European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

AN EFFICIENT MODEL FOR CLASSIFICATION OF CANCER TYPES ON GENE EXPRESSION DATA

P. Avila Clemenshia[1], Dr.B. Mukunthan[2]

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science, (Autonomous), Coimbatore. Email: avilajsph@gmail.com

[2]Associate Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science, (Autonomous), Coimbatore.

Abstract: In recent times, cancer subtype classification has been regarded for its capability of considerably advancing the disease prognosis and progressing the customized administration of patients. Cancer subtypes identification is considered to be the most complicated process, as it faces the challenge of inadequate methods with regards to the accurate determination of gene expressions. Previously, the classification process of cancer subtypes preferred a Deep Flexible Neural Forest (DFN Forest) strategy which is an incorporation of Flexible Neural Tree (FNT) approach. However, it possesses the minimum accuracy due to the inability of choosing the appropriate features and consumption of extended time for classification process. This problem has surpassed by proposing the approach that has developed as Artificial Bee Colony (ABC) with Deep Fuzzy Flexible Neural Forest (DFFN Forest) approach. The proposed research intended to measure the cancer subtypes through introducing novel strategies of feature selection and classifier, as these two processes play a significant role during the cancer diagnosis. During the feature selection process, the Artificial Bee Colony (ABC) algorithm has been used to diminish the miss rate of the classifier, besides the chosen feature has performed in the Deep Fuzzy Flexible Neural Forest (DFFN Forest), in which the Fuzzy function has presented, concerning the advancement of DFN Forest classifier's outputs. The DFFN Forest method included the fuzzy logic for updating the classifier's weight values during the prediction of cancer subtype. The evaluation of proposed algorithm has processed by considering the factors, such as accuracy, precision, recall, f-measure, and error ratio of the classifier.

Keywords: Cancer subtype, Flexible Neural Tree (FNT), Artificial Bee Colony (ABC) and Deep Fuzzy Flexible Neural Forest (DFFN Forest)

1. INTRODUCTION

In general, cancer is a term which collectively indicates the bunch of syndromes allied with unusual growth of a cell that possesses the features of aggression and metastasis [1]. According to the statement from WHO, nearly 8.2 million deaths caused by cancer annually. This ratio has estimated as 13% out of overall deaths occurred around the world, which signifies that cancer is particular among the vulnerable diseases universally.

The forecast of the WHO reveals that the occurrence of cancer might cross 70% in a couple of upcoming decades. Consequently, there is a crucial necessity of analysis to be made on the biological fundamentals of cancer and required the modifications to the current approaches of clinical treatment. Nevertheless, it has recognized beyond 100 cancer types to date, out of which every single type necessitated exclusive diagnosis method and specific treatment. Every cancer type consists of numerous subtypes [2]. So, the identification of cancer subtypes plays a significant role during the process of cancer diagnosis, as well as drug discovery. Therefore, the provision of appropriate treatment and toxicity minimization have entirely relied on the proper identification of cancer subtypes.

The evolution of high-throughput genome analysis methods among the research of cancer subtypes extensively helps in the assessment and clinical treatment of numerous cancer types [3-5]. In the last few years, a huge quantity of expression data (together with genomes, transcriptome, and epigenomes) are gathered and archived in several databases. In particular, the Cancer Genome Atlas (TCGA) [6] has known for its wide-ranging project that contains around 34 cancer types and 15 expression datasets. The approach eases the attainment of genome-scale molecular data, through which the optimization of computational approaches has enabled, as regards the identification of cancer subtypes. Several strategies have developed for identifying cancer subtypes to date, out of which a few approaches rely on single expression data that includes gene expression data [7-8], copy number [9], and DNA methylation [10]. The sparse non-negative matrix factorization (SNMF) and gene expression data have deployed to detect the subtypes among 3 cancer types [11].

The gene selection, and cancer classification have been amalgamated in order to provide optimum accuracy, where allied genes are only included to process the classification. Several feature selection and classification methods have involved to be utilized with respective pros and cons. The framework can deduce the regular benefits from these strategies that comprises the constant increment of accuracy. For feature selection and classification, the data mining methods have been presented [12]. Moreover, the feature selection utilized the Genetic Algorithm (GA) based approach, where the computation of t-statistics (t-GA) has performed to choose associated genes. At this point, the decision tree-based classification has performed, concerning the high accuracy, concurrent to the sustainment of classification time stability though there is changes in the number of genes.

The remaining segments of this study has been arranged as: Section 2 abridges the different methodologies for cancer subtype classification in recent research study. Section 3 describes the proposed strategy. Section 4 refers the utilized dataset and confers the empirical findings. Finally, Section 5 determines the contributions of this research.

2. LITERATURE SURVEY

Guo et al (2017) tends to monitor the cancer subtype classification task over small-scale biology datasets, for which they developed a deep learning system that could be regarded as the altered version of original deep forest model. But still, the developed system varies from the original deep forest model through contributing the two significant approaches, i.e. one is, a multi-class scanning approach can be proposed to equip various simple binary classifiers to boost the diversity of ensemble, simultaneously the appropriate standard of each classifier has taken to account in representations learning; the another one, a boosting methodology can be suggested to focus on the essential features in the cascade forests of representations learning. Hence, the overall

advantages of the discriminative features can be distributed across the layers, concerning the enhancement of entire classification efficiency. The results of the trials on microarray and RNA-seq data sets depict that the performance of proposed approach can surpass many of the existing techniques of classification during the process of classifying the cancer subtypes [13].

Xu et al (2019) opted for the approach, namely a Deep Flexible Neural Forest (DFN Forest) that is innovative association of FNT method, concerning the assistance in cancer subtypes classification. The DFN Forest system varies from the traditional FNT approach, as it converts a multi-classification problem into several binary classification issues over every forest. The cascade structure of DFN Forest has espoused by the developed framework, concerning the derivation of in-depth FNT model devoid of additional parameter inclusion. Besides, this proposed system amalgamates the fisher ratio and neighborhood rough set in order to process the dimensionality reduction in gene expression data regarding to the attainment of efficient classification progress. The empirical findings over RNA-seq gene expression data reveals the superior accuracy (with fewer genes) of developed gene selection strategy.

González et al (2017) presented an innovative strategy, called Case Based Reasoning that accompanies the feature selection based on gradient boosting method. The authors tend to furnish the appropriate diagnosis devoid of increased set of genes, by employing the proposed strategy in the process of squamous cell carcinoma, and adenocarcinoma discrimination. The projected system's generalization ability has tested and trialed with two individual datasets. Besides, it surpasses the conventional methods like microarray analysis, in terms of attaining superior rate of accuracy, by integrating the advanced features (learning over time, adaptability, interpretability of solutions, etc.) into the Case Based Reasoning [15].

Guo et al (2019) projected to confront the cancer subtype classification over small-scale biology datasets, by developing a deep learning method, and BCD Forest. Even though this methodology has regarded as the variant of archetype deep forest model, yet it varies from the original deep forest model through contributing the two significant approaches, i) a multi-class scanning approach has presented to equip various simple binary classifiers to boost the diversity of ensemble, simultaneously the appropriate standard of each classifier has taken to account in representations learning; ii) a boosting methodology has developed to focus on the essential features in the cascade forests of representations learning. Hence, the entire advantages of the discriminative features can be distributed across the layers, concerning the augmentation of overall classification effectiveness [16].

Vasudevan and Thangamani (2018) intended for the identification of cancer subtype through prognosis-enhanced neural network classifier in multigenetic data. Previously, the max-flow/min-cut graph clustering technique helps to identify the candidate cancer subtypes. Ultimately, the classification gets advanced by the proposed prognosis-enhanced neural network classifier, which tends to examine the heterogeneity and diagnose the subtypes of glioblastoma multiforme (GBM), (in short, glioblastoma which is an adult brain tumor and known to be violent in nature) among 215 instances that accompany microRNA expression (i.e. 12042 genes). On the basis of mutations and gene expression, the samples have divided into 4 separate categories, namely mesenchymal, classical, proneural, and neural subtypes. The evaluation of results includes the parameters, like biological stability index, silhouette width, clustering accuracy, precision, recall, and f-measure [17].

3. PROPOSED METHODOLOGY

The proposed research tends to assess the cancer subtypes by providing a novel approach of feature selection, and classification, since these two processes play a significant role during the cancer diagnosis. During the feature selection process, the Artificial Bee Colony (ABC) algorithm has been used to diminish the miss rate of the classifier, besides the chosen feature has performed in the Deep Fuzzy Flexible Neural Forest (DFFN Forest).

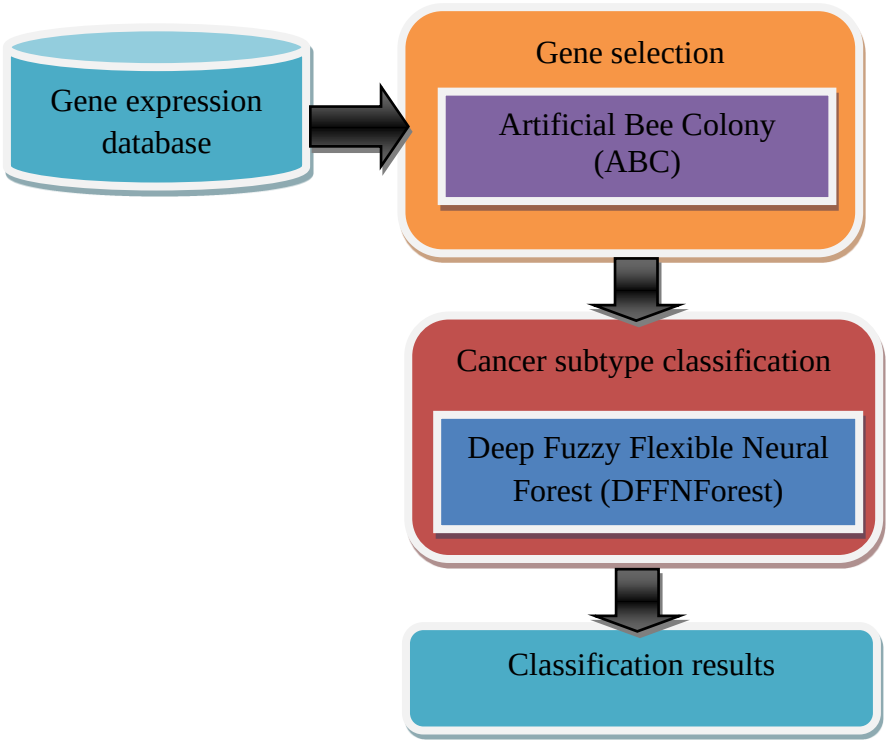

Figure 1: Flow diagram of the proposed system

3.1 Gene expression data

Even though there are thousands of genes in gene expression data, yet the obtainable quantity of samples is inadequate. Some genes only genuinely allied to the cancer subtypes out of numerous features in gene expression data, while the others contemplated being redundant/noisy features. Hence, the process of gene selection can significantly handle the issue of dimensionality reduction, since its efforts to choose the essential genes, at the same time sustain the accuracy in the classification of original genes.

3.2 Gene selection using Artificial Bee Colony (ABC) algorithm

Throughout this research, the Artificial Bee Colony (ABC) algorithm has been used to diminish the miss rate of the classifier. This algorithm is known to be a population-based stochastic optimization that replicates the activities of real honey bees, concerning their characteristic of foraging, in which the food source of bees represented as solutions. There are 3 kinds of bees in ABC, i.e. the employed, the onlooker, and the scout, where the number of employed bees and onlooker bees is same. Employed bee hunts for food sources; returns to the hive; transfers information to onlooker bee through dancing in the dance zone. Onlooker bee observes the

dances; identifies the food sources according to the dance movements. If food sources have forsaken, the employed bees turn out to be a scout and again begins hunting for the further food source.

During the first phase, in preference to the food source, the genes from the gene expression data have been considered as an input to get initialized by the ABC. The D-dimensional vector has expressed as Xi (i=1, 2…, SN) for every solution, in which D represents the number of parameters that requires optimization. The population of the positions (search progress of the employed, onlooker and scout) has iterated up to the convergence of the Maximum Cycle Number (MCN), C=1, 2…MCN.

Eq. (1) helps employed bee to alter the position. During this research, nectar amount has been taken as the accuracy of classification. The employed bee modifies the source position in its memory and explores a different gene position, by which the high classification accuracy has obtained when compared to the former one. Subsequently, the bee replaces its old position with the new one and remembers it, while ignores the previous position. Else, the bee retains the previous position in its memory.

$$v_{ij} = x_{ij} + \emptyset_{ij}(x_{ij} - x_{kj}) \quad (1)$$

in which, the arbitrarily selected indices denoted by k∈ {1, 2,…,SN} and j∈{1,2,…,D}, where k has arbitrarily estimated and expected to be varied from i, while ∅ij is arbitrarily produced number within [-1 and 1].

Post-completion of search process the entire employed bees start exchanging the information, by which the details of genes and their position have been communicated with the onlooker bee. Subsequently, the onlooker bee justifies the accuracy of classification, then determines the genes that accompanies a probability pi associated to the validated classification accuracy, as follows (Eq. 2),

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (2)$$

in which, the fitness value of the solution i has notated by fiti, and the number of genes in the gene expression data have been represented by SN. The position has modified by the employed bee; besides it validates the correctness of the candidate source classification. The onlooker bee remembers the new position, and neglects the previous position if its accuracy found to be lower than the new one.

As expressed in Eq. (3), the scouts swap the gene with forsaken accuracy by the new genes, if the position does not get further improvement. The parameter "limit" has considered as the control parameter for concluding the genes' forsakenness among the predefined number of cycles.

$$x_i^j = x_{min}^j + rand(0,1)(x_{max}^j - x_{min}^j) \quad (3)$$

Algorithm 1: Gene selection using ABC

Input: Number of genes in Gene expression data

Output: Optimal genes

1. Activate the number of genes x_i, i = 1…SN
2. Estimate the veracity of genes' classification
3. Define cycle to 1
4. Repeat
5. FOR each assigned bee
6. Create additional solutions v_i using (1)
7. Determine the veracity of classification
8. Execute the process of greedy selection
9. Estimate the probability p_i for the solution x_i using (2)
10. FOR each onlooker bee
11. Take a solution x_i according to p_i
12. Create novel solutions v_i
13. Determine the veracity of classification
14. Execute the process of greedy selection
15. If impulsive solution occurred for scout bees,
16. then,
17. Switch it by a novel solution
18. Remember the finest solution (optimal genes) attained up till now
19. cycle = cycle + 1
20. Up to cycle = MCN
21. Terminate

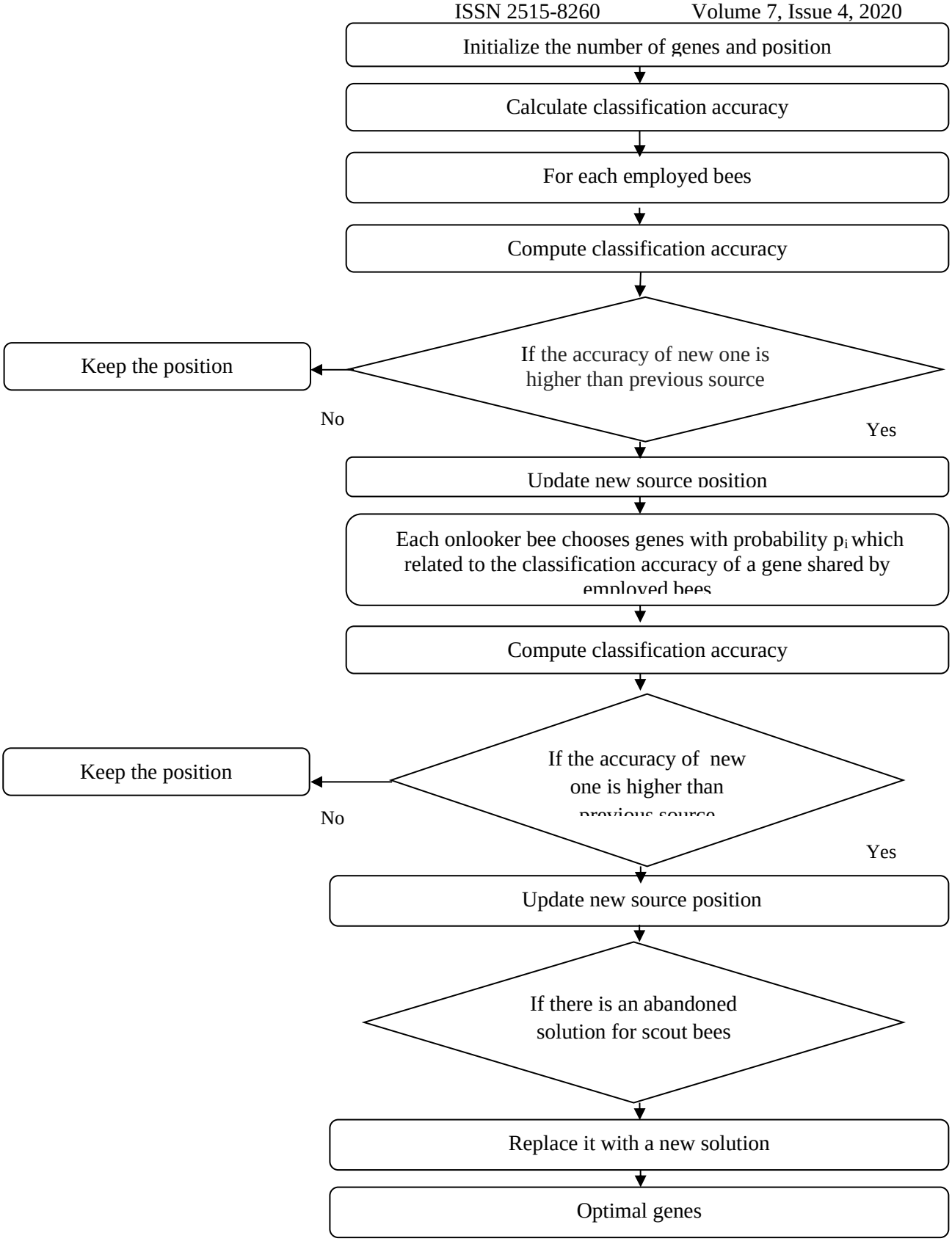

Figure 2: Flow diagram of gene selection approach

3.3. Deep Fuzzy Flexible Neural Forest (DFFN Forest) based classification

Afterwards, the Deep Fuzzy Flexible Neural Forest (DFFN Forest) has been executed with the chosen feature, in which the Fuzzy function has been presented, concerning the enhancement of DFN Forest classifier's outputs. On the other hand, the DFFN Forest method exploits the fuzzy for updating classifier's weight values during the prediction of cancer subtype.

Flexible Neural Tree

The FNT model has been produced using the function set F and terminal instruction set T, as expressed by

$$S=F \cup T =\{ +_2, +_3, \ldots\ldots,,, +_N\} \cup \{x_1, \ldots\ldots, x_n\} \quad (4)$$

in which, instructions of non-leaf nodes with i parameters indicated by $+_i$(i = 2, 3, 4,…,N), whereas the leaf nodes devoid of parameters represented by x1, x2 …, xn. Let the non-terminal instruction be $+_i$(i =2, 3, 4,….,N) to create a flexible neural tree, where the arbitrary generation of i values have processed for non-leaf node and the weights within children have concatenated. The pursuing flexible activation function can possibly be deliberated for the flexible neural tree,

$$f(x)=(1 + e^{-x})^{-1} \quad (5)$$

The following expression generates the result of flexible neuron $+_n$

$$\sum_n = \sum_{j=1}^{n} w_j * x_j \quad (6)$$

in which, the inputs indicated by x_j(j = 1, 2, …,n). The output of node $+_n$ expressed by,

$$out_n = f(sum_n) = ((1 + e^{-sumn})^{-1} \quad (7)$$

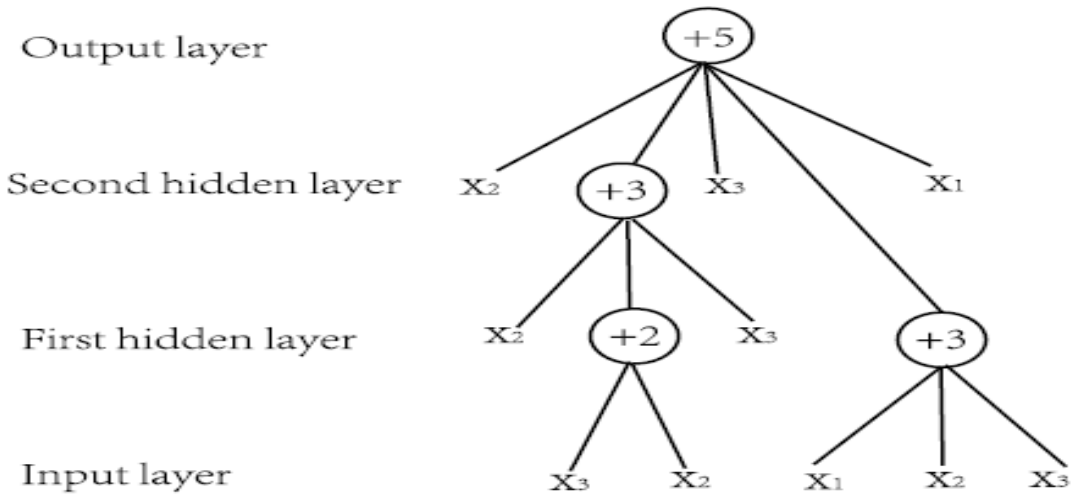

Figure 3: A typical representation of the FNT with function instruction set F={$+_2, +_3, +_4, +_5$}, and terminal instruction set T ={x_1, x_2, x_3}.

Figure 3 demonstrates a standard depiction of the FNT. Additionally, the depth-first strategy is in the place to iterate the overall output of the flexible neural tree from left to right. The FNT is known to be a sparse model that assures optimal generalization performance through consenting the over-layer correlations and identifying the structure spontaneously. We can split the FNT's optimization process into two primary phases, i.e. the optimization of the tree structure, succeeded by the optimization of the parameter.

The Deep Fuzzy Flexible Neural Forest Model

The flexible neural tree has regarded as an exclusive neural network that is capable of optimizing both structure and parameters in an automatic manner. But still, there are some challenges in this method. Primarily, it could not be the best choice to handle the problems of multi-classification, as it solely contains a single root node as an output node. Moreover, the model requires further development to provide optimum performance. Nevertheless, the utilization of this method rises the number of parameters, for which the parameter optimization algorithm has required that eventually increases the performance cost. The aforementioned challenges have conquered by the unconventional flexible neural tree associated approach, namely Deep Fuzzy Flexible Neural Forest (DFFN Forest).

The representation learning and the model complexity play a vital role to enable the deep neural networks for attaining the vast achievement on the tasks of visual and speech recognition. The representation learning defined as layer by layer process of features. The cascade forest structure has espoused by the developed framework, concerning the derivation of in-depth FNT devoid of additional parameter inclusion. The layer by layer process of features could provide the novel features that have amalgamated with original features to process it as an input to the following layer.

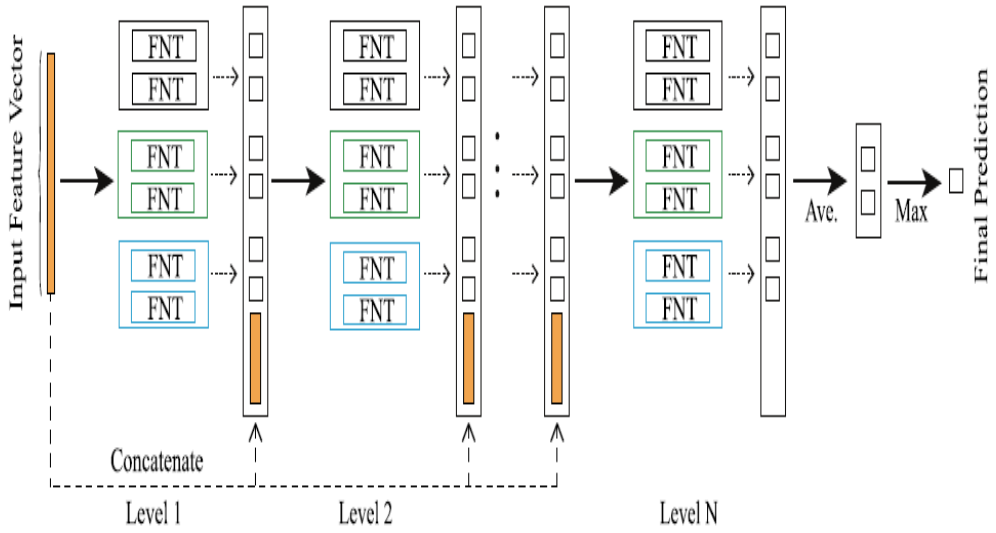

Figure 5: Illustration of the cascade forest structure

Each level of the proposed framework has been made up of FNT. Despite the utilization of decision tree in multi-grain cascade forest (gcForest), yet it contains a drawback, i.e. the direct enforcement of decision tree is not possible with continuous data which required to be divided before the decision tree applied, which may cause the information loss. FNT has considered being an optimum choice as a base classifier since the gene

expression data is known to be continuous data here. The following features of FNT have sustained by the proposed approach, i.e. i) as a sparse model the FNT enables the cross-layer association, through which the overfitting has evaded, and the optimum generalization performance has attained; ii) the structures and parameters have automatically optimized by FNT, besides the multiple FNTs enable the proposed ensemble learning to enhance the comprehensive performance. Various FNT structures have been generated by the system using several grammars, concerning the enhancement of varied ensemble learning. For more clarity, consider that there are three forests and two FNTs in each forest, then the function set F of $\{+_2, +_3, +_4\}$ will be used by first forest, subsequently $\{+_2, +_4, +_5\}$ will be used by second forest, and finally $\{+_3, +_4, +_5\}$ will be used by the last forest (see. Figure 5). Afterwards, the M-ary approach helps to rectify the multi-classification issues of FNT by converting the problem of multi-class into numerous two-class problems in a forest. It can be said in another simple way, i.e. at the instance of the four-class problem every individual forest necessitated to include k = log24 = 2 FNTs. Therefore, the classification problem enabled to define the volume of trees in the forest.

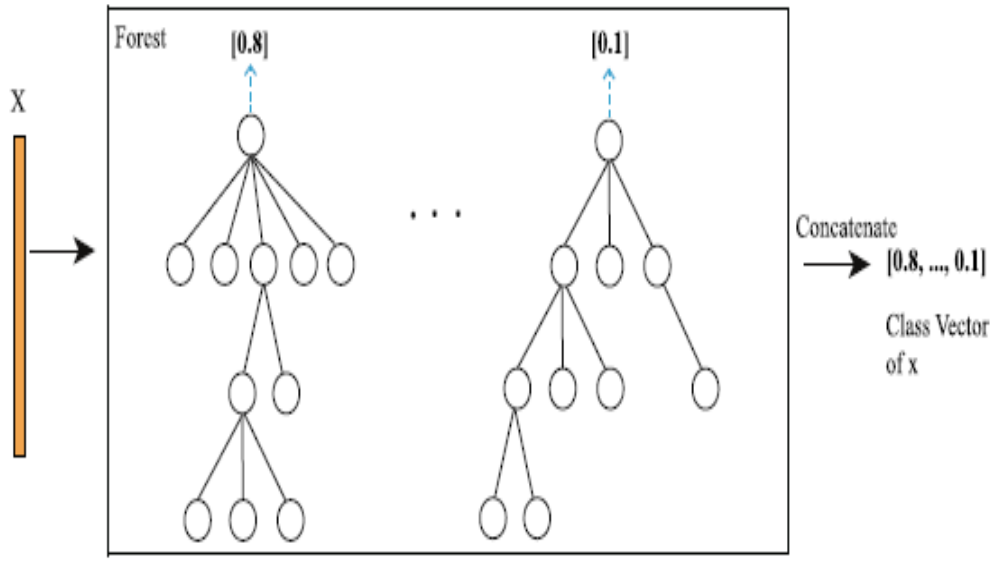

Figure 6: Illustration of class vector generation. Each FNT will generate an estimated value and then concatenate together

Figure 6 represents an instance, in which the approximate value has produced by every FNT in order to organize a class vector which concatenated to the original input feature vector as the source to process in a further phase. For instance, consider if four classes are there, consequently a two-dimensional class vector has been created by every forest; hence, the further level of cascade obtains 6 (2 ×3) enhanced features. Besides, the training set would split into two portions, in terms of assigning them in the processes of training and validation. The validation set verifies the overall cascade at the time of new level inclusion. The increment of further levels will stop when there is no progression of accuracy. Thus, the quantity of cascade levels has ascertained in an automatic way, which enables them to be utilized on various dimensions of datasets, and compatible with small-scale gene expression data.

Weight value updation using fuzzy function

The DFFN Forest method exploits the fuzzy for updating the classifier's weight values during the process of predicting the cancer subtype. A fuzzy if-then approach has considered being the condensed form of fuzzy rule-based classifier like the one that employed in fuzzy control. Let the instance be with 3 classes; then, the designating classification rules erect the weight values of the gene (feature) as follows:

IF x_i is medium AND x_j is small THEN class is 1
IF x_i is medium AND x_j is large THEN class is 2
IF x_i is large AND x_j is small THEN class is 2
IF x_i is small AND x_j is large THEN class is 3

As a numeric the two features and x_j contains the weight values, besides the linguistic values have used by the rules. In the problem, if M possible linguistic values existed for each feature, and n features, the number of possible different if-then rules of this conjunction type (AND) is M^n. A membership function indicates each linguistic value.

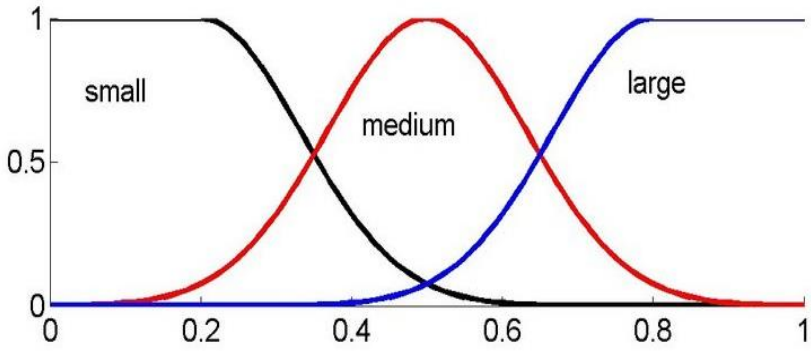

Figure 4: Membership functions for weight values for $\boldsymbol{gene\ x_i}$

An unconventional deep learning approach, namely DFFN Forest offers a replacement for deep neural networks. The tree structure optimization algorithm automatically chooses the FNT structure in every individual forest, and adaptively ascertains the levels of cascade. Though the DFFN Forest approach incorporates the FNT within, yet it rectifies the flaw of FNT in order to handle the multi-classification issues, for which the issues have transformed into several binary classification issues over every forest. Besides, the structure of cascade helps increasing the model depth deprived of any further parameter involvement.

4. EXPERIMENTAL RESULTS

The empirical assessment has been performed in MatLab environment. Here conducted cancer subtype predictions using prostate cancer dataset which is downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15484. It contains 65 samples and two classes such as grade 8 and 6. Considering the metrics, such as Accuracy, Precision, Recall, F-Measure, and Error, the efficiency of the proposed Deep Fuzzy Flexible Neural Forest (DFFNForest) strategy has differentiated from

the approach of Deep Flexible Neural Forest (DFNForest). Table. 1 demonstrates the comparison of the performance.

Table 1: Performance comparison

Performance metrics	Methods	
	DFN Forest	DFFN Forest
Accuracy	84.12	93.75
Precision	84.16	93.75
Recall	84.12	93.75
F-measure	84.14	93.75
Error	15.87	6.25

Performance metrics

4.1 Accuracy

Accuracy is known to be a basic parameter of the performance, which represents the ratio of the observation with proper prediction among the overall observations.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \qquad (8)$$

Where,
TP - True Positive
FN - False Negative
FP - False Positive
TN- True Negative

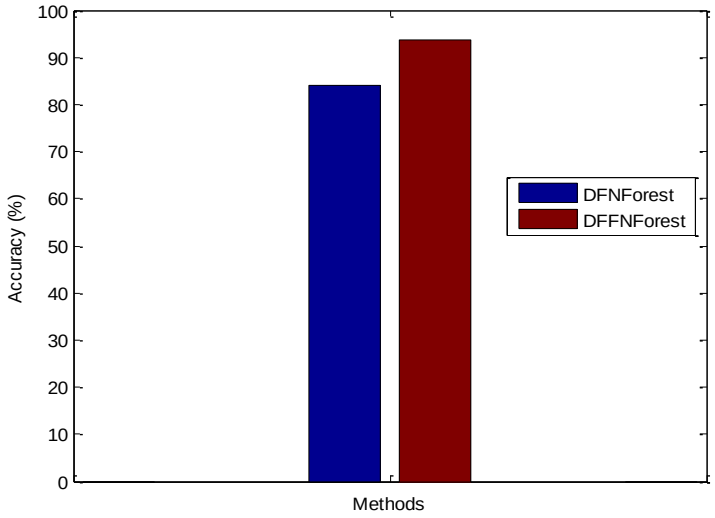

Figure 7: Accuracy comparison

Figure 7 compares the Accuracy of the proposed DFFN Forest approach and current DFN Forest approach, in which the performed methods lie on X-axis, and the Accuracy results stand on Y-axis. The Artificial Bee Colony (ABC) algorithm aids choosing the appropriate genes during this projected research; besides it enhances the rate of Accuracy. The outputs reveal the ability of the proposed strategy to obtain the Accuracy rate of 93.75%, while the DFN Forest approach holds 84.12%.

4.2 Precision

The ratio of the appropriately forecasted positive observations among the overall positive observations forecasted has termed as Precision.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (9)$$

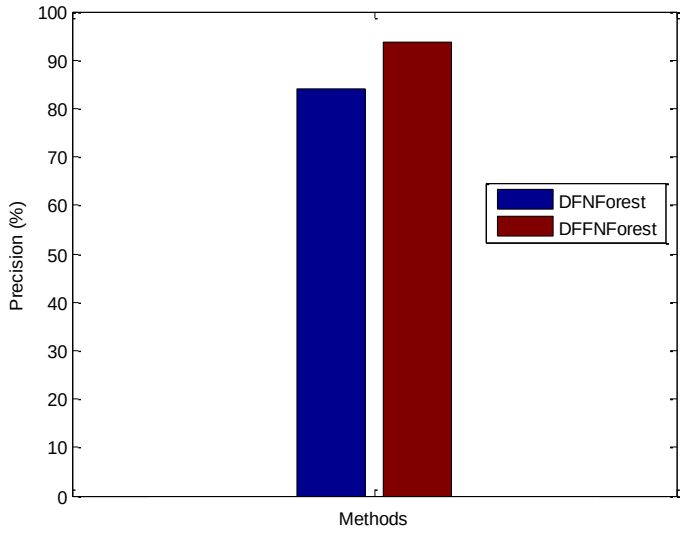

Figure 8: Precision comparison

In Figure 8, the graph illustrates the comparison of Precision obtained by the proposed DFFNForest approach and current DFN Forest approach, in which the outputs depict that the proposed strategy holds the Precision rate of 93.75%, whereas the DFN Forest approach delivers 84.16%.

4.3 Recall

Among the overall observations in actual class – yes, the ratio of appropriately forecasted positive observations has considered to be Recall.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

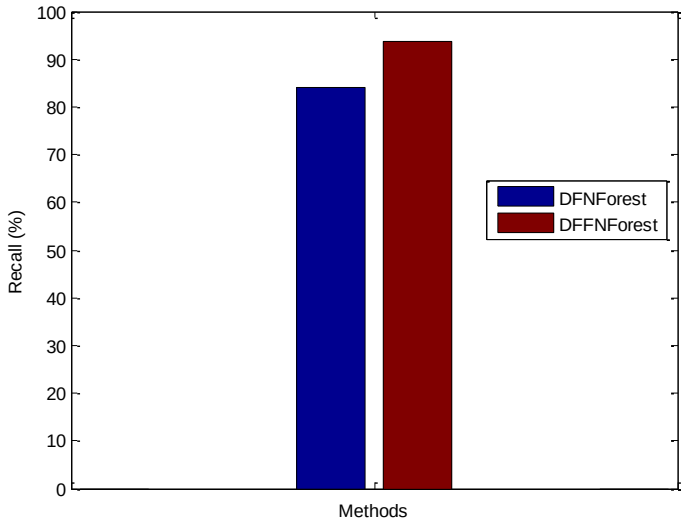

Figure 9: Recall comparison

Figure 9 demonstrates the comparison of Recall values acquired by the proposed DFFN Forest method and current DFN Forest method. Through this proposed paper, the cancer subtypes were classified using DFFN Forest approach, where the fuzzy function aids updating the classifier's weight values that enhances the rate of true positive. The outputs depict that the proficiency of the proposed strategy to reach the Recall rate of 93.75%, while the DFN Forest approach reaches 84.12%.

4.4 F-measure

The weighted average of Precision and Recall is called as F1 score, in which the score considers the false positives as well as false negatives.

$$\text{F-measure} = 2*\frac{(Recall*Precision)}{(Recall+Precision)} \quad (11)$$

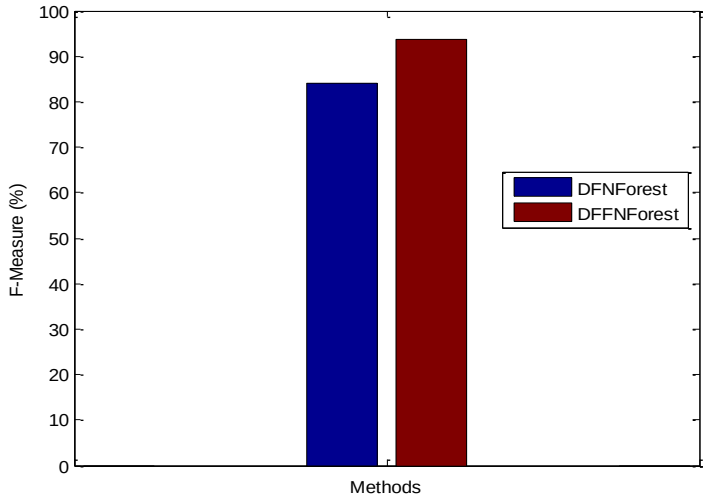

Figure 10: F-measure comparison

Figure 10 compares the F-measure of the proposed DFFN Forest approach and current DFN Forest approach, in which the X-axis represents the strategies performed, and stand on Y-axis stands for the F-Measure results. The outcomes prove that the proposed strategy's capability to procure the F-Measure of 93.75%, conversely the DFN Forest approach holds 84.14%.

4.5 Error

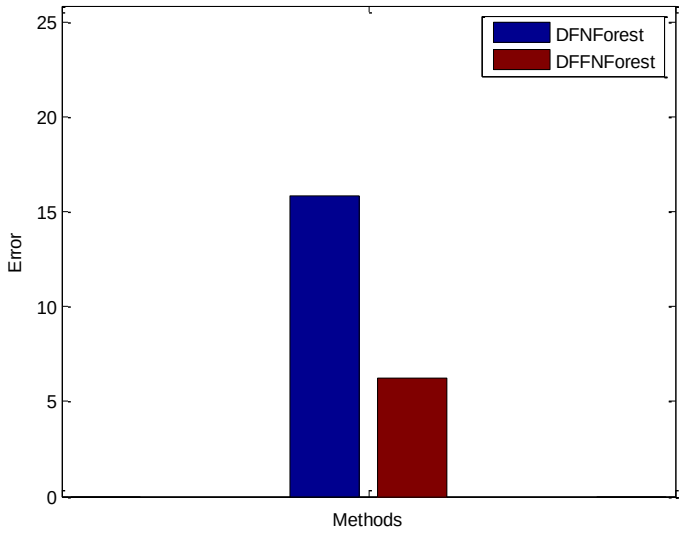

Figure 11: Error comparison

In Figure 11, the graph compares the error rate between the DFFN Forest and DFN Forest systems. The Artificial Bee Colony (ABC) algorithm has exploited to perform the appropriate gene during this research, as it diminishes the classifier's miss rate. Moreover, the above figure represents that the proposed system solely acquires the reduced error rate of 6.25%, while the DFN Forest attains higher error rate of 15.87%.

5. CONCLUSION

Throughout this proposed paper, the overall improvisation of the rate of cancer subtype classification has majorly focused on the strategies, namely Artificial Bee Colony (ABC) accompanied by Deep Fuzzy Flexible Neural Forest (DFFN Forest). Concerning the evasion of overfitting, the proposed framework has developed an Artificial Bee Colony (ABC) algorithm that chooses the informative genes, through which the dataset get rid of noise and redundancy. Accordingly, the chosen informative genes have employed in the proposed strategy in terms of classifying the cancer subtype, besides the involvement of fuzzy ensures the update of classifier's weight values, so that the accuracy of the classification enhanced. It has observed through the empirical findings that the proposed approach can deliver optimum performance than other current strategies by considering the factors, such as accuracy, precision, recall, f-measure and error.

References

1. Nidheesh, N., Nazeer, K. A., & Ameer, P. M. (2017). An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. Computers in biology and medicine, 91, 213-221.
2. Ujjwal Maulik , Anirban Mukhopadhyay and Debasis Chakraborty, "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM", 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 in breast cancer patients. Breast cancer research and treatment, 142(2), 355-364.
4. Prat, Aleix, Estela Pineda, Barbara Adamo, Patricia Galván, Aranzazu Fernández, Lydia Gaba, Marc Díez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. "Clinical implications of the intrinsic molecular subtypes of breast cancer." The Breast 24 (2015): S26-S35.
5. Thanki, K., Nicholls, M. E., Gajjar, A., Senagore, A. J., Qiu, S., Szabo, C., ... & Chao, C. (2017). Consensus molecular subtypes of colorectal cancer and their clinical implications. International biological and biomedical journal, 3(3), 105.
6. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology, 19(1A), A68.
7. Finnegan, Timothy J., and Lisa A. Carey. "Gene-expression analysis and the basal-like breast cancer subtype." (2007): pp.55-63.
8. Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome biology, 8(8), R157.
9. Wong, G., Leckie, C., & Kowalczyk, A. (2011). FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number. Bioinformatics, 28(2), pp.151-159.
10. Zhang, W., Feng, H., Wu, H., & Zheng, X. (2017). Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics, 33(17), 2651-2657.
11. Gao, Yuan, and George Church. "Improving molecular cancer class discovery through sparse non-negative matrix factorization." 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 Cancer Classification from Gene Expression Data, IEEE International Conference on Convergence Information Technology,2007
13. Guo, Y., Liu, S., Li, Z., & Shang, X. (2017, November). Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1664-1669). IEEE.
14. Xu, J., Wu, P., Chen, Y., Meng, Q., Dawood, H., & Khan, M. M. (2019). A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data. IEEE Access, 7, 22086-22095.
15. Ramos-González, J., López-Sánchez, D., Castellanos-Garzón, J. A., de Paz, J. F., & Corchado, J. M. (2017). A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. Computers in biology and medicine, 86, 98-106.
16. Guo, Y., Liu, S., Li, Z., & Shang, X. (2018). BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. BMC bioinformatics, 19(5), 118.
17. Vasudevan, P., & Murugesan, T. (2018). Cancer Subtype Discovery Using Prognosis-Enhanced Neural Network Classifier in Multigenomic Data. Technology in cancer research & treatment, 17, 1533033818790509.

AUTHOR PROFILE

P. Avila Clemenshia received Bachelor of Science in Mathematics from Bharathiar University-Coimbatore, India in 2002 and Master of Computer Applications from from Bharathiar University in the year 2005 and M.Phil from Bharathiar University in the year 2014. She is currently working as a Assistant Professor, Department of Computer Science, Nirmala College for Women(Autonomous), Coimbatore, and her research work focuses on Big Data Analytics, Data Mining. She has five years of teaching experience. She also has two years of programming experience.

Dr.B. Mukunthan pursed Bachelor of Science in Computer Science from Bharathiar University, India in 2004 and Master of Computer Applications from Bharathiar University in year 2007 and Ph.D from Anna University - Chennai in 2013. He is currently working as Associate Professor in Department of Computer Science,School of Computing, Sri Ramakrishna College of Arts and Science (Autonomous),Nava India,Coimbatore since 2017. He is a member of IEEE & IEEE computer society since 2009, a life member of the MISTE since 2010. He has published more than 25 research papers in reputed International journals. He is also Microsoft Certified Solution Developer. His main research work focuses on Algorithms, Bioinformatics, Big Data Analytics, Data Mining, IOT and Neural Networks. He also invented a Novel and Efficient online Bioinformatics Tool and filed for patent. He has 13 years of teaching experience and 11 years of Research Experience.