Applying Hybrid Classification Data Mining Techniques to Improve Lung Cancer Diagnosis

Diljot Singh, Prabhdeep Singh, Dr. Rajbir Kaur
KCET Amritsar, IKG-PTU Kapurthala, Punjabi University, Patiala
diljot.asr@gmail.com, ssingh.prabhdeep@gmail.com, Rajbir277@yahoo.co.in

Abstract: Enormous deaths are being caused by Lung cancer, one in all the harmful disease around the world. The only possible approach to improve a patient's probability for survival is the early detection of it and if it is detected beforehand, it can help to fix the disease completely. So the demand of the techniques to find the existence of cancer nodule within the early stage is expanding. Prior diagnosis of lung cancer will definitevely saves huge lives, and failing to do so will lead to severe problems resulting in unexpected fatal finish. Data mining is an incredible method to help individuals in their wellbeing, Scientific and Engineering. It uses a learning strategy to understand the data patterns. Those techniques are extracting the key information from the huge databases, which helps us to find the pattern and the relationship from the data. The Hybrid classifier is very useful in the classification of lung cancer dataset as it gives a lot higher exactness than alternative classifiers. WEKA 3.6.10 is used as a data mining tool for evaluations and results to be carried out.

Keywords: Lung Cancer. Data Mining, Classification, Weka, SMOTE

1. INTRODUCTION

Acceptance and assumption of lung malignancy in the most punctual reference point stage can be valuable to improve the survival step of patients. Be that as it may, a finding of disease is one the essential testing task for the radiologist. For distinguishing, foreseeing, and diagnosing lung disease, a intelligence PC helped analysis framework can be especially valuable for the radiologist. So as to diminish preventable instances of malignancy related passings and improve the general wellbeing framework, a progressively flexible and ground-breaking innovation is required. actually AI and information mining procedures can process wellbeing information to all the more likely gauge individuals' future wellbeing conditions and potential dangers. Be that as it may, propelled AI instruments can't be utilized without critical preparing and aptitude, which medicinal specialists don't have. Along these lines, we built up a model of a Lung Cancer expectation System at a beginning period that goes for making those strategies available in an easy to understand way so as to increase the learning of the master. The examination means to utilize different information mining systems to arrange the hazard elements of lung malignancy.

Implement information mining procedures to lung malignancy information is valuable to rank and connection disease attribute to the survival result. Further, exact result expectation can be valuable for specialists and patients to gauge survivability, yet additionally help in basic leadership to decide the best course of treatment for a patient, in light of patient-explicit traits, as opposed to depending on close to home encounters, accounts, or populace wide hazard evaluations. Analyses with a few classifiers were directed to locate that numerous meta-classifiers utilized with choice trees can give memorable outcomes, which can be additionally improved by joining the subsequent forecast probabilities from a few classifiers utilizing a troupe casting a ballot conspire

2. KNOWLEDGE DISCOVERY PROCESS AND DATA MINING.

The term Knowledge Discovery in Databases or KDD for short alludes to the full course of discovering information in information and accentuates the "abnormal state" capacity of specific information mining strategies. The joined point of the KDD procedure is to uncover information from the enormous stores. Information mining strategies to destroy out the data as indicated by the state of activities, utilizing a database alongside any fundamental preprocessing, examining, and changes of that database.The KDD procedure can be ordered as under::

A. Selection

It includes choosing data that is relevant to the task of research from the database.

B. Pre-Processing

We remove noise and inconsistency in this phase thatis fund in data and mix various data sources.

C. Transformation

This phase includes transformation of data taking place into appropriate forms to perform mining activities.

D. Data Mining

In this phase,data mining algorithm is applied for the extraction of the patterns.

E. Interpretation/Evaluation

It includes discovering consistent patterns of information hidden within the data.

2.1 Classification and Prediction

Classification is the course of finding a model that describes the data classes. The reason is to have an option to utilize the model to figure out the category of objects whose class mark is unidentified. The model that is obtained depends on the analysis of sets of training data. The obtained model can be presented in the following forms:

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

2.2 Classification

It predicts the category of objects whose class name is unidentified. Its main purpose is to find a model that describes and distinguishes data categories. The Derived Model depends on the study set of training data, i.e. the data object whose category label is accepted.

Prediction

It is used to predict the engaged or missing numerical data values instead of category labels. Regression study is normally used for prediction. Prediction can even be used for recognition of distribution trends based on the data.

3. RESULTS AND ANALYSIS

There are many research papers that include mining of the datasets that applies different algorithms with different statistics. Therefore, with the approach of Hence, with the arrival of improved and changed prediction techniques, there is a requirement for an analyst to understand which algorithm suits best for a specific set of dataset. The following steps are used in the Experiment and analysis:

3.1 DATASET COLLECTED

The dataset chosen for this work in lungcancer.arff. The dataset contains fifty six attributes, one category attribute, and thirty two instances. The remainder of the attributes indicate the areas where lung cancer starts. The dataset that we used here has been gathered from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

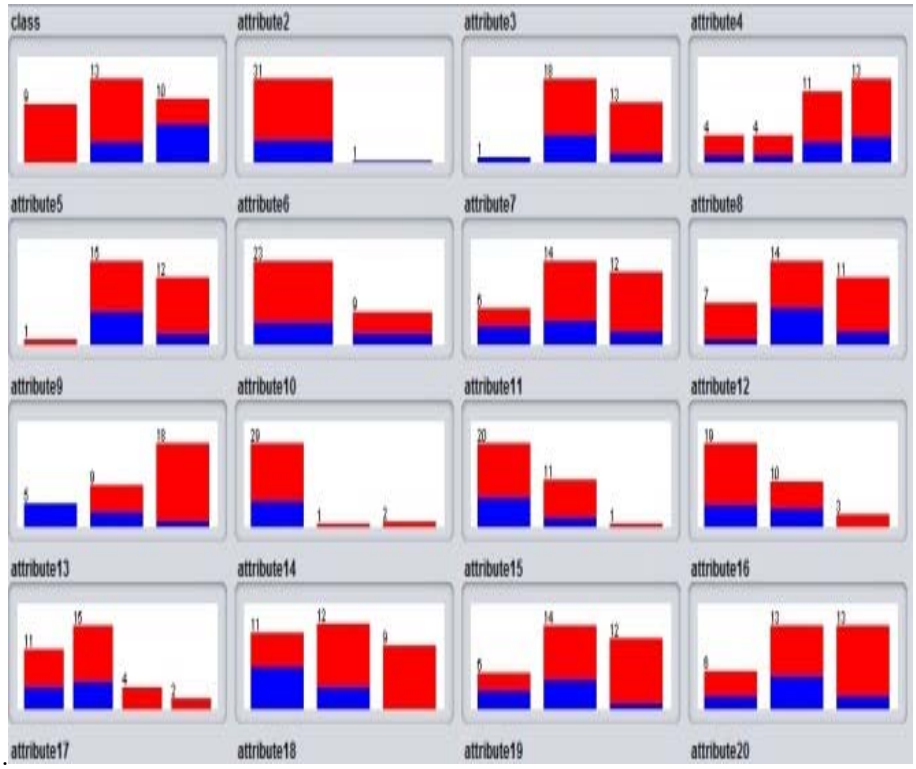

Figure 1: Data Visualization of Lung Cancer Dataset

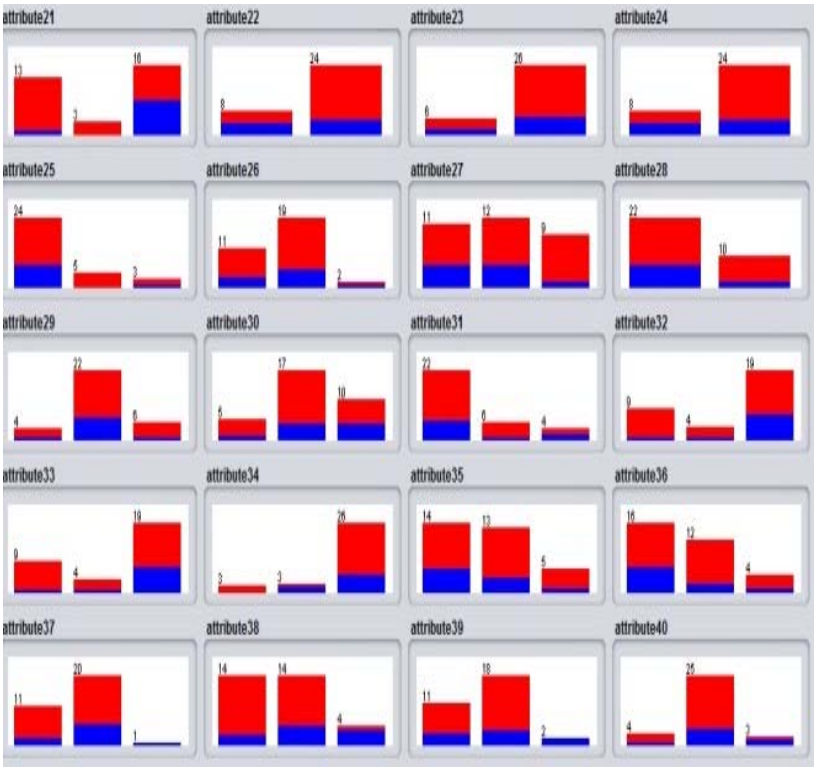

Figure 2: Data Visualization of Lung Cancer Dataset

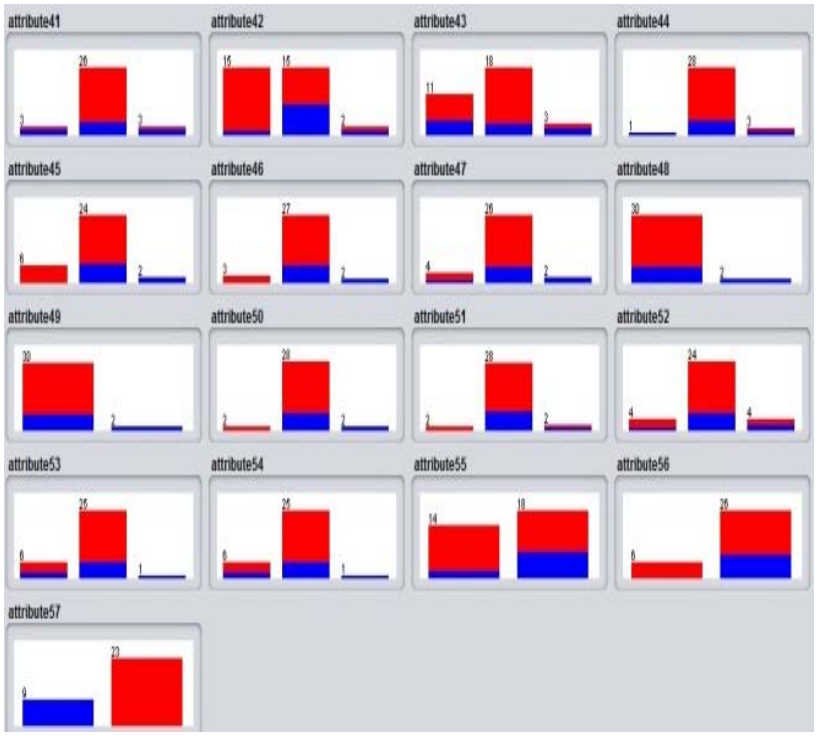

Figure 3: Data Visualization of Lung Cancer Dataset

3.2 DATA PREPROCESSING

By using the unsupervised filter, remove the useless variables. Now, remove the useless values by using the unsupervised filter. It has non-uniform category distribution among the instances due to its distribution. Therefore, Synthetic Minority Oversampling Technique (SMOTE) algorithm of supervised technique for resampling the unbalanced dataset with Randomize algorithm is used to remove the preference regading majority classes in predictions during testing. The instances of minority classes are included in the loop for seven times by applying SMOTE technique.

3.3 DATA MINING CLASSIFICATION METHODS

There are different information mining systems accessible with their appropriateness relies upon the area application. Measurements give a solid principal foundation to measurement and assessment of results. In any case, calculations dependent on measurements should be adjusted and scaled before they are connected to information mining. We currently portray a couple of Classification information mining systems with representations of their applications to social insurance. A. Guideline set classifiers Complex choice trees can be hard to understand,for example, since data around one class is typically appropriated all through the tree. C4.5 presented an elective formalism comprising of a rundown of guidelines of the structure "if A and B and C and ... at that point class X", where standards for each class are gathered.

A case is arranged by finding the principal rule whose conditions are fulfilled by the case; if no standard is fulfilled, the case is allocated to a default class. On the off chance that conditions, THEN end This sort of principle comprises of two sections. The standard antece t (the IF part) contains at least one conditions about the estimation of indicator traits while the standard resulting (THEN part) contains an expectation about the estimation of an objective quality. A precise forecast of the estimation of an objective trait will improve the basic leadership process. On the off chance that expectation principles are common in information mining; they speak to found learning at an abnormal state of deliberation. In the human services framework, it tends to be connected as pursues: (Symptoms) (Previous - history) → (Cause—of - illness). Model 1: If_then_rule incited in the conclusion of the degree of liquor in the blood. In the event that Sex = MALE AND Unit = 8.9 AND Meal = FULL, THEN Diagnosis=Blood_alcohol_content_HIGH. B. Choice Tree calculation It is an information portrayal structure comprising of hubs and branches sorted out as a tree to such an extent that, each inside non-leaf hub is named with estimations of the properties. The branches turning out from an interior hub are named with estimations of the characteristics in that hub. Each hub I named with a class (an estimation of the objective characteristic). Tree-based models, which incorporate order and relapse trees, are the regular execution of enlistment displaying. Choice tree models are most appropriate for information mining. They are cheap to build, simple to translate, simple to incorporate with a database framework, and they have similar or better precision in numerous applications. There are numerous Decision tree calculations, for example, HUNTS calculation (this is perhaps the soonest calculation), CART, ID3, C4.5 (a later form ID3 calculation), SLIQ, SPRINT The choice tree is worked from the negligible preparing set . In this table, each line compares to a patient record. We will allude to a line as an information case. The informational collection contains three indicator characteristics, specifically Age, Gender, Intensity of manifestations and one objective trait, to be specific infection whose qualities (to be anticipated from side effects) demonstrates whether the relating patient has a specific sickness or not.

Different classifiers are connected to the dataset utilizing WEKA. The outcomes demonstrate that in the event that there are in any event two classes, at that point multiclass classifier with Random backwoods calculation can furnish preferred outcomes with exactness 85.7% over double classifiers, for example, Random Forest with precision 84.5%. As the Random Forest calculation works best among all arrangement calculations, so this calculation has been decided for applying with a multiclass classifier to improve exactness and accuracy. The correlation among all calculations is appeared in table 2, which demonstrates that for the multiclass datasets multiclass classifier with the arbitrary woodland containing 10 irregular trees works best among all calculations. As appeared in the table, Accuracy of 85.7% with ROC zone is 0.997 has been accomplished by the multiclass classifier with arbitrary timberland. Figure 3 demonstrates the correlation of calculations dependent on Kappa insights. Graphically figure 2 demonstrates the presentation of Random backwoods with multiclass classifier accomplishes a high precision pace of 85.7 % among all classifiers. Figure 4 demonstrates the ROC region of classifiers, where multiclass classifier accomplishes high ROC zone.

3.4 CLASSIFICATION TECHNIQUES

We have used various classification schemes which results in the identification of top 5 classification schemes, and therefore using ensemble voting scheme to mix the prediction probabilities from the best five models. A brief decription of the meta-classifiers and classifiers that were used in the experiments are reported in this paper.

3.4.1 Naïve Bayes Classifier: The well known approach in the theory of supervised parametric classifiers is that the quadratic discriminate function which utilizes the Bayesian approach. The target here is yo purpose a standard which allows the assigning of future articles to a class when a lot of items is given for each class.

3.4.2 Logistic regression (LR)

It is a classification model with a related training algorithm that binds together Logistic Regression (LR) and decision tree learning. This is also known as Logic Model, which is utilized to display the dichotomous results of variables. Logistic model trees depend on the earlier notion of a model tree: a decision tree that haslinear regression models at its leaves to give a piecewise linear regression model.

3.4.3 Multi-Layer Perception

It is a fast decision tree learner, which builds a regression tree and a decision tree using information gain because the splitting criterion and prunes it using low-error snipping.

3.4.4 J48 decision tree.

In a decision tree classifier, the interior nodes represents different attributes, whose values would be utilized to choose the classification path, and the branches indicate the split depending on the attribute values, while the leaf node represents the final valueo of the dependent variable. While constructing the decision tree, J48 algorithm identifies the property that should be used to split the tree.

3.4.5 Random forest

The Random Forest classifier comprises of several decision trees. The final class of a case in a Random Forest is assigned by outputting the category i.e. the output mode of individual trees, which can deliever strong and exact classification, and can deal with a signifcant number of input variables. It is generally powerful against overfitting and may deal with datasets with highly imbalanced class distributions

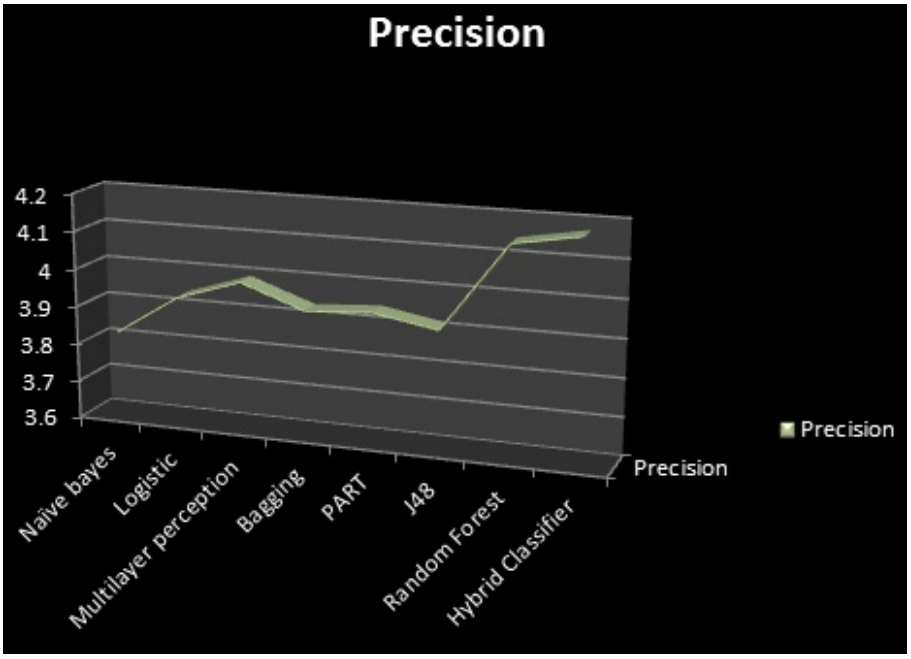

Figure 4: Precision

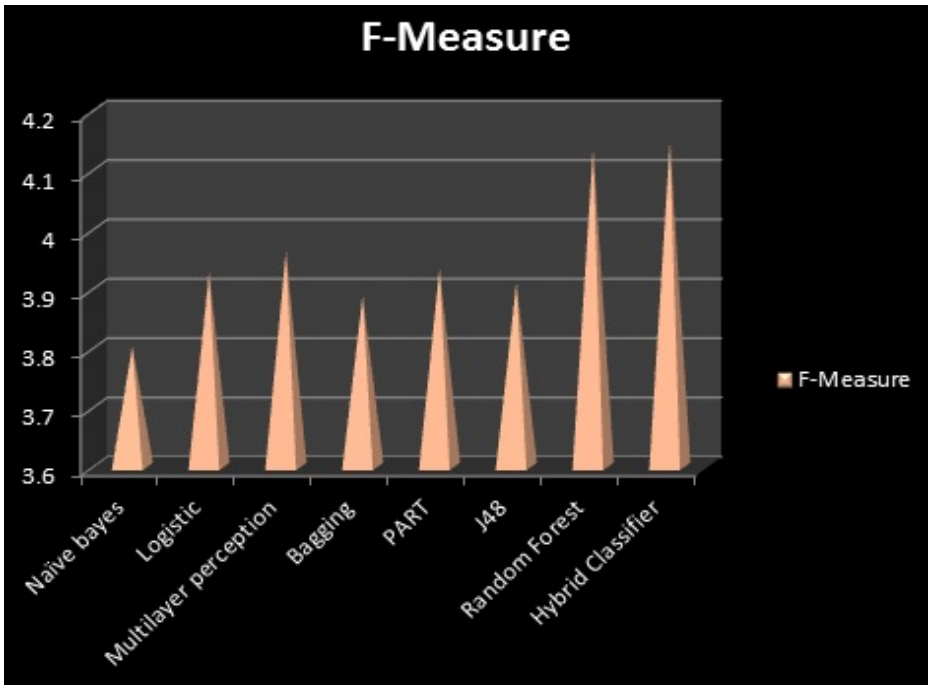

Figure 5: F- Measure

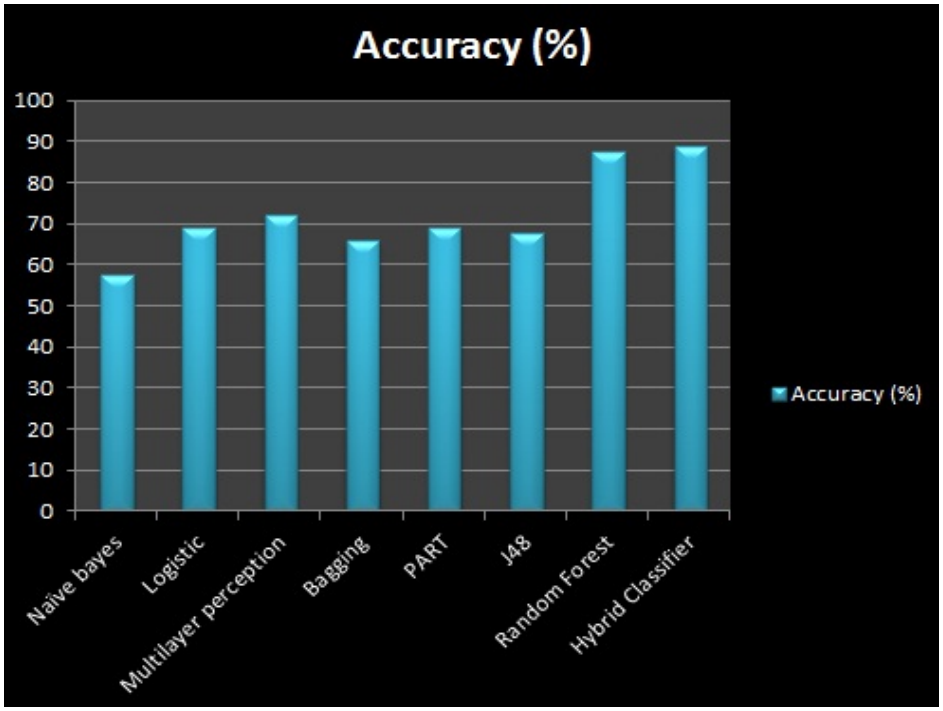

Figure 6: Accuracy (%)

3.4.6 PERFORM TESTING

The above figure depicts the predictions of Lung Cancer classes. In step 2, the testing is performed on the model selected on the test dataset by taking some number of occourences to to predict the lung cancer.

4. CONCLUSION

In this study, the matter of predicting Lung Cancer is talked about. The primary focus is on using different algorithms, and a mix of using several targets attribute for Lung cancer prediction using data mining techniques. Data mining algorithm is selected depending upon the particulars of the dataset. It is discovered that with at most two classes, binary classifiers are precise whereas with an increasing range of the category multiclass classifier with binary classifier like Random forest is more accurate. It is likewise discovered that with SMOTE strategy improves the result of selected classifier with better precision. In future, the work can be extended and improved for the automation of Lung cancer prediction.

REFERENCES

[1] Pati, Jayadeep. "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach." IEEE Access 7 (2019): 4232-4238.
[2] Wu, Jia, Peiyuan Guan, and Yanlin Tan. "Diagnosis and Data Probability Decision Based on Non-Small Cell Lung Cancer in Medical System." IEEE Access 7 (2019): 44851-44861.
[3] Tian, Suyan. "Identification of monotonically differentially expressed genes for non-small cell lung cancer." BMC Bioinformatics 20, no. 1 (2019): 177.
[4] Lakshmanaprabu, S. K., Sachi Nandan Mohanty, K. Shankar, N. Arunkumar, and Gustavo Ramirez. "Optimal deep learning model for classification of lung cancer on CT images." Future Generation Computer Systems 92 (2019): 374-382.
[5] ur Rehman, Muhammad Zia, Muzzamil Javaid, Syed Irtiza Ali Shah, Syed Omer Gilani, Mohsin Jamil, and Shahid Ikramullah Butt. "An appraisal of nodules detection techniques for lung cancer in CT images." Biomedical Signal Processing and Control 41 (2018): 140-151.
[6] Wang, Kung-Jeng, Jyun-Lin Chen, and Kung-Min Wang. "Medical expenditure estimation by Bayesian network for lung cancer patients at different severity stages." Computers in biology and medicine 106 (2019): 97-105.
[7] Garcia-Gathright, Jean I., Nicholas J. Matiasz, Carlos Adame, Karthik V. Sarma, Lauren Sauer, Nova F. Smedley, Marshall L. Spiegel et al. "Evaluating Casama: Contextualized semantic maps for summarization of lung cancer studies." Computers in biology and medicine 92 (2018): 55-63.
[8] Zhang, Guobin, Shan Jiang, Zhiyong Yang, Li Gong, Xiaodong Ma, Zeyang Zhou, Chao Bao, and Qi Liu. "Automatic nodule detection for lung cancer in CT images: A review." Computers in biology and medicine (2018).
[9] Junior, José Raniery Ferreira, Marcel Koenigkam-Santos, Federico Enrique Garcia Cipriano, Alexandre Todorovic Fabro, and Paulo Mazzoncini de Azevedo-Marques. "Radiomics-based features for pattern recognition of lung cancer histopathology and metastases." Computer methods and programs in biomedicine 159 (2018): 23-30.
[10] Huidrom, Satish Chandra, Yambem Jina Chanu, and Khumanthem Manglem Singh. "Automated Lung Segmentation on Computed Tomography Image for the Diagnosis of Lung Cancer." Computación y Sistemas 22, no. 3, (2018).
[11] Zhang, Zhichao, Yuan Zhang, Lina Yao, Houbing Song, and Anton Kos. "A sensor-based wrist pulse signal processing and lung cancer recognition." Journal of biomedical informatics 79 (2018): 107-116.
[12] Yoo, Illhoi, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. "Data mining in healthcare and biomedicine: a survey of the literature." Journal of medical systems 36, no. 4 (2012): 2431-2448.
[13] Cubillas, Juan José, M. Isabel Ramos, Francisco R. Feito, and Tomás Ureña. "An improvement in the appointment scheduling in primary health care centres using data mining." Journal of medical systems 38, no. 8 (2014): 89.

[14] Itani, Sarah, Fabian Lecron, and Philippe Fortemps. "Specifics of medical data mining for diagnosis aid: A survey." Expert Systems with Applications (2018).
[15] Dreisbach, Caitlin, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data." International journal of medical informatics (2019).
[16] Francis, Bindhia K., and Suvanam Sasidhar Babu. "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach." Journal of medical systems 43, no. 6 (2019): 162.
[17] Ekin, Tahir, Greg Lakomski, and Rasim Muzaffer Musal. "An unsupervised Bayesian hierarchical method for medical fraud assessment." Statistical Analysis and Data Mining: The ASA Data Science Journal 12, no. 2 (2019): 116-124.