International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-11 Issue-2, December 2021

Boosted Weighted Optimized Convolutional Neural Network Ensemble Classification for Lung Cancer Prediction

F. Leena Vinmalar, A. Kumar Kombaiya

Abstract: ***One of the major causes of cancer-related mortality worldwide is lung tumors. An earlier prediction of lung tumors is crucial since it may severely increase the death rates. For this reason, genomic profiles have been considered in many advanced microarray technology schemes. Amongst, an Improved Dragonfly optimization Algorithm (IDA) with Boosted Weighted Optimized Neural Network Ensemble Classification (BWONNEC) has been developed which extracts most suitable features and fine-tunes the weights related to the ensemble neural network classifiers. But, its major limitations are the number of learning factors in neural network and computational difficulty. Therefore in this article, a Boosted Weighted Optimized Convolutional Neural Network Ensemble Classification (BWOCNNEC) algorithm is proposed to lessen the number of learning factors and computation cost of neural network. In this algorithm, the boosting weights are combined into the CNN depending on the least square fitness value. Then, the novel weight values are assigned to the features extracted by the IDA. Moreover, these weight values and the chosen features are processed in different CNN structures within the boosted classifier. Further, the best CNN structure in each iteration i.e., CNNs having the least weighted loss is selected and ensemble to predict and diagnose the lung tumors effectively. Finally, the investigational outcomes exhibit that the IDA-BWOCNNEC achieves better prediction efficiency than the existing algorithms.***

Keywords: ***Lung tumor prediction, IDA, BWONNEC, Deep learning, Boosted CNN, Loss function***

I.INTRODUCTION

DNA microarray technology can identify expression of genes at the same time in a single assessment. Genomic testing is designed to gather the proper information in many sectors of molecular biology [1]. With the evolution of technology, it is becoming highly problematic to identify, evaluate, and diagnose a disease, especially lung disease. Lung tumor treatment is one of the most important fields of healthcare study. Lung tumor forecasting is a key problem, and precise forecast will be extremely useful in providing patients with better care.

Manuscript received on April 20, 2021.
Revised Manuscript received on December 27, 2021.
Manuscript published on December 30, 2021.
* Correspondence Author
F. Leena Vinmalar*, Research Scholar, Department of Computer Science, Chikkanna Government Arts College, Tirupur (Tamil Nadu), India. E-mail: anuleena7@gmail.com
Dr. A. Kumar Kombaiya, Assistant Professor, Department of Computer Science Chikkanna Government Arts College, Tirupur (Tamil Nadu), India. E-mail: kumar_kombaiya@rediffmail.com

Data mining techniques are crucial aspects and a more widely accessed technique for achieving a vital function in genomic categorization [2-3]. Genomic features, which represent the biochemical function of the cell, have promising opportunities as a clinical analytical biomarker. Conversely, in the training samples, a quite small classification dataset is usually recommended in comparison to the fraction of proteins associated. Certain classifiers procedures are hampered by restraints on training samples [4]. Feature extraction methods are applied to appropriately obtain the genetic variants that impact detection rate by isolating irrelevant and repetitive attributes [5]. To find the prominent attributes from microarray data, the Nested Genetic Algorithm (NGA) has been suggested [6]. The NGA was split into 2 parts: outer GA (OGA) and inner GA (IGA), wherein genomic databases and DNA methylation databases have been collected to extract the features, which had been included in Random Subspace (RS), Artificial Neural Network (ANN), and Sequential Minimal Optimization (SMO) for diagnosing the lung tumor. Nevertheless, NGA seems to have a greater computational burden and a random velocity issue, which influence the efficiency of screening lung tumor. As a result, an IDA [7] has been developed to select the most pertinent attributes for diagnosing the pulmonary tumor. A Brownian velocity has been applied in IDA to simulate the search strategy for a best possible dragonfly solution. Furthermore, the particle best (pbest) and global best (gbest) principles of Particle Swarm Optimization (PSO) were considered to modify the feature space and minimize unnecessary convergence. For lung tumor forecasting, the extracted features have been applied into the RS, ANN, and SMO classifiers. Although this SMO classification outperforms the RS and ANN in terms of classification results, it falls weak in terms of loss function and precision. Therefore, an IDA-BWONNEC algorithm [8] has been developed for further minimizing the loss function and increasing the precision. At the beginning, the most appropriate attributes were extracted by the IDA and these attributes were fed to the BWONNEC classification process which defines the tunable weights associated with the choice of every ensemble classification in a dynamic manner. This was based on the ensemble classification results and the correlation among the results of each ensemble classification. Then, a weak classification having a less weighted loss was computed and a novel component depending on the loss value was measured.

At last, a resultant ensemble learning classification was employed to the novel component for predicting and diagnosing the lung tumors effectively. On the other hand, it has unique problems related to the number of learning factors in neural network and its computation cost. Hence in this paper, a BWOCNNEC algorithm is proposed to lessen the number of learning factors and computation cost of neural network. In this algorithm, the boosting weights are combined into the CNN depending on the least square fitness value. Then, the novel weight values are assigned to the features extracted by the IDA. Moreover, these weight values and the chosen features are processed in different CNN structures within the boosted classifier. Further, the best CNN structure in each iteration i.e., CNNs having the least weighted loss is selected and ensemble to predict and diagnose the lung tumors effectively. The rest of the paper is organized as follows: Section 2 present the works related to the classification of the microarray data. Section 3 explains the methodology of BWOCNNEC algorithm and Section 4 displays its efficiency. Section 5 summarizes the entire work.

II.LITERATURE SURVEY

A novel categorization method [9] has been suggested for classifying the tumor from genomic data by applying a Modified K-Nearest Neighbor (MKNN) in 2 different cases: least MKNN and largest MKNN. The major concept was to apply strong adjacent from training sample using a novel weighting mechanism. But, it takes long training time for large-scale datasets and also the efficiency was depending on the number of samples in the dataset.

A new algorithm called C-HMOSHSSA [10] has been designed to select the gene using multi-objective spotted hyena optimizer and salp swarm algorithm. The major goal was to design a tumor categorization model for predicting the appropriate and novel tumor biomarkers by choosing the optimal genes. But, it does not handle the large-scale issues with high complex medical databases.

A computational gene choice framework [11] has been developed to categorize the mircroarray data using adaptive hypergraph embedded dictionary learning. First, a dictionary was trained from the attribute space of actual high-dimensional genomic data and this trained dictionary was used for representing the actual genes with a restoration coefficient matrix. After, $l_{2,1}$-norm regularization was used for imposing the row sparsity on the coefficient matrix to choose discriminate genes. Besides, a hypergrpah was adaptively trained and embedded into this framework for obtaining the local multiple geometrical structures of actual genomic data in a high-order fashion. Further, an iterative upgrading method was developed to solve the optimization issue. But, its efficiency was less when using large-scale genomic dataset having more samples and gene sizes.

A novel bacterial colony optimization with multi-dimensional population [12] has been suggested to choose the features and categorize the microarray genomic data. The main goal was to prevent the local optima and speedup the convergence to the optimal solution during iteration by applying various topological transfer structures among interior sets. Originally, a population with the multi-dimensional forms was generated for representing different sizes of attribute subgroups. Then, bacteria from the population were split into a sequence of sets and the attributes were chosen by their contributions to the most possible solutions in the overall population. After, a search was performed for the optimal attribute subgroups with different sizes. But, it has high computational burden.

An improved ensemble method [13] has been designed depending on teaching learning-based optimization and gravitational search algorithm for deciding the genes to categorize the tumors. First, the least redundancy highest similarity and an improved ensemble method were applied to select the significant genes and the informative genes from the genomic datasets. Also, a novel encoding method using gravitational search algorithm was incorporated for transmuting the continuous exploring space to binary exploring space. Moreover, a naive bayes classification was applied as an objective factor for deciding the very sensible genes which support to categorize the tumor precisely. But, its accuracy was not effective for all biological databases.

An unsupervised attribute choice method [14] has been designed to categorize the multi-class tumors from genomic information. First, the genomic profile information was gathered and the appropriate genes were extracted by merging GA. After that, these genes were fed to the extreme learning machine classifier for multi-class tumor categorization. But, an appropriate search space was not ensured since the amount of features was very high.

A novel ensemble multi-population adaptive GA [15] has been developed to neglect inappropriate genes and categorize tumor precisely. First, an ensemble gene choice scheme was applied for filtering the noisy and unwanted genes in high-dimensional data through merging multi-layer and F-score schemes. After, an adaptive GA was conducted depending on multi-population method with support vector machine and naive bayes classifications as an objective factor for choosing the very judicious genes from the reduced datasets. But, it solves only a single objective issue and applied to only optimization issues in the continuous domain.

III.PROPOSED METHODOLOGY

In this section, the IDA-BWOCNNEC is explained briefly for predicting the pulmonary tumor from genomic data. Primarily, the gene microarray data is gathered and the IDA is performed to choose the appropriate attributes in the gathered genomic data. After, the chosen attributes are fed to the BWOCNNEC as its input to predict the pulmonary tumors effectively. The systematic representation of IDA-BWOCNNEC algorithm for lung tumor prediction is illustrated in Figure 1.

Retrieval Number: 100.1/ijeat.D25200410421
DOI: 10.35940/ijeat.D2520.1211221
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

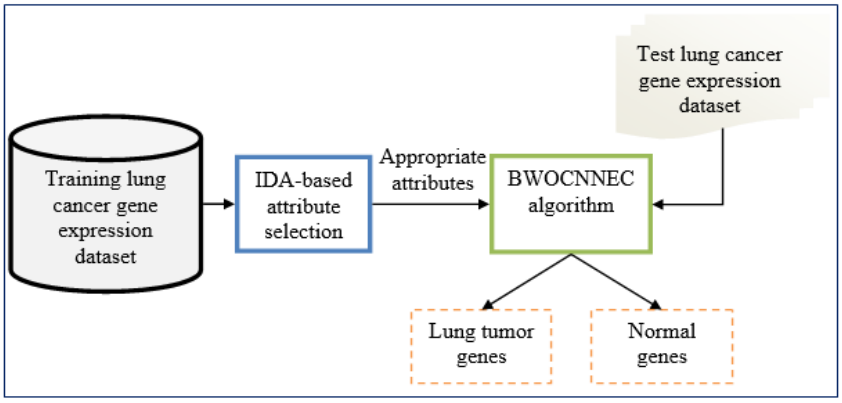

Figure 1. Systematic Representation of IDA-BWOCNNEC-based Lung Tumor Prediction using Genomic Data

3.1 Attribute Selection using IDA

First, a lung tumor genomic dataset is obtained and IDA i.e., an enhanced version of dragonfly optimization is applied to choose the most appropriate attributes from this dataset. In IDA, 2 significant attributes are considered for enhancing the efficiency of standard dragonfly optimization. One attribute is an interior memory which forecasts the promising outcomes i.e., the most appropriate attributes which have probable to converge to global optimum. The other attribute is iteration range ensembling PSO which executes on this set of stored outcomes. Also, it utilizes the socio-cognitive characteristics of PSO to achieve quick convergence and compute the global best outcome.

Meanwhile, it has the searching ability of dragonfly optimization for applying the attribute/search space in an effective manner. Based on this IDA, the most appropriate attributes are chosen successfully and they are given as input to the BWOCNNEC algorithm for categorizing the genomic data effectively.

3.2 Boosted Weighted Optimized CNN Ensemble Classifier for Lung Tumor Prediction

After extracting the most appropriate attributes from the genomic data, they are classified by the BWOCNNEC to predict the lung tumors with better accurateness and less error. First, a multiclass classification is a mapping $F: \mathcal{X} \rightarrow \{1 \dots M\}$ that maps a data x_i to its class $z_i \in 1 \dots M$. Because this is not a continuous mapping, a classifier $F(x)$ is typically learned by training a predictor $f: \mathcal{X} \rightarrow \mathbb{R}^d$ for few d and $F(x)$ is executed by

$$F(x) = \underset{k=1\dots M}{\operatorname{argmax}} \langle y_k, f(x) \rangle \quad (1)$$

In Eq. (1), y_k denotes the unit vector that refers to the label of k^{th} class and $\langle \cdot, \cdot \rangle$ denotes the dot product. For every k, a predictor $f_k(x): \mathcal{X} \rightarrow \mathbb{R}$ is learned for discrimination between data of that class versus others. To categorize a fresh data $\hat{x}$, $f_k(\hat{x})$ is determined for all $k = 1 \dots M$ and the class of major predictor is declared as the label. This process is equal to (1) by describing $f(x) = [f_1(x) \dots f_M(x)] \in \mathbb{R}^M$ and $y_k = 1_k \in \mathbb{R}^M$ i.e., k^{th} element is 1 and the remaining are 0. Generally, the selection of labels is not limited to the canonical basis in $\mathbb{R}^M$ and it is promising to utilize any group of M different unit vectors $y_1 \dots y_M \in \mathbb{R}^d$. Consider that $d = M$ and $y_k = 1_k$.

Multiclass boosting is a process that integrates many multiclass predictors $g_i: \mathcal{X} \rightarrow \mathbb{R}^d$ for creating a robust group $f(x)$ of classifiers i.e., $f(x) = \sum_{t=1}^{N} \alpha_t g_t(x)$ where g_t and α_t are the weak leaner and coefficient chosen at t^{th} boosting iteration. In this work, gradient descent multiclass boosting approach is utilized which learns a boosted predictor $f(x)$ by reducing risk of classifier as:

$$\mathcal{R}[f] = E_{X,Z}\{L(z, f(x))\} \approx \frac{1}{\|\mathcal{D}\|} \sum_{(x_i, z_i) \in \mathcal{D}} L(z_i, f(x_i)) \quad (2)$$

$$Where\ L(z_i, f(x_i)) = \sum_{j=1, j \neq z_i}^{M} e^{-\frac{1}{2}\left[\langle y_{z_i}, f(x_i) \rangle - \langle y_j, f(x_i) \rangle\right]} \quad (3)$$

In Eq. (2), $\mathcal{D}$ refers to the group of training data. The reduction is through gradient descent in function space. It initiates with $f(x) = 0 \in \mathbb{R}^d, \forall x$ and iteratively determines the directional derivative of the risk (2) to modify $f(x)$ along the direction of $g(x)$ as:

$$\delta\mathcal{R}[f; g] = \frac{\partial \mathcal{R}[f + \varepsilon g]}{\partial \varepsilon}\Big|_{\varepsilon=0} = \frac{1}{2\|S\|} \sum_{(x_i, z_i) \in \mathcal{D}} \sum_{j=1}^{M} g_j(x_i) w_j(x_i) \quad (4)$$

$$\text{Where } w_k(x_i) = \begin{cases} -e^{-\frac{1}{2}\left[f_{z_i}(x_i) - f_k(x_i)\right]}, & k \neq z_i \\ \sum_{j=1 | j \neq k}^{M} e^{-\frac{1}{2}\left[f_{z_i}(x_i) - f_k(x_i)\right]}, & k = z_i \end{cases} \quad (5)$$

After, it trains a weak learner g^* which reduces (4) and the optimum step size along g^* is determined as:

$$g^* = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, \delta\mathcal{R}[f; g] \quad (6)$$

$$\alpha^* = \underset{\alpha \in \mathbb{R}}{argmin}\, \mathcal{R}[f + \alpha g^*] \quad (7)$$

At last, the boosted predictor $f(x)$ is modified as:

$$f = f + \alpha^* g^* \quad (8)$$

An integration of boosting and CNNs is done by using CNNs as weak learners in gradient descent-multiclass boosting algorithm. In this scenario, the weak learner $g(x) \in \mathbb{R}^M$ is a deep network like AlexNet without the final softmax loss layer. By using CNNs as weak learners needs learning CNNs for reducing (4) in every iteration of boosting. But, the learning algorithms for CNNs are normally depending on reducing the loss value i.e., log-likelihood or softmax and these fitness values are independent of the boosting weights and varied from (4).

A promising solution is to change the softmax loss layer with a layer for directly optimizing (4) in the back-propagation algorithm. But, this is not practical and diverges fastly since (4) is unrestrained e.g., scaling $g(x)$ will make it infinite.

To solve this problem, observe that $\delta\mathcal{R}[f;g]$ of (4) is equal to

$$\delta\mathcal{R}[f;g] = -\frac{1}{2\|\mathcal{D}\|}\sum_{(x_i,z_i)\in\mathcal{D}}\langle g(x_i)w(x_i)\rangle \quad (9)$$

In Eq. (9), $\langle\cdot,\cdot\rangle$ denotes the Euclidean dot product. This indicates (4) is a summation of dot products between the network outcome $g(x_i)$ and boosting weights $w(x_i)$.

As a result, (4) estimates the relevance between those vectors and so the optimum network outcome $g^*(x_i)$ has to be aligned with the boosting weights i.e.,

$$g^*(x_i) = \beta w(x_i), \beta > 0 \quad (10)$$

Observe that the accurate range of β is not essential during the network learning since $g^*(x_i)$ will be scaled properly by the optimum α in (7). So, with no loss of generality, consider that $\beta = 1$ and the optimum network outcome has to reproduce boosting weights. This is equal to learn a network $g(x) = [g_1(x) \dots g_M(x)] \in \mathbb{R}^M$ for reducing the square loss as:

$$\mathcal{L}_{sl}(w,g) = \sum_{x_i\in\mathcal{D}}\sum_{j=1}^{M}\left(g_j(x_i) - w_j(x_i)\right)^2 \quad (11)$$

By using $\mathcal{L}_{sl}$ to learn a CNN, the back-propagated derivatives are

$$-\frac{\partial\mathcal{L}_{sl}}{\partial g_k(x_i)} = 2\left(w_j(x_i) - g_k(x_i)\right) \quad (12)$$

Also, the weights are determined in each iteration of boosting based on (5). But, determining $f(x)$ for every data needs executing each network in the current ensemble and is extremely costly. To solve this challenge, observe that if $w^t(x_i)$ are the weights at iteration t, then using (5) and (8),

$$w_k^{t+1}(x_i) = \begin{cases} -e^{-\frac{1}{2}v\alpha^*\left[g^*_{z_i}(x_i) - g^*_k(x_i)\right]}w_k^t(x_i), & k \neq z_i \\ -\sum_{j=1|j\neq k}^{M} w_j^{t+1}(x_i), & k = z_i \end{cases} \quad (13)$$

In Eq. (13), $v \in (0,1]$ is the shrinkage factor which serves as a regularizer for boosting algorithm, $g^*(x)$ and α^* are the network and coefficient trained at iteration t. Likewise, discovering a precise α^* in (7) is computationally costly since it uses a line search and so binary search is used for solving $\frac{\partial\mathcal{R}[f+\alpha g^*]}{\partial\alpha} = 0$. At last, observe that for learning a CNN in every boosting iteration, it is promising to start it by the random variables or variables of the network trained in the prior iteration.

After, this boosted CNN is not constrained to a single category of base-learner and its mixture of weak learners can involve networks with various structures. At every boosting iterations, these networks are separately trained for estimating boosting weights using (11) and the network which tends to high minimization in the boosted classification risk (2) is included to the ensemble.

By ensembling weak learner classification, the pulmonary tumors are predicted with the maximum precision and the minimum loss. Figure 2 shows the overall flow diagram of IDA-BWOCNNEC-based lung cancer prediction.

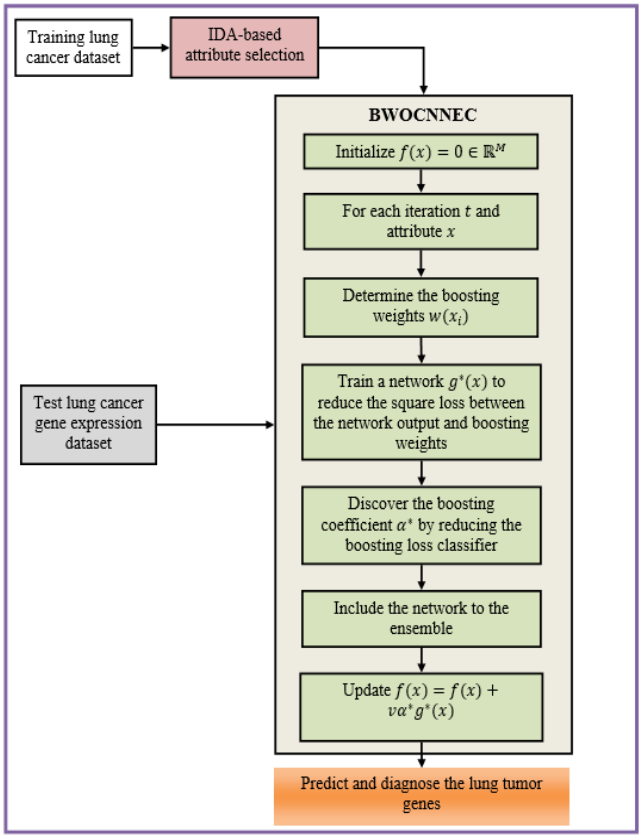

Figure 2. Overall Flow Diagram of IDA-BWOCNNEC-based Lung Cancer Prediction

The algorithm for IDA-BWOCNNEC is given below.

Algorithm: IDA-BWOCNNEC

Input: Genomic dataset $\mathcal{D} = \{(x_1,z_1),\dots(x_n,z_n)\}$ where $z_i \in \{1,\dots M\}$ is a class of data x_i, number of classes M, number of boosting, iterations t_{max}, shrinkage variable v

Output: Predictor $f(x)$ i.e., lung tumor genes or normal genes

Begin

Choose the most appropriate attributes from $\mathcal{D}$ using IDA;

Initialize $f(x) = 0 \in \mathbb{R}^M$;

for $(t = 1{:}\, t_{max})$

Determine the boosting weights $w(x_i)$ for each x_i using (5);

Train a network $g^*(x)$ to reduce the square loss between the network output and boosting weights using (11);

Discover the boosting coefficient α^* by reducing the boosting loss in (7);

Include the network to the ensemble and update $f(x) = f(x) + v\alpha^* g^*(x)$;

end for

Predict and diagnose the lung tumor genes;

End

IV.EXPERIMENTAL RESULTS

This section analyzes the efficiency of IDA-BWOCNNEC on the lung tumor dataset collected from the open source website: http://grafia.cs.ucsb.edu/autodecoder/dataset.html using MATLAB 2017b. This dataset comprises 12625 genes and 56 samples.

Retrieval Number: 100.1/ijeat.D25200410421
DOI: 10.35940/ijeat.D2520.1211221
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The examples in this dataset range between AD2 and AD384. Also, the efficiency is compared with the existing algorithms based on classifier evaluation metrics.

4.1 Accuracy

The percentage of accurately predicted lung tumor genes and normal genes is called as accuracy.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + False\ Negative\ (FN)} \times 100\%$$

TP is the number of lung tumor genes correctly predicted as lung tumor whereas TN is the number of normal genes correctly predicted as normal. Similarly, FP is the number of lung tumor genes incorrectly predicted as normal and FN is the number of normal genes incorrectly predicted as lung tumor genes. Table 1 lists the accuracy of IDA-BWOCNNEC, IDA-WONNEC and IDA-SMO algorithms.

Table.1 Accuracy of IDA-SMO, IDA-BWONNEC and IDA-BWOCNNEC

Algorithms	Accuracy (%)
IDA-SMO	93
IDA-BWONNEC	94.8
IDA-BWOCNNEC	96.2

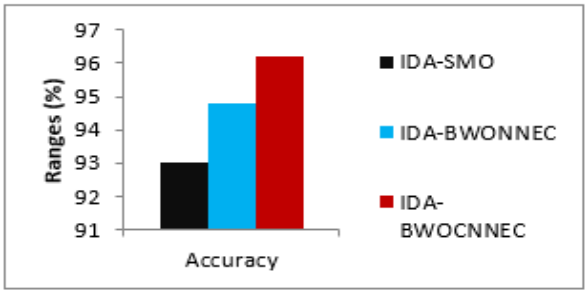

Figure 3. Accuracy of IDA-SMO, IDA-BWONNEC and IDA-BWOCNNEC

Figure 3 depicts the accuracy (in %) of IDA-SMO, IDA-BWONNEC and IDA-BWOCNNEC algorithms for predicting the lung tumor genes. This scrutiny outcomes indicates that the IDA-BWOCNNEC increases the accuracy compared to the other algorithms i.e., the accuracy of IDA-BWOCNNEC is 3.44% greater than the IDA-SMO and 1.48% greater than the IDA-BWONNEC algorithms. This is because of optimizing weight values of CNNs and also ensembling the weak learners effectively.

4.2 Precision, Recall and F-measure

The percentage of correctly predicted lung tumor genes at TP and FP rates is called as precision.

$$Precision = \frac{No.of\ correctly\ predicted\ lung\ tumor\ genes}{No.of\ correctly\ predicted\ lung\ tumor\ genes + No.of\ incorrectly\ predicted\ lung\ tumor\ genes}$$

The percentage of correctly predicted lung tumor genes at TP and FN rates is called as recall.

$$Recall = \frac{No.of\ correctly\ predicted\ lung\ tumor\ genes}{No.of\ correctly\ predicted\ lung\ tumor\ genes + No.of\ incorrectly\ predicted\ normal\ genes}$$

The harmonic average of precision and recall is called as f-measure.

$$F - measure = 2 \times \frac{(Precision \cdot Recall)}{(Precision + Recall)}$$

Table 2 lists the precision, recall and f-measure of IDA-BWOCNNEC, IDA-WONNEC and IDA-SMO algorithms.

Table.2 Precision, Recall & F-measure of IDA-SMO, IDA-BWONNEC & IDA-BWOCNNEC

Algorithms	Precision	Recall	F-measure
IDA-SMO	0.937	0.937	0.93
IDA-BWONNEC	0.945	0.952	0.946
IDA-BWOCNNEC	0.963	0.965	0.964

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

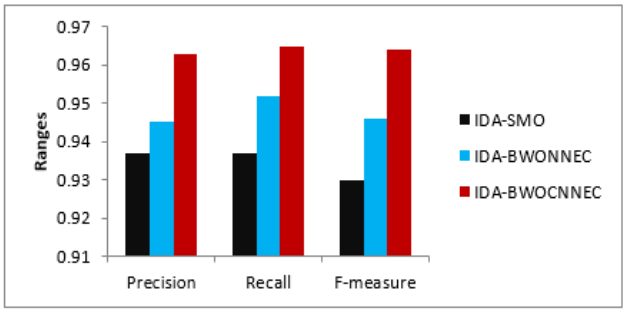

Figure 4. Precision, Recall & F-measure of IDA-SMO, IDA-BWONNEC & IDA-BWOCNNEC

Figure 4 displays the precision, recall and f-measure of IDA-SMO, IDA-BWONNEC and IDA-BWOCNNEC algorithms for lung tumor prediction from the genomic data. This testing outcomes observes that the IDA-BWOCNNEC can enhance the prediction efficiency than the other predictors i.e., the precision of IDA-BWOCNNEC is 2.77% larger than the IDA-SMO and 1.9% larger than the IDA-BWONNEC algorithms. Also, the recall of IDA-BWOCNNEC is 2.99% higher than the IDA-SMO and 1.37% higher than the IDA-BWONNEC algorithms. Similarly, the f-measure of IDA-BWOCNNEC is 3.66% greater than the IDA-SMO and 1.9% greater than the IDA-BWONNEC algorithms due to the ensembling of weak learners by optimizing the weight values of CNNs in an effective manner.

V. CONCLUSION

In this article, an IDA-BWOCNNEC algorithm is proposed for predicting the lung tumors from the gene expression data. At first, the lung tumor genomic data is collected and the most appropriate attributes are extracted from the collected dataset by using IDA. Then, the selected attributes are given to the BWOCNNEC in which the boosting weights are combined into the CNN depending on the least square fitness value. After, the novel weight values are assigned to the features extracted by the IDA. Moreover, these weight values and the chosen features are processed in different CNN structures within the boosted classifier. Further, the best CNN structure in each iteration i.e., CNNs having the least weighted loss is selected and ensemble to predict and diagnose the lung tumors effectively. To conclude, the findings are revealed that the IDA-BWOCNNEC has an accuracy of 96.2% which is higher than the IDA-BWONNEC and IDA-SMO algorithms for lung tumor prediction from the gene expression data.

REFERENCES

1. Azzawi, H., Hou, J., Xiang, Y., & Alanni, R. (2016). Lung cancer prediction from microarray data by gene expression programming. *IET Systems Biology*, *10*(5), 168-178.
2. Omar, Y., Tasleem, A., Pasquier, M., & Sagahyroon, A. (2018). Lung cancer prognosis system using data mining techniques. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, *5*, 361-368.
3. Ahmad, A. S., & Mayya, A. M. (2020). A new tool to predict lung cancer based on risk factors. *Heliyon*, *6*(2), 1-9.
4. Pati, J. (2018). Gene expression analysis for early lung cancer prediction using machine learning techniques: an eco-genomics approach. *IEEE Access*, *7*, 4232-4238.
5. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*, 1-13.
6. Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Systems with Applications*, *121*, 233-243.
7. Vinmalar, F. L., & Kombaiya, K. (2020). An improved dragonfly optimization algorithm based feature selection in high dimensional gene expression analysis for lung cancer recognition. *International Journal of Innovative Technology and Exploring Engineering*, *9*(8), 896-908.
8. Vinmalar, F. L., & Kombaiya, K. (). Boosted weighted optimized neural network ensemble classification based lung cancer prediction.
9. Ayyad, S. M., Saleh, A. I., & Labib, L. M. (2019). Gene expression cancer classification using modified k-nearest neighbors technique. *Biosystems*, *176*, 41-51.
10. Sharma, A., & Rani, R. (2019). C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. *Computer Methods and Programs in Biomedicine*, *178*, 219-235.
11. Zheng, X., Zhu, W., Tang, C., & Wang, M. (2019). Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning. *Gene*, *706*, 188-200.
12. Wang, H., Tan, L., & Niu, B. (2019). Feature selection for classification of microarray gene expression cancers using bacterial colony optimization with multi-dimensional population. *Swarm and Evolutionary Computation*, *48*, 172-181.
13. Shukla, A. K., Singh, P., & Vardhan, M. (2020). Gene selection for cancer types classification using novel hybrid metaheuristics approach. *Swarm and Evolutionary Computation*, *54*, 1-16.
14. García-Díaz, P., Sánchez-Berriel, I., Martínez-Rojas, J. A., & Diez-Pascual, A. M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics*, *112*(2), 1916-1925.
15. Shukla, A. K. (2020). Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique. *Computational Intelligence*, *36*(1), 102-131.

AUTHOR PROFILE

A. Kumar Kombaiya, Assistant Professor, 4SG, Jain Vengateswara Apartment Ramanujam Nagar, Uppalipalayam (PO), Coimbatore – 641 015. Email Id: kumar_kombaiya@rediffmail.com

F. Leena Vinmalar, Education: B.Sc(Computer Science) 2007-2010, Tirupur kumaran college for women-Tirupur, M.B.A(Finance/HR) 2010-2012, Vivekanandha Institute Of Information And Management Studies-Tiruchencode, M.Sc(Computer Science), 2012-2014, St'Joseph College For Women – Tirupur, B.ed(Computer Science) 2014-2015, St.Peter's College Of Education-Karumathampatti, M.Phil(Computer Science), 2015-2016, Chikkanna Government Arts College-Tirupur, Ph.D(Computer Science) 2017-2021, Chikkanna Government Arts College-Tirupur.
Teaching Experience: Teacher, Wisdom Mat. Hr. Sec. School. Tiruppur. 2015., Assistant Professor, Park's College-Tiruppur. 2016-2017

Retrieval Number: 100.1/ijeat.D25200410421
DOI: 10.35940/ijeat.D2520.1211221
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)