IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS USED TO DETECT LUNG CANCER

[1]ANJALI RAJ, [2]AMBILY JACOB

[1]Msc Scholar, [2]Assistant Professor

[1,2]Department of Computer Science

[1,2]St.Joseph's College (Autonomous), Irinjalakuda, Thrissur, India

ABSTRACT: Lung cancer is a dangerous disease for human beings and it causes death. Nowadays, so many people are suffering from cancer as well as covid related problems. Various studies were conducted to detect cancer at earlier stages using Machine Learning, Artificial Intelligence. So that we can cure the patients and save their life to some extent. In this paper, we provide an overview of the lung cancer prediction approaches. Various types of Machine Learning Algorithms like Multinomial Naïve Bayes, Ridge classifier, Random Forest classifier & SGD classifier have been applied in the healthcare sector for the detection and prognosis of lung cancer.

KEYWORDS: Lung Cancer, Machine Learning, Multinomial Naïve Bayes, Ridge classifier, Random Forest classifier , SGD classifier, Decision Making Trees.

INTRODUCTION:-

Human health is an important factor in the economic development of any country. Cancer is a disease in which cells instigate to develop more out of control and it is different from tumor. Cancer is a disease in which cells begin to divide uncontrollably and tumor is the mass of infected cells occurring in solid tissue i.e. bones, organs etc. Among many causes for the loss of human lives is cancer, because this disease is hard to detect in early stages. Normally, cancer is detected at the latter stages but if it is correctly detected, then we can cure it and have the chance to survive. Individuals who smoke are supposed to be at more threat of having lung cancer, through it may occur in people who are never smoked. The symptoms of lung cancers might contain continuous cough, Coughing up blood, Smallness of inhalation, Chest pain, Throatiness, Trailing weight deprived of trying, Bone pain, Headache etc. Other symptoms of the lung cancer are the over usage of tobacco, radon gases, air pollutants and chemicals in workspaces. Detection of the lung cancer is important in early stages, for that we can use machine learning. Machine learning techniques like Artificial Neural Networks, , Support Vector Machine etc.. are widely used to detect the cancer

As per the report of 2021 in India total cancer patients were 1.9 million that are suffering from common major 6 types Breast, Lung, Pancreas, Ovary , Colon-Rectum and Stomach. The most common patients with cancer were analyzed are shown in the figure 1 below as graph. The most of the patients with cancer were analyzed at the nearby radical phase for:-

1. Breast Cancer – 63%
2. Lung Cancer – 53%
3. Ovarian Cancer – 45.70%
4. Stomach Cancer - 25%
5. Pancreas Cancer – 22%
6. Colon-Rectum – 18%

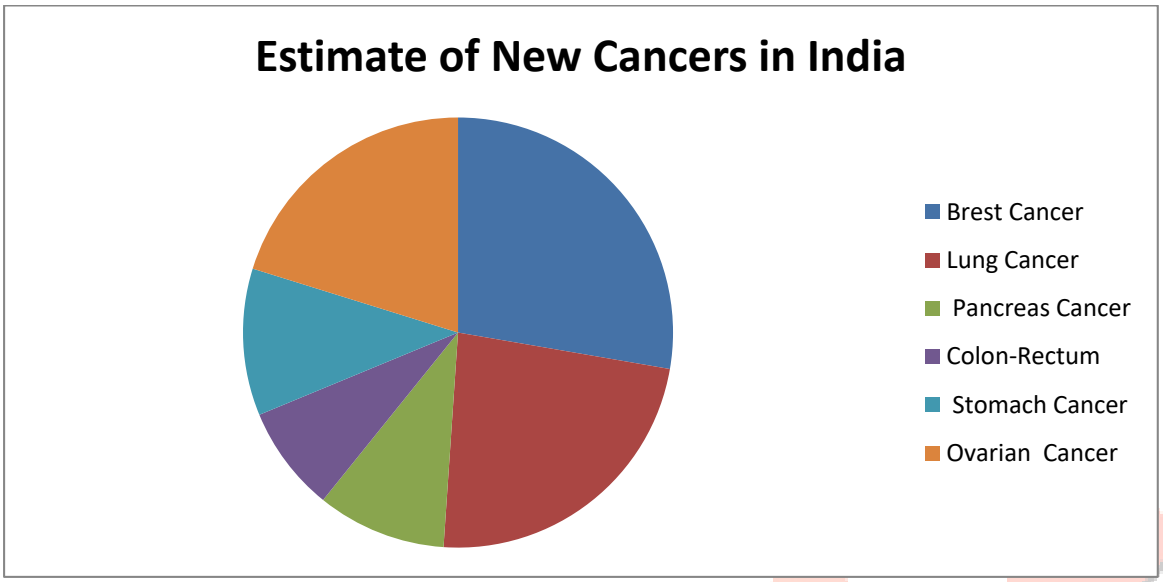

Fig.1. Percentage of person suffering from innumerable cancer disease in India in 2021

As per the report of recent 2021 globally an approximate 21.9 million different cancer cases were reported and nearly 10 million cancer death were happened in 2021. According to the study Lung cancer is the second most dangerous cancer among the other ones.

RELATED WORKS: Lot of works has been carried out to predict Lung cancer using machine learning algorithms.

Table 1. Related Works

Paper Details	**Techniques/Algorithm used**	**Conclusion drawn**
P. Nanda, Pranamita and N. Duraipandian	Naive Bayes, SVM, Decision tree, Random forest & Improved random forest were used. Their accuracies are equated.	Improved random forest gives highest accuracy of 98% among all others.
Banerjee, Nikita and Subhalaxmi Das	To check whether tumor is malignant or benign, algorithm used for ANN, Random forest and SVM.	Among all ANN is more accuracy and has correctly recognized maximum number of malignant tumor.
Raoof, Syed Saba, M. Jabbar, Syed Aley Fathima	It was a review paper that covered some important research written on detecting Lung cancer by means of ML Algorithm. They also described the SVM, KNN and ANN.	Deep learning algorithms can also be used to increase the accuracy in detecting Lung cancer.

Palani, D., and K. Venkatalakshmi	IoT centered predictive model by means of Fuzzy C mean Clustering, Neural Network for predicting disease & association rule mining. Decision Tree for classification was used.	Their proposed model got 85% accuracy as per the existing work.
Bankar, Atharva, Kewal Padamwar, and Aditi Jahagirdar	They have used Decision Tree, Random Forest & XGBoost for identifying the underlying data patterns for calculating some of important features for detecting cancer disease faster.	Blood in Cough, Finger nail clubbing, Genetic Risk; impervious Smoking & Snoring are top symptoms in Lung cancer disease.
Patra, Radhanath	RBF, KNN, Naive Bayes and J48 algorithms were used on data set on their proposed model.	81.25% of accuracy was found of RBF classifier.
Roy, Kyamelia, et al.	They have used Grayscale transformation, adjusting the contrast of the image, Saliency Enhancement, LR & SVM methods have been applied to CT scan pictures to get the area of interest.	Accuracy of proposed model may improve by using a plenty sets of images, using SVM logistic regression.
Gunaydin, Ozge, et al.	They have used KNN, SVM, Decision Tree ANN for calculating accuracy, precision, recall (before and after PCA) & F-measure on the data taken.	Decision Tree has maximum accuracy among all. They have not used Naïve Bayes and Feed Forward Neural Network to their images because of huge data size.

APPLICATIONS OF MACHINE LEARNING IN HEALTHCARE: Machine Learning is an application of Artificial Intelligence, where human automated the machines to learn deprived of our involvement. The Data provides to the computer and it learns from that data. Due to the advantages, Machine Learning is the new and involved technology in worldwide. It also proven that this technology is a boom in the medical applications. The most common applications in healthcare are programming medical billing, progress of clinical caution strategies and medical judgment support. Machine learning algorithms helps in predicting and diagnosis of disease. Figure 3 shows the application of Ml in healthcare areas.

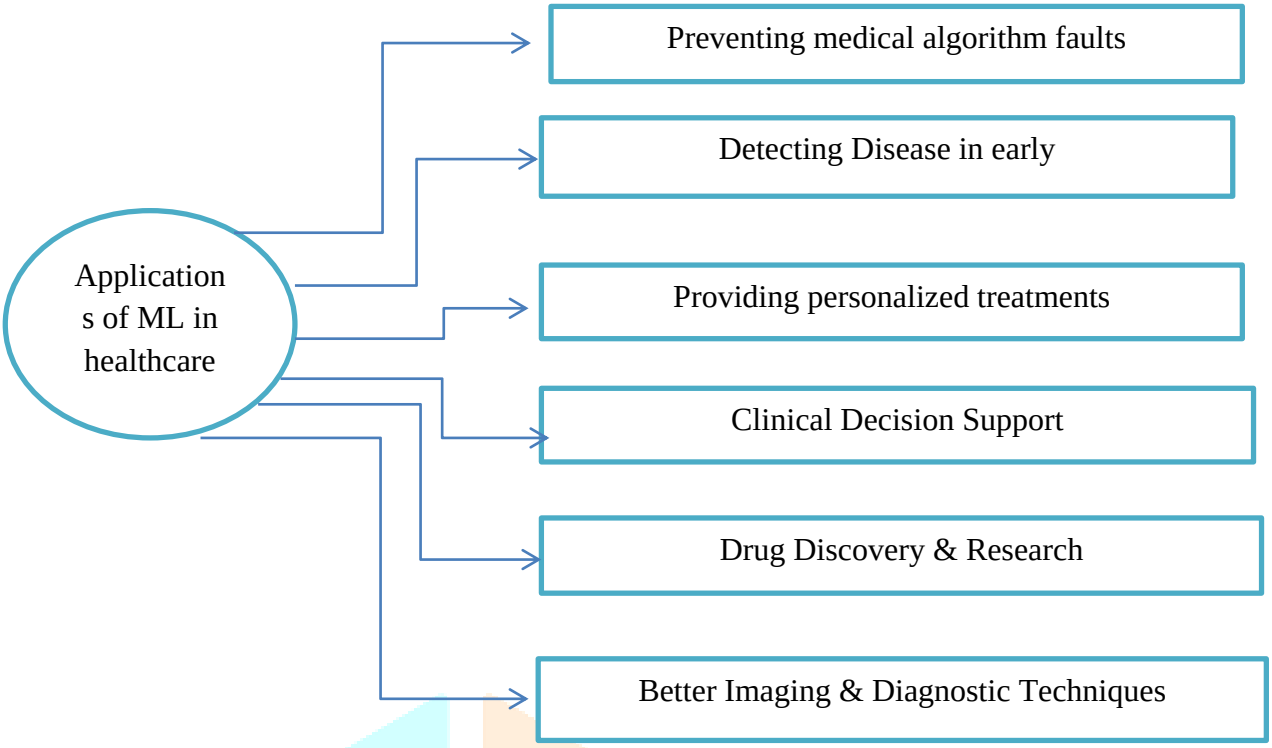

Fig.3. Uses of Machine Learning in Healthcare

MACHINE LEARNING ALGORITHMS:

- Ridge Classifier:-

It is used to evaluate multicollinearity. While documents are produced in multicollinearity, the slightly squares assessment is a neutral approximation, which has a huge variants and is then possible to be distant from the correct rate. It declines the standard errors and yields a more consistent evaluation by totaling a defined degree of eccentricity to the regression valuation. Finally it finds a linear solution i.e.

$$\arg min_{w,w0} \frac{1}{n}\sum(y_i - (W^T x_i + W_0)) \wedge 2 + \lambda \parallel W \wedge 2 \parallel \quad (1)$$

Where, x_i indicates the i^{th} individual and y_i represents the phenotype target value.

- Multinomial Naïve Bayes:-

It is used to define word frequency. Multinomial Naïve Bayes classifier can be expressed as

$$P(p/n) \propto P(p)\prod_{1\leq k\leq nd} P(t_k/p) \quad (2)$$

Where 'n' is the news article, p is the polarity. Here $P(t_k /p)$ signifies the conditional possibility that whether t_k happens in a news article with polarity p which can be calculated as:-

$$p(t_{k/p}) = \frac{count(t_{k/p})+1}{count(t_p)+|V|} \quad (3)$$

Here count (t_k/p) represents the number of intervals t_k happens in news articles which have polarity p & count (t_p) represents the entire number of tokens present in the news article of polarity p.

- Stochastic Gradient Descent Classifier:-

SGD is up-front however tremendously creative mode to contract with discriminative learning of linear classifiers beneath curved misfortune capabilities, i.e. Support Vector Machines and Logistic Regression. This classifier is efficiently applied to large-scale and scarce machine learning topics commonly skilled in DC.

By solving the equation (1) using SGD, the corresponding equation becomes

$$\omega_t = \omega_{t-1} - \eta_t S^{-1}(\lambda\omega_{t-1} + \phi_1'({\omega_{t-1}}^T X_t, Y_t)X_t) \quad (4)$$

Where, $\phi_1'(p, y) = \frac{\delta}{\delta p}\phi(p, y)$.

- Random Forest Classifier:-

It is a Machine Learning technique that is used to solve regression and classification problems. This classifier consists of many decision trees. The 'Forest' generated by the Random Forest (RF) is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of ML algorithms. This algorithm establishes the outcome based on the predictions of the decision trees and it predicted by taking the average mean of the output from various trees. Increasing in the number of trees increases the precision of the outcome.

Decision Trees are the building block of Random Forest algorithms. A Decision Tree (DT) consists of components, they are Decision nodes, Leaf nodes & a Root node. The nodes in the DT represent attributes that are used for predicting the outcome and these nodes provide a link to the leaves.

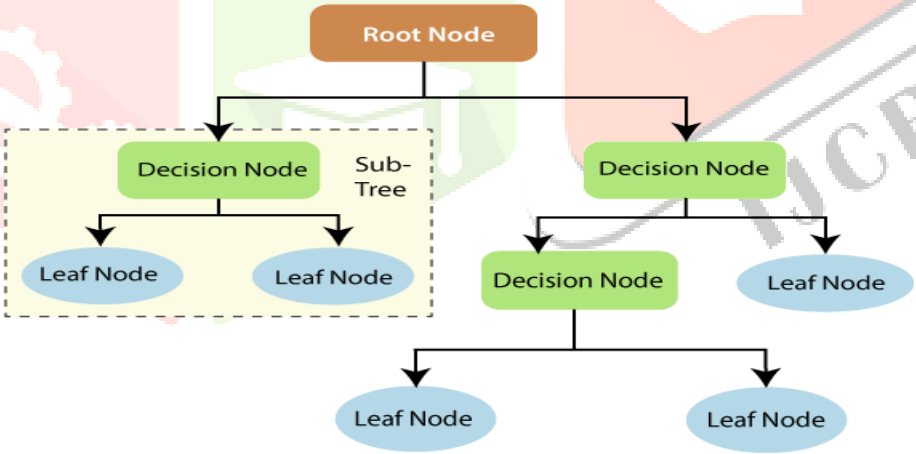

Fig.4. Three types of nodes in a DT

The difference between the DT algorithm and RF algorithm is that the establishing of the root nodes and the segregating nodes are done randomly in the latter. The RF follows the bagging method to generate the required prediction. The leaf node of each DT is the final output produced. In Random Forest classifier the selection of the final output follows the majority-voting system. That is the output chosen by the majority of DT becomes the final output of the RF system.

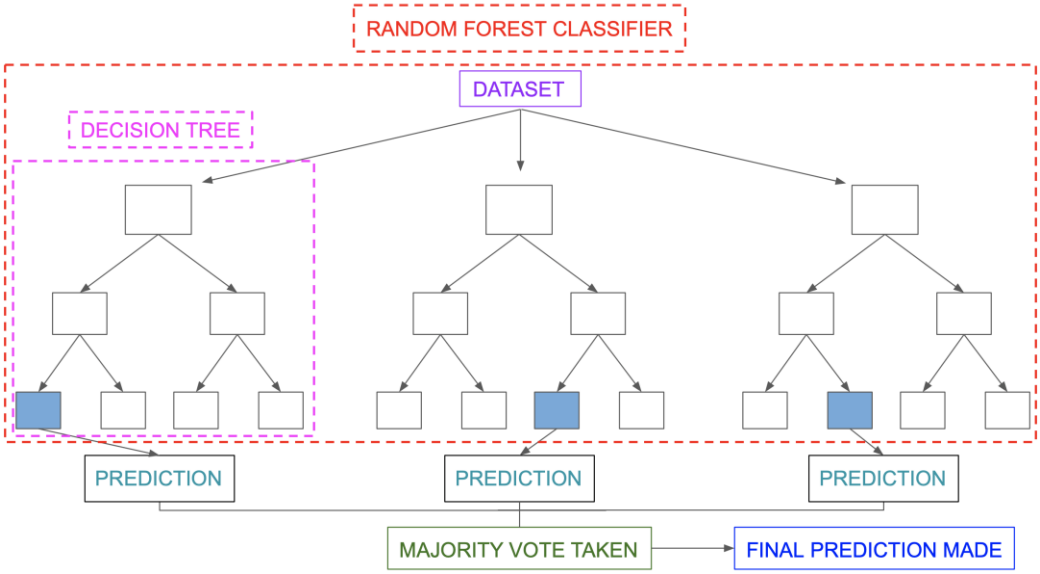

Fig.5. Diagram of Random Forest classifier

PROPOSED METHODOLGY:-

We performed 4 Machine Learning algorithms on the data set of lung cancer. The flowchart of experiment can be summarized as follows

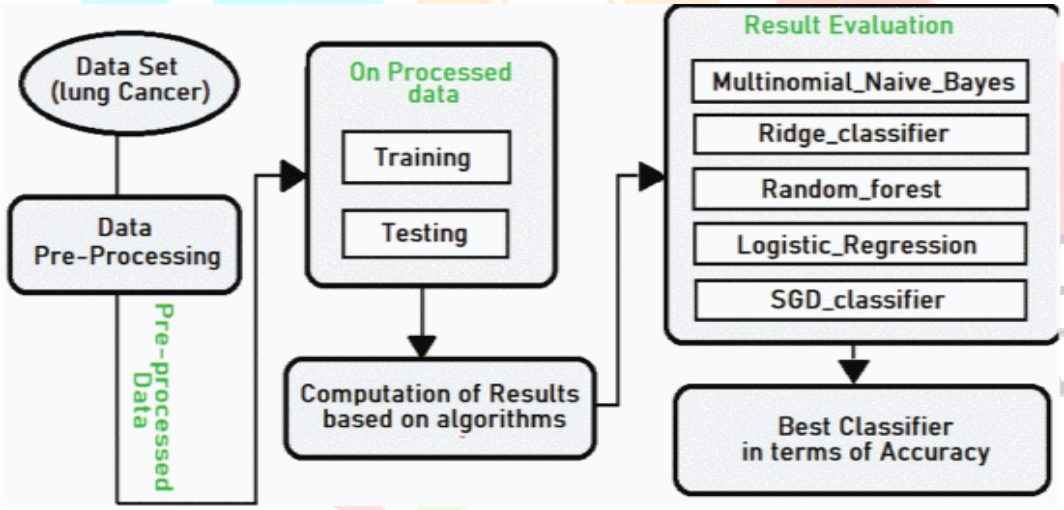

Fig.6. Proposed Methodology Flowchart

The steps involved in this methodology are:-

Step 1: Mining of the cancer datasets from the online source Kaggle.

Step 2: The pre-processing and cleaning of data on the given attributes.

Step 3: Training & Testing of data. 75% data was trained and then 25% data were tested.

Step 4: Compute the results after applying 4 ML algorithms on data. Mis-classification, Sensitivity and Specificity were calculated on the cancer dataset.

Step 5: The best algorithm is selected based on the accuracy.

RESULT EVALUATION:-The proposed work was implemented by taking 4 Machine Learning algorithms and Mis-classification, Sensitivity and Specificity were done. The following table shows the Accuracy (An assessment metric that defines the no. of right estimates made by the model), Precision (It responses the query, out of the no. of times a model forecast positive), Recall (It signifies that out of total authentic positive values), F-measure and Support on 4 algorithms. From these we can conclude that Random Forest gives the best accuracy, sensitivity, Specificity & Mis-classification among the other algorithms used in this work.

Table 2. Details of performing matrix of used algorithms

Classifier Algorithms	Accuracy (%)	Precision	Recall	F-measure	Support
Ridge classifier	92.01	0.93 0.85 0.97	0.85 0.90 1.00	0.89 0.88 0.98	78 78 94
Multinomial Naïve Bayes	76.02	0.82 0.65 0.82	0.64 0.65 0.97	0.72 0.65 0.89	78 78 94
Stochastic Gradient Descent Classifier	87.20	0.72 1.00 0.98	1.00 0.59 1.00	0.84 0.74 0.99	78 78 94
Random Forest Classifier	99.09	1.00 1.00 1.00	1.00 1.00 1.00	1.00 1.00 1.00	78 78 94

The details of sensitivity, Specificity & Mis-classification calculated on these four algorithms as follows:

Table 3. Details of Mis-classification, Sensitivity & Specification calculated on used algorithms

Name of Algorithm	Mis-classification	Sensitivity	Specification
Ridge classifier	0.11	0.98	0.85
Multinomial Naïve Bayes	0.26	0.32	0.68
Stochastic Gradient Descent Classifier	0.19	0.61	1.0
Random Forest Classifier	0.0	1.0	1.0

The following figure shows the accuracy graph of these 4 algorithms used here:-

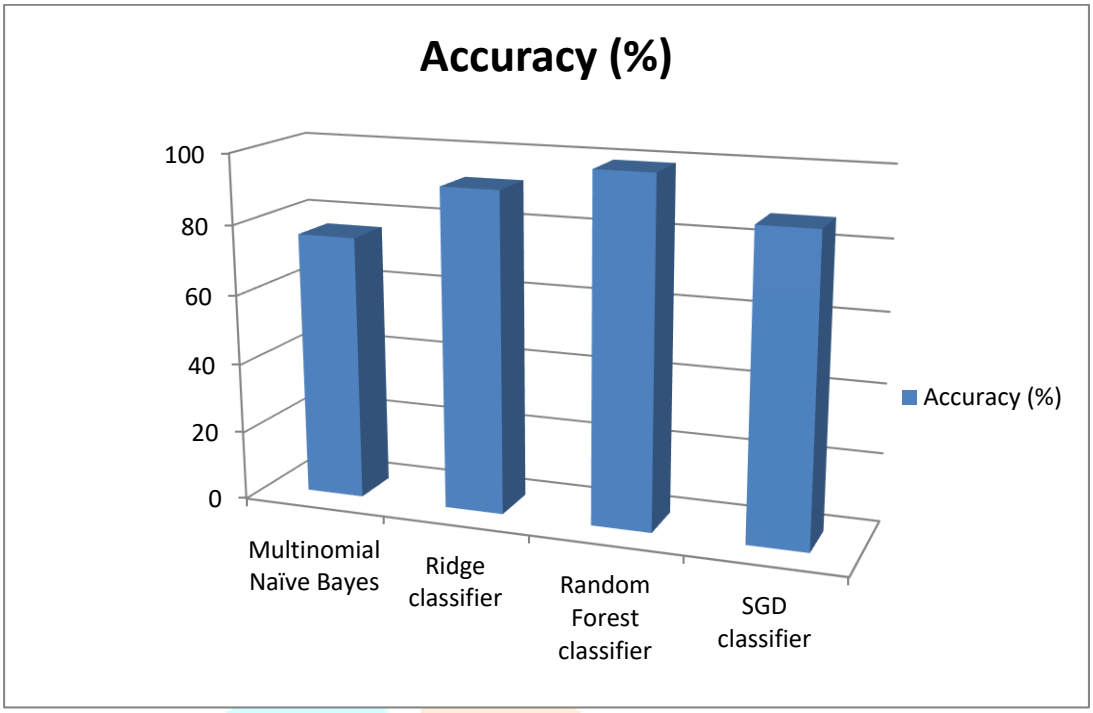

Fig.7. Accuracy of the 4 algorithms

CONCLUSION:-

The core emphasis of this paper was to find out the best machine learning algorithm to detect the chance of lung cancer in less time and more accurate way. There are some other factors that also helps in deciding whether the disease in cancer or not. Our main purpose was to explore the cancer features with the help of some initial features and by using the machine learning algorithms. For this, we have implemented four machine learning algorithms i.e. Multinomial Naïve Bayes, Ridge classifier, Random Forest classifier & SGD classifier on the dataset of lung cancer disease and have calculated Accuracy, Precision, Recall, F-measure, Support, Mis-classification, Sensitivity & Specification. From this experiment we can conclude that Random Forest has the highest accuracy of 99.09% followed by Ridge classifier having accuracy 92.01%.

REFERENCE:-

1. Syed Saba Raoof, M. Jabbar and Syed Aley Fathima, "Lung Cancer Prediction using Machine Learning: A Comprehensive Approach", *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE*, 2020.

2. S. Luxmi, J Kaur Sandhu and N Goyal, "Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach" in Advances in Communication & Computational Technology, Singapore: Springer, pp. 1033-1042, 2021.

3. Nikita Banerjee and Subhalaxmi Das, "Prediction Lung Cancer-In Machine Learning Perspective", *2020 International Conference on Computer Science Engineering & Applications (ICCSEA). IEEE*, 2020.

4. Prashant Mathur et al., "Cancer statistics 2020: report from national cancer registry programme India", *JCO Global Oncology*, vol. 6, pp. 1063-1075.

5. Sung Hyuna et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: A Cancer Journal for Clinicians*.

Show in Context

6. Pu-Yuan Xing et al., "What are the clinical symptoms and physical signs for non-small cell lung cancer before diagnosis is made? A nation-wide multicenter 10-year retrospective study in China", *Cancer medicine*, vol. 8.8, pp. 4055-4069, 2019.

Show in Context

7. G. Singh et al., "Comparison between multinomial and Bernoulli naïve Bayes for text classification", *2019 International Conference on Automation Computational and Technology Management (ICACTM). IEEE*, 2019.

Show in Context

8.D. Li, Q. Ge, P. Zhang, Y. Xing, Z. Yang and W. Nai, "Ridge Regression with High Order Truncated Gradient Descent Method", *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 252-255, 2020.

9.F. Kabir, S. Siddique, M. R. A. Kotwal and M. N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier", *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*, pp. 1-4, 2015.

Show in Context

10 . Radhanath Patra, "Prediction of Lung Cancer Using Machine Learning Classifier", *International Conference on Computing Science Communication and*, 2020.

11. Atharva Bankar, Kewal Padamwar and Aditi Jahagirdar, "Symptom Analysis using a Machine Learning approach for Early Stage Lung Cancer", *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE*, 2020.

12. Pranamita Nanda and N. Duraipandian, "Prediction of Survival Rate from Non-Small Cell Lung Cancer using Improved Random Forest", *2020 International Conference on Inventive Computation Technologies (ICICT). IEEE*, 2020.

13. Kyamelia Roy et al., "A Comparative study of Lung Cancer detection using supervised neural network", *2019 International Conference on Opto-Electronics and Applied Optics (Optronix). IEEE*, 2019.

14. Özge Günaydin et al., "Comparison of lung cancer detection algorithms", *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). IEEE*, 2019.

15. D. Palani and K. Venkatalakshmi, "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification", *Journal of medical systems*, vol. 43.2, pp. 21, 2019.

16. Hanwu Luo et al., "Logistic regression and random forest for effective imbalanced classification", *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, 2019.

17. P. R. Radhika, Rakhi AS Nair and G. Veena, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms", *2019 IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT). IEEE*, 2019.

18. W. Rahane et al., "Lung Cancer Detection Using Image Processing and Machine Learning HealthCare", *Proc. International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pp. 1-5, 2018.

19. J. Alam, S. Alam and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classification", *Proc. International Conference on Computer Communication Chemical Material and Electronic Engineering*, pp. 1-4, 2018.

20. K. V. Bawane and A. V. Shinde, "Diagnosis Support System for Lung Cancer Detection Using Artificial Intelligence", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 6, no. 1, January 2018.

21. H Sathyan and J.V Panicker, "Lung Nodule Classification Using Deep ConvNets on CT Images", *9th International Conference on Computing Communication and Networking Technologies ICCCNT*, 2018.

22. Muhammad Imran Faisal et al., "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer", *2018 3rd International Conference on Emerging Trends in Engineering Sciences and Technology (ICEEST). IEEE*, 2018.

23. Qing Wu and Wenbing Zhao, "Small-cell lung cancer detection using a supervised machine learning algorithm", *2017 international symposium on computer science and intelligent controls (ISCSIC). IEEE*, 2017.