**collation of cancer in lungs utilizing MLOPs strategies**

**Hitesh Kumar1, Nikita Narwat2, Madhulika Bhatia3**

1,2,3Amity University, Noida

**ABSTRACT** **—** After witnessing COVID-19, the gravity of the issue of early detection of a disease is apparent. Lung cancer has proven to be a prominent factor for death worldwide, with an alarming rate of about five million fatal cases per year. In lung cancer, the cancerous cells in the lungs multiply uncontrollably. Three popularly used classifiers for data analytics viz. Logistic Regression, Decision Tree and Neural Networks, were used to analyze lung cancer detection. This work revolves around the objective of investigating the effectiveness of classification algorithms in the early identification of lung cancer. This work has made use of a lung cancer dataset from an online cancer detection system in the public domain. After efficient data mining, classification models have been trained for lung cancer prediction. The experiment is performed on Azure Machine Learning Studio, Microsoft’s development environment.

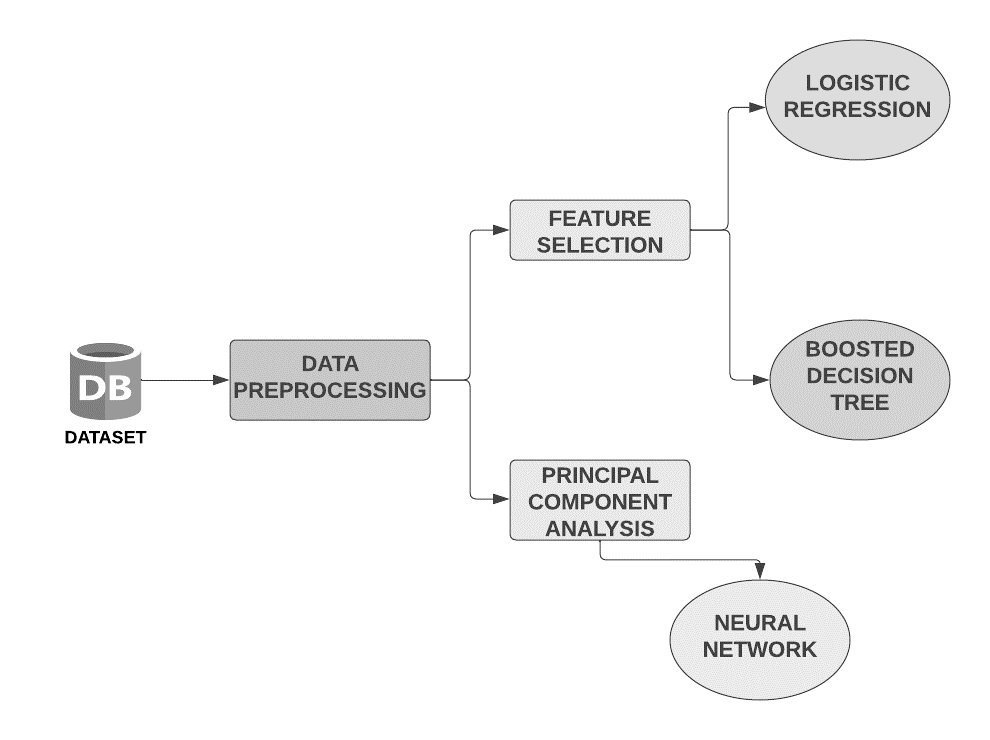
**Keywords —** lung cancer, machine learning, classification, neural network, decision tree, logistic regression

1. **INTRODUCTION**

Lung cancer has been major contributor for the mortality rate around the globe. Windpipe, major airway, or lungs are the initiation sites for the attack by lung cancer. Unregulated proliferation of malignant cells from the lungs. People diseased with emphysema and chronic bronchitis, or a history of chest difficulties are often accompanied with lung cancer. Smoking is a prevailing factor for lung cancer in Indian men; however, smoking is less observed in women of India, implicating that there are other causes that contribute to the spread of lung cancer. The workplace subjection to Radon gas, air pollution, and toxins, are all risk factors. The incubation period of lung cancer is long and are often diagnosed within the age bracket 55 to 65 [1]. As per the studies, work habits and social habits are also a major factor in the aggravation of the disease. A greater awareness of risk factors can aid in the prevention of lung cancer. Early discovery of the disease eases the process of treatment thereby increasing the survival rate of the patient. Early discovery plays a vital role in decreasing mortality rate by utilizing machine learning (ML) approaches, and if this can be used to make the diagnosis more methodical for radiologists, it will lead to early detection.

Lungs are the major site for primary lung cancer whereas in secondary lung cancer, cancerous cells developed in lung are propagated to other body organs. The measure and expansion of the tumor in body establishes its stage.

Kaggle Repository is the source of the dataset employed in this analysis. After preprocessing of the data, and the dataset is split into train and test data. The classification models such as logistic regression, boosted decision trees, and neural networks are applied over the lung cancer dataset having 2 class as the target variable values. To determine the weights of the model, they are trained over the training data. Then, they are scored on different performance measures like F1 score, recall for accuracy evaluation, by testing them over testing data. Finally, using the accuracy metrics, we can compare the different classifiers for the optimal selection of the classification model for the given dataset. For overview of the experiment, refer to *fig 1.*



*Fig 1: Overview of Experiment*

1. **RELATED WORK**

Many researchers have contributed to numerous lung cancer prediction and classification studies. Using predictive data mining methods, Danjuma [2] evaluates algorithms like Decision Trees, Naive Bayes, and Artificial Neural Networks to determine how long lung cancer patients can expect to live following surgery. The aforementioned algorithms were subjected to a stratified 10-fold cross-validation analysis, and the models' accuracy metrics were assessed.

With the collected lung cancer dataset, Zehra et al. [3] produced various outcomes for each classifier. Following the implementation of the classifiers, comparable accuracy rates for KNN, SVM, NN, and Logistic Regression were discovered. With 99.3% accuracy, Support Vector Machine comes in first. Doctors were able to make more accurate judgements thanks to the proposed method's application to the medical dataset. Ada et al. [4] Naive Bayes, the Hidden Markov Model, and other segmentation techniques were discussed. Several segmentation algorithms used to find lung cancers are properly explained in terms of how and why they work.

Yu et al. [5] used the C4.5 classifier [6] and several feature selection procedures to categorise the type of lung cancer in the Weka environment [7]. Badjio et al. [8] used different feature selection techniques with K-nearest neighbour classifiers and it was implemented as IBK [9] in the Weka environment. Avci et al. [10] suggested a general discriminant analysis (GDA) and most tiny square support vector machine (LS-SVM) based classifier to handle this low sample, high-dimension classification challenge. For classifying lung cancer, Tan et al. [11] integrated the idea of the smallest message length with an indirect decision tree inference process.

With the emergence of deep learning, it has been discovered that autoencoders and other techniques can be used to determine the underlying structure of data. Syed et al. [12] provide a deep autoencoder classification process that first learns deep features before inputting these features into training an artificial neural network. According to experimental findings, when taught with identical training samples and all attributes, the deep learning classifier beats all other classifiers. Performance enhancement is also shown to be statistically significant.

1. **MODEL DEPLOYMENT ENVIRONMENT**
   1. **AZURE MACHINE LEARNING STUDIO**

The technology used for the comparative analysis is Azure ML Studio. Thanks to Azure Machine Learning, data scientists and developers can design, deploy, and manage high-quality models quickly and confidently. With industry-leading machine learning operations (MLOps), open-source interoperability, and integrated tools, it reduces time to value. This reputable platform was created for ethical AI machine-learning applications [13]. On Azure Studio, the models were trained for the same dataset using Two-Class Boosted Decision Tree, Two-Class Logistic Regression, and Two-Class Neural Network. The results were then compared.

1. **METHODOLOGY**
2. **Dataset Description:**

This study made use of a Lung Cancer dataset from the Kaggle Repository. This dataset has already been utilized in several lung cancer prediction and analysis algorithms. The dataset has 16 total number of attributes and 284 total number of instances. Table 1 describes the dataset used and Table 2 gives the information about the attributes of the table.

***Table 1:*** *Description of dataset*

Table

Description automatically generated

***Table 2:*** *Description of attributes*

A picture containing text, document, receipt

Description automatically generated

1. **CLASSIFIERS**

**1) Logistic Regression**

It is a supervised learning technique to calculate the likelihood of a binary event. A binary logistic regression has a dependent variable with two possible values: lose/draw, pass/fail, spam/not spam, true/false, and the like. Mathematical equation for logistic regression is,

|  |  |  |
| --- | --- | --- |
|  |  | **(1)** |

Text

Description automatically generated

An illustration of logistic regression is the use of computer learning to forecast whether a person will likely develop a cold. It is referred to as binary categorization since there are only two viable answers to this question: yes, they are infected, or no, they are not. Although logistic regression can occasionally be difficult, doing regression analysis is made simple by clever statistics tools. It explains the relationship between a single dependent variable and one or more nominal independent variables.

Several underlying assumptions guide logistic regression. Binary logistic regression initially requires a binary dependent variable, but ordinal logistic regression just requires an ordinal dependent variable. Second, logistic regression requires independent observations. The observations should not be based on matched or repeated measurements, to put it another way. Third, for logistic regression, there should be little to no multicollinearity among the independent variables. This implies that the independent variables shouldn't have a high degree of correlation.

**2) Decision Tree**

It is a non-parametric supervised learning method that may be applied to both classification and regression problems. A decision tree is a diagram that shows the possible results of a number of connected decisions. It enables a person or organisation to contrast alternative courses of action according to their costs, likelihoods, and returns. These can be used to start exploratory discussions or to create an algorithm that predicts the best choice analytically.

A root node, branches, internal nodes, and leaf nodes make up its hierarchical tree structure. There are three different sorts of nodes in a decision tree: choice, chance, and end nodes. The decision nodes, which are squares and signify a decision that needs to be made, are present. The chance node displays numerous outcomes that are unknown and is represented by circles. Triangles serve as the end node representations and denote an outcome. Schematic diagram of the decision tree is shown in *Fig 2.*

Diagram

Description automatically generated

*Fig 2: Decision Tree*

A boosted decision tree is an ensemble learning approach. This encompasses that the successor trees correct the errors of all the predecessor trees cumulatively. Predictions thus depend on the ensemble of trees [14].

**3) Neural Network**

The neural network classification approach is a supervised learning method. An input layer, one or more hidden layers, and an output layer are all components of neural networks. With own weight and threshold, each node is connected to the others.  If a node's output surpasses a certain threshold value, that node is triggered and begins sending data to the network's next layer. If the threshold is greater than the output of the node, no data is then forwarded to the next network layer.

A neural network is a collection of algorithms with the objective to discover the fundamental relationships among a collection of data by employing a similar procedure to that of human brain follows. *Fig 3* shows the architecture of the neural network.

Diagram

Description automatically generated

*Fig 3: Schematic diagram of Neural Network*

To improve the model's performance, Principal Component Analysis (PCA) is applied. Being a statistical method, it summarises the information in large data tables by employing a smaller set of "summary indices" that could be more easily shown and analysed. PCA is used for dimensionality reduction.

For early detection of lung cancer, machine learning algorithms are implemented in the experiment: Binary Logistic Regression, Binary Boosted Decision Tree, and Binary Neural Network. For better accuracy, feature selection is also performed prior to model training.

1. **FEATURE SELECTION**

Feature selection is a method for reducing the number of features in an image by deleting unimportant, redundant, or noisy features from the image. Better learning performance, i.e., higher accuracy, less expensive computation, and easier model interpretation, might result from feature selection. A variety of feature selection algorithms have recently been presented by researchers in the fields of computer vision, text mining, and other fields. They have demonstrated the effectiveness of their works through theory and experiment [15]. We have used a filter-based feature selection technique in our model, which is discussed below:

**Filter Based Feature Selection**

Feature based filter selection is the module in Azure Machine Learning Studio, which offers various correlation coefficients. In a correlation analysis, the correlation coefficient is the sensitive indicator that evaluates the degree of the linear relationship between two variables. In a correlation report, the coefficient is represented by r. Different coefficients of correlation are discussed below:

***i) Pearson’s Correlation Coefficient***

Pearson's correlation coefficient is a statistical test used to figure out the statistical association between two continuous variables. It is the best method for quantifying the relationship between variables of importance because it is based on the covariance method. It denotes the magnitude of the correlation, as well as the direction of the relationship [16]. In this work, the Pearson coefficient of correlation was used to train our model. Equation for Pearson’s Correlation Coefficient is,

|  |  |  |
| --- | --- | --- |
|  | Pearson Correlation Coefficient Formula | **(2)** |

Text

Description automatically generated

***ii) Chi Squared Correlation Coefficient***

By studying the relationship between the characteristics, the chi-square test aids in feature selection. It is sensitive to lower frequencies in table cells. In general, when the predicted value in a table cell is smaller than 5, chi-square can lead to incorrect results. Correlation tests can be used to choose features in machine learning. A chi-squared test can be used to determine whether the input variables are relevant to the output variable in classification issues where the output variable is categorical, and the input variables are similarly categorical. [17]

|  |  |  |
| --- | --- | --- |
|  |  | **(3)** |

Text

Description automatically generated

***iii) Kendall Correlation Coefficient***

When data is ranked by quantity, Kendall rank correlation is employed to investigate for similarities. Kendall's correlation coefficient uses pairs of data to determine the strength of link according to the pattern of concordance and discordance between the pairings [18]. Kendall's is frequently employed when data does not meet one of Pearson's correlation conditions. Kendall's method is non-parametric, which means it does not demand for the two variables to follow a bell curve. It is defined as,

|  |  |  |
| --- | --- | --- |
|  |  | **(4)** |

where,

= no. of concordant pairs

= no. of discordant pairs

***iv) Spearman Correlation Coefficient***

Spearman's Correlation shows the degree and direction of the monotonic relationship between your two variables [19]. It has the advantage of being easier to compute, although in a data science context, you're unlikely to be doing anything by hand, and both approaches are computationally light in comparison to many other jobs you'll be undertaking.

|  |  |  |
| --- | --- | --- |
|  |  | **(5)** |

Text

Description automatically generated

***v) Fisher Score Correlation Coefficient***

Fisher score is a supervised feature selection method based on filters and feature weights. Fisher score models have several benefits connected to the application of supervised learning for selecting features, such as simplified calculations, improved accuracy, and enhanced operability, which will minimise time-space complexity [20]. Equation of Fisher Z-score is evaluated as,

|  |  |  |
| --- | --- | --- |
|  |  | **(6)** |

where,

r = Pearson’s correlation coefficient

Zr = Fisher Z transformation

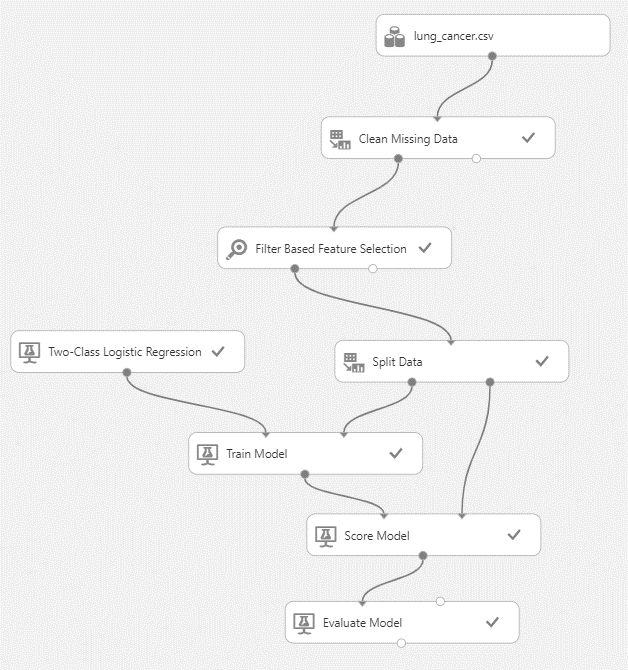
1. **Experiment Procedure**
   1. **Sub Experiment 1 – using Linear Regression**

After data pre-processing, for classification of lung cancer, two class logistic regression have been performed. During the feature selection we used Pearson correlation featuring scoring method with five number of desired features, refer to table 3 for the list of highly correlated features with the attribute lung cancer.

***Table 3****: List of highly correlated features*

|  |
| --- |
| HIGHLY CORRELATED FEATURES |
| Allergy |
| Alcohol Consumption |
| Swallowing Difficulty |
| Wheezing |
| Coughing |

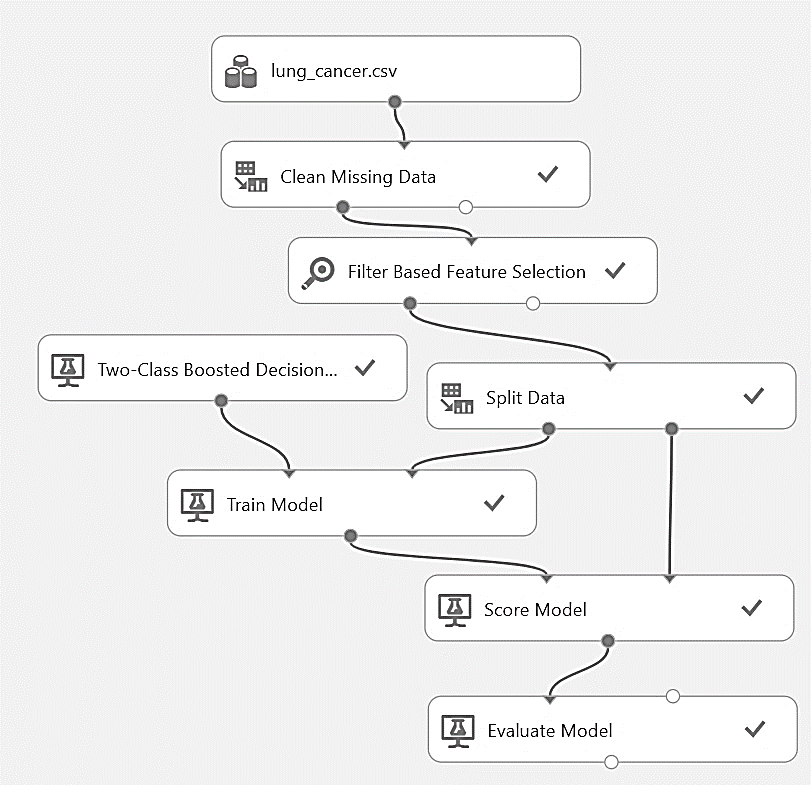
After feature selection, 70% of data is used for training of the model and 30% of data is taken for testing of the model. After training the model for logistic regression, score model module of the environment is used for scoring the trained model for classification or regression. *‘Evaluate model*’ module of the environment is then used for evaluating the trained model.



*Fig 4: Schematic view of Logistic Regression Model in Azure ML Studio*

* 1. **Sub Experiment 2 – using Boosted Decision Tree**

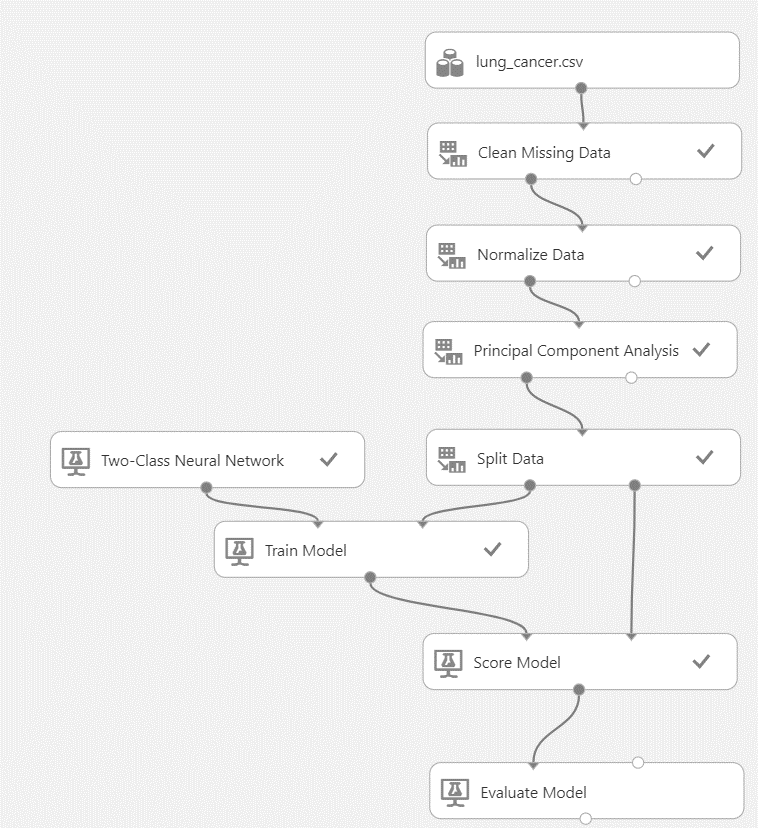
The final score of the model is recorded and results for the boosted decision tree algorithm are determined using a similar approach employing two class boosted decision tree modules.



*Fig 5: Schematic view of Boosted Decision Tree Model in Azure ML Studio*

* 1. **Sub Experiment 3 – using Neural Network**

For the training of model using two class neural network module, firstly we normalize the data using Z-Score transformation method. The Z-score is a statistical measurement that quantifies a value's relationship to the mean of a set of values. After this PCA is implemented with three as number of dimension reduction. After that, the score of the model is calculated and the model is evaluated.



*Fig 6: Schematic view of Neural Network Model in Azure ML Studio*

**Diagram

Description automatically generated**

*Fig.7 shows the flowchart that outlines the experiment explained with each Azure module representing a single step.*

1. **RESULTS AND DISCUSSION**

Performance measures are used to evaluate the performance for ML models. The most common and easiest way to describe the performance of a classification problem is to form a confusion matrix.

***Table 4:*** *Confusion Matrices of the 3 employed classifiers*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(a)** | **Actual** | | |  | | **(b)** | **Actual** | | |  | | **(c)** | **Actual** | | |
| **Predicted** |  | 1 | 0 |  | | **Predicted** |  | 1 | 0 |  | | **Predicted** |  | 1 | 0 |
| 1 | 84  **TP** | 9  **FP** |  | | 1 | 133  **TP** | 18  **FP** |  | | 1 | *133*  **TP** | 8  **FP** |
| 0 | 0  **FN** | 0  **TN** |  | | 0 | 3  **FN** | 0  **TN** |  | | 0 | 3  **FN** | 10  **TN** |
| (a) Confusion matrix for logistic regression model | | | | | (b) Confusion matrix for boosted decision tree model | | | | | | (c) Confusion matrix for neural network model | | | | |

***Graph 1*** *shows a visualization of confusion matrix in the form of column graph*

Table

Description automatically generated

A classification report also considers performance metrics like accuracy precision, recall, F1 score and AUC.

**Table 5:** *Evaluation of classifiers on different performance measures*

|  |  |  |  |
| --- | --- | --- | --- |
| Performance Measures | *Logistic Regression* | *Boosted Decision Tree* | *Neural Network* |
| Accuracy | *0.903* | *0.864* | *0.929* |
| Precision | *0.903* | *0.881* | *0.943* |
| Recall | *1.000* | *0.978* | *0.978* |
| F1 Score | *0.949* | *0.927* | *0.960* |
| AUC | *0.925* | *0.730* | *0.963* |
| Positive Label | *YES* | *YES* | *YES* |
| Negative Label | *NO* | *NO* | *NO* |

***Graph 2*** *shows a visualization of performance metrics in the form of bar graph*

*Chart

Description automatically generated*

From the *tables 4 & 5* and *graphs* *1 & 2*, it is evident that for this dataset, neural network is the optimal model having higher accuracy of 0.929, precision of 0.943 and F1 score of 0.978 among the classifiers taken for the comparison study. Azure ML Studio provides a scalable environment with lots of customization, and as such, this experiment could further expand its scope by adding more classifiers for comparison or training with different feature set derived from some other correlation metrics.

**References**

1. Y. Qiang, Y. Guo, X. Li, Q. Wang, H. Chen, and D. Cuic, “The diagnostic rules of peripheral lung cancer preliminary study based on data mining technique,” Journal of Nanjing medical university, vol. 21, no. 3, pp. 190–195, 2007.
2. KwetisheJoroDanjuma, “Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients” Department of Computer Science, ModibboAdama University of Technology, Yola, Adamawa State, Nigeria.
3. Zehra Karhan1, Taner Tunç2, “Lung Cancer Detection and Classification with Classification Algorithms” IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 22788727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77.
4. Ada, RajneetKaur, ” A Study of Detection of Lung Cancer Using Data Mining Classification Techniques ” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
5. L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in ICML, vol. 3, 2003, pp. 856–863.
6. J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The weka data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.
8. F. Badjio and F. Poulet, “Dimension reduction for visual data mining,” in international symposium on applied stochastic models and data analysis (ASMDA-2005), 2005.
9. F. Badjio and F. Poulet, “Dimension reduction for visual data mining,” in International symposium on applied stochastic models and data analysis (ASMDA-2005), 2005.
10. E. Avci, “A new expert system for diagnosis of lung cancer: Gdals svm,” Journal of medical systems, vol. 36, no. 3, pp. 2005–2009, 2012.
11. P. J. Tan and D. L. Dowe, “Mml inference of oblique decision trees,” in AI 2004: Advances in Artificial Intelligence. Springer, 2005, pp. 1082–1088.
12. S. M. Salaken, A. Khosravi, A. Khatami, S. Nahavandi and M. A. Hosen, "Lung cancer classification using deep learned features on low population dataset," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, 2017, pp. 1-5, doi: 10.1109/CCECE.2017.7946700.
13. Frogglew. (n.d.). What is azure machine learning? - azure machine learning. Azure Machine Learning | Microsoft Learn. Retrieved February 26, 2023, from https://learn.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning
14. Likebupt. (n.d.). Two-class boosted decision tree: Component reference - azure machine learning. Two-Class Boosted Decision Tree: Component Reference - Azure Machine Learning | Microsoft Learn. Retrieved February 26, 2023, from https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/two-class-boosted-decision-tree
15. Jianyu Miao, Lingfeng Niu, A Survey on Feature Selection, Procedia Computer Science, Volume 91, 2016, Pages 919-926, ISSN 1877-0509,
16. Haomiao Zhou, Zhihong Deng, Yuanqing Xia and Mengyin Fu, A new sampling method in particle filter based on Pearson correlation coefficient, Neurocomputing, http://dx.doi.org/10.1016/j.neucom.2016.07.036
17. Plackett, R. L. (1983). Karl Pearson and the Chi-Squared Test. International Statistical Review / Revue Internationale de Statistique, 51(1), 59–72. https://doi.org/10.2307/1402731
18. O’Gorman, T. W., & Woolson, R. F. (1995). Using Kendall’s τb Correlations to Improve Variable Selection Methods in Case-Control Studies. Biometrics, 51(4), 1451–1460. https://doi.org/10.2307/2533275
19. Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126(5):p 1763-1768, May 2018. | DOI: 10.1213/ANE.0000000000002864
20. Lin Sun, Tianxiang Wang, Weiping Ding, Jiucheng Xu, Yaojin Lin, Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification, Information Sciences, Volume 578, 2021, Pages 887-912, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2021.08.032.