COLLATION OF CANCER IN LUNGS UTILIZING MLOPS STRATEGIES

Hitesh Kumar[1], Nikita Narwat[2], Madhulika Bhatia[3]

[1,2,3]Amity University, Noida

ABSTRACT — After witnessing COVID-19, the gravity of the issue of early detection of a disease is apparent. Lung cancer is one of the major reasons for death worldwide among men and women, with an alarming rate of about five million fatal cases per year. In lung cancer, the cancerous cells in the lungs multiply uncontrollably. Three classification techniques, namely Logistic Regression, Decision trees, and neural networks, were used to analyze lung cancer prediction. The primary goal of this study is to investigate 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 detection. 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 is the leading cause of cancer-related death. Lung cancer can begin in the windpipe, major airway, or lungs. It is caused by unregulated proliferation and spread of cells from the lungs. People with lung illness, such as emphysema, and a history of chest difficulties are more likely to be diagnosed with lung cancer. Smoking is the most prevalent risk factor for lung cancer in Indian men; however, smoking is less common in Indian women, india cating that other factors contribute to lung cancer. Exposure to radon gas, air pollution, and toxins in the workplace are all risk factors. It often takes long time to develop, and most people are diagnosed with the disease 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. The

key to improving survival rates is early detection utilizing machine learning approaches, and if this can be used to make the diagnosis process more efficient and effective for radiologists, it will be a significant step towards the aim of enhanced early detection.

Primary lung cancer begins in the lung, whereas secondary lung cancer begins in the lung and spreads to other regions of the body. The stage of cancer is determined by the size of the tumor and how far it has spread. An early-stage cancer is a minor cancer that is detected in the lung, whereas an advanced cancer has progressed into surrounding tissue or another section of the body.

The Lung Cancer dataset for this study was retrieved from the Kaggle Repository. First, the data is preprocessed, and the dataset is separated into training and testing data. The model is then subjected to classification algorithms such as two-class logistic regression, two-class boosted decision trees, and two-class neural networks. To determine the accuracy of models, classification models are trained over the training data and associated models are tested over testing data. Finally, the conclusion was drawn after comparing the accuracy rates of each classification model. For overview of the experiment, refer to *fig 1.*

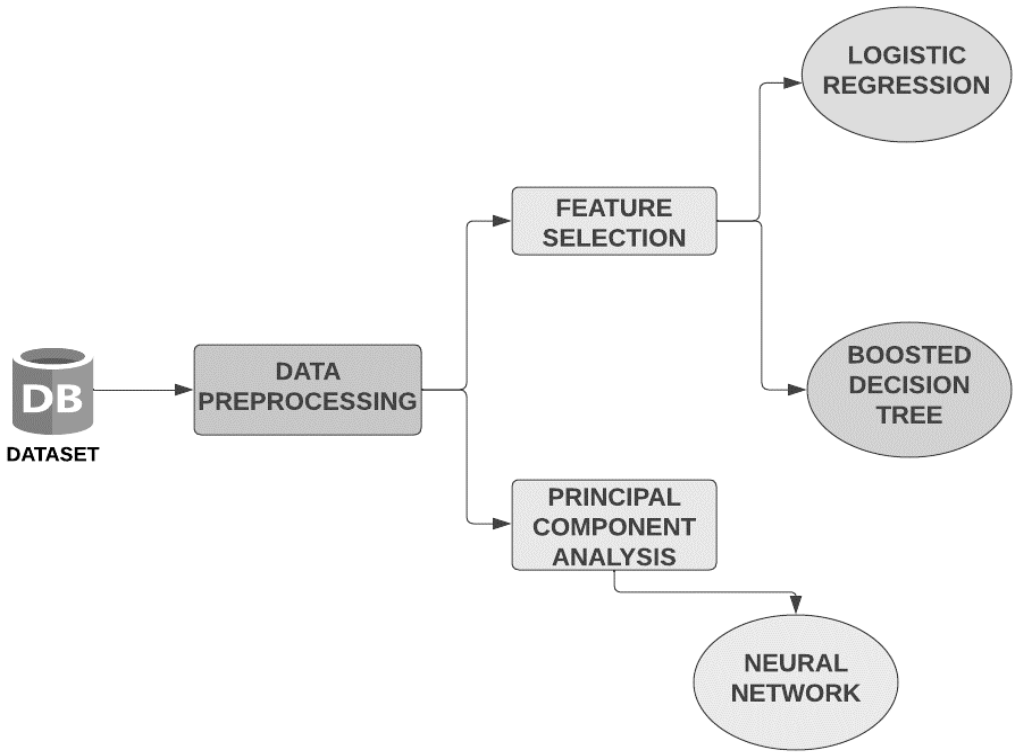


Fig 1: Overview of Experiment

2 | RELATED WORK

Many researchers have contributed to numerous lung cancer prediction and classification studies. Danjuma [2] predicts post-operative life expectancy in lung cancer patients using predictive data

mining algorithms to compare algorithms such as Decision Trees, Naive Bayes, and Artificial neural networks. A stratified 10-fold cross-validation comparative analysis was conducted on the above algorithms, and accuracy was calculated for each classifier.

Zehra et al. [3] produce different results for each classifier on the lung cancer dataset obtained. The classifiers such as KNN, SVM, NN, and Logistic Regression were implemented, and corresponding accuracy rates were obtained. Support Vector Machine has the highest accuracy, with 99.3%. The proposed method was applied to the medical dataset, which helped doctors to make more correct decisions. Ada et al. [4] Various segmentation algorithms were discussed, which include Naïve Bayes, Hidden Markov Model, etc. A proper explanation is given about how and why different segmentation algorithms are used in the detection of Lung tumors.

Yu et al. [5] classified the kind of lung cancer in the Weka environment [6] using a variety of feature selection strategies and the C4.5 classifier [7]. K-nearest neighbor classifiers were used with a variety of feature selection algorithms by Badjio et al. [8] (implemented as IBk [9] in the Weka environment). To address this low sample, high-dimension classification problem, Avci et al. [10] proposed a general discriminant analysis (GDA) and most miniature square support vector machine (LS-SVM) based classifier. Tan et al. [11] combined the shortest message length concept with an indirect decision tree inference mechanism for categorizing of lung cancer.

With the advent of deep learning, it has been found to identify the underlying structure of data through autoencoders and other techniques. Syed et al. [12] propose a deep autoencoder classification mechanism, which first learns deep features and then trains an artificial neural network with these learned features. Experimental results show that the deep learning classifier outperforms all other classifiers when trained with all attributes and the same training samples. It is also demonstrated that performance improvement is statistically significant.

3 | MODEL DEPLOYMENT ENVIRONMENT

5.1 | AZURE MACHINE LEARNING STUDIO

Azure ML Studio is the tool used for the comparative study. Azure Machine Learning empowers data scientists and developers to build, deploy, and manage high-quality models faster and with confidence. It accelerates time to value with industry-leading machine learning operations (MLOps), open-source interoperability, and integrated tools. This trusted platform is designed for responsible AI applications in machine learning. [13]. The models were trained for the same dataset using Two-Class Logistic Regression, Two-Class Boosted Decision Tree and Two-Class Neural Network on Azure Studio and then compared the results.

4 | METHODOLOGY

4.1 | 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*

Data Set Characteristics	*Multivariate*
Attribute Characteristics	*Integer, Categorical*
Associated Tasks	*Classification*
Number Of Instances	*284*
Number Of Attributes	*16*
Missing Values	*No*
Area	*Life*

Table 2: *Description of attributes*

ATTRIBUTE	***DESCRIPTION***
Gender	*M(male), F(female)*
Age	*Age of the patient*
Smoking	✓ , ✗
Yellow Fingers	✓ , ✗
Anxiety	✓ , ✗
Peer Pressure	✓ , ✗
Chronic Disease	✓ , ✗
Fatigue	✓ , ✗
Allergy	✓ , ✗
Wheezing	✓ , ✗
Alcohol	✓ , ✗
Coughing	✓ , ✗
Shortness Of Breath	✓ , ✗
Swallowing Difficulty	✓ , ✗

Chest Pain	✓ , ✗
Lung Cancer	✓ , ✗

4.2 | CLASSIFIERS

1) Logistic Regression

Logistic regression is an example of supervised learning. It is used to compute or forecast the likelihood of a binary (yes/no) event occurring. 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,

$$\boldsymbol{P = \frac{1}{1 + e^{-(a+bx)}}} \quad \textbf{(1)}$$

where,

P is the probability of a 1 (the proportion of 1s, the mean of Y),
e is the base of the natural logarithm (about 2.718)
a and b are the parameters of the model

A logistic regression example would be using machine learning to determine whether or not a person is likely to be infected with a cold. This is known as binary classification since there are only two potential answers to this question: yes, they are infected or no they are not infected. Although logistic regression might be challenging at times, intelligent statistics tools make it simple to undertake regression analysis. It explains the link between one or more nominal independent variables and a single dependent variable.

Some assumptions that logistic regression follows include First, binary logistic regression requires a binary dependent variable, whereas ordinal logistic regression requires an ordinal dependent variable. Second, logistic regression necessitates that the observations be independent of one another. To put it another way, the observations should not be based on repeated measurements or matched data. Third, logistic regression requires that the independent variables have little or no multicollinearity. This implies that the independent variables should not be overly correlated.

2) Decision Tree

A decision tree is a non-parametric supervised learning approach that can be used for classification as well as regression applications. A decision tree is a diagram that depicts the potential outcomes of a set of related choices. It enables an individual or organization to compare potential actions based on

their price, likelihood, and returns. They can be used to spark preliminary talks or to develop an algorithm that analytically predicts the optimal option.

It has a tree structure that is hierarchical and consists of a root node, branches, internal nodes and leaf nodes. A decision tree consists of three types of nodes, Decision nodes, Chance nodes and end nodes. The decision nodes are represented by squares and indicates a decision to be made. The chance node is represented by circles and shows multiple uncertain outcomes. The end node is represented by triangles and indicates an outcome. *Fig 2* shows the schematic diagram of the decision tree.

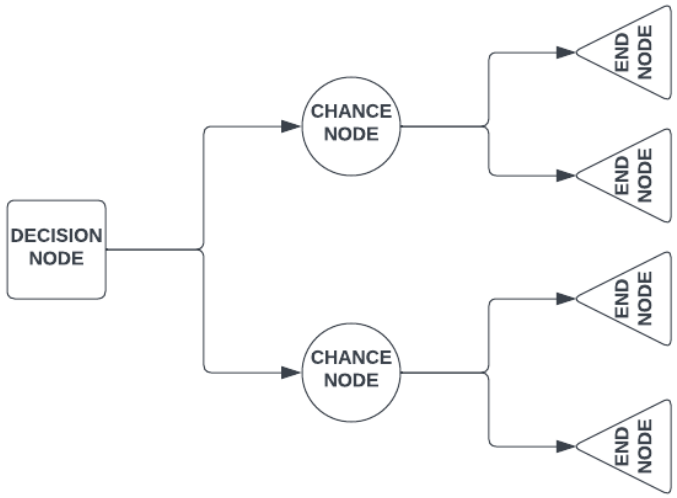


Fig 2: Schematic diagram of Decision Tree

A boosted decision tree [14] is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction.

3) Neural Network

The neural network classification approach is a supervised learning method. Neural networks include an input layer, one or more hidden layers, and an output layer. Each node is linked to another and has its own weight and threshold. If the output of any particular node exceeds the given threshold value, that node is activated and begins transferring data to the network's next tier. Otherwise, no data is sent to the next network layer.

A neural network is a set of algorithms that attempts to recognise underlying relationships in a batch of data using a technique similar to how the human brain works. Consider *Fig 3* for schematic working of the neural network.

The Principal Component Analysis (PCA) has been used to increase the model's efficiency. PCA is a statistical approach allowing you to summarise the information in huge data tables using a smaller collection of "summary indices" that may be more readily displayed and studied. PCA is used for dimensionality reduction.

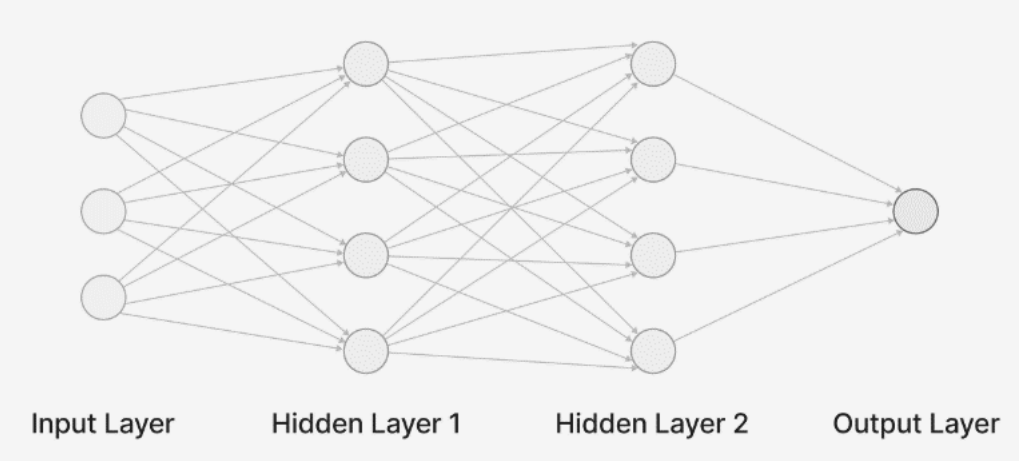


Fig 3: Schematic diagram of Neural Network

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.

4.3 | FEATURE SELECTION

Feature selection, as a dimensionality reduction technique, aims to choose a small subset of the relevant features from the original by removing irrelevant, redundant, or noisy features. Feature selection usually can lead to better learning performance, i.e., higher learning accuracy, lower computational cost, and better model interpretability. Recently, researchers from computer vision, text mining and so on have proposed a variety of feature selection algorithms and show the effectiveness of their works in terms of theory and experiment. [15] We have used 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 ML Studio, which provides different correlation coefficients. The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The coefficient is symbolized with the r in a correlation report. Different coefficients of correlation are discussed below:

i) Pearson's Correlation Coefficient

Pearson's correlation coefficient is a test statistic used to determine the statistical link, or association, between two continuous variables. Because it is based on the covariance method, it is the best approach to quantifying the relationship between variables of interest. It indicates the size of the association, or correlation, and 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,

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}} \quad \textbf{(2)}$$

where,

r = Pearson Coefficient
n = number of the pairs of the stock
$\sum xy$ = sum of products of the paired stocks
$\sum x$ = sum of the x scores
$\sum y$= sum of the y scores
$\sum x^2$ = sum of the squared x scores
$\sum y^2$ = sum of the squared y scores

ii) Chi Squared Correlation Coefficient

The chi-square test aids in feature selection by examining the relationship between the features. Chi-Square 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 or not 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]

$$\boldsymbol{\chi^2} = \sum \frac{(\boldsymbol{O_i} - \boldsymbol{E_i})^2}{\boldsymbol{E_i}} \quad \textbf{(3)}$$

where,

O_i = observed value (actual value)
E_i = expected value

iii) Kendall Correlation Coefficient

Kendall rank correlation is used to look for similarities in data ordering when it is ranked by quantity. Kendall's correlation coefficient employs pairs of data to establish the strength of association based on the pattern of concordance and discordance between the pairings, whereas other types of correlation

coefficients use observations as the basis of the correlation. Kendall's is frequently employed when data does not meet one of Pearson's correlation conditions [18]. Kendall's method is non-parametric, which means it does not demand for the two variables to follow a bell curve. Mathematically it is defined as,

$$\boldsymbol{\tau = \frac{n_c - n_d}{n(n-1)\,/\,2}} \quad \textbf{(4)}$$

where,

n_c = number of concordant pairs
n_d = number of discordant pairs

iv) Spearman Correlation Coefficient

Spearman's Correlation determines the strength and direction of your two variables' monotonic relationship [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.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad \textbf{(5)}$$

where,

ρ = Spearman's rank correlation coefficient
n = number of observations
d_i = difference between the two ranks of each observation

v) Fisher Score Correlation Coefficient

Fisher score is a filter-based supervised feature selection method with feature weights. As a feature relevance criterion, Fisher score models have many advantages associated with the use of supervised learning for feature selection, such reduced calculations, higher accuracy, and stronger operability, which can efficiently reduce time–space complexity. In recent years, Fisher score technologies have progressed in terms of feature selection. [20] Equation of Fisher Z-score is evaluated as,

$$\boldsymbol{Z_r = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2}} \quad \textbf{(6)}$$

where,

r = Pearson’s correlation coefficient

Z_r = Fisher Z transformation

5 | Experiment Procedure

5.1 | 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 the data is spilt into 70:30 training and testing data. Then the model is trained for logistic regression and then the model is scored using the score model module which is used to score a trained classification or regression model. Finally, the model is evaluated using the evaluate model module.

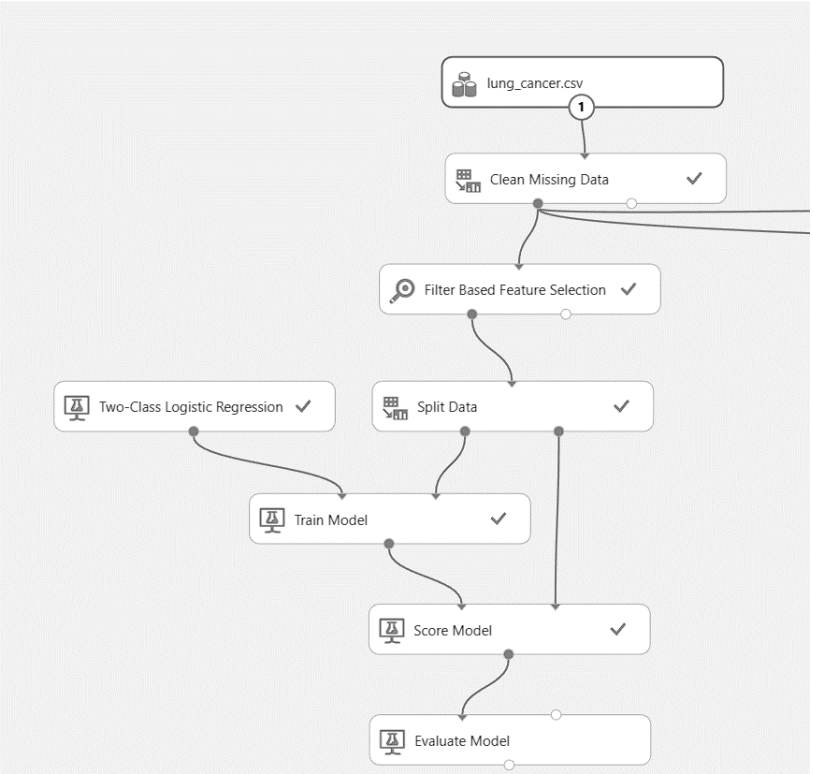


Fig 4: Schematic view of Logistic Regression Model in Azure ML Studio

5.2 | Sub Experiment 1 – using Boosted Decision Tree

A similar approach is used for determining the results for boosted decision tree algorithm by using two class boosted decision tree module and the final score of the model is recorded.

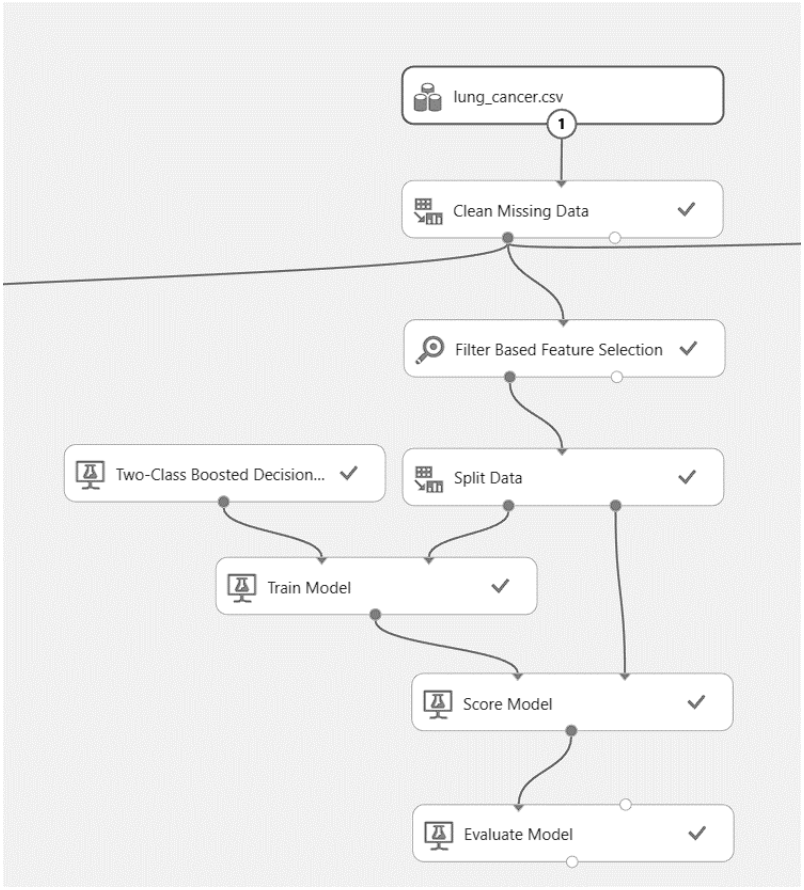


Fig 5: Schematic view of Boosted Decision Tree Model in Azure ML Studio

5.3 | Sub Experiment 1 – 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 measure that describes the relationship of a value to the mean of a group 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.

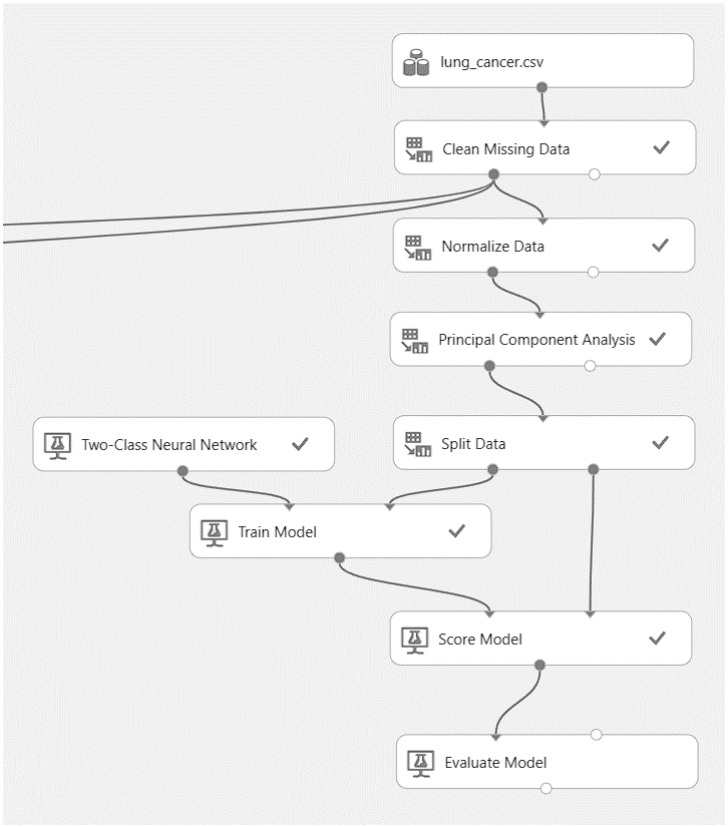


Fig 6: Schematic view of Neural Network Model in Azure ML Studio

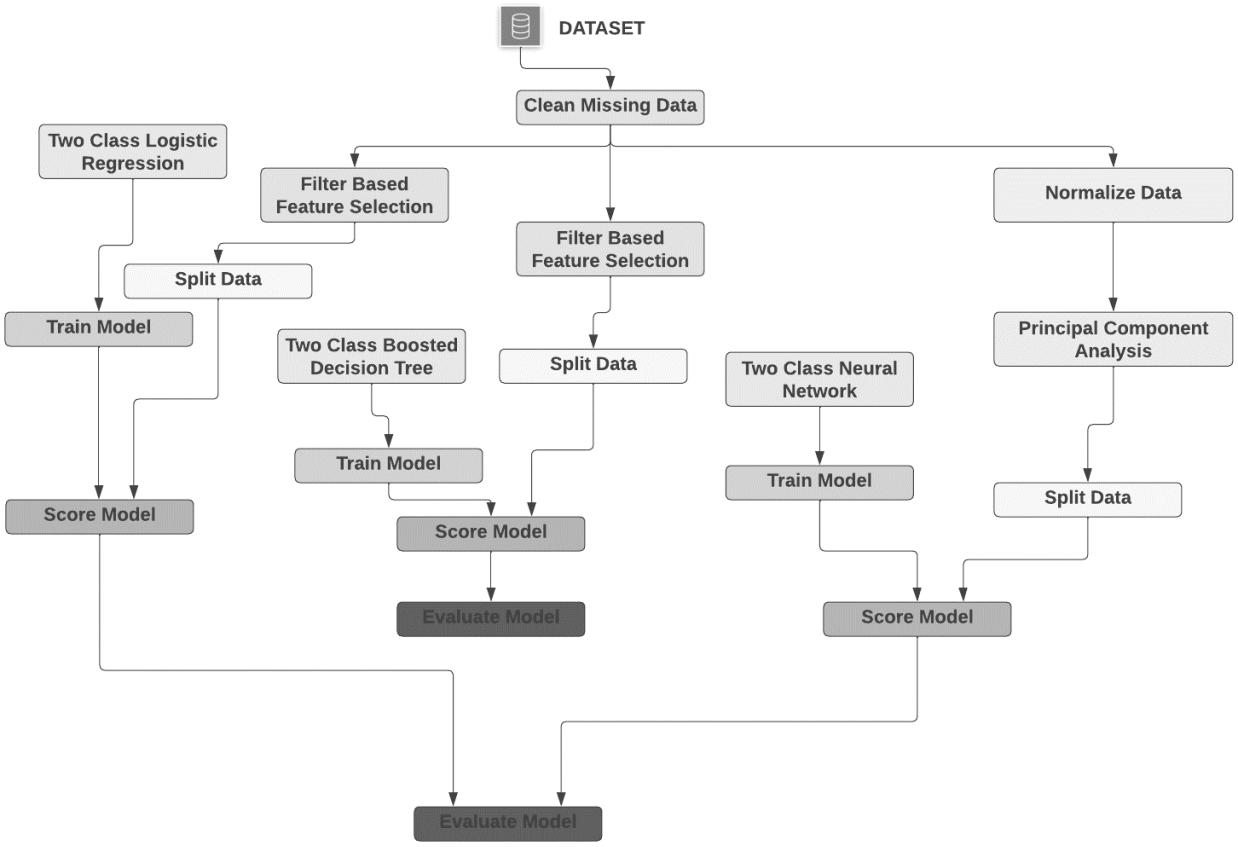


Fig.7 shows the flowchart that outlines the experiment explained with each Azure module representing a single step.

6 | RESULTS AND DISCUSSION

Performance measures are used to evaluate the performance for machine learning 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)

Predicted \ Actual	1	0
1	84 **TP**	9 **FP**
0	0 **FN**	0 **TN**

(a) Confusion matrix for logistic regression model

(b)

Predicted \ Actual	1	0
1	133 **TP**	18 **FP**
0	3 **FN**	0 **TN**

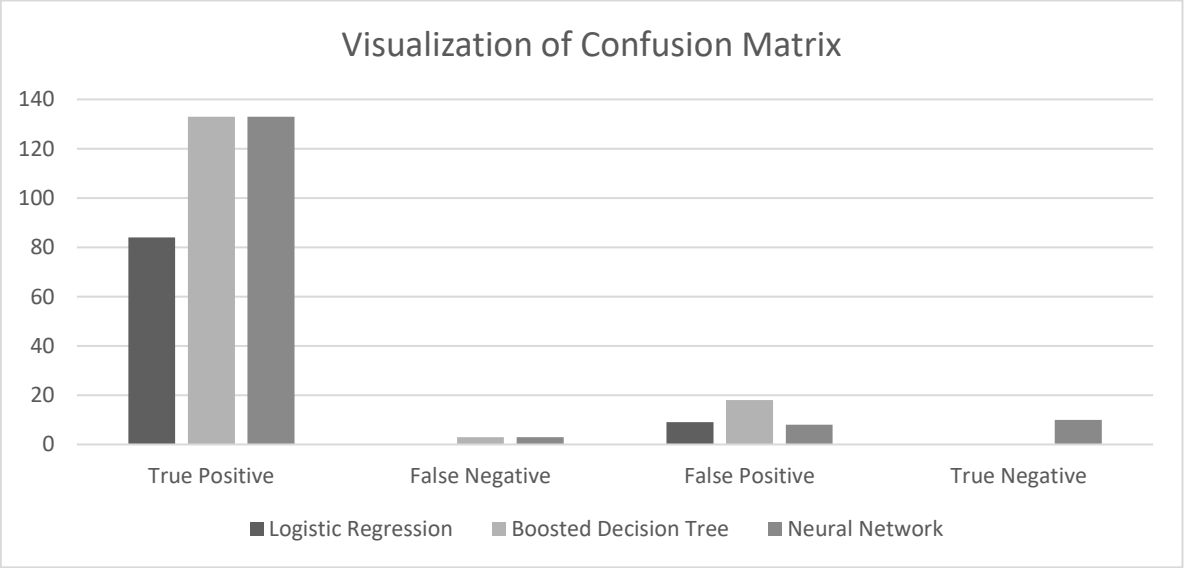
(b) Confusion matrix for boosted decision tree model

(c)

Predicted \ Actual	1	0
1	*133* **TP**	8 **FP**
0	3 **FN**	10 **TN**

(c) Confusion matrix for neural network model

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



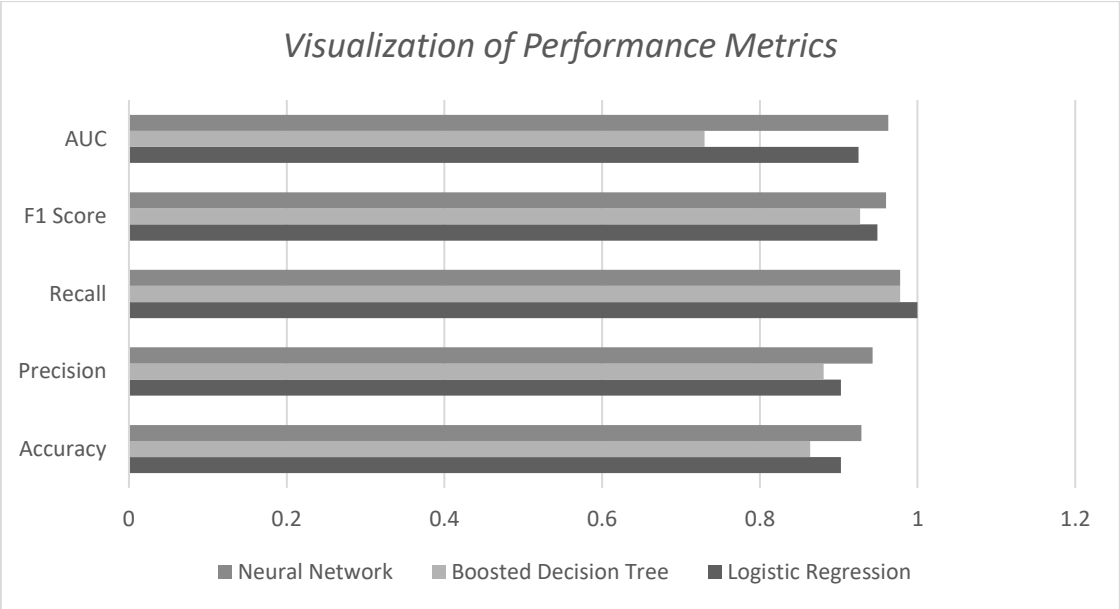
A classification report also considers performance metrics like accuracy precision, recall, F1 score and AUC.

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*



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, precision and F1 score 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] 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.

[7] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.

[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.