iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44537

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Survival Rate PredictionPost Thoracic Surgery

Tentu Meghana[1], Vasireddy Shravya[2], Velmula Amulya[3], Dr. A. Venkata Ramana[4]
[1, 2, 3]*Graduate Scholar,* [4]*Associate Professor, Department of Electronic and Computer Engineering, Sreenidhi Institute of Science and Technology*

Abstract: ***In order to improve quality initiatives, healthcare administration, and consumer education, it is critical to track health outcomes. The data obtained from patients who had large lung resections for primary lung cancer is referred to as thoracic surgery. Attribute ranking and selection are critical components of successful health outcome prediction when using machine learning algorithms. Researchers used several procedures, such as early-stage examinations, to determine the type of cancer before symptoms appeared. The most relevant attributes are identified using attribute ranking and selection, and the duplicated and unnecessary attributes are removed from the dataset. The goal of our study is to look at patient mortality over the course of a year after surgery. More precisely, we're looking into the patients' underlying health issues, which could be apowerful predictor of surgical-related mortality.***

Keywords: Thoracic Surgery, Attribute Selection, Machine Learning.

I. INTRODUCTION

The introduction of computer applications into the medical enterprise has had an instantaneous effect on doctors' productiveness and accuracy in current years. One of those programs is the examiner of fitness consequences. Health consequences are absolutely turning into a lot extra critical within side the shopping and control of healthcare. In maximum nations, most cancers is now one of the main reasons of mortality. Lung most cancers is presently the maximum common indication for thoracic surgery.

Massive datasets of most cancers had been gathered and made to be had to clinical experts because of the development of latest equipment within side the subject of medicine. The maximum tough challenge, however, is exactly predicting a sickness outcome. As a result, modern-day studies makes a specialty of using system gaining knowledge of strategies to find out and outline fashions, in addition to relationships among them, from huge quantities of facts. The fact is analyzed to extract beneficial data that helps sickness prediction, in addition to enhance fashions that expect healthcare consequences greater accurately.

Thoracic surgery is the most common surgery performed on patients suffering from lung cancer. The survival rate is a highly important criterion for sawbones when deciding which patients to operate on. One of the most common medical choice problems in thoracic surgical procedure is deciding on the proper affected person for surgical procedure, preserving in thoughts the dangers and benefits for the affected person within side the immediate (for example, post-operative problems, including the fatality price within side the first month) and long-time period outlook (e.g., survival for 1-5 years). In current decades, a number of machine learning techniques, as well as attribute ranking and selection methods, were carried out to disease diagnosis and prediction.

Machine learning systems' performance and accuracy are frequently harmed by large datasets. Datasets with a high number of dimensional attributes/features have a higher processing complexity and take longer to predict. A solution to complex datasets is attribute ranking and selection. The major aim of these methods is to eliminate attribute/featuresthat are unnecessary, misleading, or redundant, as these attributes/features increase the size of the search area, making it more difficult to analyse data and thus not contributing to the learning process. The process of selecting the best attributes/features from among all the attributes/features that can be used to distinguish classes is known as attribute and ranking selection.

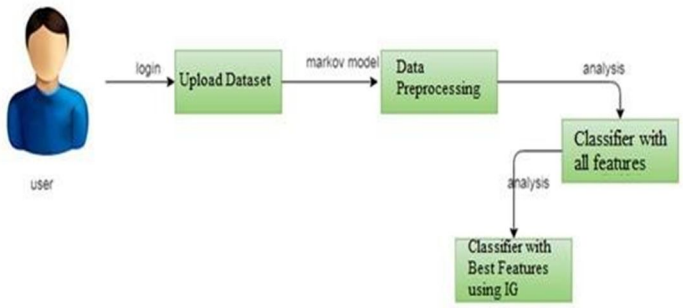


Figure 1: Block diagram

II. EXISTING SYSTEM

It has always been a difficult task to accurately predict the life expectancy post an operation. The prediction relies upon on numerous fitness elements of which a few have a far important function in comparison to the opposite elements. A famous approach used within side the beyond became to investigate the CT test photos of the lungs and expect primarily based totally at the everyday check-ups. The thirty-day mortality price is one statistic that has been used to estimate mortality charges within side the beyond. This statistic, however, won't be absolutely correct due to the fact many sufferers die or come to be very frail right now after this time period, requiring them to be transferred to any other organization earlier than passing death. As a result, a large number of these deaths gounreported.

III. PROPOSED SYSTEM

In order to improve quality initiatives, healthcareadministration, and consumer education, it is critical to trackhealth outcomes. Data from patients who have undergone extensive lung resection for primary lung cancer is called thoracic surgery. Attribute ranking and selection are important components for correctly predicting health outcomes when using machine learning algorithms. Researchers have used several techniques, such as early screening, to determine the type of cancer before symptoms appear. The most relevant attributes / characteristics are identified using attribute ranking and selection, and duplicate unwanted attributes / characteristics are removed from the dataset.

The goal of our study is to look at patient mortality over the course of a year after surgery. More precisely, we're lookinginto the patients' underlying health issues, which could be a powerful predictor of surgical-related mortality.

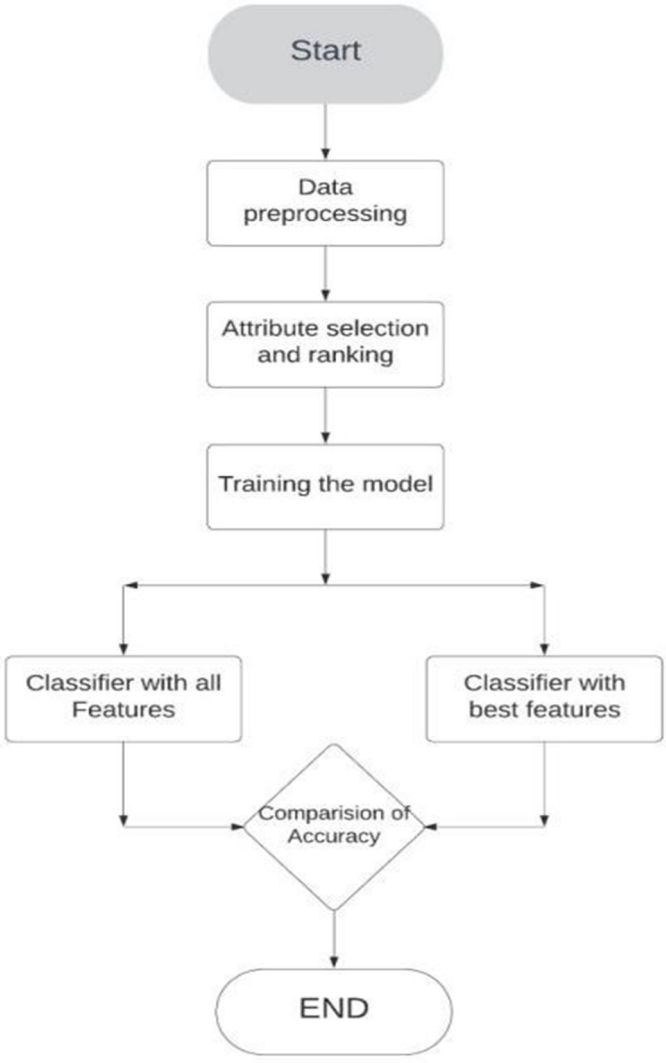


Figure 2: Flowchart

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

IV. DATA PREPARATION

Thoracic Surgery Data is provided primarily to revitalize therisk of lung surgery in real clinics for cancer patients. Data collected randomly by Marek Lubicz et al .The Wroclaw Thoracic Surgery Center for consecutive patients - elderly 21 to 87 who underwent major lung surgery for major lung cancer in 2007-2011. The facility is affiliated with the Department of Thoracic Medical Surgery University of Wroclaw and Lower-Silesian Center for Pulmonary Diseases, Poland, at some point of the studies database forms a part of the National Registry Cancer Registry, that's regulated with the aid of using the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. Database inclusive of 470 cases (70 real and four hundred false) and sixteen non-lacking symbols and binary cost stage (demise inside 365 days after surgical treatment - survival). Generally, splitting of the information is finished into parts

After performing correlation on the attributes of the dataset the following heat has been produced

Figure 3: Heat Map

V. DATA PREPROCESSING

The most important step before fitting an algorithm is to get the attribute/features in to a comprehensive format, for example changing nominal data into numeric data using dummy variables. This step usually comes after importing the dataset.

The steps in data pre-processing are as follows:

A. Dealing with Missing Data or Outliers

An easy way to handle missing or outliers is to first find out where the missing or outliers are and then replace them with the mean or median of that particular attribute.

Syntax: from sklearn.impute import SimpleImputer

Using this SimpleImputer it can replace the missing values with mean or median. It has parameters such asmissing values (np.nan), strategy (mean/median).

VI. DEALING WITH NOMINAL DATA

Models cannot directly use the nominal data to fit intoany model so it is necessary that it is to be converted into a proper numeric form. So, it can directly use pd.get_dummies() function to convert the nominal into dummy variables.

Where, pd = pandas' library

Another way to create dummy variables is to use one-hot encoder, that can be imported from sklearn.preprocessing.

```
#dummy variables for input variable
dummy = pd.get_dummies(df['DGN'])
df = pd.concat([dummy,df],axis = 1)
df = df.drop(['DGN'], axis= 1)

#dummy variables for input variable
dummy = pd.get_dummies(df['PRE6'])
df = pd.concat([dummy,df],axis = 1)
df = df.drop(['PRE6'], axis= 1)

#dummy variables for input variable
dummy = pd.get_dummies(df['PRE14'])
df = pd.concat([dummy,df],axis = 1)
df = df.drop(['PRE14'], axis= 1)
```

Figure 4: Basic syntax for dummy variables.

VII. IMPLEMENTATION

The essence of machine learning is model fitting. Theoutcomes produced by the model will not be accurate enough to be used for actual decision-making if it doesn't fit the data appropriately. Hyper parameters ina correctly fitted model capture the complicated interactions between known factors and the target variable, allowing the model to identify useful insights and generate accurate predictions. The models used inthis paper include Decision trees, KNN classifier, Logistic Regression. These models are used both for the classification with all features and classification with the best features.

Based on the heatmap the correlation between the attributes can be observed which help in selecting the attributes having a strong relationship to the output. The below are the list of the few best attribute/featureswhich increase the performance:

- PRE5 (FEV1) - Volume of the exhaled part at the end of the first-second of the forced expiration.
- PRE6 – Zubrod scale (Performance status).
- PRE9 – Dyspnoea before the surgery was done.
- PRE11 – Weakness before the surgery was done.
- PRE17 – Diabetes mellitus (Type 2 DM).
- PRE14– Size of the tumor (fromOC11(smallest) to OC14 (Largest)).

A. Algorithms

Algorithm fitting refers to how well a machine learning model is generalized to data comparable to trained data. A well-fitted model will produce more accurate output / results. The model can be overfitted or overfitted. The overfitting model fits the data too well, and the overfitting model does not fit the data properly.

Every machine learning algorithm/method has a set of basic parameters which can be tweaked to get an improved performance. During this fitting phase, a machine learning model is created by running the algorithm on the data for which one know's the target value, also called as "labeled date". The correctness of the end result is then determined through evaluating them to actual, observed values of the target variable.

Then this data is used and tweak the algorithm's normal settings to minimize the error and improve its accuracy in detecting anomalies and relation b/w the target and the rest of its features. The procedure is done until the algorithm discovers best settings for producing valid, practical, and usable insights for yourreal-world business challenge.

The following are the algorithms used in this paper:

B. Decision Trees

A decision tree is a type of tree that is a method of supervised learning (that is, an input and the output corresponding to that particular input) in which the data is iteratively classified using specific parameters. It can be defined by two things: the decision area and the leaf. The final decision or termination of the tree is called a leaf. And the node is the decision to separate the functions.

A decision tree is a graphical representation of all solutions that are possible to a decision based on certain conditions. Tree models where the target variable can take a some set of values are known as classification trees and target variable can take continuous values are called regression trees. Decision tree is used as one of the algorithm in this project.

The example of decision tree could be easily explained by the use of the above binary tree. Supposeone wants to know if a person wants to be given a diet, exercise, etc. Decisions here are questions such as 'How old are you?', 'Do you exercise?', 'Have you eaten more no of pizzas'? etc. The end-nodes, which are effects such as 'suitable', or 'unsuitable'. In this case it is a binary split i.e, a Yes/no type problem. There are mainly two different types of decision trees Classification trees and Regression trees.

C. Logistic Regression

It is a statistical primarily based totally version that during its easy shape makes use of a logistic feature to create a based variable that is binary, aleven though many tough extensions exists. In regression analysis, log-it regression (or Logistic regression) is calculating the parameters of a shape of binary regression. It is a prediction set of rules the usage of unbiased variables to get based variables, including Linear Reversible, however the distinction is that the based variables must be labeled variable.

Logistic regression is a statistical based model that uses conditional probability. It can calculate the conditional probability of the dependent variable and independent variable using the formula below for binary regression.

P(Y=0/X) or P(Y=1/X)

This can be read as the conditional probability of Y equals to 1, given X / conditional probability of Y equals to 0, given X.
P(Y/X) is approximated as sigmoid function which is applied to linear combination of different features.
Logistic Regression also can be used for multi class classification or for binary classification. In multi- class classification, It has more than one outcomes like some may have the everyday fever or flu, or normal cold or corona virus. Binary classification is when there are possible results like someone is infected with corona virus or isn't infected with corona virus.

D. K-Nearest Neighbour Classifier

K-Nearest Neighbors Classifier (KNN) is the simplest algorithm used in machine learning for both classification and regression problems. The ANN algorithm takes data and classifies new points based on their similarity to the point (such as Euclidean distance). To do this, the distance is taken and summed with the nearest neighbor. The image below shows the classification of different classes based on distance..

In KNN classifier, 'K' is the no of nearest/closestneighbours. The no of neighbours is the mostimportant aspect of the classifier. 'K' is usually is not taken as an even number when the number of classes are two. When the value of K is one, then it is called as the nearest neighbour algorithm. It is the simplest case among all others.

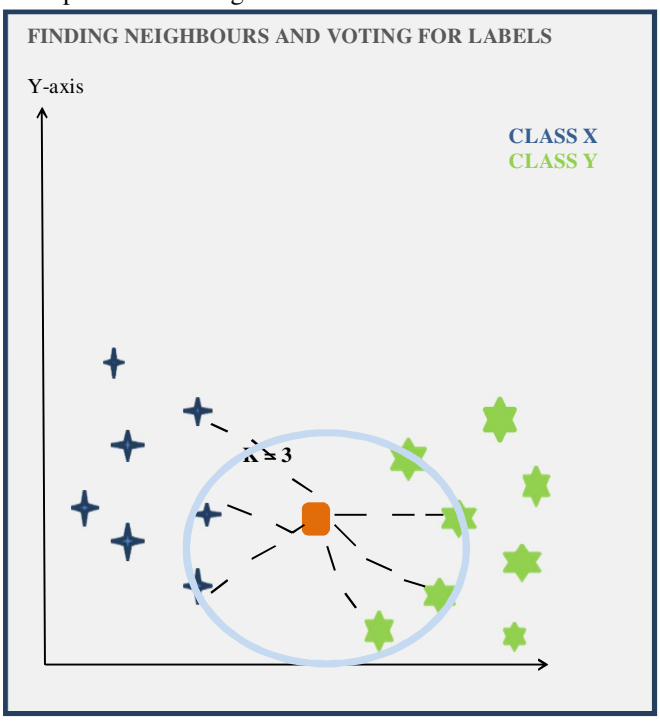


Figure 6: KNN Classifier Example

VIII. RESULT ANALYSIS

In order to calculate the accuracy of various techniques it can calculate the true positives, true negatives, false positives, and false negatives produced by the algorithm or techniques.

- True Positive (TP) consists of the numberof patients who survived and were correctly classified.
- True Negative (TN) consists of the number of patients who did not survive and were correctly classified.
- False Positive (FP) consists of the number of patients who survived but were not correctly classified.
- False Negative (FN) consists of the number of patients who did not survive and were not correctly classified.

A. Classification Accuracy

The ratio of (sum of true positive and true negative) to the (sum of all four values of confusion matrix) which is multiplied by 100 to get the percentage.

Classification accuracy = (TP+TN) *100/ (TP+TN+FP+FN)

Accuracies of each model with all features and withbest features is shown as a table below,

Algorithms/Model	Classifier with all features	Classifier with best features
Decisiontress	76.2	80.5
Logistic Regression	81.3	82.2
KNN Classifier	81.3	83.05

Table 1: Comparison between the accuracies of each model

IX. CONLUSION

In this study, the quality of the three classification methods/ algorithms have been tested to improve the prognosis for the life of patients with thoracic thoracic cancer after thoracic cancer surgery. Three ways to train the machine before and after using the quality level options are compared with their improved versions. The results show that boosting is not always the best solution, where the level of responsibility and choice can make better at improving predictive accuracy. Other qualifications and machine learning strategies can be introduced in the future work to find the best performance of the data forecast model. The results indicate that decision tree, logistic regression and KNN give us better performance than other data mining algorithms, we also have monitored and mentioned the performance of each algorithm. From this study we deducted that *FEV1, Dyspnoea before surgery, Size of the tumour, Weakness before surgery and performance status* are the attributes/features that are mainly responsible for the survival of patients after the thoracic surgery.

This study asserts the use of data mining algorithms in medical field as the results are better than the existing system which are confirmed by statistical analysis.

X. FUTURE SCOPE

This study does have few limitations. The results/outputs which were obtained are particular to a nation or an organization which collected the data set. Results obtained may be time-limited (2007–11). The dataset used as part of this project has very less records and may impede the accuracy of few algorithms that are used. In any case, this dataset can serve as a starting point to raise a better understanding of thoracic surgery patients. These tests can be further extended.

This analysis only makes and uses three data mining methods. Therefore, some more machine learning methods could be used to get more knowledge about the data-set as a future work.

REFERENCES

[1] V. Sindhu, S. A. S. Prabha, S. Veni and M. Hemalatha. (2014), "Thoracic surgery analysis using data mining techniques", International Journal of Computer Technology & Applications, Vol. 5 pp.578-586.

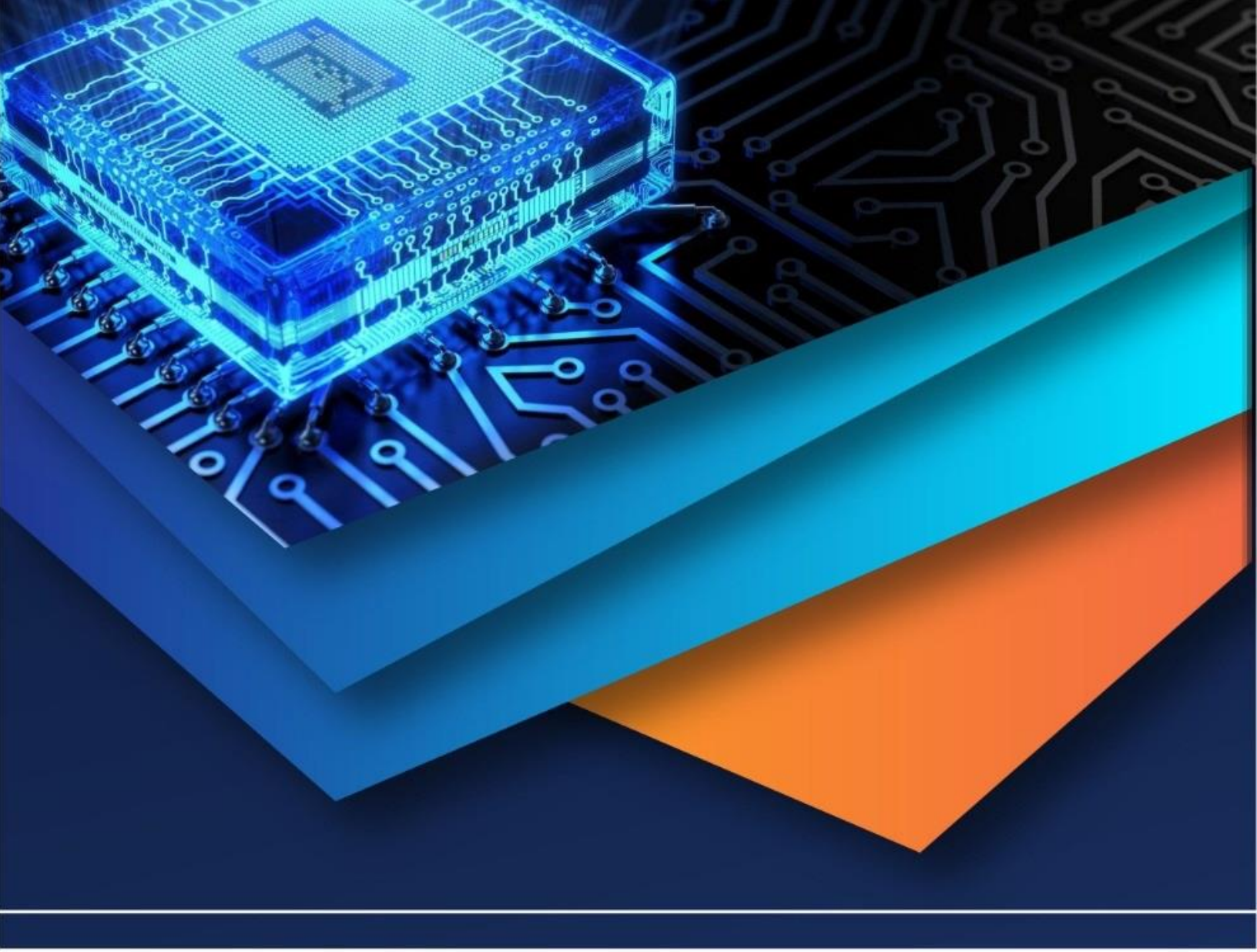
[2] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadisa. (2015), "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, Vol. 13, pp.8-17.

[3] Kwetishe Joro Danjuma. (2015), "Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients", IJCSI International Journal of Computer Science Issues, Vol. 12, No. 2, pp.189-199.

[4] Joseph A. Cruz, David S. Wishart. (2006), "Applications of machine learning in cancer prediction and prognosis", Cancer Informatics, Vol. 2, pp.59-77 2006.

[5] Mehdi Naseriparsa, Amir-Masoud Bidgoli and Touraj Varaee. (2013), "A hybrid attribute/feature selection method to improve performance of a group of classification algorithms", International Journal of Computer Applications, Vol. 69, No. 17, pp.28-35.

[6] Desuky, A.S. and El Bakrawy, L.M., 2016. Improved prediction of post- operative life expectancy after Thoracic Surgery. Advances in Systems Science and Applications, 16(2), pp.70-80.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY