancer types [1]. Recently, cancer is the major reason behind most of the deaths around the world. Generally, about 1 death from 6 total deaths is because of a cancer [2]. Thus, by 2030, the number of new cancer estimated cases per year may increase up to 25 million [3]. However, timely diagnosis of a cancer may save countless lives and billions of dollars. The early identification and prediction of a cancer type is very critical for patient's health and in cancer research [4]. When a cancer is diagnosed at an early stage, treatment is very effective. Previously, classification of a cancer relies on the morphological and clinical techniques [5].Through gene expression data, significant improvement in accurate observation of thousands of cancer genes is made [6]. This approach provided a large amount of data to researchers by whom they can explore a lot of knowledge but it has some issues [7].

The major issues of microarray data are they are noisy, have high dimensionality and low sample size [8]. To select most significant genes related to cancer and to classify cancer type more precisely and accurately is the main challenge of research nowadays [9, 10]. The selected genes provide understanding the disease, enhance the performance of a cancer classification process and reduce the expense of medical diagnosis [11].Gene expression data generally comprises of the huge number of genes, several researchers analyzed and evaluated the cancer classification problem using various data mining, statistical and machine learning- based approaches [12]. However, there are still some issues with these approaches that make the cancer classification a nontrivial task [14–16]. Microarray techniques will lead to a more complete understanding of the molecular variations among tumors, hence to a more reliable classification [17]. A drawback of traditional machine learning (ML) algorithms is that they require pre-engineered organization of raw input data into structured datasets [18]. The inability of certain ML algorithms to use unstructured data has limited their utility in the cancer classification task [19]. Popular algorithms suitable to analyze unstructured data are based on deep learning method of designing neural networks [20]. Deep learning is a branch of machine learning that uses layered architecture to build sophisticated models with the capability to understand complex data [21]. These algorithms learn significant features in the training process, without pre-engineering unstructured data [22]. This ability allowed DL algorithms to outperform against traditional ML algorithms in several fields like computer vision, image classification and speech recognition, etc., [23–25]. Deep learning is very helpful in the diagnosis of cancer at an early stage. Published research of NVIDIA

indicated that deep learning decrease error rate by 85% for diagnosis of breast cancer [26].

Since deep learning is of incredible potential to support medical and paramedical professionals by decreasing the human error rate, helping in diagnosis of cancer and in analysis of complex data. In this work, CNN model is designed for the classification of cancer data. First preprocessing was performed on raw input data then we selected most relevant genes from gene expression data and after that we classified selected genes using convolutional neural network. The objective of this research is the following.

- To develop an algorithm to classify and predict the type of disease using ML.
- To develop an algorithm to classify and predict the type of disease using DL
- To compare the results with existing literature.

The rest of paper is presented as follows. Section 2 provides summarized review of the relevant literature. Section 3 presents comprehensive description of the pro- posed methodology. In Sect. 4, experimental results, analysis and findings of the proposed study are presented. Section 5 provides conclusion, recommendations and future directions in this work.

II. LITERATURE REVIEW

There exists a lot of problems associated during the cancer data classification. In [27], methodology is Cross-validation technique of k-fold methodology, the Convolutional Neural Networked. Convolutional Neural Networks were used to classify tumors without labelling them. Lung, kidney, and brain cancer datasets were used in the procedure's training and testing stages. Using the cross- validation technique of k-fold methodology, the Convolutional Neural Network has an accuracy rate of 96.43%.

In [28], paper it is based on Artificial intelligence, cancer, deep learning, gene expression, Rna-sequences methodology. In this paper a ANN RNA sequence algorithm is implemented for diagnosis and detection of cancer and achieved 95.64% accuracy

In [29], paper algorithm can achieve better classification accuracy 97.7% under low integration and has good robustness with K-Nearest Neighbor; Naïve Bayes; Decision Tree; Genetic Algorithm.

Microarray data have low size of samples and huge number of features so, selecting informative features is crucial before classification task. In [30] This research proposed a neural network algorithm is implemented for diagnosis and detection of breast cancer and achieved 87.64% accuracy. The methodology used in this papers are Machine Learning, Neural Network, Support vector machine.

In this research, a hybrid deep learning model based on Laplacian Score Convolutional Neural Network (LS-CNN) is employed for the classification of given cancer's data. The performance of the proposed system was evaluated on 10 different benchmark datasets using various performance measurement metrics such as accuracy and confusion matrix. The experimental results conclude that proposed LS- CNN model outperformed compared to traditional machine learning and recently used deep learning approaches.

In [31], This paper introduce a new Convolutional Neural Network architecture called Gene expression Network (GeneXNet), which is specifically designed to address the complex nature of gene expressions. Our proposed GeneXNet provides capabilities of detecting genetic alterations driving cancer progression by learning genomic signatures across multiple tissue types without requiring the prerequisite of gene feature selection. Our model achieves 98.9% classification accuracy on human samples representing 33 different cancer tumor types across 26 organ.

Microarray data has millions of genes while the avail- able number of samples are frequently fewer (\ 100) so, gene selection is necessary for classification purpose. In [32] paper algorithms used for classification and prediction of cancer are RNA-sequencing data from TCGA , (DT), kNN (linear SVM), (poly SVM), and (ANN) and All proposed algorithms for the classification of cancer's data achieved average accuracy ranges from (90 to 95%).

In medical field classification of cancer is a hot topic and several machine learning methods were investigated in the literature on different cancer datasets for efficient prediction of cancer. In [33], research designed a localization refinement approach for specific categories of traffic signs. It can detect all categories of traffic signs. It uses Keras with TensorFlow in backend. Algorithm used are Faster R-CNN and MobileNets giving accuracy 97%.

The existing machine learning methods are good for microarray data classification but still hybrid and improved machine learning techniques are required for efficient microarray data classification .In [34] research study, they propose a feed forward gene selection technique, wherein, two feature selection techniques are used one after the other. It is observed that feed forward method of gene selection substantially reduces the feature space compared to mRMR

and Chi-Square, thereby reducing the computational time from 6.54 sec (Chi-Square) and 4.35 sec (mRMR) to 2.54 sec.

This paper [35] presents an overview of various cancer classification methods and evaluate these proposed methods based on their classification accuracy, computational time and ability to reveal gene information. We have also evaluated and introduced various proposed gene selection method. In this paper, several issues related to cancer classification have also been discussed.

From the literature, it is concluded that it is observed that there are many machine learning algorithms are used for classification of cancer based on genes data and observed that RF algorithm gives efficient results also the performance improves using XGboost algorithm. So this project is focused on the XGboost algorithm and CNN prediction technique.

III. PROPOSED METHODOLOGY

The existence We developed a computational tool for metagenomics-based prediction tasks based on machine learning classifiers (i.e., random forests (RFs), XGboost). The tool uses as features quantitative microbiome profiles including species-level relative abundances and presence of strain-specific markers. The framework is fully automatic, including model and feature selection, permitting a systematic and non-over fitted analysis of large metagenomics datasets. Two main kinds of analysis are implemented, i.e., cross-validation (to evaluate the prediction strength of metagenomics data) and cross-study (to evaluate the generalization of the model between different studies). Additionally, the most relevant features are detected for biomarker discovery tasks. Finally, a set of tools is provided to evaluate classification performances in different ways including i) evaluation metrics such as overall accuracy (OA), precision, recall, F1, and area under the curve (AUC); ii) receiver operating characteristic (ROC) curve plots; iii) confusion matrices. Proposed model is implemented with the help machine learning and deep learning algorithms. First step is to import the datasets and process the data. Then split the data into Test data (30%) and Train data (70%) sets. We apply RF and XGboost algorithm for classification of given disease and train the model to distinguish the cancer genes from set of mixed genes dataset. Also apply CNN algorithm of Deep Learning. We use these algorithms to evaluate the prediction accuracy of cancer and finally we compare the result of algorithms.

3.1 Experimental datasets

We have considered 2424 samples from eight studies and six different diseases to assess the independent prediction accuracy of models built on shotgun metagenomics data and to compare strategies for practical use of the microbiome as a prediction tool. Most of these datasets were used in the relevant literature so, we preferred to choose these datasets.

Fig. 1 Phases of proposed classification system

Also these datasets contain multiple records and are quite different from each other.

Fig 3 outlines the particulars of the experimental datasets.

3.1.1 Data preprocessing

Preprocessing is an essential part to properly analyze the input data. Generally, we have solved three problems of the input data during the preprocessing phase. Various tasks in preprocessing include, removal of categorical row and column, removal of missing values from samples and to convert it into numeric feature matrix. We have removed categorical rows (e.g., gene IDs) and categorical columns (e.g., class labels) from the raw input data. After that we generated labels from raw data and stored in separate variables. Missing values were filled with the mean value of the respective column. Finally, we converted 1D input data into a 3D Array and stored it in numeric matrix format so that it can be passed directly to CNN model. After preprocessing of data, we split it into train and test set data with 70:30 ratio.

IV. CLASSIFICATION METHODS

4.1.1 Random Forest (RF)

As Random forest (RF): A random forest classifier is well known as an ensemble classification technique that is used in

the field of machine learning and data science in various application areas. This method uses "parallel ensembling" which fits several decision tree classifiers in parallel, as shown in Fig. 2, on different data set sub-samples and uses majority voting or averages for the outcome or final result. It thus minimizes the over-fitting problem and increases the prediction accuracy and control. Therefore, the RF learning model with multiple decision trees is typically more accurate than a single decision tree based model. To build a series of decision trees with controlled variation, it combines bootstrap aggregation (bagging) and random feature selection. It is adaptable to both classification and regression problems and fits well for both categorical and continuous values.

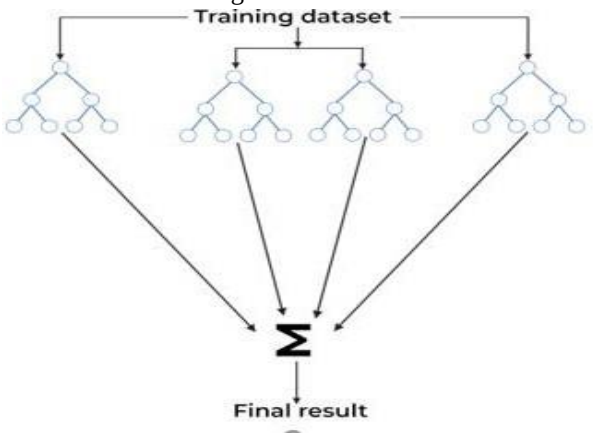

Fig. 2 Proposed RF classification system.

4.1.2 Extreme gradient boosting (XGBoost):

Gradient Boosting, like Random Forests [19] above, is an ensemble learning algorithm that generates a final model based on a series of individual models, typically decision trees. The gradient is used to minimize the loss function, similar to how neural networks [41] use gradient descent to optimize weights. Extreme Gradient Boosting (XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model. It computes second-order gradients of the loss function to minimize loss and advanced regularization (L1 and L2), which reduces over-fitting, and improves model generalization and performance. XGBoost is fast to interpret and can handle large-sized datasets well.

main features provided by XGBoost are:

Parallel Processing: XG Boost provides Parallel Processing for tree construction which uses CPU cores while training.

Cross-Validation: XG Boost enables users to run cross-validation of the boosting process at each iteration, making it easy to get the exact optimum number of boosting iterations in one run.

Cache Optimization: It provides Cache Optimization of the algorithms for higher execution speed.

.

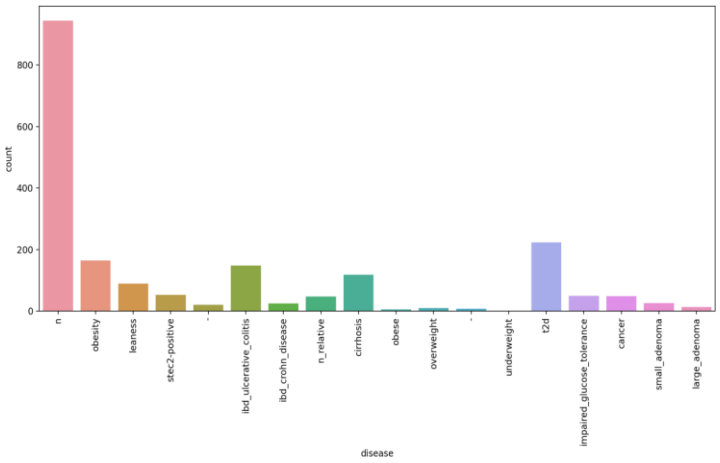

Fig. 3 Experimental dataset

4.2 Convolutional neural network (CNN)

A CNN is a feedforward neural network, which was ini- tially, established in late 90s but at that time for problems like pattern recognition it was considered inefficient. However, with the development in fast and parallel processing, now this technique is better than other traditional machine learning techniques in understanding and classification of raw input data. Generally Convolutional Neural Network (CNN) architecture is divided into three main parts: A convolutional layer is a layer which extracts features from an input image and also helps with operations like noise reduction, blurring, edge detection, sharpening or others that help the algorithm to learn specific characteristics of an image.

To increase the model performance highly non-linear transformation like ReLU is used. A pooling layer helps to reduce the image dimension without losing important features or patterns. A fully connected layer is a layer in which the convolution process is fed through one or more neural layers to generate a prediction. Lastly, Softmax is a probability distribution which is output layer. vision, image classification, and cancer classification tasks.

4.2.1 Proposed CNN architecture

Preprocessing: First, the input data is preprocessed using the LabelEncoder and to categorical functions from the Keras library. The input features are one-hot encoded using the get dummies function from the Pandas library, and the labels are encoded as integers using the LabelEncoder function. The labels are then converted to one-hot encoded vectors using the to categorical function.

Building the model: The model architecture is defined using the Sequential class from the Keras library. The model consists of two 1D convolutional layers with 32 and 64 filters

,respectively, followed by max pooling layers with pool size 2. The output from the second convolutional layer is then flattened and passed through a fully connected layer with 64 units, followed by a final output layer with softmax activation to produce the class probabilities.

Compiling the model: The model is compiled using the categorical cross entropy loss function, the Adam optimizer, and the accuracy metric.

Training the model: The model is trained using the fit function with the input features and labels, and the number of epochs and batch size specified. During training, the weights of the model are updated using backpropagation to minimize the loss function. Evaluating the model: The trained model is used to make predictions on the test data using the predict function. The predicted class probabilities are converted to class labels using the argmax function. The performance of the model is evaluated using the classification _report function from the sklearn .metrics library, which calculates precision, recall, and F1-score for each class, as well as the overall accuracy. Finally, the plot confusion matrix function from the sklearn .metrics library is used to plot the confusion matrix for the model.

4.2.2 Proposed CNN training options

We specified training option after outlining the structure of network. The mentioned network is trained with Extreme Gradient boost having batch size of 32 and max 10 number of epochs. The accuracy is monitored during the training of network by setting validation frequency to 30 for validation data. During validation, data does not update the weights of the network.

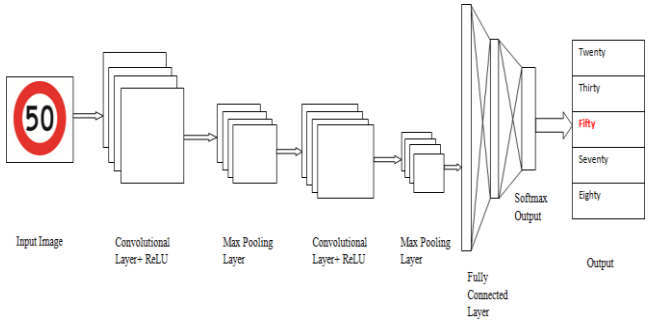

Fig. 4 CNN architecture

V. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 XGBoost :

First we create a list called cols containing the names of all the columns in the DataFrame then creates a list called species that contains only the columns that start with the prefix 'k_'. These are likely to represent microbial species. Now we selects only the columns in the species list from the pd_abundance conv DataFrame, converts them to floating point numbers, and assigns the result back to pd_abundance_conv. This ensures that the species columns are in a numeric format that can be used for analysis and ensures that all the necessary information is included in the same DataFrame. Now we create a dictionary called data_sets that maps disease names to a list of study names. This will be used to identify which samples belong to which disease. Now we separate the dataset with control and no control labeled samples. Then we drop any rows from pd_disease that correspond to a disease in not_disease to ensures that only the relevant diseases are included in the analysis. After selecting only four diseases dataset we split the data in training and testing dataset and apply XGBoost algorithm to classify and predict the disease. The result of this experiment with seed 41 is given in Table 1 and Table 2.

The average ROC AUC score for cancer averaged over 5 folds: 0.9145.

F1 Score : 0.6364, Recall : 0.5833 , Precision : 0.7000

Fold	ROC AUC score (valid set)	F1 score (valid set)	ROC AUC score (test set)	F1 score (test set)	ROC AUC score (full test set)	F1 score (full test set)
1	0.9916	0.8333	0.8914	0.6667	0.8002	0.2222
2	0.8487	0.5455	0.9287	0.6667	0.8044	0.1772
3	0.8447	0.3077	0.9167	0.6364	0.8101	0.1795
4	0.9567	0.7143	0.8969	0.6364	0.8038	0.1750
5	0.9697	0.7273	0.9167	0.5714	0.7682	0.1379

Table. 1 Experimental result for XGBoost.

	ROC AUC Score	F1 Score	Recall	Precision
Cancer	0.9145	0.6364	0.5833	0.7000
Cancer vs. Obesity	0.8726	0.5833	0.5833	0.5833
Cancer vs. Cirrhosis	0.8084	0.3333	0.5833	0.2333
Cancer vs. T2D	0.7890	0.2692	0.5833	0.1750

Table. 2 Prediction result for XGBoost.

5.2 Random Forest :

This section illustrates the results RF algorithm applied for given dataset. This line creates an instance of the Random

Forest Classifier with 30 decision trees and using the default value of 'auto' for the max_features parameter, which means that the square root of the number of features will be used as the maximum number of features considered at each split. The random_state=101 parameter is used to ensure that the classifier is initialized in a reproducible manner. Then we prints a classification report that summarizes the performance of the classifier. The classification report contains various metrics such as precision, recall, F1-score, and support for each class label, as well as the overall accuracy of the classifier. The result is shown in Table 3.

	precision	recall	f1-score	support
cirrhosis	1.00	0.91	0.95	43
WT2D	0.63	0.86	0.73	59
cancer	1.00	0.50	0.67	16
obesity	0.64	1.00	0.78	54
accuracy			0.71	597
macro avg	0.29	0.31	0.29	597
weighted avg	0.57	0.71	0.63	597

Table. 3 Experimental result for RF.

VI. CONCLUSION

In this paper we used XGBoost algorithm, RF algorithm and CNN for classification and prediction of diseases. We use metagenonimes genes dataset for our experiment. Finally, when we compare the result of all machine learning algorithm with deep learning algorithm. Our model is proposed for the classification of different diseases data. The major objective was to analyze the effects of the feature selection methods on the final accuracy of convolutional neural network. We have used classification accuracy and confusion matrix for evaluation parameters for the comparison and better understanding of experimental results. The experimental result showed that accuracy obtained by XGBoost with different seed value ranges from 90-95% while with RF accuracy is 71%. And with CNN accuracy is in the range of 90-100%. Thus we conclude that deep learning algorithm gives higher accuracy as compare to ML algorithm. In future, proposed CNN model can be applied on multi-class image datasets in order to achieve better accuracy results.

REFERENCES

[1]. NIH (2019) National Cancer Institute (NCI), cancer statistics. Available from: https://www.cancer.gov/. Accessed 23 April 2019

[2]. World Health Organization, Cancer (2018) Available from: https://www.who.int/news-room/fact-sheets/detail/cancer. Acces- sed 23 April 2019

[3]. Babu M, Sarkar K (2016) A comparative study of gene selection methods for cancer classification using microarray data. In: 2016 second international conference on research in computational intelligence and communication networks (ICRCICN). IEEE

[4]. Arslan MT, Kalinli A (2016) A comparative study of statistical and artificial intelligence based classification algorithms on cen- tral nervous system cancer microarray gene expression data. Int J Intell Syst Appl Eng. https://doi.org/10.18201/ijisae.267094

[5]. Bolo´n-Canedo V, Sa´nchez-Marono N, Alonso-Betanzos A, Ben´ıtez JM, Herrera F (2014) A review of microarray datasets and applied feature selection methods. Inf Sci 282:111–135

[6]. Hu H, Niu Z, Bai Y, Tan X (2015) Cancer classification based on gene expression using neural networks. Genet MoRes 14:17605– 17611

[7]. Bhola A, Tiwari AK (2015) Machine learning based approaches for cancer classification using gene expression data. Mach Learn Appl Int J 2(3/4):01–12

[8]. Singh RK, Sivabalakrishnan M (2015) Feature selection of gene expression data for cancer classification: a review. Procomput Sci 50:52–57

[9]. Go¨lcu¨k G (2017) Cancer classification using gene expression data with deep learning. Paper presented at Department of Electronics, Informatics and Bioengineering Polytechnic University of Milan, Italy, 20 Dec 2017. http://hdl.handle.net/10589/138427

[10]. Khan MZ, Harous S, Hassan SU, Khan MUG, Iqbal RMumtaz S (2019) Deep unified model for face recognition based on con- volution neural network and edge computing. IEEE Access 7:72622–72633

[11]. Guillen P, Ebalunode J (2016) Cancer classification based on microarray gene expression data using deep learning. In: 2016 international conference on computational science and computational intelligence (CSCI). IEEE

[12]. Bhat RR, Viswanath V, Li X (2017) DeepCancer: detecting cancer via deep generative learning through gene expressions. In: 2017 IEEE 15th international conference on dependable, auto- nomic and secure computing, 15th international conference on pervasive intelligence and computing, 3rd international confer- ence on big data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberS-ciTech). IEEE

[13]. Danaee P, Ghaeini R, Hendrix DA (2017) A deep learning approach for cancer detection and relevant gene

identification. In: Pacific symposium on biocomputing 2017. World Scientific

[14]. Wenyan Z, Xuewen L, Jingjing W (2017) Feature selection for cancer classification using microarray gene expression data. Biostat Biom Open Access J 1(2):55555.

[15]. Dang S, Wen M, Mumtaz S, Li J, Li C (2020) Enabling multi- carrier relay selection by sensing fusion and cascaded ANN

[16]. NIH (2019) National Cancer Institute (NCI), cancer statistics. Available from: https://www.cancer.gov/. Accessed 23 April 2019

[17]. World Health Organization, Cancer (2018) Available from: https://www.who.int/news-room/fact-sheets/detail/cancer. Acces- sed 23 April 2019

[18]. Babu M, Sarkar K (2016) A comparative study of gene selection methods for cancer classification using microarray data. In: 2016 second international conference on research in computational intelligence and communication networks (ICRCICN). IEEE

[19]. Arslan MT, Kalinli A (2016) A comparative study of statistical and artificial intelligence based classification algorithms on cen- tral nervous system cancer microarray gene expression data. Int J Intell Syst Appl Eng. https://doi.org/10.18201/ijisae.267094

[20]. Bolo´n-Canedo V, Sa´nchez-Marono N, Alonso-Betanzos A, Ben´ıtez JM, Herrera F (2014) A review of microarray datasets and applied feature selection methods. Inf Sci 282:111–135.

[21]. Hu H, Niu Z, Bai Y, Tan X (2015) Cancer classification based on gene expression using neural networks. Genet Mol Res 14:17605–17611

[22]. Bhola A, Tiwari AK (2015) Machine learning based approaches for cancer classification using gene expression data. Mach Learn Appl Int J 2(3/4):01–12

[23]. Singh RK, Sivabalakrishnan M (2015) Feature selection of gene expression data for cancer classification: a review. Proc Comput Sci 50:52–57

[24]. Go¨lcu¨k G (2017) Cancer classification using gene expression data with deep learning. Paper presented at Department of Electronics,Informatics and Bioengineering Polytechnic University of Milan,Italy, 20 Dec 2017. http://hdl.handle.net/10589/138427

[25]. Khan MZ, Harous S, Hassan SU, Khan MUG, Iqbal R, Mumtaz S (2019) Deep unified model for face recognition based on con- volution neural network and edge computing. IEEE Access 7:72622–723

[26]. Guillen P, Ebalunode J (2016) Cancer classification based on microarray gene expression data using deep learning. In: 2016 international conference on computational science and compu- tational intelligence (CSCI). IEEE27.

[27]. Hatim Z Almarzouki (2022) Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile

[28]. Surabhi Gupta1, Manoj K. Gupta. Mohammad Shabaz (2022) Deep learning techniques for cancer classification using microarray gene expression data.

[29]. Huijuan Lu, Huiyun Gao, Minchao Ye, Xiuhui Wang (2021) A Hybrid Ensemble Algorithm Combining AdaBoost and Genetic Algorithm for Cancer Classification with Gene Expression Data.

[30]. Pushkar Sathe, Moiz Bombay, Gayathri Sudalai Mani, Dilna Kalathil, Avadhut Phadtare (2020) Cancer Detection using Machine Learning.IRJET.

[31]. Shamveel Hussain Shah1 • Muhammad Javed Iqbal1 • Iftikhar Ahmad2 • Suleman Khan3 • Joel J. P. C. Rodrigues (2020) Optimized gene selection and classification of cancer from microarray gene expression data using deep learning.

[32]. Tarek khorshed, (member, ieee), mohamed n. moustafa , (member, ieee), and ahmed rafea (2020)Deep Learning for Multi- Tissue Cancer Classification of Gene Expressions (GeneXNet)

[33]. Yi-Hsin Hsu; Dong Si (2020) Cancer Type Prediction and Classification Based on RNA-sequencing Data.

[34]. Saima Rathore, 1,3Muhammad Aksam Iftikhar, 1 Mutawarra Hussain (2014) A novel approach for automatic gene selection and classification of gene based colon cancer datasets.

[35]. Salem H, Attiya G, El-Fishawy N (2017) Classification of human cancer diseases by gene expression profiles. Appl Soft Comput 50:124–134

[36]. Liu J, Wang X, Cheng Y, Zhang L (2017) Tumor gene expression data classification via sample expansion-based deep learning. Oncotarget 8(65):109646.

[37]. Lee K, Man Z, Wang D, Cao Z (2013) Classification of bioin- formatics dataset using finite impulse response extreme learning machine for cancer diagnosis. Neural Comput Appl 22(3–4):457– 468

[38]. Wu Q, Boueiz A, Bozkurt A, Masoomi A, Wang A, DeMeo DL, Weiss ST, Qiu W (2018) Deep learning for predicting disease status using genomic data. PeerJ Preprints

[39]. Liu Y, Zhang N, He Y, Lun L (2015) Prediction of core cancer genes using a hybrid of feature selection and machine learning methods. Genet Mol Res 14(3):8871–8882

[40]. He X, Cai D, Niyogi P (2006) Laplacian score for feature selection. In: Advances in neural information processing systems

[41]. Mandal S, Banerjee I (2015) Cancer classification using neural network. Int J Emerg Eng Res Technol 3(7):172–178

[42]. Liu B, Wei Y, Zhang Y, Yang Q (2017) Deep neural networks for high dimension, low sample size data. In: IJCAI

[43]. Kim B-H, Yu K, Lee PC (2020) Cancer classification of single- cell gene expression data by neural network. Bioinformatics 36(5):1360–1366.