Applied Soft Computing Journal 80 (2019) 579–591

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Boosted neural network ensemble classification for lung cancer disease diagnosis

Jafar A. ALzubi [a], Balasubramaniyan Bharathikannan [b], Sudeep Tanwar [c,*], Ramachandran Manikandan [d], Ashish Khanna [e], Chandrasekar Thaventhiran [d]

[a] School of Engineering, Al-Balqa Applied University, Jordan
[b] School of Computer Science and Engineering, Golgotha's University, Greater Noida, India
[c] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, India
[d] School of Computing, SASTRA Deemed University, India
[e] Department of Computer Engineering, Maharaja Agrasen Institute Technology, GGSIP University, India

HIGHLIGHTS

- To increase the performance of lung cancer diagnosis accuracy for big data as compared to state-of-the-art works.
- To minimize the classification time for early LCD diagnosis, integrated Newton–Raphsons MLMR preprocessing model is used.
- To reduce the error (i.e. false positive rate) and also improve the disease diagnosis accuracy of big data with higher classification efficiency and lower classification time as compared to conventional methods.

ARTICLE INFO

Article history:
Received 3 December 2018
Received in revised form 19 April 2019
Accepted 27 April 2019
Available online 2 May 2019

Keywords:
Machine learning
Lung cancer disease
Weighted optimized
Neural network
Maximum likelihood boosting

ABSTRACT

Accurate diagnosis of Lung Cancer Disease (LCD) is an essential process to provide timely treatment to the lung cancer patients. Artificial Neural Networks (ANN) is a recently proposed Machine Learning (ML) algorithm which is used on both large-scale and small-size datasets. In this paper, an ensemble of Weight Optimized Neural Network with Maximum Likelihood Boosting (WONN-MLB) for LCD in big data is analyzed. The proposed method is split into two stages, feature selection and ensemble classification. In the first stage, the essential attributes are selected with an integrated Newton–Raphsons Maximum Likelihood and Minimum Redundancy (MLMR) preprocessing model for minimizing the classification time. In the second stage, Boosted Weighted Optimized Neural Network Ensemble Classification algorithm is applied to classify the patient with selected attributes which improves the cancer disease diagnosis accuracy and also minimize the false positive rate. Experimental results demonstrate that the proposed approach achieves better false positive rate, accuracy of prediction, and reduced delay in comparison to the conventional techniques.

1. Introduction

In the present era, one of the foremost causes of death in developing countries is lung cancer which is increasing rapidly with the dramatic upsurge in cigarette smoking. According to the survey conducted by big data research in Non-Small Cell Lung Cancer (NSCLC) [1], the deep learning can be used to improve the rate of diagnostic accuracy by means of prediction and decision in the medical system. Moreover, Artificial Intelligence techniques were used to solve the prediction and decision for big data in NSCLC. However, the image and diagnostic parameters were integrated via machine learning algorithm. Therefore, combining image and diagnostic parameters was an efficient method for doctors to solve patient's diagnosis in large data (i.e., Healthcare 4.0) sets. However, the time consumed to diagnosis for big data was not concentrated.

Boosted Support Vector Machine (SVM) method for imbalanced data (BSI) proposed by Zięba et al. [2] to solve the issues related to the imbalanced data. They have combined the advantages of applying ensemble classifiers for uneven data with the cost-sensitive SVM. Three steps were carried out with an input dataset. In the first step, the information gain criterion was used to select the effective and required features. Followed by a feature selection step, the problem to predict postoperative life expectancy was analyzed according to Gmean criterion, where

* Corresponding author.
E-mail addresses: j.zubi@bau.edu.jo (J.A. ALzubi), bbharathikannan@gmail.com (B. Bharathikannan), sudeep149@rediffmail.com (S. Tanwar), srmanimt75@gmail.com (R. Manikandan), ashishkhanna@mait.ac.in (A. Khanna), thaventhiran@gmail.com (C. Thaventhiran).

https://doi.org/10.1016/j.asoc.2019.04.031
1568-4946/

the rules were said to be extracted. In the third step, the accuracy and coverage measure were evaluated for the extracted rules which results in the improved prediction accuracy. However, the prediction accuracy was achieved, but less focus was made on the error aspect.

Considering the aforementioned issues, in this paper, a new combination approach for classifier ensembles using the Newton–Raphson's MLMR preprocessing model is proposed, where the essential features are extracted to reduce the time for LCD diagnosis. Newton–Raphson's Maximum Likelihood model is applied to the MLMR attributes is proposed. Moreover, the first and the second derivative results of maximum relevance minimum redundant attributes are used to select the most relevant attributes. To achieve it, we explore the features of the MLMR model, Newton–Raphson's Maximum Likelihood in the combination process of an ensemble. Then, Boosted Weighted Optimized Neural Network Ensemble Classification algorithm is proposed to minimize the error (i.e., false positive error rate) and improve diagnosis accuracy. Optimized weights related to the decision of each ensemble classifier are defined dynamically, according to the ensemble classifier outputs and the relation among the outputs of all ensemble classifiers. In order to evaluate the feasibility of the proposed approach, an empirical analysis of ensemble performance using Thoracic Surgery Dataset, comparing its performance with ensemble classifier using traditional methods.

1.1. Motivation

Healthcare is one of the essential sources in big data. Accurate analysis of healthcare data is highly in demand for diagnosing the disease at early stage. Recently, many research works have been designed for identifying disease in the big data with higher quality. But, there is a requirement for novel classification technique to increase the diagnosis accuracy with time. Moreover, ML algorithms are designed to increase the prediction accuracy in big data. However, error rate still not exploited to its full potential. Therefore, this research work motivates optimized machine learning algorithms to improve the diagnosis accuracy with lower time and error.

1.2. Research contributions

Contributions of this paper are as follows.

- To increase the performance of lung cancer diagnosis accuracy for big data as compared to state-of-the-art works, WONN-MLB method is used with Weight Optimized Neural Network to have Maximum Likelihood Boosting for Lung Cancer Disease.
- To minimize the classification time for early lung cancer disease diagnosis, integrated Newton–Raphsons MLMR preprocessing model is used to select the relevant attributes, to obtain higher diagnosis accuracy.
- To reduce the error (i.e. false positive rate) and improves the disease diagnosis accuracy with higher classification efficiency and lower classification time, Boosted Weighted Optimized Neural Network Ensemble Classification algorithm is designed in WONN-MLB method.

1.3. Organization

LCD Section 2 describes the related works on various LCD diagnosis. In Section 3, the ensemble classification method along with a preprocessing model for LCD diagnosis is investigated, the maximum relevance method along with the maximum likelihood function is explored in detail, and the effects of the extracted relevant attributes on the ensemble classification performance of the WONN-MLB method are also studied. In Section 4, the performance of the proposed approach is compared with the state-of-the-art approaches to demonstrate its effectiveness for LCD diagnosis and Section 5 concludes the paper.

2. Related works

With the invention of the microarray technique, scientists and researchers have immense opportunity to evaluate the expression levels of thousands of genes concurrently in a single experiment. In Ghorai et al. [3], the Nonparallel Plane Proximal Classifier (NPPC) was proposed for cancer classification in a Computer Aided Diagnosis (CAD) framework to ensure high classification accuracy and to minimize the computation time. But, Valvular heart disorders were considered to be one of the most difficult classification problems. Sengur et al. [4] used three powerful and popular ensemble learning representative called, bagging, boosting, and random subspaces to early detect Valvular heart disorders. However, the classification time was minimized using methods, but the rate at which the accuracy was said to be attained remained unaddressed. In Costaa et al. [5], three Generalized Mixture (GM) functions were applied via dynamic weights to improve the classification accuracy of the classification system. Though the function handles single-label classification, multi-label classification problem was not addressed.

A case study for brain tumor diagnosis using global optimization based hybrid wrapper-filter feature selection with ensemble classification methods was proposed by Huda et al. [6]. It increases the classification accuracy, but the classification time was not minimized. Approximately 40% of the world's population is affected by cancer. A Proportion SVM was used by Huseeinet al. [7] for efficient categorization of Lung Nodules, which results in the improved diagnosing accuracy. The proportion of SVM failed to minimize the error rate in disease categorization. Another method to early detect lung cancer was proposed by Abetiba et al. [8] using Radial Basis Function Neural Network with Affine Transforms which in turn achieved high classification accuracy and low mean square error. But, the performance of feature extraction was not improved. A review of feature selection and parallel classification systems was carried out by Jain et al. [9] to enhance the classification accuracy for disease perdition, but classification time was not minimized.

A Critical assessment of ANN was carried out in Dande et al. [10] which results in an increase in the efficacy and specificity of the diagnostic techniques, but it fails to minimize the computational complexity. Tumor tissue based on pathological evaluation is considered to be one of the most pivotal for early diagnosis in cancer patients. However, the automated image analysis methods have the potential to improve the accuracy of disease diagnosis and to minimize human errors. Khosravia et al. [11] proposed different computational methods using convolutional neural networks (CNN), where a stand-alone pipeline was constructed in an effective manner to classify several histopathology images across different types of cancer. But, it fails to minimize the computation cost while classifying the various types of cancer.

Sharma et al. [18] proposed a two-stage hybrid ensemble classification technique to increase the prediction accuracy of chronic kidney disease with ML technique. It improves the disease diagnosis, but the multistage classification was not performed with minimum time. Early diagnoses of lung cancers and differentiation between the tumor types and non-tumor types have been required to improve the patient survival rate. In Hosseinzadeh et al. [13], a diagnostic system with structural and physicochemical attributes of proteins via feature extraction, feature selection, and prediction models was designed. Then, the ML models were

Table 1
Comparison of the proposed approach with the state-of-the-art approaches.

Author	Year	Approach	Objective	Pros	Cons
Das et al. [4]	2010	Ensemble learning methods	To classify the Valvular heart disease	Minimize the classification time	Classification accuracy rate remained unsolved
Ghorai *et al.* [3]	2011	Nonparallel Plane Proximal Classifier (NPPC)	Perform cancer classification with higher accuracy in a Computer Aided Diagnosis	Provides better classification accuracy with lesser computation time	Valvular heart disorders classification was difficult
Baz et al. [12]	2012	Computer-aided diagnosis (CAD) system	Lung cancer diagnosis	Achieve better detection and diagnosis of lung nodules	Accurate feature selection was not performed
Hosseinzadeh et al. [13]	2013	Machine learning models	Predict and detect the type of lung tumors	Provide more accurate results in lung tumor detection	The false positive rate was not minimized
Adetiba et al. [8]	2015	Radial Basis Function Neural Network with Affine Transforms	Classifies the Lung Cancer	Improve classification, accuracy and achieves, and low mean square error	Performance of feature extraction was not improved
Kumar et al. [14]	2016	Evolutionary algorithms	Lung cancer detection	Detect the lung cancer accurately with minimum time	The error rate was not minimized
Huda et al. [6]	2016	Global optimization based hybrid wrapper-filter feature selection with ensemble classification	Tumor classification with the imbalanced healthcare data	Increases the imbalanced healthcare data classification	Classification time was not minimized
Podolsky et al. [15]	2016	Machine learning algorithm	Lung cancer diagnosis	Increase the accuracy of predicting cancer susceptibility and Minimize false positive	Classification time remained unsolved
Zhou et al. [16]	2017	Multi-modality and multi-classifier radiomics predictive models	Extract numbers of quantitative features and disease prediction	Increase the disease prediction accuracy with the features	Failed to minimize the disease prediction time.
Kang et al. [17]	2017	Multi-view convolutional neural networks (MV-CNN)	lung nodule classification	Increases the classification accuracy and minimizes the time	Failed to attain accurate disease prediction with features
Dande et al. [10]	2017	Artificial Neural Network	Diagnosis and evaluation of medical conditions	Increase the efficacy and specificity of disease diagnosis	Failed to minimize the computational complexity
Costaa et al. [5]	2018	Generalized mixture (GM) functions	Increase the classification accuracy of a classification system	Handles single-label classification problems	Multi-label classification problem remained unaddressed
Hussein et al. [7]	2018	Proportion-Support Vector Machine (SVM)	Categorizes the Lung Nodules	Improve the diagnosing accuracy	Failed to minimize the error rate
Jain et al. [9]	2018	Feature selection and parallel classification systems	Enhancing the accuracy of classification systems	Classification systems for effective disease prediction	The classification time was not minimized
Khosravia et al. [11]	2018	Deep convolutional neural networks (CNN)	Classifying the various cancer tissues	Increase the precision of diagnosis and minimizes the error	Computation cost was not minimized
Sharma et al. [18]	2018	Two-stage hybrid ensemble technique	Classifying the chronic kidney disease	Accurate diagnosis of the disease with a feature set	The multi-stage diagnosis was not performed with minimum time
Rabbani et al. [19]	2018	Machine learning (ML) method Combining artificial intelligence approaches	Extracting and analyzing several quantitative features from medical images	Improves diagnosis, treatment and outcomes	The ML algorithms used for feature extraction was not attained the accurate results
Baranidharan et al. [20].	2016	Image-based features selection method	Classify the lung cancer images	Increase the true positive rate	Error rate was not effectively minimized
Proposed	–	Weight Optimized Neural Network with Maximum Likelihood Boosting (WONN-MLB) technique	Lung Cancer Disease diagnosis with big data	Increase diagnosing accuracy and minimizes the false positive rate, classification time	–

applied to both original and newly created database to predict the lung cancer type of tumors which results in improved accuracy. However, the model reduces the processing time, but the false positive rate was not minimized. Evaluation of ML algorithm for lung cancer diagnosis was carried out by Podolsky et al. [15]. It accurately predicts cancer vulnerability as well as minimizes the false positive rate. But, the classification time was not exploited which can be helpful for early lung cancer detection.

A narrative review based on radiomic features to help diagnose lung cancer in an early stage was proposed by Rabbani et al. [19], where the ML algorithms were combined with artificial intelligence approaches. The objective of radiomics remains in extracting and analyzing several quantitative features from medical images. Moreover, they focused on highly promising in staging, diagnosing, and predicting outcomes of cancer treatments. However, the machine learning algorithms used, but the feature extraction does not provide accurate results. Zhou et al. [16]

proposed a multi-modality and multi-classifier radiomics predictive models to address the aforementioned issues using a new reliable classifier fusion strategy. Here, the training of modality-specific classifiers was first made, followed by an analytic evidential reasoning (ER) rule, which was used to combine the output score from each modality to build an optimal predictive model towards disease diagnosis. This model failed to minimize the disease prediction time.

A systematic review of mortalities and survival rate of lung cancer with evolutionary algorithms was conducted by Dubey et al. [14] to identify a better method for early lung cancer diagnosis and to achieve higher accuracy rate with deep learning techniques. It does not minimize the error rate. Liu et al. [17] proposed a MultiView Convolutional Neural Networks (MV-CNN) for efficient lung nodule classification, to improve the accuracy, and the classification time. Here, accurate detection was not performed with the features. Baz et al. [12] explored some crucial challenges and methodologies with CAD system for lung cancer. It increases the detection and diagnosis of lung nodules, but the accurate feature selection was not performed to minimize the detection time.

Deep feature fusion and hand-crafted features for lung nodule classification was developed by Wang et al. [21]. But, classification performance was not accurate. CAD was introduced for enhancing the performance of nodule candidate classification by Chen et al. [22]. However, classification time was not minimized. In order to effectively classify the lung nodules, deep features were extracted in CT images with higher accuracy by Kumar et al. [23]. But, the error rate was remained unaddressed. Image-based features selection method was developed for classifying the lung cancer images with higher accuracy Baranidharan et al. [20]. In this method, novel fusion-based selection was used to select the features for classification. During the feature selection, the redundant features were unable to be removed thus introduced an error in classification process. To overcome this problem, the proposed WONN-MLB method used Newton–Raphson's Maximum Likelihood mode, where MLMR are used to choose the most relevant attributes. Then, the boosting classifier is applied to classify the attributes for LCD diagnosis, which reduces the error rate in the classification process.

Data analysis of population statistics and data mining techniques were used in [24] to determining the cancer morbidity and mortality data in a regional cancer registry. However, false positive rate was not minimized. Multiple aspects of large scale knowledge mining was covered in [25] for medical and diseases examination. A new image-based features selection method was planned in [26] to categorize the lung computed tomography images with a higher accuracy. But, the feature selection rate was not improved.

Table 1 presents a comparison of the proposed approach with state-of-the-art approaches. The main aim of this paper is to design diagnosis for LCD using ensemble classification algorithm with an objective to reduce the classification time and false positive rate as compared to the state-of-the-art approaches.

3. Materials and methods

In this paper, we proposed a WONN-MLB method to increase the performance of LCD diagnosis. The WONN-MLB is designed with an implementation of Newton–Raphsons MLMR preprocessing model and Boosted Weighted Optimized Neural Network Ensemble Classification algorithm. To validate the proposed WONN-MLB method, the Thoracic Surgery Data Dataset Wroclaw Thoracic Surgery Centre is used [26]. The patient data contains underwent major lung resections for primary lung cancer in the years 2007–2011. The center is linked through the Thoracic Surgery of the medical university of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland. In order to conduct the experiments, a different number of patient data is taken, i.e., 10,000 patient data from Thoracic Surgery data dataset. In this data set, the information related to forced vital capacity, pain before surgery, Haemoptysis before surgery, Dyspnoea before surgery, cough before surgery, weakness before surgery, peripheral arterial diseases , smoking , asthma, age at surgery, and year survival period were collected. Based on this information, the LCD classification was made in the proposed approach.

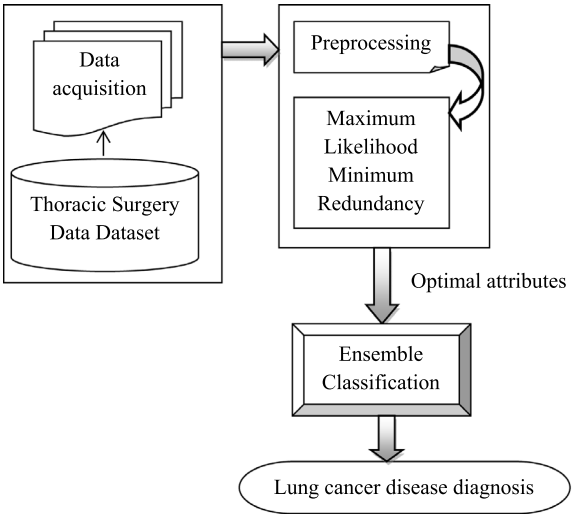

Fig. 1. Architecture of proposed approach for LCD diagnosis.

3.1. Proposed approach

This section describes the proposed approach and the proposed architecture with WONN-MLB method for LCD, as shown in Fig. 1. The different phases to implement and utilize the proposed approach are shown in Fig. 1. These include the data acquisition (Thoracic Surgery Data Dataset) Zięba et al. [2], feature selection or preprocessing (reducing big data feature dimensionality), and ensemble classification (using WONN-MLB) and are comprehensively discussed in the next subsections.

3.1.1. Data acquisition

The data is obtained for classification problem related to lung cancer patients from the Thoracic Surgery Domain (TSD) archive in the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, from UCI Machine Repository. The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for 1200 patients who underwent major lung resections for primary lung cancer in the years 2007–2011. We have used these predictors for lung cancer prediction from the online UCI repository acquired from Zięba et al. [2].

3.1.2. Newton–Raphsons Maximum Likelihood and Minimum Redundancy preprocessing

To overcome the time complexity, accuracy problems in big data classification, initially preprocessing step is needed to extract the relevant attributes. While extracting the relevant attributes the redundant attribute removal is unable to be performed in conventional techniques. This produces the misclassification results in LCD diagnosis. Therefore, Newton–Raphsons

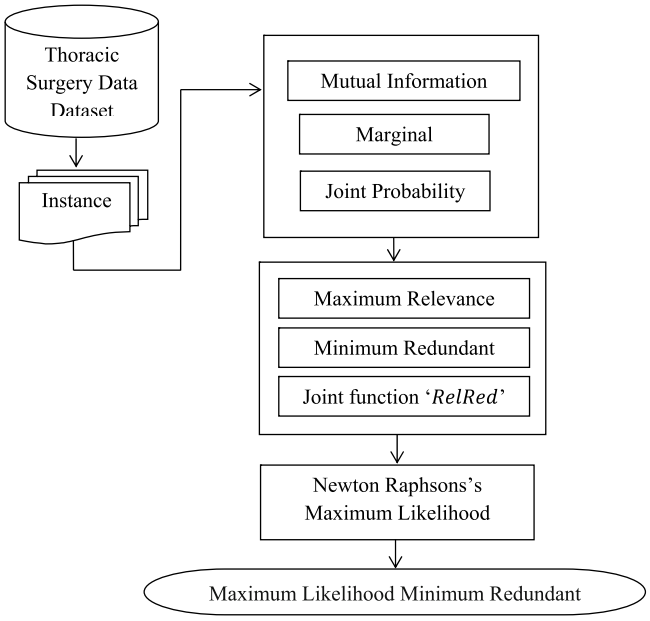

Fig. 2. Flow of MLMR preprocessing model.

Maximum Likelihood and Minimum Redundancy pre-processing techniques is developed to perform relevant attribute extraction through removing redundancy.

A large-scale ML classifier based on boosted classifiers [2] was used for the classification of biomedical lung cancer data. Moreover, an iterative process was carried out by updating the boosting coefficient value to minimize the weighted error function. Despite to minimize the weighted error, less focus was made on the time consumed for lung cancer diagnosis. In this work, an integrated Newton–Raphsons MLMR preprocessing model is applied to the data acquired from the Thoracic Surgery Data Dataset Zięba et al. [2] with an objective not only to reduce the weighted error, but also to minimize the time consumed. The preprocessing proposed model is based on the Newton–Raphson's method with the maximum likelihood, to obtain more robust results than other well-known algorithms such as SVMs Zięba et al. [2].

MLMR preprocessing model is employed to find the most relevant and least redundant attributes in the set of class. At first, the maximum relevancy is identified between set of attributes and class based on the mutual information. The results often contained most relevance but redundant. In order to solve this issue, minimum redundancy between attributes is measured in MLMR preprocessing model. These two conditions are equally important and these are combined into a single criterion function in MLMR. In WONN-MLB method, additive combination is used to integrate the maximum relevancy and minimum redundancy. Lastly, maximization is performed on resultant attributes using Newton–Raphsons's Maximum Likelihood function thus minimizes the time required to diagnosis the lung cancer. Fig. 2 shows the flow diagram of proposed MLMR preprocessing model.

As shown in Fig. 2, let us assume a standard feature selection problem by means of instance '$eis = (e1s, e2s, \ldots, ens, eCs)$', where '*eis*' represents the '*ith*' attribute value of the '*sth*' sample and '*eCs*' represents the value of the output class '*C*'. Moreover, let us assume a training dataset '*D*' with '*m*' examples consists of a set '*Attr*' with '*n*' attributes. The main objective of MLMR preprocessing model is to identify the maximum dependency between a set of attributes '*Attr*' and the class '*C*', using mutual information, denoted by '*MI*'. The value of '*MI*' is obtained using the marginal probabilities (i.e., with a pair of attributes) '*prob* (*x*)' and '*prob* (*y*)',(where '$x \in Attr and y \in Attr$') and the joint probability '*prob* (*x*, *y*)' as given in Eq. (1).

$$MI\,(X, Y) = \sum_{x \in X} \sum_{y \in Y} prob\,(x, y) \log \left[\frac{prob\,(x, y)}{prob\,(x)\, prob\,(y)} \right] \tag{1}$$

However, with big data in consideration for lung cancer analysis, maximum relevance, and minimum redundancy is measured. Maximum relevance '*Rel*' consists to search attributes with higher relevancy factor and is formulated as follows:

$$Max \rightarrow Rel\,(Att, C) \rightarrow Rel = \frac{1}{Att} * MI\,[Att_i, C] \tag{2}$$

With reference to Eq. (2), the maximum relevance between attributes '*Attr*' in class '*C*' is obtained according to the mutual information factor '*MI*', while to select the attributes based on the maximum relevance criterion results in larger amount of redundancy. To minimize it, the minimum redundancy '*Red*' criterion is used and is formulated as follows:

$$Min \rightarrow Red\,(Att) \rightarrow Red = \frac{1}{Att} * MI\left[Att_i, Att_j\right] \tag{3}$$

From Eq. (3), the minimum redundant attributes '*Att*' is obtained between the set of attributes 'Att_i' and 'Att_j', respectively. From Eqs. (2) and (3), the integration and optimization of both maximum relevancy '*Rel*' and the minimum redundancy '*Red*' results in maximum relevance minimum redundancy called as '*RelRed*'. The maximum relevance minimum redundancy is calculated as follows:

$$\begin{aligned} RelRed(RR) &= \frac{1}{Att} * MI\,[Att_i, C] - \frac{1}{Att} * MI\left[Att_i, Att_j\right] \\ &= MI\,[Att_i, C] - MI\left[Att_i, Att_j\right] \end{aligned} \tag{4}$$

Followed by maximum relevance minimum redundancy attributes obtained for LCD diagnosis with an objective to minimize the time consumed, in this work, a Newton–Raphsons's Maximum Likelihood function is used to the resultant attributes. The log-likelihood function for Eq. (4) is formulated as follows:

$$\ln\,(C|X) = -\sum_{i=1}^{n} \log\left(1 + [Att_i - C]^2\right) \tag{5}$$

In the log-likelihood function the first derivative and second derivative are formulated as follows:

$$\frac{\partial \ln\,(C|X)}{\partial C} = 2\sum_{i=1}^{n} (Att_i - C)\left(1 + [Att_i - C]^2\right)^{-1} \tag{6}$$

$$\begin{aligned} \frac{\partial^2 \ln\,(C|X)}{\partial^2 C} = \Bigg[& 2\sum_{i=1}^{n} (Att_i - C)^2 \left(\left(1 + [Att_i - C]^2\right)^{-2}\right) \\ & \times \left(\left(1 + [Att_i - C]^2\right)^{-1}\right) \Bigg] \end{aligned} \tag{7}$$

The log-likelihood function is used to maximize the maximum relevance and minimum redundant attributes. From that, the most relevant attributes are taken for classification process which effectively reduces the time required to lung cancer disease diagnosis. The pseudo-code of the proposed Maximum Likelihood Minimum Redundant preprocessing is given in algorithm 1.

The Maximum Likelihood Minimum Redundant Preprocessing is described in algorithm 1, where for each training dataset (i.e., big data), all the attributes are not essential. In this work, for lung cancer diagnosis with big data as the input dataset, maximum relevance, and the minimum redundant attributes are selected. Then, Newton–Raphson's Likelihood Estimation is evaluated with respect to the first and the second derivative with an objective to minimize the time consumed for lung cancer diagnosis.

Input: $D, Att = Att_1, Att_2, \ldots, Att_n$
Output: Maximum Likelihood Minimum Redundant attributes selected 'RR'
1: **Begin**
2: **For** D with Att
3: Find $MIAtt$
4: Determine Rel attribute
5: Minimize Red attribute
6: Combine $RelRed$
7: Formulate $\ln(C \mid X)$
8: Obtain $\frac{\partial \ln(C \mid X)}{\partial C}$ and $\frac{\partial^2 \ln(C \mid X)}{\partial^2 C}$
9: **End for**
10: **End**

Algorithm 1: Maximum Likelihood Minimum Redundant preprocessing

3.1.3. Weighted optimized neural network with maximum likelihood boosting

Once the Maximum Likelihood Minimum Redundant attributes are obtained, then an ensemble classification model is used to improve the lung cancer diagnosis accuracy for big data. In this work, an ensemble of WONN-MLB attributes is applied to achieve the objective of lung cancer diagnosis accuracy with minimum time and error.

The given 'n' training data (i.e., attributes) '$\{(Att_1, b_1), (Att_2, b_2), \ldots.., (Att_n, b_n)\}$', '$Att_i$' consists of a vector corresponding to an input sample data, associated with 'RR' input attributes, and 'b_i' represents the target variable with a class label of either '$1 or -1$'. To start with, in the proposed model, a weak classifier is trained using distribution 'D_i', where '$D_i \in RR$'.

An artificial neuron consists of 'n' synapses related to the input attributes $(Att_1, Att_2, \ldots, Att_n)$ and each input attribute has the corresponding weight 'w_i'. Here, the signal at input i is multiplied by the weight wi, then the summation of weighted inputs and a linear combination of the weighted inputs are obtained. Moreover, a bias 'bs' is summed to the linear combination and a weighted sum 'ws' is obtained as follows:

$$ws = bs + w_1 Att_1 + w_2 Att_2 + \cdots + w_n Att_n \tag{8}$$

Then, a nonlinear activation function 'f' is applied to the weighted sum 'ws' as given in Eq. (9)which results in an output 'b':

$$b = f\,(ws) \tag{9}$$

Then, a weak classifier with low weighted error is selected and is formulated as follows:

$$\varepsilon_i = z = Prob_{D_i}\,[b + w_1 Att_1 + w_2 Att_2 + \cdots + w_n Att_n] \tag{10}$$

$$= Prob_{D_i}\,[f\,(ws)] \tag{11}$$

From Eqs. (10) and (11), the low weighted error 'ε_i' is obtained based on the probability of distribution function '$Prob_{D_i}$' for a linear combination of weighted inputs (i.e., attributes) '$f\,(ws)$'. Finally, a new component 'k_i' based on error function is calculated as follows:

$$k_i = \frac{1}{n}\sum_{i-1}^{n}(Actual\ error - Observed\ error)^2 \tag{12}$$

Upon successful completion of all of the boosting iterations, final ensemble learning classifier which possesses weighted error that is better than chance, is evaluated by combining all weak classifiers with an optimal weight Mana et al. [27]. This is formulated as follows:

$$f\,(WS) = SIGN\left(\sum_{t=1}^{T} k_t f\,(s)\right) \tag{13}$$

From Eq. (13), the final ensemble learning classifier is measured as a weighted majority vote of the weak classifiers '$f\,(s)$', where each classifier is assigned by weighting 'k_t'. The pseudo code of ensemble classification is given in algorithm 2.

The Boosted Weighted Optimized Neural Network Ensemble Classification Algorithm is introduced to classify the LCD with minimum error, which is given in algorithm 2. In first step for each maximum likelihood minimum redundant attributes weights are initialized. Then, a weight initialization and the weighted sum value is obtained. Moreover, a conditional checking is performed to see whether the weighted sum is less than or equal to the optimal weight Mana et al. [3]. Upon unsuccessful checking, the weighted sum value with different weights being initialized is obtained. Then, the process is continued by applying a boosting technique. Here, three steps are carried out. In first step, a weak classifier with low weighted error is measured. Then, in second step, a new component based on error function is obtained. Finally, in third step, final ensemble learning classifier is applied to the new component. Hence, the lung cancer disease diagnosis accuracy is said to be improved with minimum error rate.

4. Experimental settings and results discussion

To evaluate the performance of proposed WONN-MLB approach the Thoracic Surgery Data Set [2]is used. The proposed WONN-MLB approach is implemented in JAVA platform using Weka tool. The Thoracic Surgery Data Dataset is dedicated to classification problem related to the post-operative life expectancy in the lung cancer patients. The data was collected retrospectively at Wroclaw Thoracic Surgery Centre. The patient data includes those who underwent major lung resections for primary lung cancer in the years 2007–2011. The Centre mainly concentrates on the Pulmonary Diseases which is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre, Poland. However, the research database constitutes a part of the National Lung Cancer Registry. The Lung Cancer Registry is administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. Specifically, the preprocessing is first performed on the attributes in Thoracic Surgery Data dataset including, maximum relevancy, minimum redundancy and maximum likelihood to obtain the relevant features. With the Maximum Likelihood Minimum Redundant attributes, the next process of ensemble classification is performed for improving diagnosing accuracy with minimum error and time.

The experimental work of proposed approach is performed for many instances with respect to various numbers of patient data with an objective to analyze its performance. The effectiveness of proposed approach is compared with Non-Small Cell Lung Cancer

Input: Maximum Likelihood Minimum Redundant attributes 'RR', bs, $w = w_1, w_2, \ldots, w_n$, iteration $i = 1, 2, \ldots, n$, Optimal weight 'β'
Output: Improved lung cancer diagnosis accuracy
1: **Procedure**
2: **Initialize** w
3: **For** each RR and iteration t
4: Measure ws
5: **If** '$ws \leq \beta$'**then**
6: Compute ε_i
7: Obtain k_i
8: Obtain $f(WS)$
9: **End if**
10: **Else**$ws > \beta$
11: Go to step 4
13: **End for**
14: **End**

Algorithm 2: Boosted Weighted Optimized Neural Network Ensemble Classification algorithm

(Big data research in NSCLC) by Wu et al. [1], Boosted Support vector machine (BSVM) method by Zięba et al. [2], Nonparallel Plane Proximal Classifier (NPPC) by Ghorai et al. [3], and Multi-View Convolutional Neural Networks (MV-CNN) by Liu et al. [17] For the better understanding among the readers, the discussion on obtained results of the proposed approach is explained with different parameters such as-diagnosing accuracy, false positive rate or error rate, and classification time, F1-score.

4.1. Scenario 1: Impact of diagnosing accuracy

It is considered as one of the important parameters for early disease diagnosis. Higher the diagnosing accuracy, early disease diagnosis is said to be achieved and therefore the method is also said to be efficient. It provides evidence on how well a method precisely recognizes the disease and informs upcoming decisions about treatment for physicians or patients. It is given as follows:

$$DA = \sum_{s=1}^{n} \frac{CD_{disease}}{s} * 100 \tag{14}$$

From Eq. (14), the diagnosing accuracy 'DA' is arrived at based on the number of data correctly diagnosed as disease '$CD_{disease}$' to the total samples 's' considered for experimentation. It is measured in percentage. The values obtained through Eq. (14) are represented as shown in Fig. 3 for different patient data using the proposed WONN-MLB approach and compared it with the NSCLC and BSVM approaches. The sample calculation to measure the diagnosing accuracy using the aforementioned three methods is given as follows:

Sample calculation:

- **Proposed WONN-MLB**: With '1000' patient data considered for experimentation and number of data correctly diagnosed as disease being '930', the diagnosing accuracy is calculated as follows:

$$DA = \frac{930}{1000} * 100 = 93\%$$

- **NSCLC**: With '1000' patient data considered for experimentation and number of data correctly diagnosed as disease being '890', the diagnosing accuracy is calculated as follows:

$$DA = \frac{890}{1000} * 100 = 89\%$$

- **BSVM**: With '1000' patient data considered for experimentation and number of data correctly diagnosed as disease being '860', the diagnosing accuracy is calculated as follows:

$$DA = \frac{860}{1000} * 100 = 86\%$$

- **NPPC:** With '1000' patient data considered for experimentation and number of data correctly diagnosed as disease being '800', the diagnosing accuracy is calculated as follows:

$$DA = \frac{800}{1000} * 100 = 80\%$$

- **MV-CNN:** With '1000' patient data considered for experimentation and number of data correctly diagnosed as disease being '740', the diagnosing accuracy is calculated as follows:

$$DA = \frac{740}{1000} * 100 = 74\%$$

Fig. 3 shows the diagnosing accuracy comparison between proposed approach and existing NSCLC and BSVM, respectively. It is found that the diagnosing accuracy of lung cancer is improved using WONN-MLB because of measurement of the weak classifier with low weighted error and new component based on error function through ensemble classification. The results confirm that with an increase in the number of patient data, the diagnosing accuracy increases for minimum patient data, then reduces with an increase in the number of patient data. This happens because with an increase in the number of patient data, many irrelevant attributes are also present. Moreover, preprocessing performed in the WONN-MLB method, the certain error is occurred, which results in certain amount of irrelevant attributes even after preprocessing. However, the comparison made with the existing methods NSCLC, BSVM, NPPC and MV-CNN shows an improvement is observed by using the WONN-MLB method. This happens because of the application of ensemble classification that not only minimizes the error by updating the weak classifier, but also minimizes the time by boosting the updated results. This in turn improves the diagnosing accuracy using WONN-MLB method by 7%, 11%, 19% and 28% as compared to NSCLC, BSVM, NPPC, and MV-CNN, respectively.

4.2. Scenario 2: Impact of false positive rate

The second important parameter used to measure the early diagnosing of lung cancer is the rate of false positive or error, while to conduct multiple comparisons in a statistical framework, the false positive rate refers to the probability of falsely rejecting the null hypothesis for a specific test. In other words, the false positive rate is measured as the ratio between the number of negative events (i.e., not diagnosed with lung cancer) wrongly categorized as positive (i.e., diagnosed with lung cancer) and the

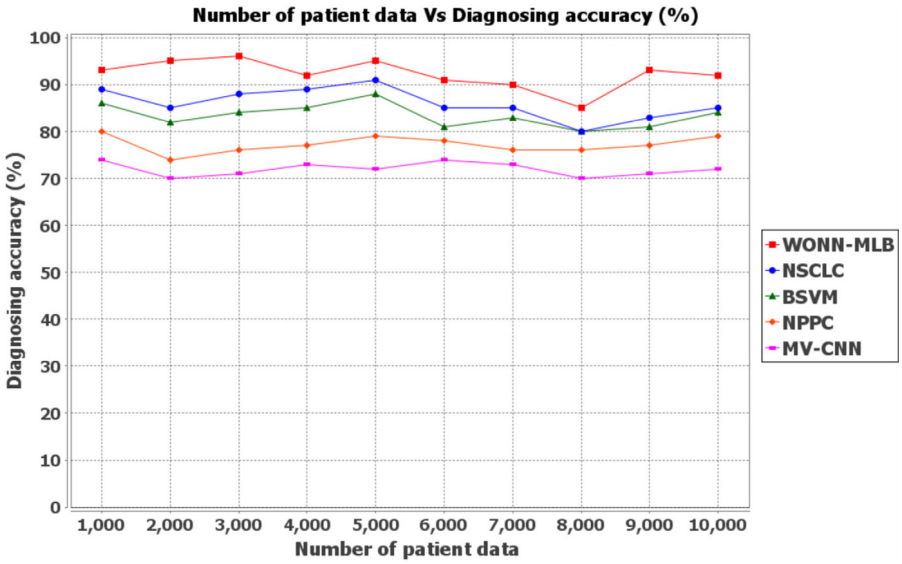

Fig. 3. Diagnosing accuracy with 10,000 patient data.

total number of actual negative events (i.e., not diagnosed with lung cancer). It is formulated as follows:

$$FPR = \frac{ICD_{disease}}{s} * 100 \tag{15}$$

From Eq. (15), the false positive rate '*FPR*' refers to the ratio of number of patient data incorrectly diagnosed as disease '$ICD_{disease}$' to the total samples '*s*' considered for experimentation. It is measured in terms of percentage (%). The values obtained through Eq. (15) are represented as shown in Fig. 5 for different patient data using the proposed WONN-MLB approach and compared it with the NSCLC and BSVM. The sample calculation for measuring false positive rate using the three methods is given as follows:

Sample calculation

• **Proposed WONN-MLB**: With '1000' number of patient data considered as samples and '90' number of patient data incorrectly diagnosed with lung cancer disease, the false positive rate is as given as follows:

$$FPR = \frac{90}{1000} * 100 = 9\%$$

• **NSCLC**: With '1000' number of patient data considered as samples and '120' number of patient data incorrectly diagnosed with lung cancer disease, the false positive rate is given as follows:

$$FPR = \frac{120}{1000} * 100 = 12\%$$

• **BSVM**: With '1000' number of patient data considered as samples and '140' number of patient data incorrectly diagnosed with lung cancer disease, the false positive rate is as given as follows:

$$FPR = \frac{140}{1000} * 100 = 14\%$$

• **NPPC:** With '1000' number of patient data considered as samples and '160' number of patient data incorrectly diagnosed with lung cancer disease, the false positive rate is as given as follows:

$$FPR = \frac{160}{1000} * 100 = 16\%$$

• **MV-CNN:** With '1000' number of patient data considered as samples and '170' number of patient data incorrectly diagnosed with lung cancer disease, the false positive rate is as given as follows:

$$FPR = \frac{170}{1000} * 100 = 17\%$$

From Eq. (15), the false positive rate for different number of patient data in the range of 1000 to 10,000 is measured. The results of experimental evaluations conducted to measure the false positive rate as shown in Table 1. The false positive rate obtained using the proposed WONN-NLB approach offers comparable values than the state-of-the-art methods.

Fig. 4 shows the performance analysis of false positive rate for disease diagnosis for big data. As illustrated in Fig. 4, when 1000 number of patient data is considered as samples, 90 patient data were incorrectly diagnosed with lung cancer using WONN-MLB, 120 patient data were incorrectly diagnosed with lung cancer using NSCLC, 140 patient data are incorrectly diagnosed with lung cancer using BSVM, 160 patient data were incorrectly diagnosed using NPPC, 170 patient data were incorrectly diagnosed using MV-CNN. The false positive rate using WONN-MLN is minimized by 25%, 36%, 44% and 47% as compared to NSCLS, BSVM, NPPC, and MV-CNN, respectively. This result is achieved with Newton–Raphsons MLMR preprocessing model. The advantage of applying MLMR preprocessing model is that instead of using all the attributes in the dataset, only the maximum likelihood and relevancy attributes are considered for disease diagnosis. With the application of log-likelihood function, the attribute availability also gets changed and reflected in the maximum relevance minimum redundancy coefficient. This adaptive change made through maximum relevance minimum redundancy coefficient in terms minimizes the incorrect lung cancer diagnosis using the WONN-MLN method. The resultant attributes are then used to classify the patients as lung cancer and normal patient which in turn minimizes the false positive rate by 39%, 53%, 58% and 61% as compared to NSCLS, BSVM, NPPC, and MV-CNN, respectively.

4.3. Scenario 3: Classification time

The third parameter considered for the early diagnosis of lung cancer is the classification time. The classification time refers to the time taken to classify the patient data as diagnosed with lung cancer or not diagnosed with lung cancer. The classification time is calculated as follows:

$$CT = s * Time\,(f\,(WS)) \tag{16}$$

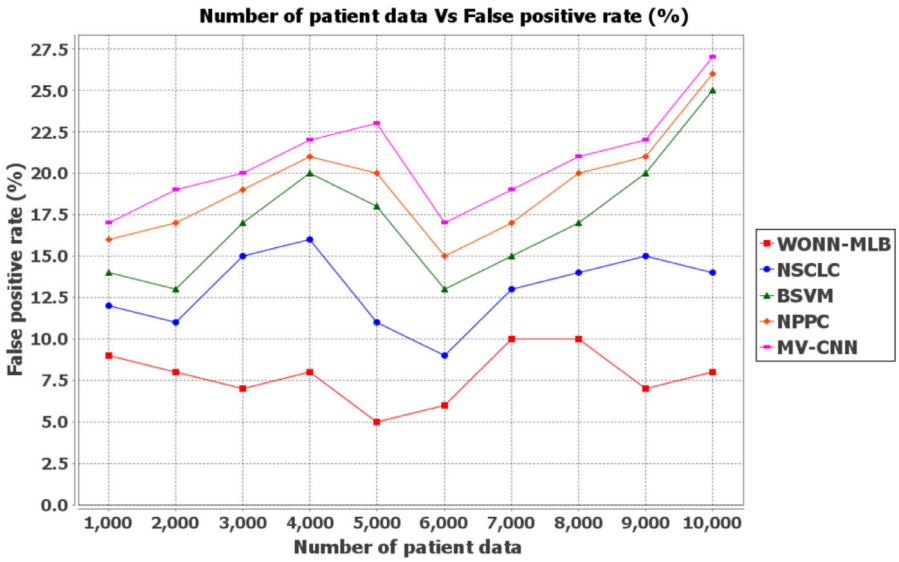

Fig. 4. Performance measure of false positive rate.

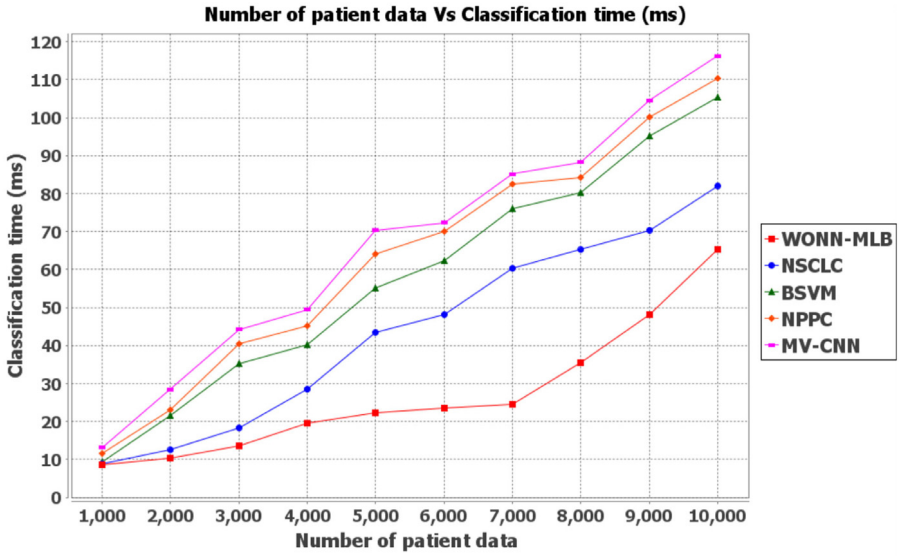

Fig. 5. Performance measure of classification time.

From Eq. (16), the classification time '*CT*' is calculated according to the samples '*s*' and the time consumed to perform ensemble classification '*Time* (*f* (*WS*))'. Lower the classification time, early the lung cancer diagnosis is said to be. It is measured in terms of milliseconds (ms). The values obtained through Eq. (16) are represented in Fig. 4 with the proposed WONN-MLB approach, existing NSCLC and BSVM. The sample calculation for classification time using the three methods is given as follows:

Sample calculations:

• **Proposed WONN-MLB**: With the time taken for classification of single patient data being '0.0085 ms', with '1000' number of patient data considered as samples, the classification time is calculated as follows:

$$CT = 1000 * 0.0085 \text{ ms} = 8.5 \text{ ms}$$

• **NSCLC**: With the time taken for classification of single patient data being '0.0089 ms', with '1000' number of patient data considered as samples, the classification time is given as follows:

$$CT = 1000 * 0.0089 \text{ ms} = 8.9 \text{ ms}$$

• **BSVM**: With the time taken for classification of single patient data being '0.0093 ms', with '1000' number of patient data considered as samples, the classification time is given as follows:

$$CT = 1000 * 0.0093 \text{ ms} = 9.3 \text{ ms}$$

• **NPPC:** With the time taken for classification of single patient data being '0.0115 ms', with '1000' number of patient data considered as samples, the classification time is given as follows:

$$CT = 1000 * 0.0115 \text{ ms} = 11.5 \text{ ms}$$

• **MV-CNN:** With the time taken for classification of single patient data being '0.0132 ms', with '1000' number of patient data considered as samples, the classification time is given as follows:

$$CT = 1000 * 0.0132 \text{ ms} = 13.2 \text{ ms}$$

Fig. 5 shows the measure of classification time to classify the patient data with diagnosed as disease or not, the proposed approach is implemented in Java Language using various numbers of patient data in the range of 1000 to 10,000. The experimental

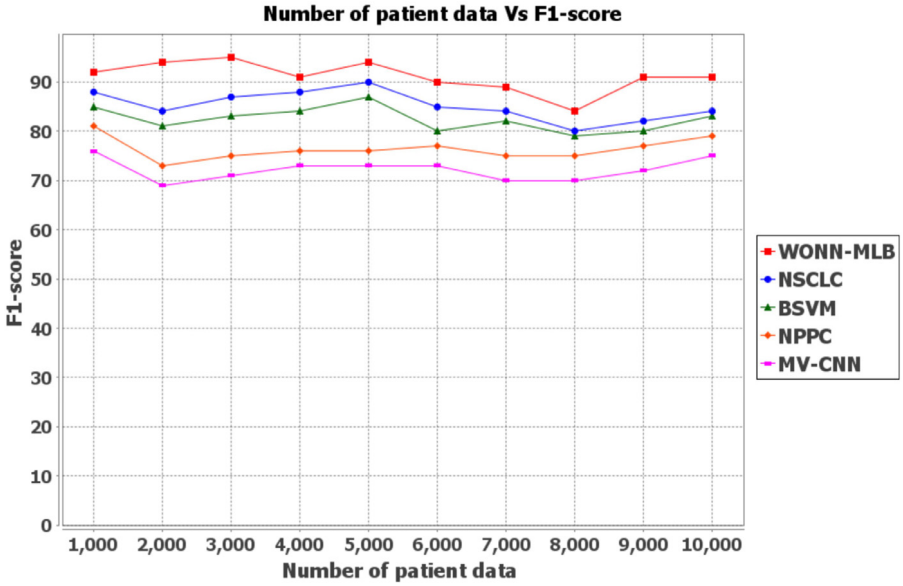

Fig. 6. Performance measure of F1-score.

result of classification time using proposed method is compared with existing NSCLC and BSVM. When considering 1000 number of patient data for the experimental work, the proposed method consumed 8.5 ms to classify, whereas the existing NSCLC, BSVM, NPPC and MV-CNN consumed 8.9 ms, 9.3 ms, 11.5 ms, and 13.2 ms, respectively. Thus, it is clear that the classification time using proposed approach is less as compared to other existing methods [1,2]. However, with an increase in the number of patient data and increase in the number and size of the attributes, the classification time is also increases using all the three methods. Comparative analysis shows that the classification time using proposed approach is less than the [1–3] and [17] methods. This is because of the application of the Newton–Raphson's Maximum Likelihood model in addition to the maximum relevance minimum redundancy factor, which applies the first derivate and the second derivate to extract the most relevant attributes. With this most relevant attributes extracted, the classification time is reduced using proposed approach by 34%, 51%, 56%, and 59% as compared to NSCLC by Wu et al.. [1], BSVM by Zięba et al. [2], NPPC by Ghorai et al. [3], and MV-CNN by Liu et al. [17], respectively.

4.4. Scenario 4: F1-score

The fourth parameter taken for classifying lung cancer diagnosis is F1-score. F1-score is a single measure of performance test for the positive class. It is defined both precision and recall of the test. Precision is the number of correct positive results divided by the number of all positive results returned by the classifier and recall is the number of correct positive results divided by the number of all relevant samples. The F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

From Eq. (17), the classification time '$F1 - score$' is measured using average mean of precision and recall value. Higher the F1-score, early the lung cancer diagnosis is said to be. The sample calculation for F1-score using the five methods is given as follows:

Sample calculations:

• **Proposed WONN-MLB:** With '1000' patient data considered for experimentation and precision is value is identified as 93 and recall value is 91, the F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{93 \times 91}{93 + 91} = 92\%$$

• **NSCLC**: With '1000' patient data considered for experimentation and precision is value is identified as 89 and recall value is 87, the F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{89 \times 87}{89 + 87} = 88\%$$

• **BSVM**: With '1000' patient data considered for experimentation and precision is value is identified as 86 and recall value is 85, the F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{86 \times 85}{86 + 85} = 85\%$$

• **NPPC:** With '1000' patient data considered for experimentation and precision is value is identified as 80 and recall value is 82, the F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{80 \times 82}{80 + 82} = 81\%$$

• **MV-CNN:** With '1000' patient data considered for experimentation and precision is value is identified as 74 and recall value is 78, the F1-score is calculated as follows:

$$F1 - score = 2 \times \frac{74 \times 78}{74 + 78} = 76\%$$

Fig. 6 illustrates the measure of F1-score to classify the patient data with higher accuracy. In order to conduct the experiments, 1000 to 10,000 patient data is considered. The performance analysis of F1-score using proposed method is compared with existing NSCLC, BSVM, NPPC, and MV-CNN . When considering 1000 number of patient data for the performance analysis, the proposed method provides the F1-score of 92% whereas the existing NSCLC, BSVM, NPPC and MV-CNN produced 88%, 85%, 81%, and 76%, respectively. From the discussion, it is clear that the F1-score using proposed method is higher as compared to other existing methods. While increasing the number of patient data, the value of F1-score is increased in all methods. Comparatively, F1-score using proposed method is high than the [1–3] and [17] methods.

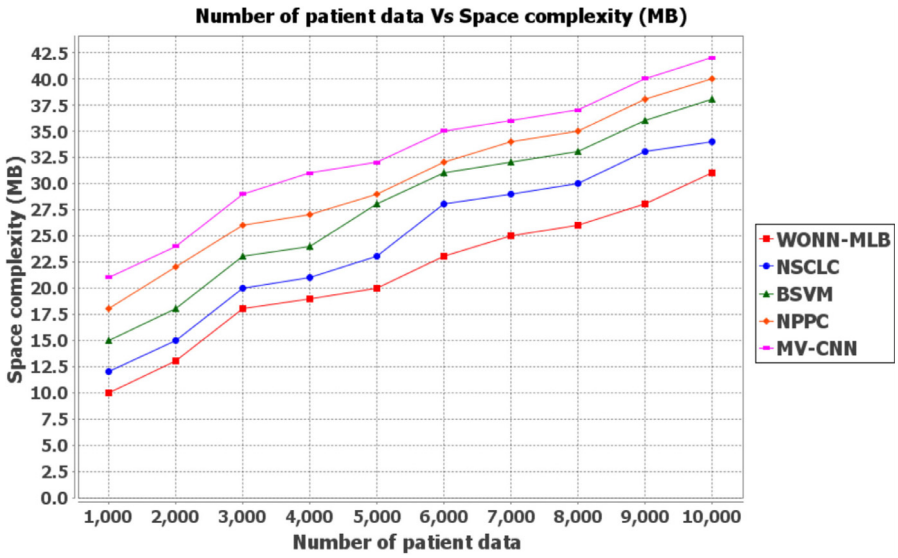

Fig. 7. Performance measure of space complexity.

This is due the application of weighted optimized neural network with maximum likelihood boosting which classifies the patient data with higher accuracy. Therefore, F1-score is improved using proposed WONN-MLB approach by 7%, 11%, 19%, and 26% as compared to NSCLC by Wu et al. [1], BSVM by Zięba et al. [2], NPPC by Ghorai et al. [3], and MV-CNN by Liu et al. [17], respectively.

4.5. Scenario 5: Space complexity

Space complexity is defined as an amount of storage space required to store the patient data in big healthcare data analytics. It is measured in terms of megabyte (MB). The mathematical formula for space complexity is measured as follows,

$$SC = n * space\ (storing\ the\ one\ patient\ data) \tag{18}$$

In (18), '*SC*' denotes a space complexity and 'n' denotes the number of the patient data. The sample calculation for space complexity using the five methods is given as follows:

Sample calculations:

• **Proposed WONN-MLB:** With '1000' patient data considered for experimentation and space for storing one patient data is 0.01 MB, the space complexity is calculated as follows:

$$SC = 1000 * 0.01\ \text{MB} = 10\ \text{MB}$$

• **NSCLC**: With '1000' patient data considered for experimentation and space for storing one patient data is 0.012 MB, the space complexity is calculated as follows:

$$SC = 1000 * 0.012\ \text{MB} = 12\ \text{MB}$$

• **BSVM**: With '1000' patient data considered for experimentation and space for storing one patient data is 0.015 MB, the space complexity is calculated as follows:

$$SC = 1000 * 0.015\ \text{MB} = 15\ \text{MB}$$

• **NPPC:** With '1000' patient data considered for experimentation and space for storing one patient data is 0.018 MB, the space complexity is calculated as follows:

$$SC = 1000 * 0.018\ \text{MB} = 18\ \text{MB}$$

• **MV-CNN:** With '1000' patient data considered for experimentation and space for storing one patient data is 0.021 MB, the space complexity is calculated as follows:

$$SC = 1000 * 0.021\ \text{MB} = 21\ \text{MB}$$

Fig. 7 shows the measure of space complexity to store the patient data with minimum space. To conduct the experiments, 1000 to 10,000 patient data is considered. From Fig. 7, the performance analysis of space complexity using WONN-MLB approach is compared with existing NSCLC, BSVM, NPPC, and MV-CNN. While considering 1000 number of patient data for analyzing the performance, the proposed WONN-MLB approach provides the 10 MB of space complexity whereas the existing NSCLC, BSVM, NPPC and MV-CNN offers 12 MB, 15 MB, 18 MB, and 21 MB, respectively. From the above discussion, space complexity using proposed WONN-MLB approach is lower as compared to other existing [1–3] and [17] methods. This is because of the application of boosted weighted optimized neural network ensemble classification algorithm in proposed WONN-MLB approach. This algorithm classifies the patient data with higher accuracy and it is further stored for diagnosing the cancer diseases. Therefore, space complexity is reduced using proposed WONN-MLB approach by 13%, 24%, 31%, and 36% as compared to NSCLC by Wu et al. [1], BSVM by Zięba et al. [2], NPPC by Ghorai et al. [3], and MV-CNN by Liu et al. [17], respectively.

4.6. Scenario 6: Feature selection rate

Feature selection rate is defined as the ratio of number of relevant features that are correctly selected to the total number of features. It is measured in terms of percentage (%). The mathematical formula for feature selection rate is measured as follows,

$$FSR = \frac{number\ of\ correctly\ selected\ features}{Total\ number\ of\ features} * 100 \tag{19}$$

In (19), '*FSR*' denotes a Feature Section Rate. The sample calculation for feature selection rate using the five methods is given as follows:

Sample calculations:

• **Proposed WONN-MLB:** With '20' features considered for experimentation and the number of features correctly selected is 18, then the feature selection rate is calculated as follows:

$$FSR = \frac{18}{20} * 100 = 90\%$$

• **NSCLC**: With '20' features considered for experimentation and the number of features correctly selected is 17, then the

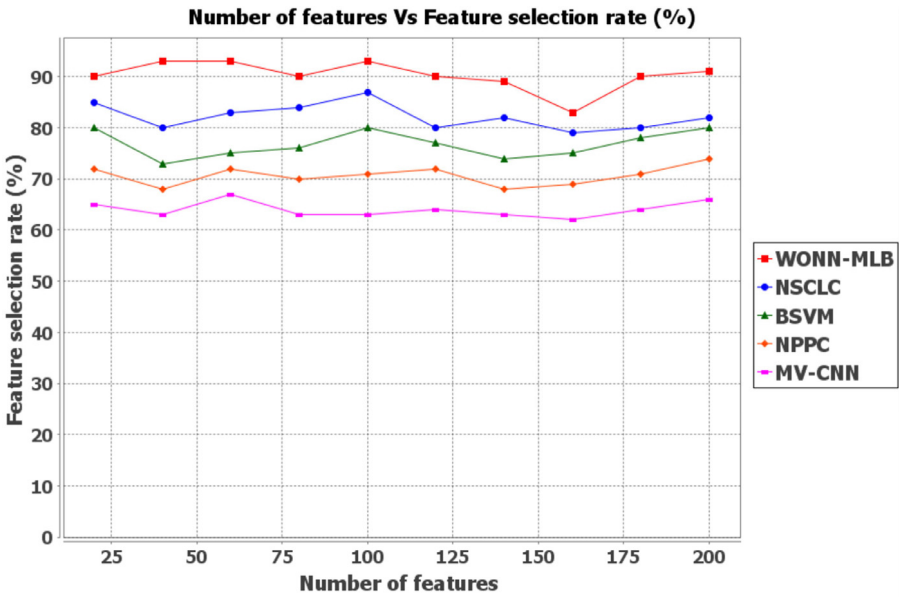

Fig. 8. Performance measure of feature selection rate.

feature selection rate is calculated as follows:

$$FSR = \frac{17}{20} * 100 = 85\%$$

- **BSVM**: With '20' features considered for experimentation and the number of features correctly selected is 16, then the feature selection rate is calculated as follows:

$$FSR = \frac{16}{20} * 100 = 80\%$$

- **NPPC**: With '20' features considered for experimentation and the number of features correctly selected is 14, then the feature selection rate is calculated as follows:

$$FSR = \frac{14}{20} * 100 = 72\%$$

- **MV-CNN**: With '20' features considered for experimentation and the number of features correctly selected is 13, then the feature selection rate is calculated as follows:

$$FSR = \frac{13}{20} * 100 = 65\%$$

Fig. 8 depicts the feature selection rate comparison between proposed approach and existing NSCLC, BSVM, NPPC, and MV-CNN, respectively. In order to conduct the experiments, 20 to 200 features are considered. The performance analysis of feature selection rate using proposed WONN-MLB approach is compared with existing NSCLC, BSVM, NPPC, and MV-CNN. When considering 20 number of features for the performance analysis, the proposed WONN-MLB approach provides the feature selection rate of 90%, whereas the existing NSCLC, BSVM, NPPC and MV-CNN obtains 85%, 80%, 72%, and 65%, respectively. From the discussion, it is clear that the feature selection rate using proposed WONN-MLB approach is higher as compared to other existing [1–3] and [17] methods. This is due the application of identifying maximum relevancy between set of attributes and reducing minimum redundancy attributes in preprocessing. This helps to selects the accurate features for cancer disease diagnosis. Therefore, feature selection rate is improved using proposed WONN-MLB approach by 10%, 18%, 28%, and 41% as compared to NSCLC by Wu et al. [1], BSVM by Zięba et al. [2], NPPC by Ghorai et al. [3], and MV-CNN by Liu et al. [17], respectively.

5. Conclusion

An effective Weight Optimized Neural Network with Maximum Likelihood Boosting for LCD in big data is investigated to improve the LCD diagnosis accuracy and to minimize the false positive rate as well as classification time. To achieve these, the preprocessing the model using Newton–Raphson's MLMR attributes retrieved and remove the irrelevant features is used. Therefore, the classification time gets minimized. With the most relevant attributes, an ensemble classification model called Weighted Optimized Neural Network and Boosting is applied for early lung cancer diagnosis with a higher accuracy rate. Here, not only the weighted sum function is considered, but also the most optimal values are obtained. The final ensemble technique finds the weak classifier with less error value and new component update based on the error function. This process attains higher disease diagnosing accuracy with the minimum false positive rate. Experimental evaluation is conducted with different parameters such as-disease diagnosing accuracy, false positive rate, and classification. The experimental results show that the proposed approach achieved accurate results for big data processing as compared to existing methods. Proposed WONN-MLB approach is tested with different dataset, but still there is huge amount of data points are presented which need to be tested with the proposed approach in future.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.04.031.

References

[1] Jia Wu, Yanlin Tan, Zhigang Chen, Ming Zhao, Decision based on big data research for non-small cell lung cancer in medical artificial system in developing country, Comput. Methods Programs Biomed. 159 (2018) 87–101, [Big data research in Non-Small Cell Lung Cancer – Big data research in NSCLC].

[2] Maciej Zięba, Jakub M. Tomczak, Marek Lubicz, Jerzy Świątek, Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients, Appl. Soft Comput. 14 (Part A) (2014) 99–108, [Boosted Support vector machine (SVM) method for Imbalanced data (BSI)].

[3] Santanu Ghorai, Anirban Mukherjee, Sanghamitra Sengupta, Pranab K. Dutta, Cancer classification from gene expression data by NPPC ensemble, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (3) (2011) 659–671.
[4] Resul Das, Abdulkadir Sengur, Evaluation of ensemble methods for diagnosing of valvular heart disease, Expert Syst. Appl. 37 (7) (2010) 5110–5115.
[5] Valdigleis S. Costaa, Antonio Diego S. Fariasa, Benjamín Bedregala, Regivan H.N. Santiagoa, AnneMagaly de P. Canutoa, Combining multiple algorithms in classifier ensembles usinggeneralized mixture functions, Neuro Computing 313 (2018) 402–414.
[6] Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, Michael Buckland, A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis, IEEE Access 4 (2016) 9145–9154.
[7] Sarfaraz Hussein, Pujan Kandel, Candice W. Bolan, Michael B. Wallace, Ulas Bagci, Supervised and unsupervised tumor characterization in the deep learning era, IEEE Trans. Med. Imaging 2 (2018) 1–11.
[8] Emmanuel Adetiba, Oludayo O. Olugbara, Improved classification of lung Cancer using radial basis function neural network with affine transforms of voss representation, PLoS One J. 10 (12) (2015) 1–25.
[9] Divya Jain, Vijendra Singh, Feature selection and classification systems for chronic disease prediction: A review, Egyptian Inf. J. 19 (3) (2018) 179–189.
[10] Payal Dande, Purva Samant, Acquaintance to artificial neural networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review, Tuberculosis 108 (2018) 1–9.
[11] Pegah Khosravia, Ehsan Kazemic, Marcin Imielinskid, Olivier Elemento, Iman Hajirasouliha, Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images, EBioMedicine 27 (2018) 317–328.
[12] Ayman El-Baz, Garth M. Beache, Georgy Gimel'farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman, Behnoush Abdollahi, Computer-aided diagnosis systems for lung Cancer: Challenges and methodologies, Int. J. Biomed. Imaging 2013 (2012) 1–46.
[13] Faezeh Hosseinzadeh, Amir Hossein KayvanJoo, Mansuor Ebrahimi, Bahram Goliaei, Prediction of lung tumor types based on protein attributes by machine learning algorithms, Springer Plus 2 (238) (2013) 1–14.
[14] Ashutosh Kumar Dubey, Umesh Gupta, Sonal Jain, Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis, Chin. J. Cancer 35 (71) (2016) 1–13.
[15] Maxim D. Podolsky, Anton A. Barchuk, Vladimir I. Kuznetcov, Natalia F. Gusarova, Vadim S. Gaidukov, Segrey A. Tarakanov, Evaluation of machine learning algorithm utilization for lung Cancer classification based on gene expression levels, Asian Pac. J. Cancer Prevent. 17 (2) (2016) 835–838.
[16] Zhiguo Zhou, Zhi-Jie Zhou, Hongxia Hao, Shulong Li, Xi Chen, You Zhang, Michael Folkert, Jing Wang, Constructing multi-modality and multiclassifier radiomics predictive models through reliable classifier fusion, IEEE Comput. Soc. (2017) 1–13.
[17] Kui Liu, Guixia Kang, Multiview convolutional neural networks for lung nodule classification, Int. J. Imaging Syst. Technol. 27 (1) (2017) 12–22.
[18] Sahil Sharma, Vinod Sharma, Atul Sharma, A two stage hybrid ensemble classifier based diagnostic tool for chronic kidney disease diagnosis using optimally selected reduced feature set, Int. J. Intell. Syst. Appl. Eng. 6 (2) (2018) 113–122.
[19] Mohamad Rabbani, Jonathan Kanevsky, Kamran Kafi, Florent Chandelier, Francis J. Giles, Role of artificial intelligence in the care of patients with nonsmall cell lung cancer, Eur. J. Clin. Invest. 48 (4) (2018) 1–7.
[20] Thangavel Baranidharan, Thangavel Sumathi, Vadivelraj Chandra Shekar, Weight optimized neural network using metaheuristics for the classification of large cell Carcinoma and adenocarcinoma from lung imaging, Curr. Signal Transduct. Therapy 11 (2) (2016) 91–97.
[21] Changmiao Wang, Ahmed Elazab, Jianhuang Wu, Qingmao Hu, Lung nodule classification using deep feature fusion in chest radiography, Comput. Med. Imaging Graph. 57 (2017) 10–18.
[22] Sheng Chen, Kenji Suzuki, Heber MacMahon, Development and evaluation of a computer-aided diagnostic scheme for lung nodule detection in chest radiographs by means of two-stage nodule enhancement with support vector classification, Int. J. Med. Phys. Res. Pract. 38 (4) (2011) 1844–1858.
[23] Devinder Kumar, Alexander Wong, David A. Clausi, Lung nodule classification using deep features in CT images, in: 2015 12th Conference on Computer and Robot Vision, 3-5 2015, pp. 133–138.
[24] I. Varlamis, I. Apostolakis, Sifaki-Pistolla, Dey, Georgoulias, C. Lionis, Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of crete, Greece, Comput. Methods Progr. Biomed. 145 (2017) 73–83.
[25] Md. Sarwar Kamal, Nilanjan Dey, Amira S. Ashour, Large scale medical data mining for accurate diagnosis: A blueprint, in: Handbook of Large-Scale Distributed Computing in Smart Healthcare, Springer, 2017, pp. 157–176.
[26] Thoracic Surgery Data Data Set: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data.
[27] Zhihong Mana, Kevin Lee, Dianhui Wang, Zhenwei Cao, Suiyang Khoo, An optimal weight learning machine for handwritten digit image recognition, Signal Process. 93 (6) (2013) 1624–1638.