Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.47750/jptcp.2023.30.12.028

A Deep Learning Model for Classification of Cancer Types on Gene Expression Data

P. Avila Clemenshia[1*], C. Deepa[2]

[1]PhD Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts & Science, India.

[2]Associate Professor & Head, Department of CS (AI & DS), Sri Ramakrishna College of Arts & Science, India.

***Corresponding author:** P. Avila Clemenshia, PhD Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts & Science, India.

Submitted: 22 March 2023; Accepted: 17 April 2023; Published: 12 May 2023

ABSTRACT

Cancer Classification have great significance in cancer diagnostics and inventions of new drugs. Earlier researches in this area mostly focused on clinical aspects with low diagnostic capabilities. Classifying cancers using gene expression data have the ability to address most preliminary issues associated with diagnosis of cancers or inventions of drugs. Advancements in DNA micro-array approaches have opened the way to monitor thousands of gene expressions. This enterprising quality of gene expression data has been the impetus of this study which examines the feasibility of identifying cancers from gene expression data. For analysis of tumor types, this work proposed DFN Forest (Deep Flexible Neural Forest) model and introduced an improved model for cancer types classification. In this work, Principal Component Analysis (PCA) algorithm is used for dimensionality reduction. Feature selection is done with the help of ICA (Imperialist Competitive Algorithm). DFFN Forest (Deep Fuzzy Flexible Neural Forest), which uses fuzzy logic to update the weight values, is used in this work to classify cancer subtypes. Results from experiments show that the proposed model is effective in terms of metrics like precision, recall, accuracy, and error rate.

Keywords: *Feature selection, Local Optimum, Dimensionality reduction, weight values and Fuzzy Logic*

INTRODUCTION

Cancer is one of the primary reasons behind mortality all over the world. It is an unified set of diseases and every kind of cancer is labelled using the primary part of the body in which the cancer cells develop [1]. Another set of causal genes results in every kind of cancer and the disease develops as a result of the fusion of different mutations that these genes go through. The planning of the treatment to cancer is done in accordance with the mutations that drive [2].

The consequence of unknown or incorrect analysis of these mutations can be wrong treatments and this constitutes one among the important issues that cancer patients face [3]. Genomic data can be used for diagnosing the disease and to identify the various kinds. Genomic tests disclose the gene mutations, which may be the influential factor behind how cancer behaves. This information is useful to the physicians when the specific treatment for the patient has to be determined [4].

Driving mutations are found through an extensive study on genomic data [5]. The entire genome sequences and variant calling are used to analyze the mutation [6]. Both coding and non-coding areas of the DNA are evaluated to discover the mutational signatures of the types of cancer [7,8]. In addition to an elaborate statistical study, machine learning methods may be effective in identifying the driving mutations. For the categorization of cancer, gene expression data [9,10] is a preferred data format. Gene expression data have been used in several publications, and the categorization of the various types of cancer has received attention. The lower size of the samples that are high dimensional is a significant issue with the use of expression data.

Every sample may be thousands of genes but just some are efficient on the target disease, and a majority them have no relevance [11,12]. Gene selection techniques are typically used before classification with the goal of getting rid of the high dimensionality problem. However, the feature selection stage may remove genes, which poses light impact on the development of the disease typically, when they are still importance for diagnosing the specific kinds of cancer for few patients. Moreover, noise is added by the unwanted genes and the classifier performance becomes poor for machine learning techniques.

To overcome this issue in existing work proposed DFN Forest model. In which ABC is used for feature selection. However, ABC based feature selection is time-consuming and produces the local optimum features subset by which it decreases the accuracy of the classifier. To avoid the above-mentioned problems in this work introduced an improved model for cancer types classification. In this work dimensionality reduction is performed based on PCA. Feature selection is done employing ICA. Cancer subtypes classification is done by using DFFN Forest, in which fuzzy logic is utilized for updating the weight values.

PROPOSED METHODOLOGY

This section discusses the proposed cancer types classification model in detail. Proposed model consists of three phases. First one is dimensionality reduction using PCA, second one is gene selection using ICA and the third phase is DFFN Forest based cancer types classification. Figure 1 depicts the overall proposed architecture of this work.

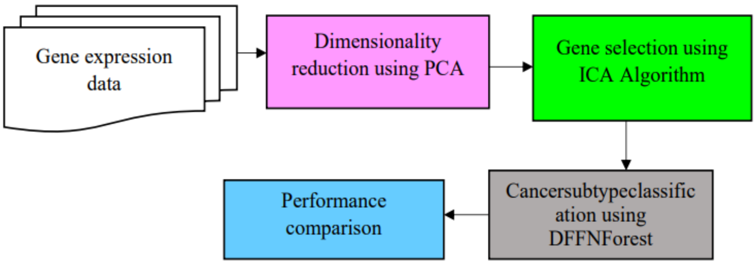
Gene expression data
Dimensionality reduction using PCA
Gene selection using ICA Algorithm
Cancersubtypeclassific ation using DFFNForest
Performance comparison

FIGURE 1: Overall architecture of the proposed model

Dimensionality reduction using Principal Component Analysis (PCA)

Input gene data has more dimensions, and for minimizing the dimensions of gene data in this study, using PCA. In PCA, real variables are transformed into fresh sets of variables called principle components which are created by linear combinations of actual variables, using PCA. The reduced dimensionalities of the dataset are dependent on principal component counts. The mathematical expression of reducing dimensionalities using PCA is given below:

Datasets having n observations and p variables indicate n x p data matrices X. The objective of PCA is to modify actual variables into fresh sets of k variables (principal components) which are capable of acquiring most important changes in data. Principal components are actually linear combinations of original variables and expressed using:

$$PC\_\ 1 = a\_11 * X\_1 + a\_12 * X\_2 + \ldots + a\_1p * x\_p \qquad (1)$$

$$PC\_\ 2 = a\_21 * X\_1 + a\_22 * X\_2 + \ldots\ldots + a\_2p * x\_p \qquad (2)$$

$$\ldots\ldots$$

$$PC\_\ K = a\_K1 * X\_1 + a\_K2 * X\_2 + \ldots\ldots + a\_Kp * x\_p \qquad (3)$$

where aij implies loads or weights of variables xj on principal components PCi, and xj represents jth variable in data matrices X. In the arrangements of Principal components, first components PC1 acquire most important data changes followed by PC2 which acquired second most important changes and so on. This work uses k principal components implying dataset dimensions are reduced to k components.

Gene selection using ICA

After dimensionality reduction it needs to do gene selection. Throughout this research, the ICA has been utilized for reducing the classifier's miss rate. Figure 2 illustrates the flowchart pertaining to the imperialist competitive algorithm. Similar to other evolutionary or population-based algorithms, ICA begins with a first level population. The best nations are chosen to be imperialist nations and the remaining become the imperialist colonies. All the colonies in the first level population are split among the imperialist's nations in accordance with their power [13].

On partitioning of the colonies, they start moving towards controlling colonies where overall controls are held by all settlers who control their colonies. This is indicated by controllers together with proportions of normal controls of their colonies.

On starting colonialist fights between realms, any realm that cannot remain in competition and expanding controls will be eliminated from radical bats. The colonialists fight lead to controls of strongest domains to extend and controls of weaker realms will reduce. The weakest domains mostly have their controls misplaced and, in the long run, break away from one after another. All countries eventually shape states in which fair domains exist in the globe and rest of the countries are colonies of that domain [14,15].

The expression for the establishment of empires (Initialization) are given by [16]:

$$Country\ =\ [p_1, p_2, \ldots, p_{Nvar}] \quad (4)$$

$$Cost\ =\ f\ (Country)\ =\ f\ (p_1, p_2, \ldots, p_n) \quad (5)$$

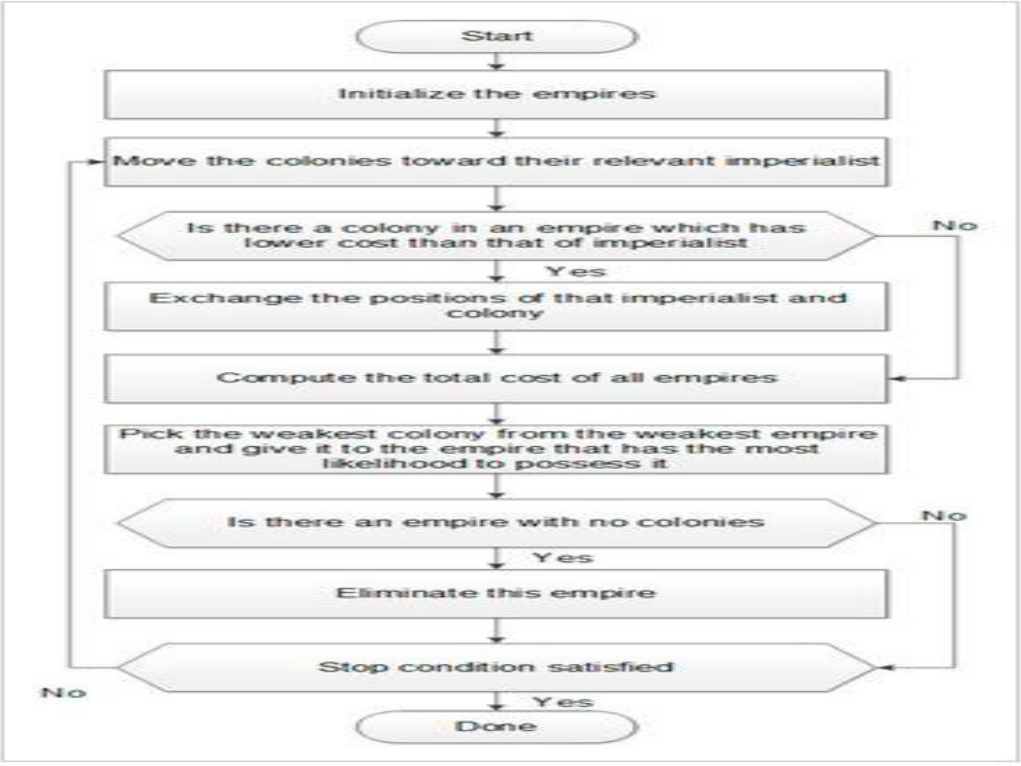
Start
Initialize the empires
Move the colonies toward their relevant imperialist
Is there a colony in an empire which has lower cost than that of imperialist
No
Yes
Exchange the positions of that imperialist and colony
Compute the total cost of all empires
Pick the weakest colony from the weakest empire and give it to the empire that has the most likelihood to possess it
Is there an empire with no colonies
No
Yes
Eliminate this empire
Stop condition satisfied
No
Yes
Done

FIGURE 2: Imperialist competitive algorithm flowchart

$N_{coI} = N_{pop} - N_{imp}$ (6)

$C_n = \underset{i}{max}\{c_i\} - c_n$ (7)

$p_n = \frac{C_n}{\sum_{i=1}^{N_{tmp}} C_i}$ (8)

$NC_n = Round\{p_n N_{col}\}$(9)

Expression (4)Country indicates a nation, N_{var} signifies the number of variables under consideration and p_i refers to ith variable's value, expression(5) indicates costs of countries, expression(6) is critical to discover colonies counts within populaces, expression(7) assists in obtaining normalized costs of imperialists, where cn stands for the nth imperialist's costs, expression(8) computes controls of colonialists where pn alludes to controls that nth imperialists have and expression(9) shows colonies counts of imperialists.

The expressions to move the colonies in the direction of their imperialist nation (assimilation) are given as:

$x - U(0, \beta d)$ ((10)

$\theta - U(-\gamma, \gamma)$(11)

Equation (9) refers to a colony, which goes through distances x towards imperialists, β stands for numbers in the range (1,2) and d indicates distances between colonies and imperialists, Equation (10) is useful in the search carried out for diverse positions surrounding the imperialist, where γ indicates a parameter where a huge value provides the prediction for a global search and a smaller value influences the local search.

The expression for the entire power that an empire holds is as given [16,17]:

$TC_n = Cost(imp) + \xi mean\{Cost(Col)\}$(12)

This expression signifies the overall cost incurred by n-th empire and ξ implies numbers between 0 to 1. Where smaller values of ξ, impact imperialist powers to decide on overall powers that they possess, and a higher value of ξ, impact average powers possessed by colonies in computations of empire's full power.

The expression for the imperialist battle is as below [16]:

$NTC_n = \underset{i}{max}\{TC_i\} - TC_n$(13)

$Pp_n = \frac{NTC_n}{\sum_{i=1}^{N_{tmp}} NTC_i}$ where $\sum_{i=1}^{N_{tmp}} Pp_n = 1$(14)

equation (12) starts with imperialist battles where total costs are calculated while (13) computes the likelihood of having colonies. The primary stages involved within algorithms are summarized within pseudo codes [18,19].

Classification using DFFN Forest

After selecting features they are classified to identify cancer subtypes based on DFFN Forest. FNT (adaptable neural tree) address architectural issues neural networks. Tree structure optimizations could not address multi-class problems, but FNT architecture automatically selected models and increasing the depth of FNT enhances performances. However, they also increase parameter optimization costs. Hence, this work uses DFN Forest framework to overcome the issue.

The cascade arrangement helps to increase the depth of FNT with no addition of other parameters. Figure. 3 implies that the cascade arrangement implies that features are processed in sequential layers, layers can get fresh inputs and new features are merged with unprocessed features as inputs to subsequent levels. Though DFN Forest is based on deep forests, base classifiers are dissimilar DT (Decision trees) are used by deep forests, whereas DFN Forest uses FNT which is considered as the basic classifier in place of DT which are incapable of processing continuous data and may lose important information. Since, biological information have continuity, FNT has been used as the base classifier. Performances of ensembles rely largely on accuracies and versatilities of base classifiers. Assuming three forests and two FNTs are used, Fig. 5 depicts first forest making use of function sets F {+2, +3, +4}, second forest makes use of {+2, +4, +5}, and final one utilizes {+3, +4, +5}. For forests, M-ray methods transform multi-classifications into multiple binary classifications.

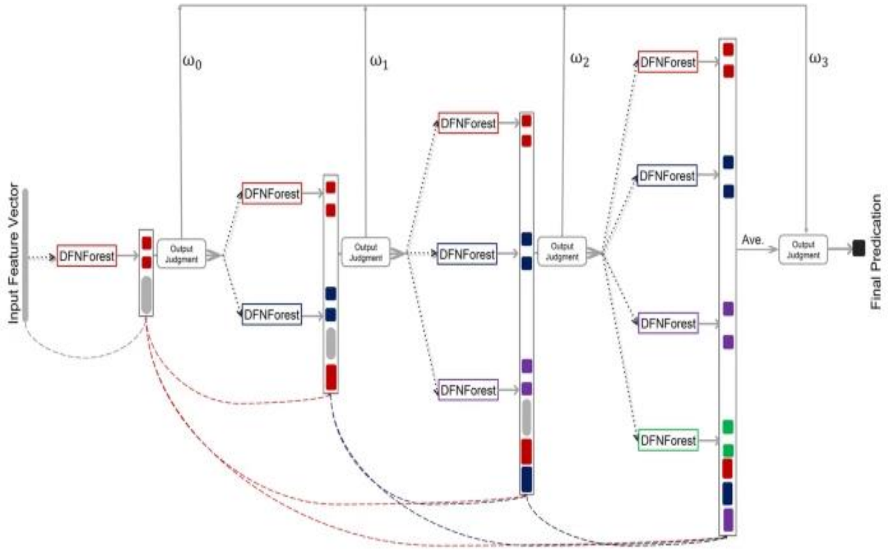

FIGURE 3: Cascading flexible neural forest model

With a sample provided, every FNT yields a predicted value. The concatenation of the values predicted of FNTs in forests in the form of class vectors. Concatenations of class vectors belonging to forests in layers and raw inputs are fed to subsequent layers. Data sets are partitioned into three namely training, validation set, and test sets. Validation sets assist in performance validations of schemas when layers counts vary. The cascades layers or their counts are automatically decided based on variable sizes.

Weight value updating using fuzzy function

This DFFN forest method using the fuzzy model for weight value updating. Assume that the input data presented with 3 classes namely X, Y and Z then the weight values for the gene (feature) will be calculated as given:

$IF\ x1\ is\ A1\ AND\ \ldots xm\ is\ An\ THEN\ C\ is$ X

$IF\ x1\ is\ A1\ AND\ \ldots xm\ is$ Am $THEN\ C\ is$ Y

$IF\ x1\ is$ C1 $AND\ \ldots xm\ is$ An $THEN\ C\ is$ Y

$IF\ x1\ is$ B1 $AND\ \ldots xm\ is$ Am $THEN\ C\ is$ Z.

EXPERIMENTAL RESULTS

In this section the experimental outcomes of the proposed model are discussed. Proposed model is implemented in MATLAB. To show the effectiveness proposed DFFN Forest model which is compared with DFN Forest in terms of Precision, Recall, Accuracy and F-Measure. Proposed work uses prostate cancer dataset that is available on https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15484 and this data has It has 65 samples along with two classes such as grade 8 and 6.Table.1.Produces the performance comparison results.

TABLE 1: Results of performance comparison

Performance metrics	Methods	
	DFN Forest	DFFN Forest
Accuracy	85	93.25
Precision	86.3035	94.48
Recall	86.10	90.05
F-measure	86.20	92.21
Error	15	6.75

Performance metrics

Accuracy

Accuracy refers to the ratio of predictions model that are obtained correct and it is calculated as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (15)$$

Where,

TP signifies True Positive FN-False Negative FP - False Positive TN-True Negative

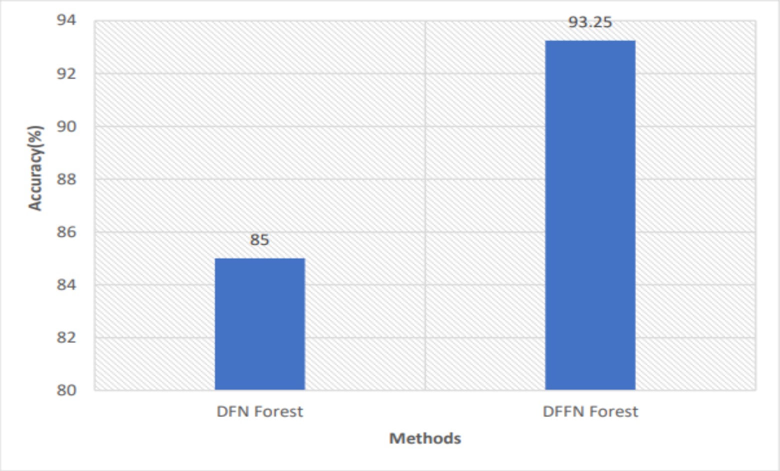

FIGURE 4: Accuracy results

Figure 4. demonstrates the results of the performance comparison between the proposed DFFN Forest and the available DFN Forest in terms of accuracy. In the above figure, the techniques are listed along the X-Axis and the accuracy values are listed along the Y- Axis. From the above figure it is concluded that the proposed framework produces a high accuracy rate. Proposed DFFN Forest model achieves 93.25% accuracy and the existing DFN Forest model produces the 85%.

Precision

Precision is how good the model is at predicting a specific category and it is calculated as

$$Precision = \frac{TP}{TP + FP} \qquad (16)$$

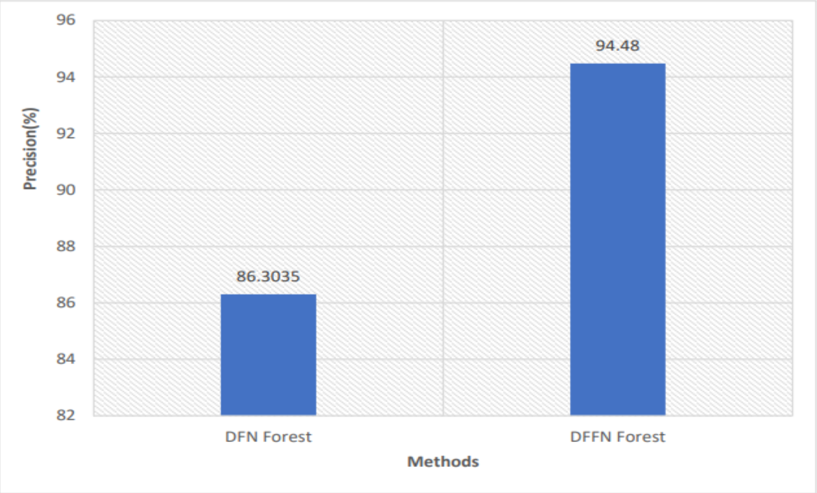

FIGURE 5: Precision results

The results of the Performance comparison between the proposed DFFN Forest and the available DFN Forest in terms of precision is shown in figure 5. In the above figure, the techniques are listed along the X-Axis and the precision values are listed along Y- Axis. From the above figure it is concluded that the proposed model produces high precision rate. Proposed DFFN Forest model achieves 94.48% precision and the existing DFN Forest model produces the 86.3035%.

Recall

Recall measures the completeness of positive predictions and it is calculated as

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (17)$$

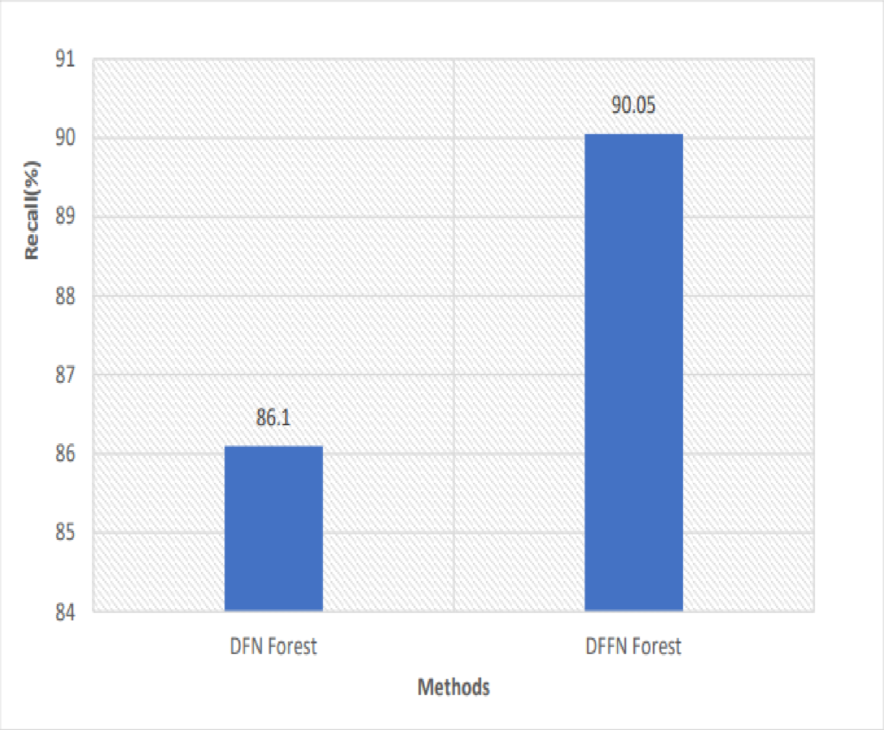

FIGURE 6: Recall results

Figure 6 shows the results of the recall metric performance comparison for the proposed DFFN Forest and the available DFN Forest. In the above figure, the methods are listed along the v and the Y- Axis represents the recall values. From the above figure it is concluded that the proposed model produces the high recall rate. Proposed DFFN Forest model achieves 90.05% recall and the existing DFN Forest model produces the 86.10%.

F-measure

The F-measure is measured by the harmonic mean of precision and recall, with the same weight given to each.

F-measure=2*$(Recall*Precision)$ (18)

$(Recall+Precision)$

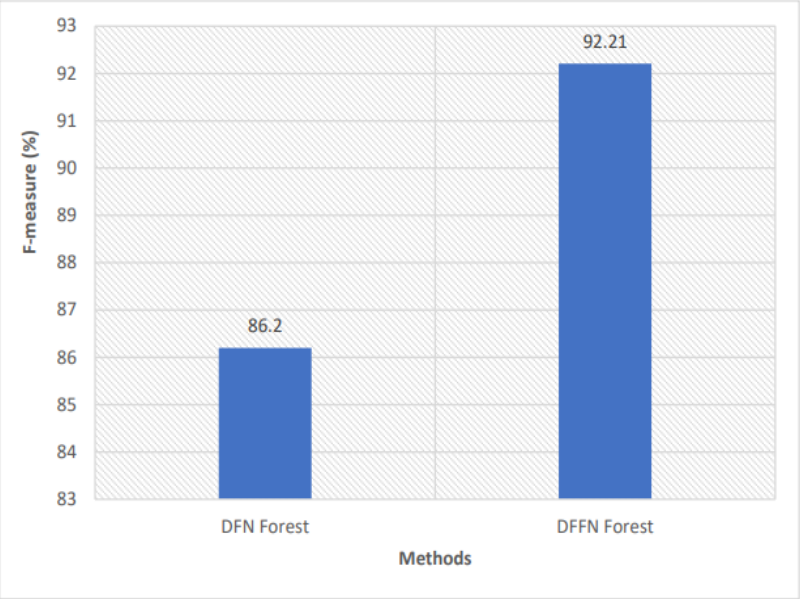

FIGURE 7: F -measure results

Classifier such as proposed DFFN Forest and the existing DFN Forest performance comparison is shown in figure 7. In the above figure, the techniques are listed along the X-Axis and the f -measure values are listed along the Y- Axis. From the above figure it is concluded that the proposed model produces high f -measure results. Proposed DFFN Forest model achieves 92.21% f -measure and the existing DFN Forest model produces the 86.2%.

Error rate

Error rate measures the number of wrongly predicted values and it is calculated as

Error rate = 100- accuracy (19)

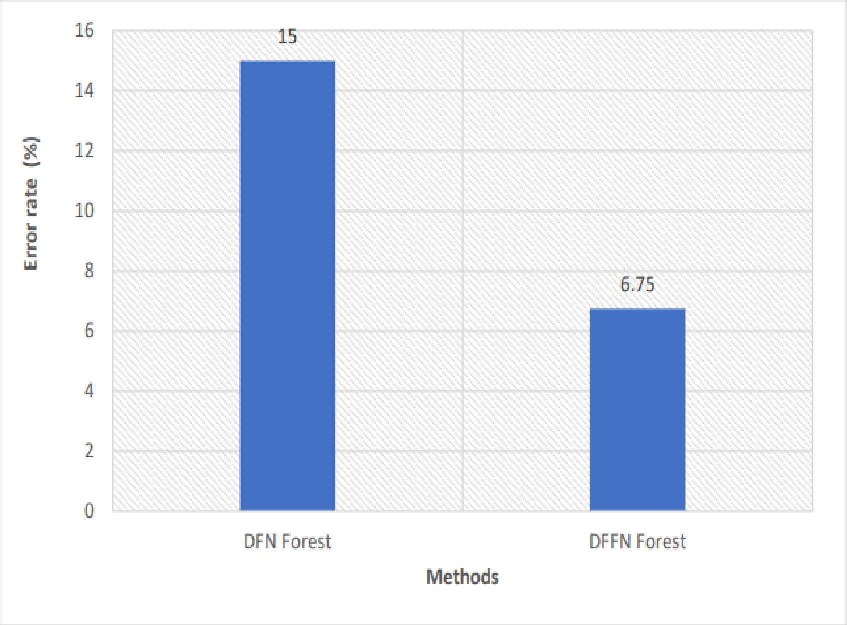

FIGURE 8: Error rate results

Proposed DFFN Forest and the existing DFN Forest is shown in figure 8. In the above figure, the techniques are listed along the X-Axis and the error rate values are listed along the Y- Axis. From the above figure it is concluded that the proposed model produces lower error rate results. Proposed DFFN Forest model achieves 6.75% error rate and the existing DFN Forest model produces the 15%.

CONCLUSION AND FUTURE WORK

Microarray experiments are anticipated to make a substantial contribution towards the improvement in cancer treatments with exact and early stage diagnosis facilitated. This work aimed to provide an improved model for cancer types classification. PCA is applied in this work for dimensionality reduction. Feature selection is computed using ICA. Cancer subtypes classification is carried out applying DFFN Forest. Experimental results show that the proposed model produces the 95.90% accuracy and 95.82% precision which are better than other existing methods. However, input data may have missing values and it affects the classifier performance. Future work planned to use an effective method for pre-processing.

REFERENCES

1. Nidheesh, N., Nazeer, K. A., & Ameer, P. M. (2017). An enhanced deterministic K-Means clusteringalgorithmforcancersubtypepredictionfromgeneexpressiondata. Computersinbiologyandmedicine,91,213-221.
2. Ujjwal Maulik, Anirban Mukhopadhyay andDebasis Chakraborty, “Gene-Expression-Based CancerSubtypesPredictionThroughFeatureSelectionandTransductiveSVM”,Vol.60,issues(4),2013
3. De Kruijf, E. M., Engels, C. C., van de Water, W., Bastiaannet, E., Smit, V. T., van de Velde, C. J., ... &Kuppen, P. J. (2013). Tumor immune subtypes distinguish tumor subclasses with clinical implications inbreastcancer patients.Breastcancerresearchand treatment,142(2), 355-364.
4. Prat,Aleix,EstelaPineda,BarbaraAdamo,Patricia Galván,AranzazuFernández,LydiaGaba,MarcDíez,Margarita Viladot, Ana Arance, and Montserrat Muñoz. "Clinical implications of the intrinsic molecularsubtypesof breastcancer."TheBreast24 (2015): S26-S35.
5. Thanki,K.,Nicholls,M.E.,Gajjar,A.,Senagore,A.J.,Qiu,S.,Szabo,C.,...&Chao,C.(2017).Consensus molecularsubtypesofcolorectalcancerandtheirclinicalimplications. Internationalbiologicalandbiomedicaljournal,3(3),105.
6. Tomczak, K., Czerwińska, P., & Wiznerowicz,M. (2015). The Cancer Genome Atlas (TCGA): animmeasurablesourceofknowledge.Contemporary oncology,19(1A), A68.
7. Finnegan, Timothy J., and Lisa A. Carey. "Gene-expression analysis and the basal-like breast cancersubtype." (2007): pp.55-63.
8. Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007). An immune responsegeneexpressionmoduleidentifiesagoodprognosissubtypeinestrogenreceptornegativebreast cancer. Genomebiology, 8(8), R157.
9. Wong,G.,Leckie,C.,&Kowalczyk,A.(2011).FSR: featuresetreductionforscalableandaccuratemulti-classcancer subtype classificationbased oncopynumber.Bioinformatics, 28(2),pp.151-159.
10. Zhang, W., Feng, H., Wu, H., & Zheng, X. (2017). Accounting for tumor purity improves cancer subtypeclassificationfrom DNAmethylation data.Bioinformatics, 33(17),2651-2657.
11. Gao,Yuan,andGeorgeChurch."Improvingmolecularcancerclassdiscoverythroughsparsenon-negativematrixfactorization."Bioinformatics 21, no. 21 (2005): pp.3970-3975.
12. Jinn-Yi Yeh, Tai-Shi Wu, Min-Che Wu, Der-Ming Chang, Applying Data Mining Techniques for CancerClassification from Gene Expression Data, IEEE International Conference on Convergence InformationTechnology,2007
13. Atashpaz-Gargari, E.; Lucas, C. Imperialist competitive algorithm for minimum bit error rate beamforming. Int. J. Bio-Inspired Comput. 2009, 1, 125–133. [Google Scholar]
14. Rasul, E.; JavedaniSadaei, H.; Abdullah, A.H.; Gani, A. Imperialist competitive algorithm combined with refined high-order weighted fuzzy time series (RHWFTS–ICA) for short term load forecasting. Energy Convers. Manag. 2013, 76, 1104–1116. [Google Scholar]
15. Shamshirband, S.; Amini, A.; Nor Badrul, A.; Mat Kiah, M.L.; Ying, W.T.; Furnell, S. D-FICCA: A density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. J. Int. Meas. Confed. 2014, 55, 212–226. [Google Scholar] [CrossRef]

16. Atashpaz-Gargari, E.; Lucas, C. Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2007), Singapore, 25–28 September 2007; pp. 4661–4667.
17. Nourmohammadia, A.; Zandiehb, M.; Tavakkoli-Moghaddamca, R. An imperialist competitive algorithm for multi-objective U-type assembly line design. J. Comput. Sci. 2012, 4, 393–400. [Google Scholar] [CrossRef]
18. Banaei, M.; Seyed-Shenava, S.; Farahbakhsh, P. Dynamic stability enhancement of power system based on a typical unified power flow controllers using imperialist competitive algorithm. Ain Shams Eng. J. 2014, 5, 691–702. [Google Scholar] [CrossRef]
19. Hadidi, A.; Hadidi, M.; Nazari, A. A new design approach for shell-and-tube heat exchangers using imperialist competitive algorithm (ICA) from economic point of view. Energy Convers. Manag. 2013, 67, 66–74. [Google Scholar] [CrossRef]