TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

Impute, Select, Decision Tree and Naïve Bayes (ISE-DNC): An ensemble learning approach to classify the Lung Cancer

[1]Bhanumathi S, [2]Dr. S N Chandrashekara
[1]Research Scholar, S J C Institute of Technology, Chickballapur-562101, Email:sbhanureddy14@gmail.com
[2]M.Tech, Ph.D, Professor & HOD, Department of CSE, C.Byregowda Institute of Technology Kolar-563101, Email: snc_chandru@yahoo.co.in

Article Info
Volume 83
Page Number: 132 - 141
Publication Issue:
May - June 2020

Article History
Article Received*: 11August 2019*
Revised*: 18November 2019*
Accepted*: 23January 2020*
Publication*:07May2020*

Abstract:
In this work, we have introduced a hybrid novel approach to classify the lung cancer data using ensemble learning. According to this approach, first of all, we present data preprocessing model where missing values are imputed with the help of knn. Later, we incorporated filtering-based feature selection to reduce the feature dimension. Later, decision tree and Naïve Bayes classifiers are used to create the ensemble learner. Finally, voting based decisions are made to classify the data. The proposed approach is represented as ISE-DNC (Impute, Select, Decision Tree and Naïve Bayes) classifier. The proposed approach is implemented on two lung cancer public datasets which are obtained from the UCI repository. The experimental study shows that the proposed approach achieves 96.87% and 89.78% of classification accuracy for lung cancer and Thoracic Surgery Dataset.

I. INTRODUCTION

Recently, during the latter half of twentieth century, the world has noticed a drastic increase in the occurrence of health-related chronic disorders. These disorders are instigated due to the degraded environmental toxins, less healthy lifestyle, less nutritious food and inappropriate heath care. Several chronic diseases such as Cancer, cardiovascular disease and diabetes, have affected millions of people worldwide. Recently, GLOBOCAN reported that 18.1 million new cancer cases are identified, and 9.6 million deaths have been reported due to cancer [1]. In [2] authors have reported that currently around 140 million people are having diabetes and this number is expected to reach up to 300 million by 2025. Similarly, in [3], authors have reported that 4.15 hundred million adults are having diabetes. However, several techniques have been presented to diagnose these diseases but diagnosis of cancer still remains a challenging task. About 606,880 Americans are expected to die of cancer in 2019 [4]. Moreover, cancer is considered that most common cause of death in US, hence, these works focuses on cancer diagnosis and provide a new solution for efficient diagnosis.

A. Cancer: A brief overview

Cancer is known as a group of disease which is categorized as the uncontrolled or uneven growth and abnormal spreading of cells in the body. This uncontrolled growth of cells can lead towards the dysfunction of various organs and can cause death to the patient. According to a recent study presented by American Cancer Society researchers, it is found that 19% of cancers are caused due to smoking and 18% of cancers are caused due to excessive body weight, excessive alcohol consumption and poor nutrition.

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

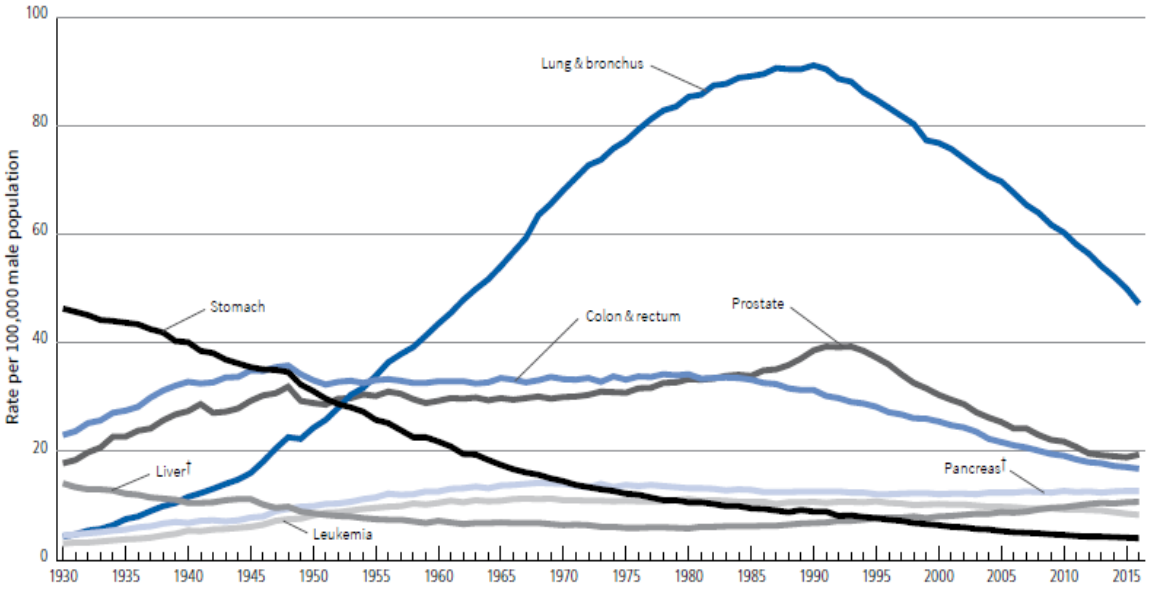

Figure 1: Cancer death rates [4]

Various types of cancer are present such liver, breast, stomach, pancreas etc. Figure 1 shows death rate due to cancer. According to this study, Lung cancer is a the most common cause of deaths. Lung cancer diagnosis is performed using different types of tests such as Imaging tests, Sputum cytology and Tissue sample (biopsy). After detecting the lung cancer, several types of treatments need to be performed such as surgery, radiation therapy, Chemotherapy, Radiosurgery, and Immunotherapy etc.

B. Cancer treatment, diagnosis and use of machine learning

Several mechanisms are present to detect and diagnose the cancer. However, recent technological advancements have enabled new paths to develop the automated system for Lung cancer detection. The imaging-based systems are widely adopted where machine learning methods are applied to classify the lung cancer. Kumar et al. [5] presented deep learning approach to classify the lung nodules, Hua et al. [6] introduced deep learning model on CT (Computed Tomography) images to classify the lung nodules, Shen et al. [7] presented multi-level convolutional neural network (CNN) based machine learning model. according to these models, the lung images are processed through the automated computer vision systems where different types of patterns are extracted and classified with the help of classifiers.

Generally, the clinical decisions are made based on the doctor's experience rather than extracting the rich hidden information in the database. hence, this leads towards the error and unwanted biases which affects the patient's health. Thus, data mining-based schemes are widely adopted to extract the rich information from the database to improve the quality of diagnosis system. Several methods have been introduced recently based on the concept of data mining such as backpropagation neural network [8], multi-class SVM [9], and back-propagation with decision tree [14] etc... The existing classifiers suffer from well-known challenge which caused due to the dataset dimensionality issue. In order to deal with dimensionality related issues, dimension reduction and feature selection techniques are introduced such as filter-based feature selection [11], wrapper feature subset selection [12], Fisher and RelieF feature selection [13]. However, the general classification algorithms suffer from various challenges where achieving the desired classification accuracy remains a tedious task. Hence, current researches have focused on hybrid classifier where optimization schemes are incorporated such as Selvanambi et al.

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

[10] presented recurrent neural network with glowworm swarm optimization, artificial bee colony for feature selection [15], particle swarm optimization [16], and cuckoo search optimization [17] etc.

The main objectives of this work are to study about Lung cancer, role of data mining and machine learning to predict the lung cancer to improve the diagnosis performance. Hence, we present a novel classification approach which uses an ensemble classification model using decision tree, Naïve Bayes and association rules. Finally, a majority vote model is also incorporated to achieve the better classification performance.

The rest of the article is organized as follows: section II presents a brief literature review study about recent techniques for lung cancer classification, section III presents proposed hybrid solution to improve the classification accuracy, section IV presents the outcome of proposed approach and comparative study, section V presents concluding remarks and future works.

II. Literature Survey

This section presents the brief overview about recent techniques of Lung cancer classification using data mining and machine learning. This study includes several aspects of data mining and machine learning such as data pre-processing, feature selection, and classifier construction. We present a brief discussion about recent techniques to improve the performance of lung cancer classification.

According to Cai et al. [18], the lung cancers are classified as: non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. In this work, authors presented an ensemble classification approach to classify LADC, SQCLC and SCLC cancer types based on machine learning techniques. This study is carried out on the cancer specific bio-markers. In order to achieve this, ensemble feature selection method is developed where incremental feature selection (IFS) strategy and Random Forest classifier is used to predict the lung cancer.

Montazeri et al. [19] presented heuristic approach for feature selection to classify the lung cancer using data mining technique. The genetic algorithm (GA) is used along with Pearson Correlation Coefficient to achieve the optimal attributes. After feature selection, the neural network classifier is implemented to classify the lung cancer.

Panthong et al. [12] introduced wrapper feature selection mechanism to overcome the dimensionality issue. The ensemble feature selection uses sequential backward selection (SBS), sequential forward selection (SFS), and evolutionary algorithms such as bagging and AdaBoost. Further, the decision tree and naïve Bayes classifier are used to classify the multidimensional data.

Rajan et al. [20] reported that the computational complexity and misclassification are the two challenging issues which affects the quality of prediction tools. In order to deal with these issues, authors presented neural network-based classification model and introduced multi-class neural network model.

Shakeel et al. [21] focused on the development of automated system to predict the lung cancer. Authors reported that overfitting and dimensionality are the challenging issues in this field of biomedical data classification. In order to overcome these issues, authors introduced a novel ensemble learning model using AdaBoost and Neural network. According to this approach, the noise from the data is removed by applying data smoothing and normalization process. Further, an optimization model is incorporated using minimum repetition and Wolf heuristic models which help obtain the optimized features by reducing the dimensionality issue.

Rani et al. [22] applied machine learning based approach on microarray data to classify the lung cancer. In microarray dataset, gene selection is considered as an important paradigm to improve the classification performance. Hence, authors presented hybrid gene selection approach using mutual information and genetic algorithm model to classify

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

the cancer data. First of all, mutual information model is applied which selects the genes which are having high mutual information related to cancer. These genes are processed through next step where genetic algorithm is applied to select the best features to improve the classification performance. Finally, Support Vector Machine (SVM) classifier model is applied to classify the data.

ALzubi et al. [23] introduced boosted neural network model to enhance the classification accuracy of lung cancer. According to this approach, an integrated Newton–Raphsons Maximum Likelihood and Minimum Redundancy (MLMR) is applied during pre-processing which helps to reduce the classification time. With the help of these attributes, an ensemble classification model is developed using weighted optimized neural network and boosting which acts as weak classifier and minimizes the classification error. Similarly, Hsu et al. [24] presented ensemble classification model. in this work, authors developed ensemble classification using neural network and decision tree classifier.

Venkataraman et al. [25] also focused on the feature selection and presented a novel hybrid model of feature selection for effective data classification. The main aim of feature selection is to reduce the classification time and increase the classification accuracy. Selvanambi et al. [26] presented a higher order neural network which is developed using recurrent neural network with Levenberg–Marquardt model. Furthermore, glowworm swarm optimization technique is also incorporated for feature selection and optimization.

This section presents a brief overview about recent techniques of lung cancer classification using machine learning techniques. From this discussion, we conclude that feature selection plays important role to improve the classification performance and reduces computational complexity. Moreover, hybrid classifier or ensemble classifier construction can help to achieve the better accuracy. However, existing schemes suffer from classification performance related issues hence in next section we present proposed solution to overcome thee issues faced in existing systems.

III. Proposed Model

In this section, we present the proposed solution to classify the lung cancer data with the help of ensemble learning. The complete approach is divided into following stages:

(a) Data preprocessing: first of all, we apply data pre-processing model where missing value imputation and data normalization is performed to minimize the redundancy in imbalanced data.
(b) Dimension reduction: in this stage, we apply dimensionality reduction model where we use filtering and correlation-based technique for optimal attribute selection.
(c) Classifier: finally, we introduce a novel ensemble classification model where decision tree and naïve Bayes are ensembled and majority voting is applied to obtain the prediction.

A. Missing value imputation

Missing value imputation is a process where the missing data is identified and replaced with a value which is closer to the exact value. This process is done based on the neighboring data. Mainly, the missing values are categorized as Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Due to increased use of data mining application, the demand of automated systems is increasing which can handle all types of missing values.

In this work, we use knn based missing value imputation model. Due to simplicity, less complexity and higher accuracy of imputation, we have adopted this in our model. Generally, the conventional methods use NN (Nearest Neighborhood) method but due to overfitting error, the existing NN model is extended to knn model. The KNN imputation is performed using following steps:

Step 1: Select set of attributes which are near to the missing values. Let the lung cancer data formulate a

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

matrix with attributes as C which contains m rows and columns. In order to impute the missing value x_{ci} of x_c , we need to find k other experiments with known attributes.

Step 2: in order to find the similarity between attributes, several techniques are presented in the literature such as Euclidean distance, Pearson correlation and variance minimization. However, outliers may affect the similarity measurement performance, hence, we consider Euclidean distance for measuring the similarity between attributes. This can be expressed as:

$$d_{ij} = dist(x_i, x_j) = \sqrt{\sum_{p=1}^{n} (x_{ip} - x_{jp})^2} \quad (1)$$

where $dist(x_i, x_j)$ is the Euclidean distance between x_i and x_j attributes, p denotes the experiment and n denotes total number of experiments

Step 3: In this step, we aim on predicting the missing value based on the average values of closest attribute values. The estimated missing value can be obtained as:

$$\tilde{x}_{gi} = \frac{\sum_{\forall x\_\alpha \in N_g} x_{\alpha i}}{k} \quad (2)$$

Where N_c is the nearest cancer data attributes, for each attribute, the weights are computed and the higher weight represents the more similar attributes.

Step 4: The weights for k nearest neighborhood can be computed as:

$$\tilde{x}_{gi} = \sum_{\forall x\_\alpha \in N_g} x_{\alpha i} W_i ; \qquad W_i = \frac{1/d_i}{\sum_{i=1}^{k} 1/d_i} \quad (3)$$

Where k is the closest attribute and d_i is the distance between i^{th} attribute and target missing value

B. Decision tree and Naïve Bayes ensemble classifier

In this section we present the proposed ensemble learning and classification approach by combining decision tree and naïve Bayes classification algorithm. First of all, we briefly describe the decision tree and naïve Bayes classifier, later, majority voting is incorporated to obtain the final classification decision.

C. Decision tree classifier

The decision tree classification approach is widely adopted in data mining task. This is considered as top-down greedy approach which provides the effective solution to the classification problem. Generally, the DT approach divides the training data into various smaller subsets until its final class is obtained. In this work, we adopted Iterative Dichotomiser (ID3) for DT classification. In this process, the root node is selected based on the highest information gain of the considered attributes. Let us consider that we have training data D and each instance of this data need to be classified as $x_i \in D$ belongs to a class $\mathbb{C}_i$, the probability that x_i belongs a class $\mathbb{C}_i$ is p_i, hence, the required average information to classify the x_i is expressed as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (4)$$

The goal of DT process is to divide the training data D into multiple subsets as $\{D_1, D_2, D_3, \dots D_n\}$ until the each D_i belongs to the class $\mathbb{C}_i$. Similarly, $Info_A(D)$ is the expected information which is required to classify correctly from D_i based on the partitioning attributes A. This can be computed as:

$$Info_A(D) = \sum_{i=1}^{n} \frac{|D_j|}{|D|} \times Info(D_j) \quad (5)$$

The information is defined based on the information required and current information which are obtained from eq. (4) and (5). Information gain can be computed as:

$$Gain(A) = Info(D) - Info_A(D) \quad (6)$$

Further, we incorporate a gini value for the dataset D as:

$$Gini(D) = 1 - \sum_{j=1}^{m} p_j^2 \quad (7)$$

Where p_j is the frequency of class $C_j \in D$

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

D. Naïve Bayes Classifier

Naïve Bayes classifier is known a probabilistic classification approach which can predict the probabilities of class membership. Let us consider that the training data is given as $\mathbb{D} = \{X_1, X_2, X_3, \ldots . X_n\}$ and each data has record as $X = \{x_1, x_2, x_3, \ldots x_n\}$, the data D contains the attributes as $\{A_1, A_2, .. A_n\}$ and each attributes has different values as $\{A_{i1}, A_{i2}, .. A_{in.}\}$. Each attribute has classes as $C = \{C_1, C_2, .. C_m\}$. in order to predict the class of attribute, the NB classifier computes the highest probability and predicts that X belongs to class C_i if following probabilistic condition is satisfied, as

$$P(C_i|X) > P(C_j|X) \; for \; 1 \le j \le m, j \ne i \quad (8)$$

To represent the X_i belongs to the class C_i, the $P(C_j|X)$ is maximized which is known as e Maximum Posteriori Hypothesis and can be expressed as:

$$P(C_i|X) = \frac{P(X|C_i)}{P(X)} \quad (9)$$

Computing the value of $P(X|C_i)$ is tedious task because it incudes estimation of exponential number of joint-probabilities of features. In order to achieve appropriate estimation, Naïve Bayes classification assumes that class labels and attributes are independent in each class. Thus, the Bayes rule can be represented defined as:

$$P(X|C_i) = \prod_{i=1}^{n} P(X_i|C) \quad (10)$$

which shows that we need to compute feature value in each class to estimate the conditional probability in the features. This helps to reduce the computations of joint-probabilities. During training phase, the NB classifier learns the patter for each class based on their estimated probabilities. Similarly, during testing phase, the largest probability value of X_i will be predicted as the output class C_i, given as:

$$P(C_i|X) \propto P(C) \prod_{i=1}^{n} P(X_i|C) \quad (11)$$

E. Ensemble classification and majority voting

In this section, we present the combined model of decision tree and NB classifier. In order to achieve this, first of all, an entropy value need to be defined which is used for characterizing the instances in the given dataset. Let us assume that we have obtained S number of attribues with V different values , then the entropy can for different attributes can be defined as:

$$E(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (12)$$

Further, the reduction in entropy due to partition is defined as

$$Gain(S, A) = E(S) - \sum_{v \in value\,(A)} \frac{|S_v|}{|S|} E(S) \quad (13)$$

$value(A)$ denotes the all possible values of A, S_v is the subset of S whose attribute A has value v.Finally, a majority voting based combining rule is incorporated to select the final values which can be given as:

$$\Delta_{ki} = \begin{cases} 1 \; if \; P(C_i|X) = \max_{j=1}^{K} P(C_i|X_i) \\ 0 \; otherwise \end{cases} \quad (14)$$

IV. Results And Discussion

In this section, we present the outcome of proposed ensemble classifier to classify the lung cancer. The complete experimental study is carried out using MATAB simulation tool. The obtained performance is compared with existing techniques to show the robust performance of proposed approach.

A. Dataset details

In this work, we have used two dataset which are obtained from UCI dataset repository. These datasets are known as

(a) UCI lung cancer dataset: this dataset contains 32 samples with 56 attributes and 1 class attribute. Three class labels are present as type A, type B and type C.

(b) Thoracic Surgery Dataset: this dataset contains the record of patients who have

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

undergone the major lung resection during the years 2007-2011.

B. Performance measurement

The performance of proposed model is measured with the help of classifier's confusion matrix. We have considered several parameters such as classification accuracy, precision, recall, F1-score, sensitivity and specificity. Table 1 show the representation of confusion matrix in classification models.

Table 1: Confusion matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Where TN = true negative, TP= true positive, FN= false negative, and FP = false positive.

With the help of this, the classification accuracy can be computed as:

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

Along with accuracy, we have several parameters which are used to analyze the robustness of classifier.

Sensitivity: it is the measurement of classifier's ability to classify the patterns in positive class. It can be computed as:

$$Sensitivity = \frac{(TP)}{(TP + FN)} \quad (16)$$

Specificity: it is the measurement of classifier's ability to classify the patterns in negative class. It can be computed as:

$$Sensitivity = \frac{(TN)}{(TN + FP)}$$

F-Measure: this is computed with the help of precision and recall (sensitivity) where precision is given as $Precision = \frac{(TP)}{(TP+FP)}$ and Recall is computed as $Recall = \frac{(TP)}{(TP+FN)}$. With the help of these parameters, the F-measure is computed as:

$$F - Measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (17)$$

C. Comparative performance for UCI lung cancer dataset

This section presents experimental study for UCI lung cancer dataset and obtained performance is compared with existing techniques. For this experiment, we divide the complete data as 70% training and 30% for testing. Below given table shows confusion matrix for UCI lung cancer data.

Table 2: Confusion matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN (28)	FP (1)
Actual Positive	FN (0)	TP (3)

The classification accuracy for this data is obtained as 96.8%. Similarly, other statistics parameters are obtained. A comparative study with different classifiers is preseted in table 3.

Table 3: Comparison of classifiers

Classifiers	Accuracy	Precision	Recall	F-Measure
Proposed Model	96.87	100	75	85.71429
Decision Tree [27]	78.33	68.75	70.57	69.65
Naïve Bayes [27]	85	77.08	79.71	78.37
Random Forest [27]	79.17	39.15	50	43.91
Neural Network [27]	71.67	63.18	66.57	64.83
SVM [27]	79.17	66.07	60.28	63.04
Gradient Boosted Tree [27]	90	87.82	83.71	85.71
MLP [27]	78.33	68.75	70.57	69.65
Majority voting [27]	88.57	84.44	76.57	80.31

The comparative study is also depicted in below given figure 2 where we have compared the performance of each algorithm in terms of F-measure, Recall, precision and accuracy.

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

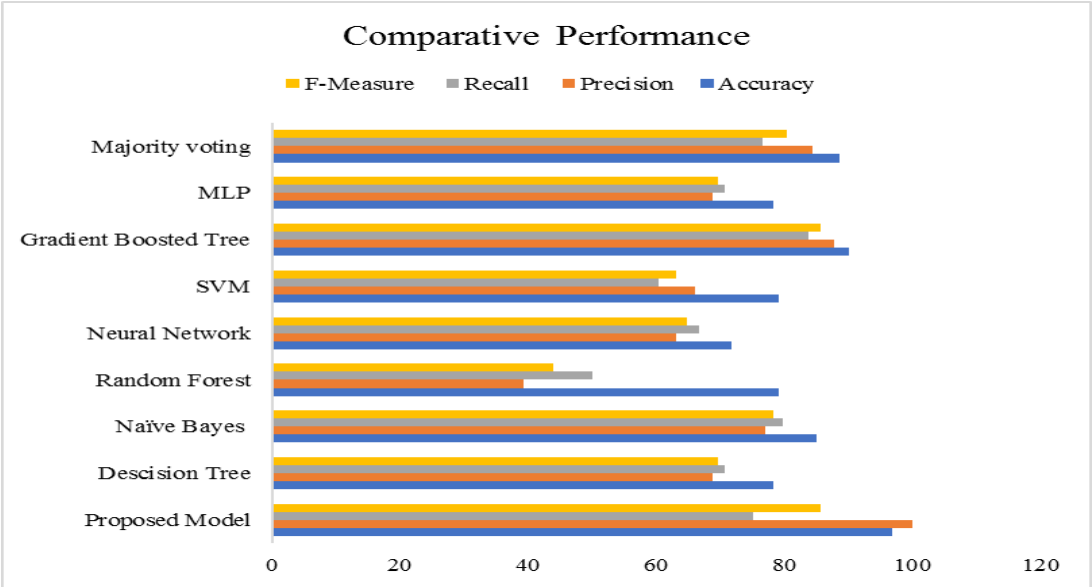

Figure 2: Comparative performance analysis

D. Comparative performance for Thoracic Surgery Dataset

Several existing techniques such as Conjunctive Rule Application, Decision Table Application, DTNB Application, JRip Application, NNge Application, OneR Application, PART Application, Ridor Application, and ZeroR Application have been reported in [28]. We compare the classification accuracy score with these techniques as reported in table 4.

Table 4: Comparative performance for Thoracic Surgery Dataset

	Correctly Classified Instances	Root Mean Squared Error	Accuracy
Conjunctive Rule	399	0.359	84.8936
Decision Table	397	0.3635	84.4681
DTNB Application	384	0.3651	81.7021
JRip Application	398	0.3586	84.6809
NNge Application	381	0.4362	81.0638
OneR Application	392	0.4074	83.4043
PART Application	372	0.4155	79.1489
Ridor Application	399	0.3887	84.8936
ZeroR Application	400	0.356	85.1064
Proposed Model	422	0.215	89.7872

This comparative study shows that proposed approach correctly classifies 422 instances out of 470. Proposed approach achieves 89.78% classification accuracy with less root mean squared error.

V. CONCLUSION

In this work, we have studied about the Lung cancer and role of data mining and machine learning to detect and classify the cancer. Several classification studies have been reported during last decade but dimensionality, computational complexity and classification accuracy remains challenging task. Moreover, missing values and optimal feature selection are also considered as important aspects of data mining. Hence, in this work, we present KNN based missing value imputation and later constructed decision tree and Naïve Bayes ensemble

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

classification model with the help of majority voting. The performance of proposed approach is measured in terms of classification accuracy, precision, recall and root mean square error. The comparative study shows the proposed approach achieves better performance when compared with the existing techniques.

REFERENCES

[1]Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., & Bray, F, "Estimating the global cancer incidence and mortality in 2018": GLOBOCAN sources and methods. International journal of cancer, 144(8), 1941-1953,2019.

[2] Ormazabal, V., Nair, S., Elfeky, O., Aguayo, C., Salomon, C., & Zuñiga, F. A. (2018)," Association between insulin resistance and thedevelopment of cardiovascular disease. Cardiovascular diabetology", 17(1), 122.

[3] Chen, P., & Pan, C. "Diabetes classification model based on boosting algorithms". BMC bioinformatics, 19(1), 109.2018.

[4] https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf

[5] Kumar, D., Wong, A., & Clausi, D. A,"Lung nodule classification using deep features in CT images", In 2015 12thConference on Computer and Robot Vision", (pp. 133-138). IEEE.

[6] Hua, K. L., Hsu, C. H., Hidayati, S. C., Cheng, W. H., & Chen, Y. J,"Computer-aided classification of lung nodules on computed tomography images via deep learning technique," OncoTargets and therapy, 2015.

[7] Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J," Multi-scale convolutional neural networks for lung nodule Classification", In International Conference on Information Processing in Medical Imaging (pp. 588-599). Springer, Cham, June 2015.

[8] Varadharajan, R., Priyan, M. K., Panchatcharam, P., Vivekanandan, S., & Gunasekaran, M.,"A new approach for prediction of lung carcinoma using back-propagation neural network with decision tree classifiers", Journal of Ambient Intelligence and Humanized Computing, 1-12,2018.

[9] Alam, J., Alam, S., & Hossan, A,"Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier", In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE, February 2018.

[10] Selvanambi, R., Natarajan, J., Karuppiah, M., Islam, S. H., Hassan, M. M., & Fortino, G,"Lung cancer prediction using higher-order recurrent neural network based on glowworm swarm optimization", Neural Computing and Applications, 1-14,2018.

[11] Lee, I. H., Lushington, G. H., & Visvanathan, M.," A filter-based feature selection approach for identifying potential biomarkers for lung cancer". Journal of Clinical Bioinformatics, 1(1), 11.2011.

[12] Panthong, R., & Srivihok, A. ," Wrapper feature subset selection for dimension reduction based on ensemble learning. Algorithm", Procedia Computer Science, 72, 162-169,2015.

[13] Xie, N. N., Hu, L., & Li, T. H. (2014)," Lung cancer risk prediction method based on feature selection and artificial neural Network", Asian Pac J Cancer Prev, 15(23), 10539-10542.

[14] Hsu, C. H., Manogaran, G., Panchatcharam, P., & Vivekanandan, S," A New Approach for Prediction of Lung Carcinoma Using Back Propagation Neural Network with Decision Tree Classifiers", In 2018 IEEE 8th International Symposium on Cloud and Service Computing (SC2) (pp. 111-115). IEEE, November 2018.

May – June 2020
ISSN: 0193-4120 Page No. 132 - 141

[15] Shunmugapriya, P., & Kanmani, S," A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid) Swarm and evolutionary computation", 36, 27-36,2017.

[16] Xi, M., Sun, J., Liu, L., Fan, F., & Wu, X," Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine", Computational and mathematical Methods in Medicine. 2016.

[17] Gunavathi, C., & Premalatha, K. (2015),"Cuckoo search optimisation for feature selection in cancer classification: a new approach", International journal of data mining and bioinformatics, 13(3), 248-265

[18] Cai, Z., Xu, D., Zhang, Q., Zhang, J., Ngai, S. M., & Shao, J," Classification of lung cancer using ensemble-based feature selection and machine learning methods", Molecular bioSystems, 11(3), 791-800.2015.

[19] Montazeri, M., Baghshah, M. S., & Enhesari, A," Hyper-Heuristic algorithm for finding efficient features in diagnose of lung cancer disease", arXiv preprint arXiv:1512.04652.2015.

[20] Rajan JR.,Chelvan A.C.,& Duela,J.S," Multi-Class Neural Networks to Predict Lung Cancer. Journal of medical systems", 43(7), 211.2019.

[21] Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z., & Jaber, M. M ,"Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks",Neural Computing and Applications, 1-14.2019.

[22] Rani, M. J., & Devaraj, D," Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification", Journal of medical systems, 43(8), 235.2019.

[23] ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C ," Boosted neural network ensemble classification for lung cancer disease diagnosis", Applied Soft Computing, 80, 579-591.2019.

[24] Hsu, C. H., Manogaran, G., Panchatcharam, P., & Vivekanandan, S ,"A New Approach for Prediction of Lung Carcinoma Using Back Propagation Neural Network with Decision Tree Classifiers", In 2018 IEEE 8th International Symposium on Cloud and Service Computing (SC2) (pp. 111-115). IEEE, November 2018.

[25] Venkataraman, S., & Selvaraj, R. ,"Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification. In Advances in Systems, Control and Automation ",(pp. 499-514). Springer, Singapore 2018.

[26] Selvanambi, R., Natarajan, J., Karuppiah, M., Islam, S. H., Hassan, M. M., & Fortino, G ," Lung cancer prediction using higher- order recurrent neural network based on glowworm swarm optimization", Neural Computing and Applications, 1-14.2018.

[27] Faisal, M. I., Bashir, S., Khan, Z. S., & Hassan Khan, F ," An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer", 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST),2018.

[28] Koklu, M., Kahramanli, H., & Allahverdi, N ," Applications of Rule Based ClassificationTechniques for Thoracic Surgery",. In Joint International Conference 2015, TIIM,2015.