Performance analysis of different machine learning algorithm for Detection Of lung cancer

Kabita Sahoo [1] Dr. Abhaya Kumar Samal[2

Research Scholar (Enrollment No. DSR01011), Biju Patnaik University of Technology Rourkela, Odisha, India. kabitasahoo789@gmail.com

[2] Professor, Dept. CSE, Trident Academy of Technology, BPUT, Odisha, India,

kabhaya1@gmail.com

[3]Sekharesh barik

Asst prof ,Ajay binod institute of technology

]

Abstract

Lung cancer is a deadly disease. It is one of the leading cause of death all over the world. Early detection may increase survival rate of the patient. Traditional Method of diagnosis results in last stage detection, necessitating the development of more advanced and accurate predictive models. In this paper we have used various machines learning algorithm for the early detection of lung cancer .Early detection is very much needed for the survival of the patient. Diagnostic for lung cancer include physical examination, imaging ,CT scan and MRI .For cost effective detection we have used ML algorithm and achieved more than 95 percent accuracy.

Introduction

Lung cancer is difficult to detect than other disease. Early detection and treatment is needed .treatment of lung cancer are based on type of cancer and the person's medical history. The primary cause of detection failure is the size of lesion. Which is called as the nodule. Cancer cell size is small in the beginning but it become malignant after a certain period .Early detection is very much needed.

Machine learning is a branch of Artificial intelligence which focuses on creating algorithm and model which have the ability to learn and adopt from extensive data sets. Machine learning model can learn can make prediction and make decision based on the past experience and the pattern. ML model can be designed to analyze meaningful information from the large data set. Model will learn from the historical data it can make prediction from the unseen data.

They can identify the pattern through the process of training. The applications of ML is vast and diverse which include natural language processing ,recommendation system ,fraud detection ,disease prediction and many more .

Deep learning is an advance form of ML which can be used to in object detection ,voice recognition and other complex data processing. It utilizes deep neural network with multiple features to extract the pattern from the data .

Lung cancer is one of the cancer where most of the people die.If predicted yearly 15 percent people with receiving therapy will survive for more than 5 years after their diagnosis. A computer can help to diagnose lung cancer. It is difficult to recognize which is benign and malignant. malignant tumor marked by the development of cellular tissue that is out of sequence. When cancer cells invade new tissues, the process is known as metastasis. Cancer tends to spread and is incurable if it goes too far; thus, it should be found as early as possible.The main problem is lung cancer shows the symptoms only at its advance stage. It is very challenging and practically impossible to treat at late stage. Image of lung cancer is captured using image captured technique like CT scan, positron emission tomography(PET) ,MRI and X ray. CT scan is the most widely used imaging method .We can use machine learning for complex data categorization and decision making. Machine learning can be used for prediction of lung cancer based upon the range of variable.

In this research paper we have used ten different machine learning classification algorithm including logistic regression, decision tree, K nearest neighbor, Gaussian naïve Bayes, multinomial naive Bayes, Support vector classifier, random forest, XG Boost,multi layer perceptron and gradiant boosting classifier to predict the lung cancer based on different variable.The data sets collected from kaggle .The different parameters are gender, age, smoking, yellow finger,anxiety,peer pressure,chronic disease,fatigue,allergy,wheezing,alcohol consuming,coughing,shortness of breath,swallowing difficulty,cheast pain and lung cancer.We have analysed the variable and used different machine learning algorithm to identify pattern and compared the accuracy .

Literature review

Roy et al used a combination of image processing and biomedical technique for the early detection of lung cancer. They have used lung representation from CT images .the scan images are pre processed and ROI is performed. They have used the random forest classifier and SVM classifier .They got the accuracy of 94.5%.

Faisal et al used machine learning classifier and multilayer perceptron ,Naïve bayes classifier , Gradiant Boosted tree and SVM.They have shown that Gradianr boosted tree out performed.

Banerjee et al have done tumor classification .They have done the mat lab simulation .The accuracy was calculated 79% ,SVM 86%,and ANN 92%.They have used Jupiter note book for machine learning classification.

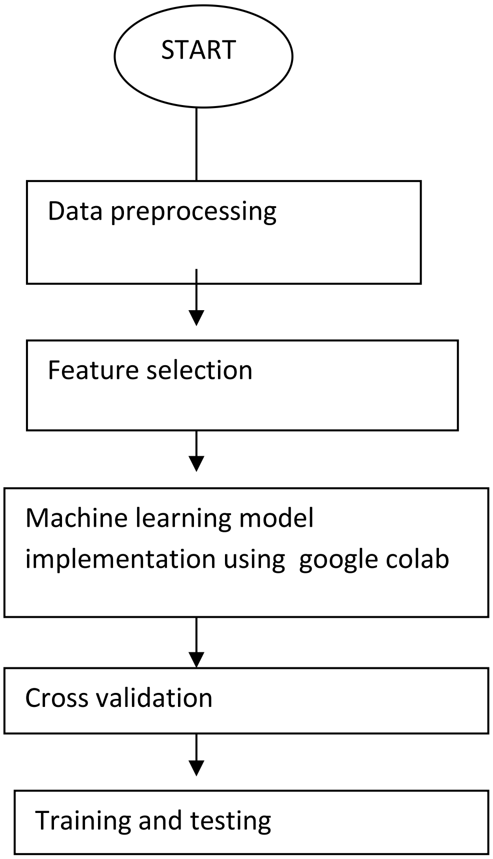
Reddy et al proposed a model using machine learning algorithm .The model combines the algorithm like K-NN ,decision tree and neural network using bagging ensemble approach to improve the overall accuracy. The accuracy score was 97% for decision tree ,94 % for KNN and 96% percent .they have used an integrated model where the accuracy is 98%.

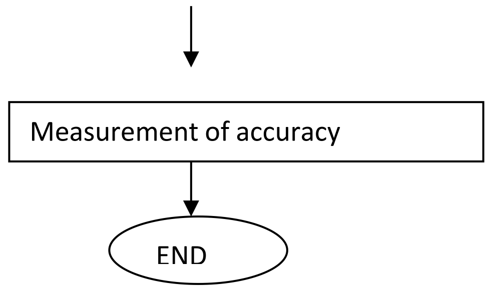
Boban et al used different ML algorithm including multi layer perceptron ,KNN and SVM classifier.They have used grey level coccurance matrix to pick the most important feature. They got the classification accuracy 98%.for SVM 70.45%.They got accuracy 99.2% for KNN.

Elnakib et al done the early lung node identification using low dose computed tomography. They have implemented different deep learning architecture including Alex ,VGG16, and VGG19. They have adapted GA for feature selection and optimization.With VGG19 they got the classification accuracy 97.5%.

Experimental frame work

The proposed work include





Data collection:

The data is collected from kaggle web site. The data set contains collection of attribute related to individual .It will give a idea how difference factor associated with lung cancer .The data set contains demographic information ,life style choice and health indicator. The data set is examined to see how different parameter affect the lungs cancer level of danger.

Data pre processing

Data preprocessing includes data cleaning, data selection and normalization. To check the missing data a reliable data format is created. It also identify the duplicate data and clean up insufficient data.

Classification method

Support vector machine

Support vector machine uses a hyper plane concept to separate the data.It is basically used for multi class problem. It is one of the strongest algorithm for machine learning .It constructs hyper plane by maximizing the wideness between support vector points and minimize the risk of misclassification example of test data set.

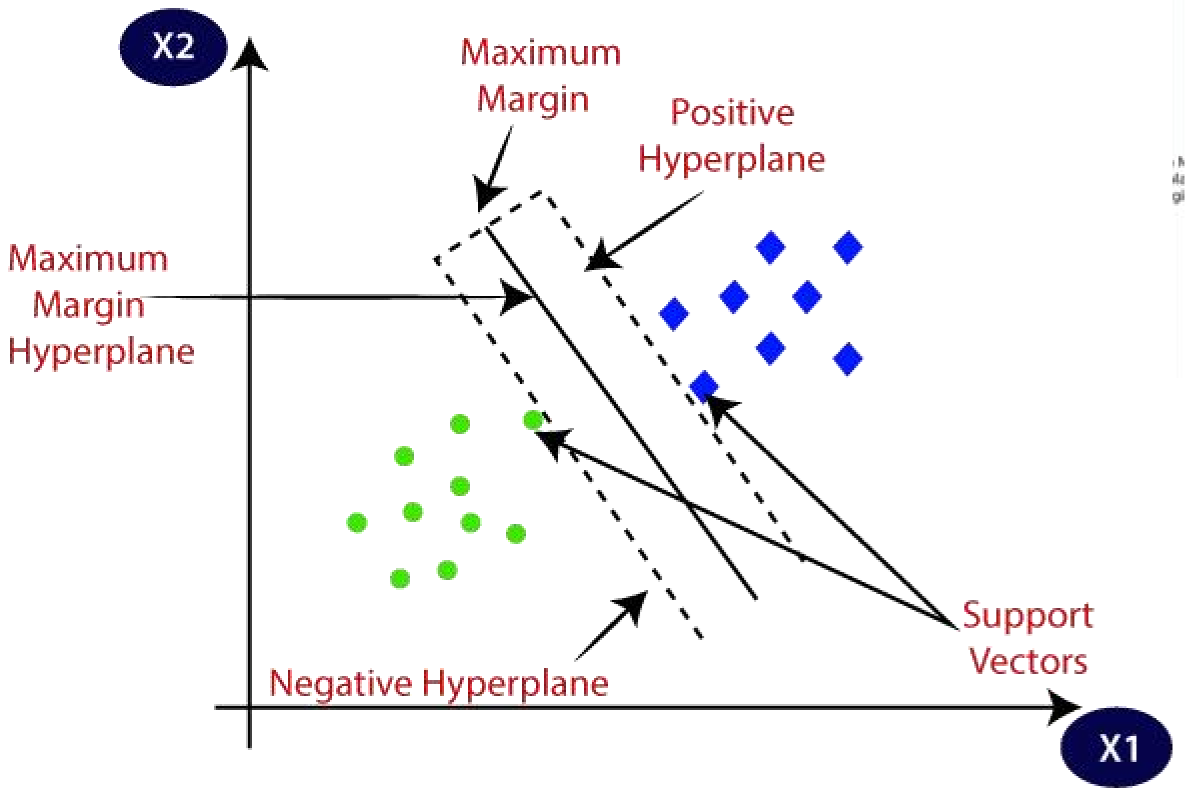
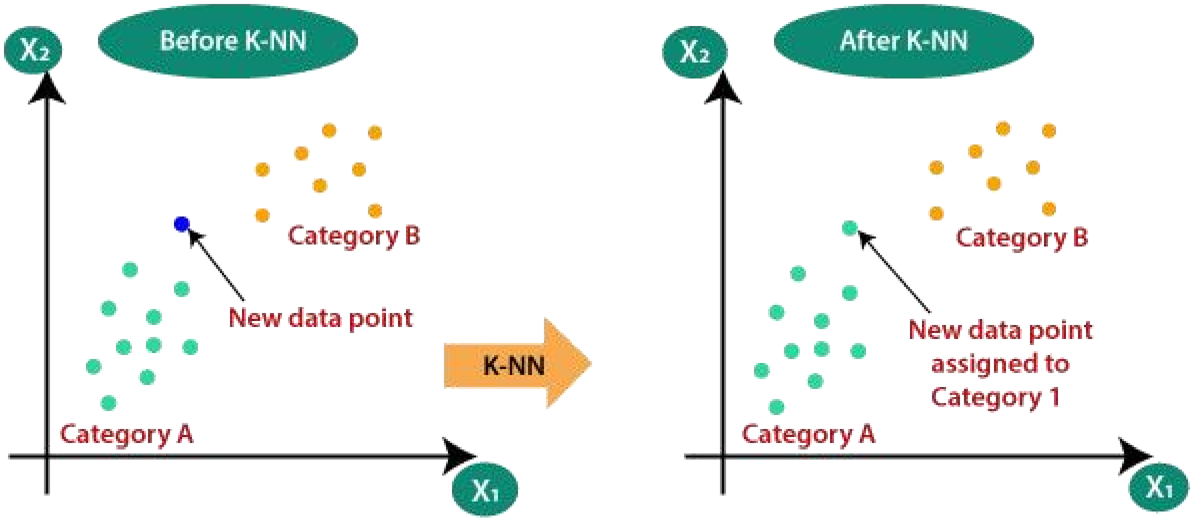


Figure represent the hyper plane, margin and support vector used to distinguish the class.

K nearest neighbor algorithm:

The algorithm is very simple in the machine learning method .In this method the classifying objects assigns to the label of nearest neighbor according to the feature vectors from the reference space. It is classified according to the distance and the determined number of k. So it is called as K-nearest neighbor algorithm.



Naïve bayes algorithm

Naïve bayes algorithm is is used in solving different types of machine learning and statistical problem.The classification process is made by taking the advantage of statistical method

Logistic Regression

Logistic Regression (LR), a popular mathematical modeling procedure used in the analysis of epidemiologic data set.

Logistic Regression method can be run in these steps [8]

1. Calculate with the logistic function.

2. Learn the coefficients for a logistic regression model.

3. Finally, make predictions using a logistic regression model

$$y = \frac{e^{(b_0 + b_1X)}}{1 + e^{(b_0 + b_1X)}}$$

Logistic regression parameters are estimated by maximizing logarithmic likelihood function using training data

XG Boost

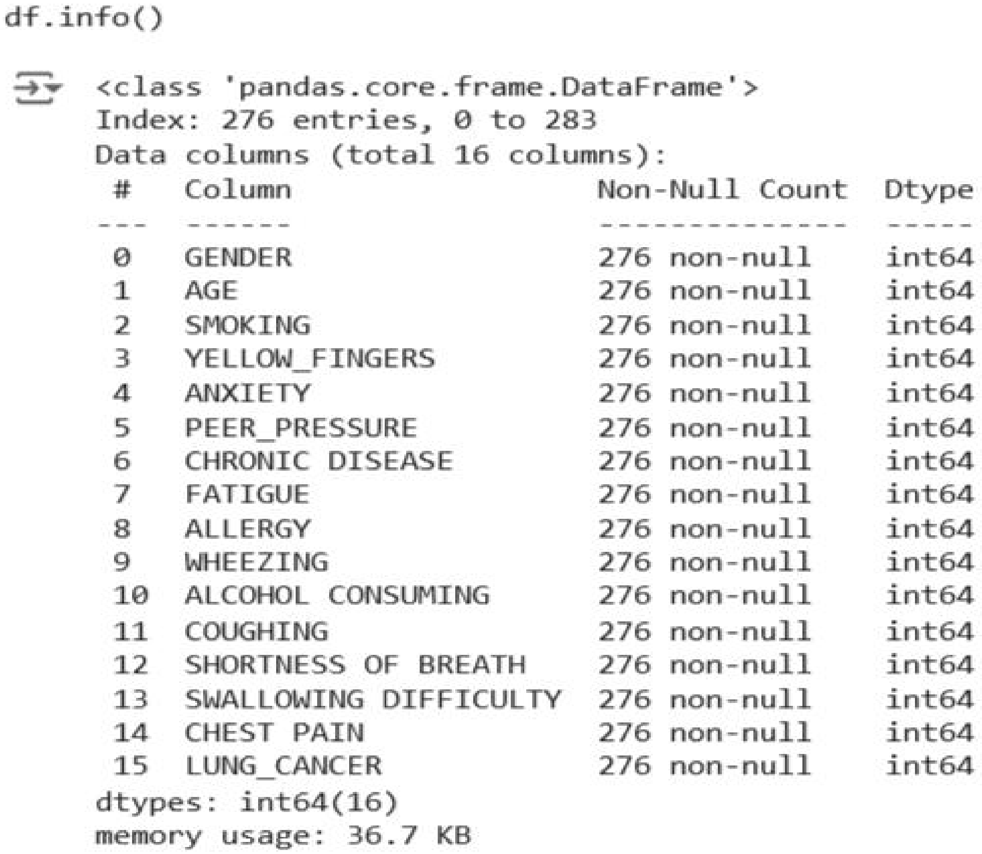
XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for "Extreme Gradient Boosting" and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

One of the key features of XGBoost is its efficient handling of missing values, which allows it to handle real-world data with missing values without requiring significant pre-processing. Additionally, XGBoost has built-in support for parallel processing, making it possible to train models on large datasets in a reasonable amount of time.

Experimental observation

The data set contains following informations.

The lung dataset consist of 309 entries and 16 number of column. For the study, the data is divided into training data and test data. The training set is used to build the model of classifier and test set is used to confirm it. In this study, as training data and test data are used in 75% and 25% , respectively. Our dependent variable has two category so we begin by considering classification problems using only two classes. Formally, each instance I is mapped to one element of the set of positive and negative class labels. A classifier model is a mapping from instances to estimated classes.The data set contains following information shown in the fig no(1)

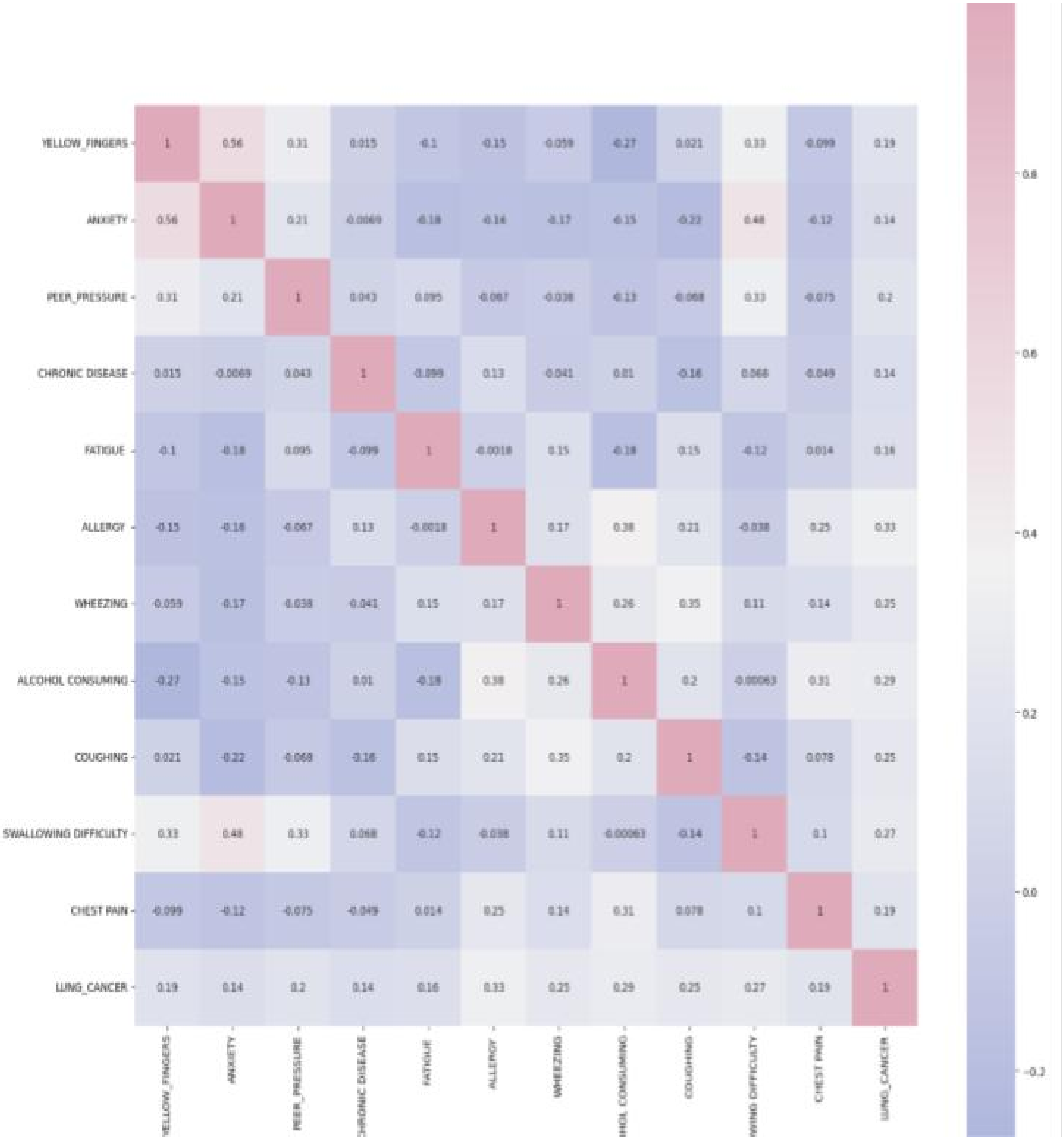


```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 276 entries, 0 to 283
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 276 non-null    int64
 1   AGE                    276 non-null    int64
 2   SMOKING                276 non-null    int64
 3   YELLOW_FINGERS         276 non-null    int64
 4   ANXIETY                276 non-null    int64
 5   PEER_PRESSURE          276 non-null    int64
 6   CHRONIC DISEASE        276 non-null    int64
 7   FATIGUE                276 non-null    int64
 8   ALLERGY                276 non-null    int64
 9   WHEEZING               276 non-null    int64
 10  ALCOHOL CONSUMING      276 non-null    int64
 11  COUGHING               276 non-null    int64
 12  SHORTNESS OF BREATH    276 non-null    int64
 13  SWALLOWING DIFFICULTY  276 non-null    int64
 14  CHEST PAIN             276 non-null    int64
 15  LUNG_CANCER            276 non-null    int64
dtypes: int64(16)
memory usage: 36.7 KB
```

Fig 1

The correlation metrics shown in the fig.2



Gender ,age ,smoking, shortness of breath do not have much impact on lung cancer . The correlation matrix shows that ANXIETY and YELLOW_FINGERS are correlated more than 50%

Performance comparison based on classification report

Classifier report for Logistic Regression shown in the given table (1)

	precision	recall	f1-score	support
0	0.96	1.00	0.98	64
1	1.00	0.95	0.97	56
accuracy			0.97	120
Macro avg	0.98	0.97	0.97	120
weighted avg	0.98	0.97	0.97	120

Classifier report for Decision Tree Classifier shown in the given table(2)

	precision	recall	f1-score	support
0	0.93	0.97	0.95	64
1	0.93	0.91	0.94	56
accuracy			0.94	120
Macro avg	0.94	0.94	0.94	120
weighted avg	0.94	0.94	0.94	120

Classifier report for KNeighborsClassifier shown in the given table (3)

	precision	recall	f1-score	support
0	0.93	1.00	0.96	64
1	1.00	0.91	0.95	56
accuracy			0.96	120
Macro avg	0.96	0.96	0.96	120
weighted avg	0.96	0.96	0.96	120

Classifier report for Classifier report for Gaussian Naive Bayes classifier in the given table (4)

	precision	recall	f1-score	support
0	0.95	0.89	0.92	64
1	0.88	0.95	0.91	56
accuracy			0.92	120
Macro avg	0.92	0.92	0.92	120
weighted avg	0.92	0.92	0.92	120

Classifier report for Classifier report for Gaussian Naive Bayes classifier in the given table(5)

	precision	recall	f1-score	support
0	0.95	0.89	0.92	64
1	0.88	0.95	0.91	56
accuracy			0.92	120
Macro avg	0.92	0.92	0.92	120
weighted avg	0.92	0.92	0.92	120

Classifier report for Gaussian MultinomialNB in the given table(6)

	precision	recall	f1-score	support
0	0.89	0.73	0.80	64
1	0.75	0.89	0.81	56
accuracy			0.81	120
Macro avg	0.82	0.81	0.81	120
weighted avg	0.82	0.81	0.81	120

Classifier report for Classifier report for Support vector classifier in the given table(7)

	precision	recall	f1-score	support
0	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56
accuracy			0.98	120
Macro avg	0.98	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Classifier report for random forest classifier is shown in table(8)

	precision	recall	f1-score	support
0	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56
accuracy			0.98	120
Macro avg	0.98	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Classifier report for MLP classifier is shown in table(9)

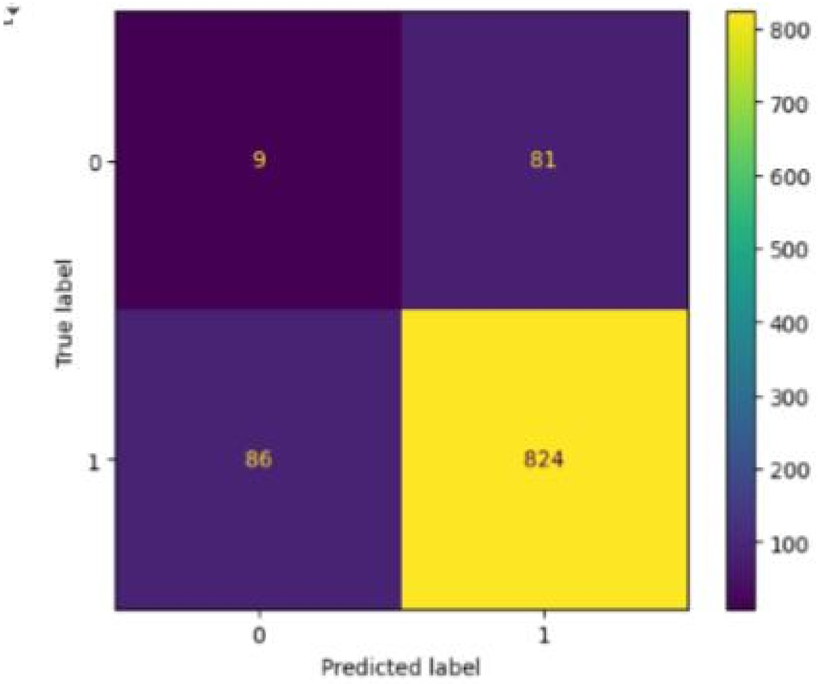
	precision	recall	f1-score	support
0	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56
accuracy			0.98	120
Macro avg	0.98	0.98	0.98	120
weighted avg	0.98	0.98	0.98	120

Classifier report for Gradient Boosting Classifier is shown in table(10)

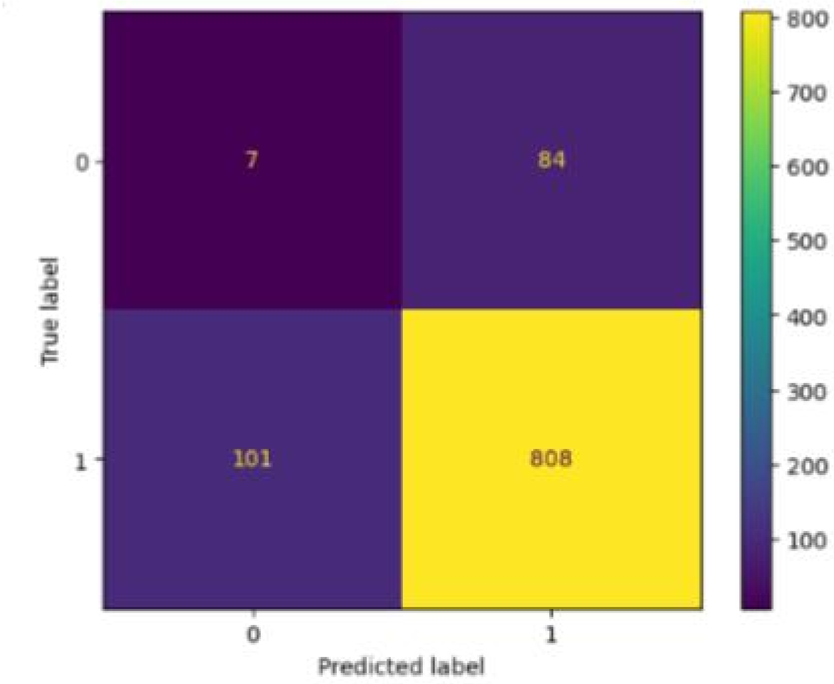
	precision	recall	f1-score	support
0	0.98	0.98	0.98	64
1	0.98	0.98	0.98	56
accuracy			0.98	120
Macro avg	0.98	0.98	0.98	120
weighted	0.98	0.98	0.98	120

avg				

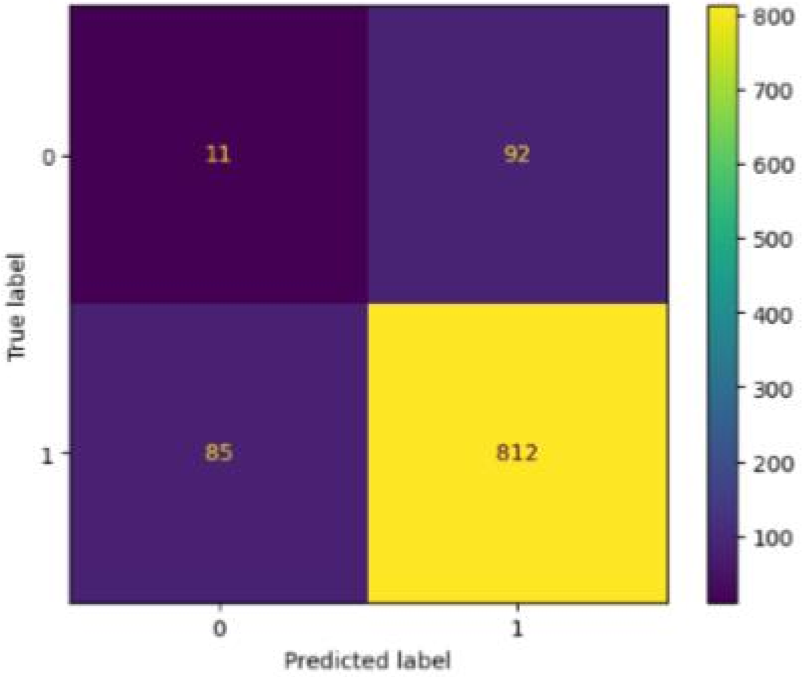
Confusion metrics for logistic regression



Confusion metrics for Decision Tree Classifier

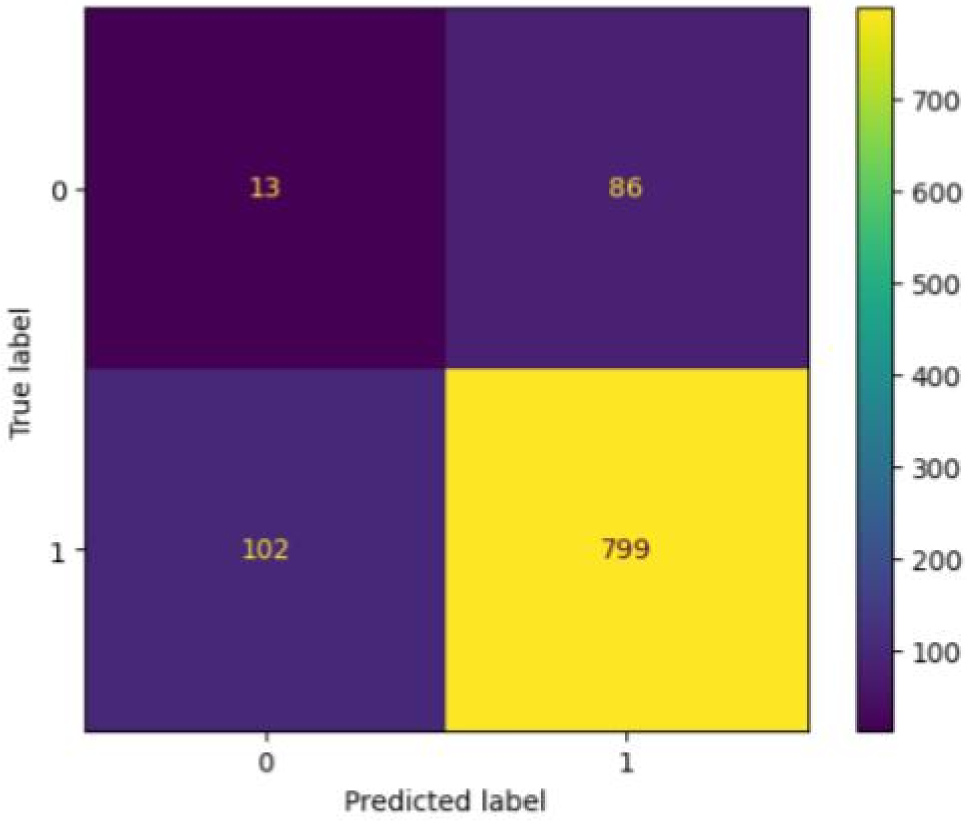


Confusion metrics for KNeighborsClassifier

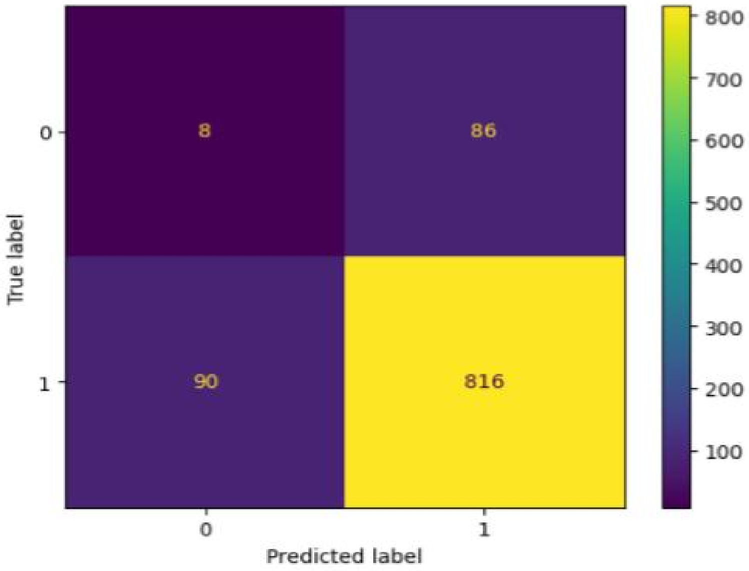


Confusion metrics for GaussianNB

Confusion metrics for GaussianNB classifier



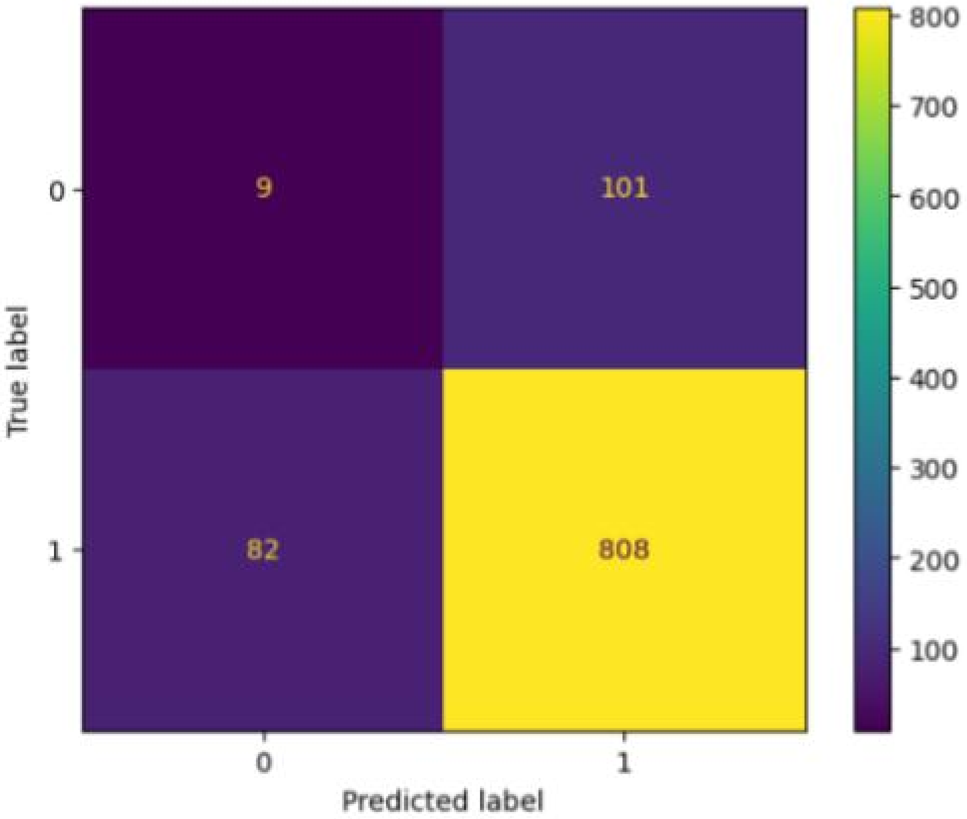
Confusion metrics for SVC classifier



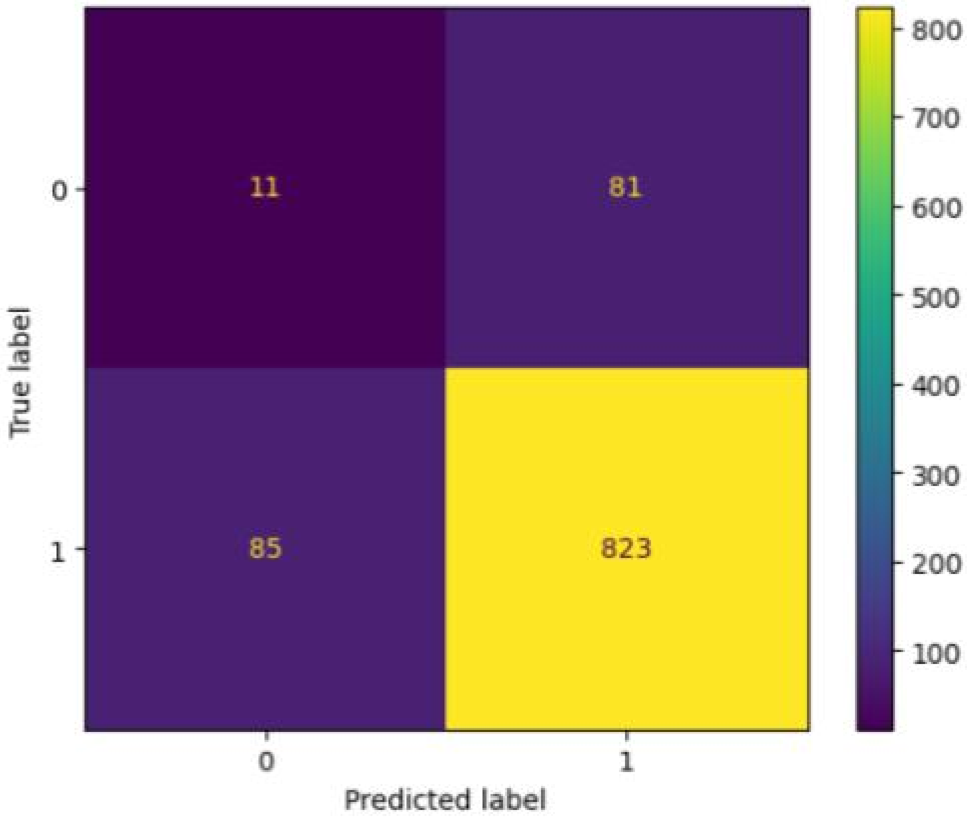
aining

1.

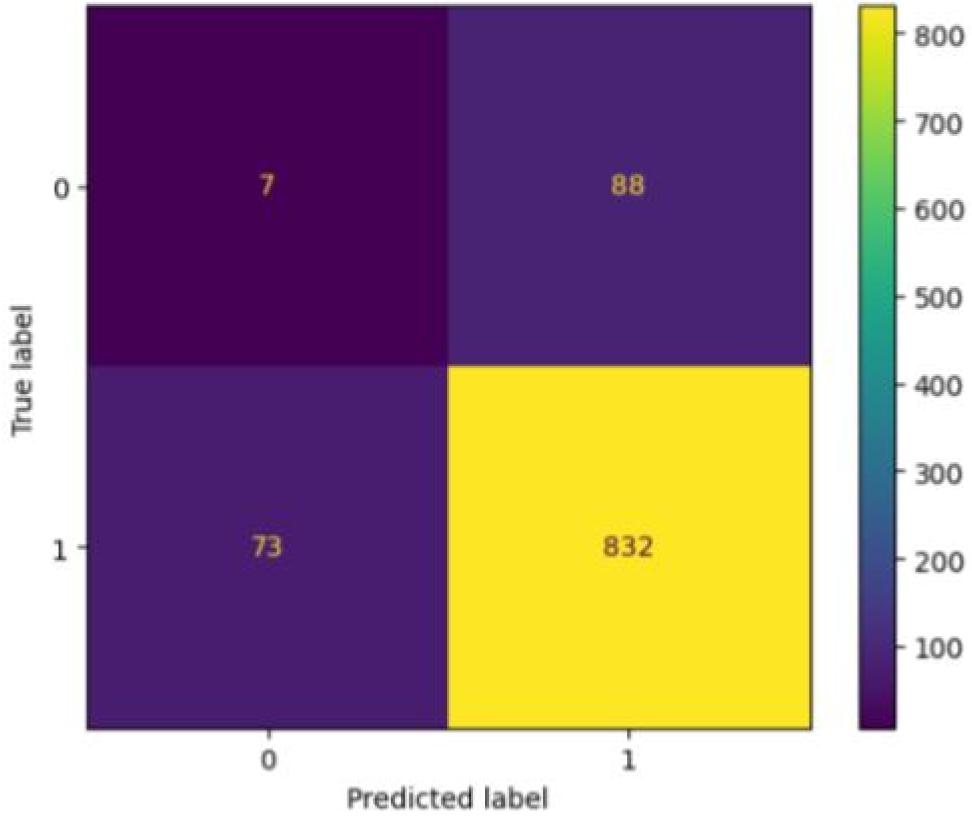
Confusion metrics for MultinomialNB



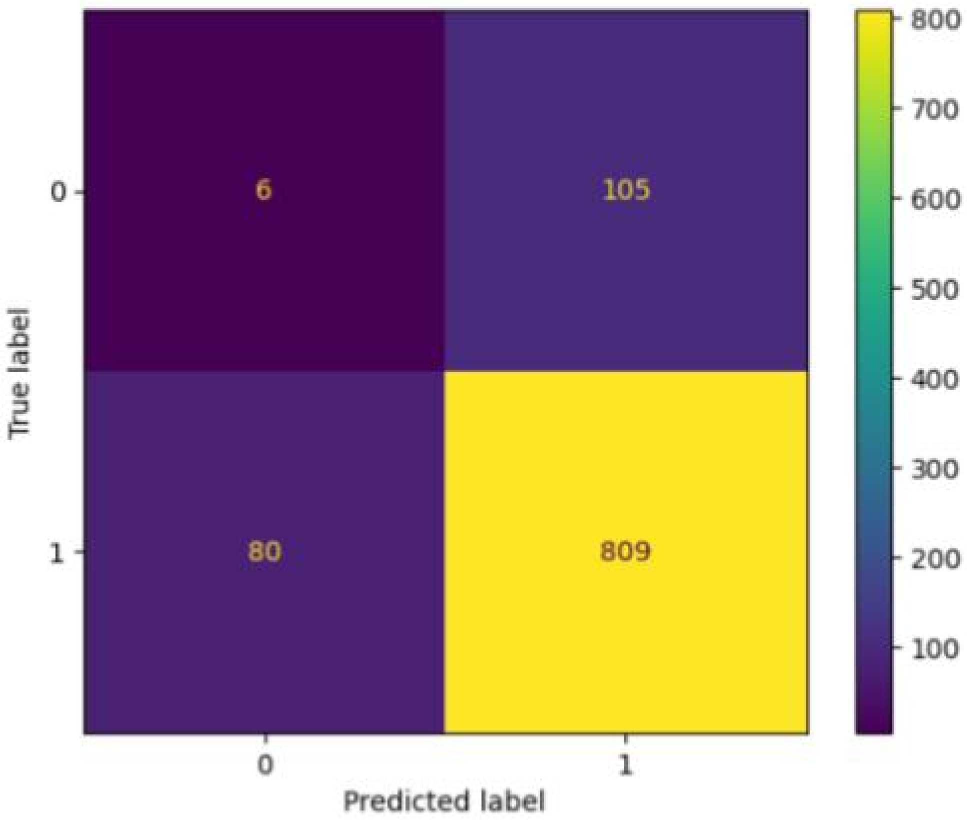
Confusion metrics for Random Forest Classifier



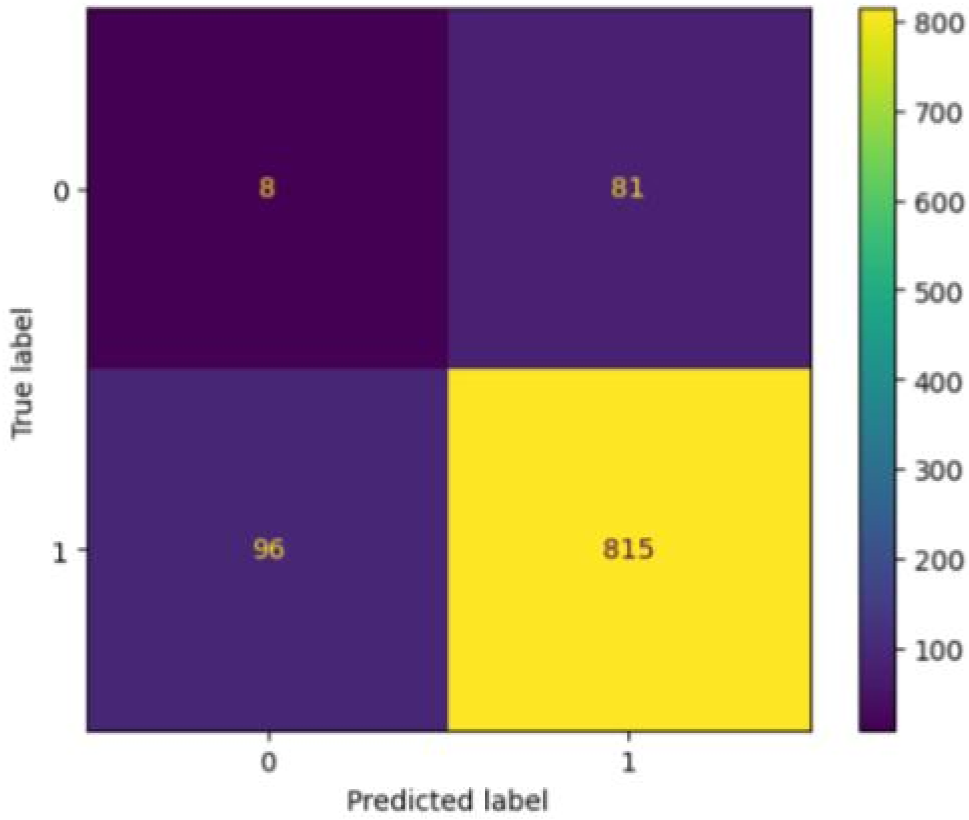
Confusion metrics for   xgboostclassifier



Confusion metrics for MLPClassifier



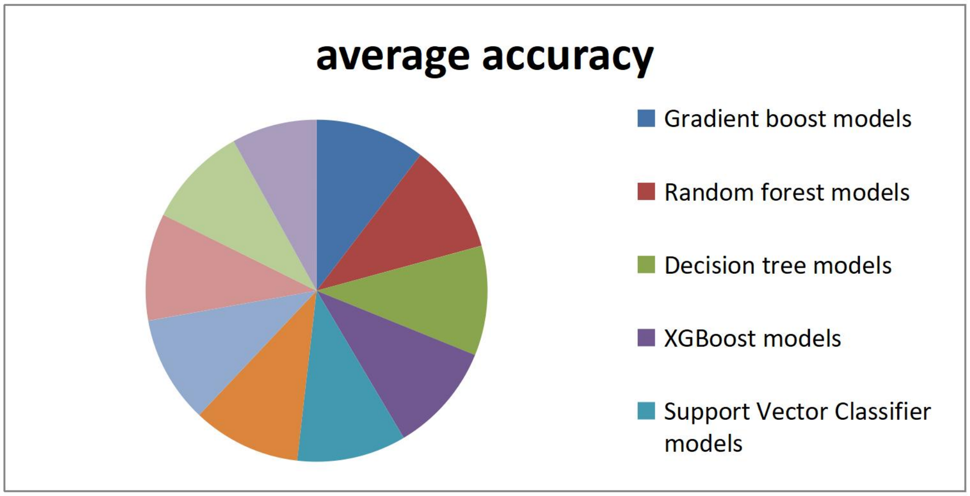
Confusion metrics for Gradient Boosting Classifier



The following table shows the accuracy of k fold cross validation.

machine learning algorithm	average accuracy
Gradient boost models	0.956072695
Random forest models	0.954033688
Decision tree models	0.953989362
XGBoost models	0.951861702
Support Vector Classifier models	0.949734043
Multi-layer perceptron model	0.941400709
KNN models	0.933067376
Logistic regression models	0.93089539
Gaussian naive bayes models	0.884618794
Multinomial naive bayes models	0.742154255

The comparison is shown in the figure. Gradient boost model showing highest average accuracy. Multinomial naïve bayes model showing lowest average accuracy.the graph is shown below



Conclusion:

lung cancer is the major global health concern due to rapid progression and high mortality rate .Early detection is very much crucial as a result the survival rate can be increased. In this research paper we we have used ten different machine learning classification algorithm including logistic regression, decision tree, K nearest neighbor, Gaussian naïve Bayes, multinomial naive Bayes, Support vector classifier, random forest, XG Boost,multi layer perceptron and gradiant boosting classifier to predict the lung cancer based on different variable, Gradient boost model showing highest average accuracy. Multinomial naïve bayes model showing lowest average accuracy. In future We will implement CT scan image to implement deep learning model for better prediction.

References

[1]Sinjanka, Y., Kaur, V., Musa, U. I., & Kaur, K. (2024). Ml-based early detection of lung cancer: an integrated and in-depth analytical framework. *Discover Artificial Intelligence*, *4*(1), 92.

2[]Li, Y., Wu, X., Yang, P., Jiang, G., & Luo, Y. (2022). Machine learning for lung cancer diagnosis, treatment, and prognosis. *Genomics, proteomics & bioinformatics*, *20*(5), 850-866.

[3]Omar, A. A. C., & Nassif, A. B. (2023, February). Lung cancer prediction using machine learning based feature selection: A comparative study. In *2023 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-6). IEEE.

[4]Shatnawi, M. Q., Abuein, Q., & Al-Quraan, R. (2025). Deep learning-based approach to diagnose lung cancer using CT-scan images. *Intelligence-Based Medicine*, *11*, 100188.

[5]Meeradevi, T., Sasikala, S., Murali, L., Manikandan, N., & Ramaswamy, K. (2025). Lung cancer detection with machine learning classifiers with multi-attribute decision-making system and deep learning model. *Scientific Reports*, *15*(1), 8565.

[6] K. Roy et al., "A Comparative study of Lung Cancer detection using supervised neural network," in 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), 2019, pp. 1–5.

[7] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 2018, pp. 1–4.

[8] Banerjee, N., & Das, S. (2020, March). Prediction lung cancer– in machine learning perspective. In 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-5). IEEE

[9] Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends, 2(01), 10-19.

[10] Karhan, Z., & Tunç, T. (2016). Lung Cancer Detection and Classification with Classification Algorithms. IOSR Journal of Computer Engineering (IOSR-JCE), 18(6), 71-7.

[11] D. Reddy, E. N. H. Kumar, D. Reddy, and P. Monika, "Integrated Machine Learning Model for Prediction of Lung Cancer Stages from Textual data using Ensemble Method," in 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 353–357.

[12] B. M. Boban and R. K. Megalingam, "Lung Diseases Classification based on Machine Learning Algorithms and Performance Evaluation," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 315– 320.

[13] Elnakib, A., Amer, H. M., & Abou-Chadi, F. E. (2020). Early lung cancer detection using deep learning optimization.

[14] Karhan, Zehra, and Taner Tunç. "Lung cancer detection and classification with classification algorithms." *IOSR Journal of Computer Engineering (IOSR-JCE)* 18.6 (2016): 71-7