Lung Care: Advanced Lung Cancer Survival Prediction System

Dr. Rohini Hanchate[1], Vaibhavi Narkhede[2], Sushil Narsale[3], Mahesh Belhekar[4]
Department of Computer Engineering [1,2,3,4]
Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra [1,2,3,4]

Abstract — This report offers a thorough comparative analysis of three prominent machine learning models—Naive Bayes, Gradient Boosting, and Ensemble Learning—in the domain of predicting the severity levels of lung cancer. Through meticulous data curation and preprocessing, a wide array of health parameters and lifestyle factors were incorporated to ensure the robustness of predictive modeling. The report delineates the rigorous methodologies employed in model training and evaluation, encompassing the utilization of diverse performance metrics to assess predictive efficacy comprehensively. By conducting extensive experimentation and comparative analysis, invaluable insights into the predictive capabilities and limitations of each model were garnered. These findings carry profound implications for healthcare professionals, furnishing them with evidence-based insights to facilitate early intervention and personalized treatment planning for patients at risk of lung cancer progression. Ultimately, this study endeavors to elevate clinical decision-making processes, fostering improved patient outcomes and more efficient allocation of healthcare resources in the management of lung cancer.

***Keywords*—**Lung cancer, Prediction, Feature Selection, Ensemble Learning, Voting Classifiers, Naïve Bayes, Random Forest, Gradient Boosting, F1-score.

I. INTRODUCTION

Lung cancer constitutes a major public health challenge globally, exerting a significant burden on healthcare systems and communities due to its high incidence and mortality rates [7]. Despite advancements in treatment modalities, the prognosis for lung cancer patients remains largely dependent on the stage at diagnosis, underscoring the critical importance of early detection and accurate prognostication. In this report, we delve into the realm of predictive analytics within the context of lung cancer severity assessment, employing machine learning algorithms to harness vast datasets encompassing patient demographics, medical histories, environmental exposures, and lifestyle factors [8]. By leveraging these multifaceted data sources, our aim is to develop robust predictive models capable of stratifying patients based on the severity of their lung cancer, thereby enabling clinicians to tailor interventions and treatment plans accordingly.

The utilization of machine learning techniques, including Naive Bayes, Gradient Boosting, and Ensemble Learning, holds immense promise in augmenting traditional clinical approaches to lung cancer management [10]. These models can identify complicated patterns and relationships in large datasets, which makes it easier to find risk factors and predictive biomarkers that indicate the course of a disease. Through this comparative analysis, we seek to elucidate the strengths and limitations of each algorithm in accurately predicting lung cancer severity levels. Moreover, by shedding light on the relative performance of these models, we endeavor to inform clinical decision-making processes, enhance risk stratification strategies, and ultimately, improve patient outcomes in the realm of lung cancer care.

II. PROBLEM DEFINITION

The project aim is to construct and evaluate machine learning algorithms capable of accurately predicting the severity levels of lung cancer. This entails leveraging a comprehensive dataset comprising various patient attributes, clinical markers, and lifestyle indicators. The objective is to enhance the prognosis and treatment planning for lung cancer patients by providing clinicians with reliable predictive tools. The goal is to create models that will help medical practitioners make decisions based on the needs of each patient by utilizing the power of cutting-edge computational approaches. This initiative seeks to contribute to improved patient outcomes, facilitating early detection, personalized treatment strategies, and ultimately, better survival rates for individuals affected by lung cancer.

III. OBJECTIVE

- The primary aim of this project is to develop and assess the effectiveness of machine learning models in predicting the severity levels of lung cancer. By leveraging patient attributes and lifestyle factors, such as age, gender, smoking habits, and environmental exposures, we aim to construct predictive models capable of accurately classifying the severity of lung cancer cases.

- In this endeavor, we plan to employ three distinct machine learning algorithms: Naive Bayes, Gradient Boosting, and Ensemble Learning. Each of these algorithms offers unique advantages and characteristics that can contribute to the predictive accuracy and reliability of the models. By utilizing a diverse set of algorithms, we can comprehensively evaluate their performance and identify the most suitable approach for lung cancer severity prediction.

- To evaluate the performance of the constructed models, we will employ various evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics will provide insights into the models' ability to correctly classify lung cancer severity levels and assess their overall predictive capabilities.
- Following the model construction and evaluation phase, we will conduct a comparative analysis to identify the strengths and weaknesses of each approach. By comparing the predictive performance of Naive Bayes, Gradient Boosting, and Ensemble Learning models, we aim to determine the most effective methodology for lung cancer severity prediction.
- We will also investigate the role that various traits have in predicting the severity of lung cancer. By analyzing feature importance metrics provided by the models, we can gain insights into the factors that contribute most significantly to disease progression and severity.
- The ultimate goal of this project's findings is to advance knowledge about the prognosis of lung cancer and assist healthcare professionals in devising personalized treatment strategies for patients. By developing accurate predictive models and identifying crucial predictive features, we can facilitate early intervention and improve patient outcomes in lung cancer management.

IV. ALGORITHM

Here's the algorithm for our model:

1. Start:
Begin the algorithm.

2. Load Dataset:
Load the dataset containing features and target variable (lung cancer level).

https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

3. Preprocess Data:
Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features.

4. Split Data:
Divide the data set into training data sets and testing data sets.

5. Train Models:
Train three different models: Naive Bayes, Gradient Boosting, and Ensemble Learning, on the training data.

6. Evaluate Models:
To test each model, use the appropriate test metrics on the test data.

7. Compare Models:
Compare the performance of Naive Bayes, Gradient Boosting, and Ensemble Learning based on their evaluation metrics.

8. Select Best Model:
Select the model with the highest performance as the final model.

9. Make Predictions:
Utilize the model(s) you have chosen to make predictions based on new/unknown data.

10. End:
End the algorithm.

This algorithm outlines the process of building, evaluating, and selecting the best predictive model for predicting the level of lung cancer based on given features, information.

V. MATHEMATICAL MODEL

In this project, we have applied three distinct machine learning algorithms to improve prediction accuracy, each with its unique mathematical framework tailored to its specific approach. These mathematical structures define the principles guiding learning and decision-making within each model. A comprehensive understanding of the mathematical foundations of these algorithms facilitates informed model selection and result interpretation, allowing for a holistic evaluation of their performance in predicting lung cancer severity.

Ensemble Learning:

Voting Mechanism -
In the ensemble learning approach, we employ a voting mechanism to aggregate individual predictions from each algorithm, determining the final prediction. Weighting individual predictions based on algorithm performance allows more accurate models to have a greater influence on the final ensemble prediction.

Naïve Bayes:

Naive Bayes, a probabilistic machine learning algorithm based on Bayes' theorem, is utilized for lung cancer survival prediction. Its mathematical model calculates the probability of survival given a set of features, incorporating prior probabilities and conditional probabilities based on observed data.
The mathematical model of Naive Bayes for the lung cancer survival prediction is stated as:

$$P(S|X) = \frac{P(X|S) * P(S)}{P(X)}$$

where,
P(S| X) is the probability of survival given a set of features X.
P(X |S) is the probability of observing features X given survival.

P(S) is the prior probability of survival. P(X) is representing probability of observing features X.

To apply Naive Bayes in predicting lung cancer survival, it's essential to gather a dataset containing information about lung cancer patients with documented survival outcomes. This dataset should encompass pertinent features related to survival, including age, gender, cancer stage, and treatment regimen. After obtaining the dataset, the Naive Bayes model can be trained by computing the conditional probabilities P(X | S) and P(S).

The conditional probabilities are also calculated using the following formula:

Understanding the dataset's characteristics, including its size, feature distribution, and target variable distribution, guides our modeling approach and interpretation of model performance. Exploratory data analysis provides insights into attribute correlations and their impact on lung cancer prediction, making the dataset pivotal in our quest to develop accurate predictive models and gain insights into lung cancer risk factors.

The link of dataset used in the project is - https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link.

$$P(X = x|S = s) = \frac{count(X = X, S = s)}{count(S = s)}$$

VII. SYSTEM ARCHITECTURE

Where, count(X = x, S = s) is the number of patients with feature X = x and survival outcome s.

count(S= s) is the number of patients with survival outcomes.

The prior probability P(S) can be produced using the following formula:

$$P(Survival|s) = \frac{count(Survival = s)}{total\ patients}$$

Where, count (S= s) is number of patients with the survival outcome s. Total patients represents total number of patients in the dataset

.

Gradient Boosting:

Gradient Boosting constructs an ensemble of weak learners sequentially to minimize a loss function, enhancing prediction accuracy through iterative refinement of predictions. It sequentially builds an ensemble of weak learners to minimize a loss function:

Fi(y) =Fi-y(y) +α.hi(y)

Fi(y) = represent the current ensemble prediction after ith iterations.

Fi- y(y) =is prediction from the earlier iteration.

α=named as learning rate.

hi(y) =is prediction from the weak learner that has been newly trained.

Random Forest:

Random Forest, another ensemble learning method, combines multiple decision trees to predict lung cancer patient survival. It mitigates overfitting and handles complex datasets effectively, providing clinicians with interpretable insights for informed decision-making.

VI. DATASET

In our project, the dataset forms the core foundation for developing predictive models for lung cancer. It encompasses various demographic, lifestyle, and health-related attributes such as age, gender, environmental exposures (e.g., air pollution, dust allergy), lifestyle choices (e.g., smoking, alcohol consumption), and symptoms (e.g., coughing of blood, shortness of breath). Preprocessing steps ensure data quality and consistency by handling missing values, encoding categorical variables, and standardizing numerical features.

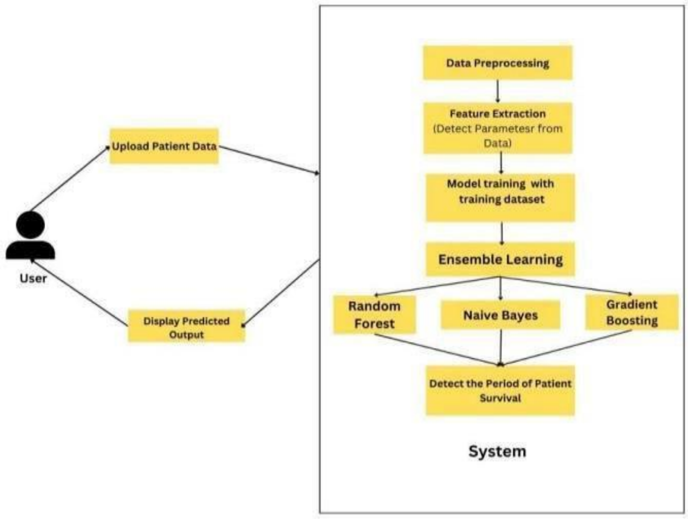


Fig.1 System Architecture

The proposed architecture for the report generation system comprises several interconnected components aimed at efficiently processing input data, detecting similarities, generating comprehensive reports, and presenting the findings to users. At the forefront is the User Interface (UI), designed to facilitate user interaction by providing intuitive forms for inputting data or uploading documents. Upon submission, the Input Processing module extracts and cleans the input data, preparing it for analysis.

Central to the system is the Plagiarism Detection Engine, which employs sophisticated algorithms and machine learning models to analyze text similarities within the input data. This engine compares the input text against a database of known sources to identify instances of similarity or potential plagiarism. The results of this analysis are then utilized by the Report Generation component to construct detailed reports, including information on detected similarities, original sources, and recommendations.

The Output Presentation module is responsible for displaying the generated reports to users through the UI, offering options for downloading or sharing the reports in various formats. Additionally, this module may provide further analysis or actions based on the report findings, empowering users to address any identified issues effectively.

The system ensures secure access and data protection through Authentication and Authorization mechanisms, implementing encryption and other security measures to safeguard user data and system integrity. Integration with external services, such as academic institutions or content providers, enables access to reference materials and verification of sources.

Compliance with legal and ethical standards regarding data privacy, copyright, and intellectual property rights is prioritized, with features implemented for data anonymization and consent management. Scalability and performance are ensured through the use of cloud services and other technologies to accommodate varying loads and maintain optimal performance.

VIII. METHODOLOGY

1. Data Collection:

We acquire datasets containing information regarding lung cancer patients, encompassing their demographic particulars, medical backgrounds, treatment records, and survival outcomes.

2. Data Preprocessing:

The collected data is cleansed and prepared by addressing missing values, encoding categorical variables, and normalizing numerical features, ensuring the dataset's readiness for analysis**.**

3. Feature Extraction:

Relevant features indicative of lung cancer patient survival is identified and extracted from the dataset. Techniques like dimensionality reduction may be utilized to simplify the feature space while retaining essential information.

4. Model Training using Training Dataset:

Ensemble learning methods such as Random Forest, Gradient Boosting, and Naive Bayes are employed for model training. The dataset is divided into training and testing sets, with the former utilized for model training. Hyperparameters are adjusted to enhance model performance.

5. Machine Learning Algorithms:

Random Forest:

A Random Forest classifier is trained, capable of handling intricate datasets with high dimensionality and providing robust predictions.

Gradient Boosting: A Gradient Boosting model is implemented to iteratively improve predictive accuracy, particularly in areas of initial model weakness, capturing complex data relationships effectively.

Naive Bayes: A Naive Bayes classifier is also trained, known for its computational efficiency and suitability for datasets with numerous features, serving as a baseline for comparison with more intricate models.

6. Evaluation Metrics:

Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score computed on the testing dataset. This assessment enables the determination of the models' accuracy in predicting lung cancer patient survival and facilitates comparisons of their efficacy.

7. Experimental Setup:

Experiments are conducted on the lung cancer dataset by partitioning it into training and testing subsets. The models are trained using the training dataset and subsequently evaluated on the separate testing dataset. Furthermore, comparisons are made between the performance of our ensemble models and other advanced techniques for predicting lung cancer patient survival. Proper citation and acknowledgment of relevant sources and methodologies employed in the project are ensured.

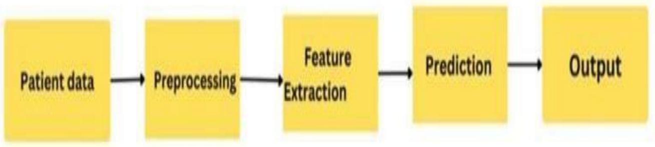


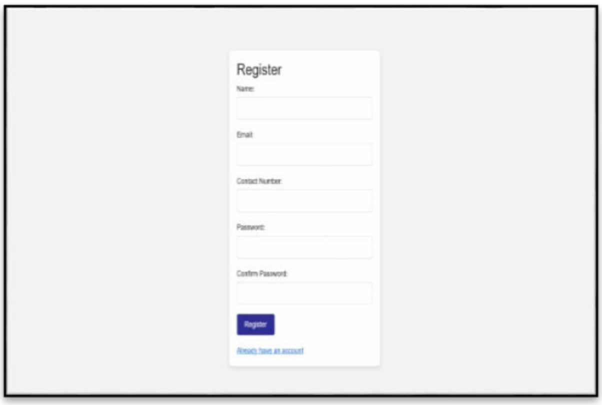
Fig.2 Processing Steps

IX. RESULT

People Care

Your Lung Cancer Survival Predictor

Fig.3 Lung cancer prediction user interface



Register

Name:

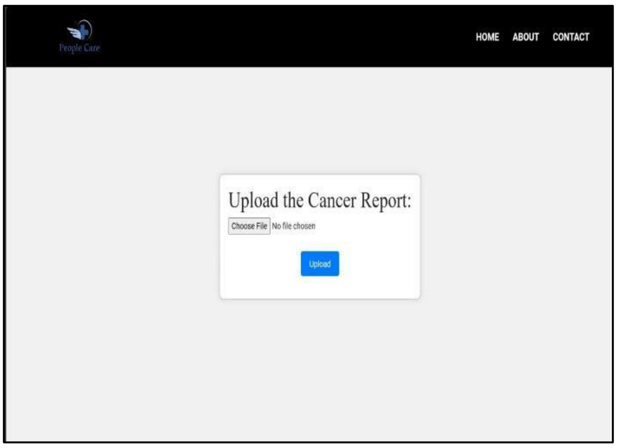
Email:

Contact Number:

Password:

Confirm Password:

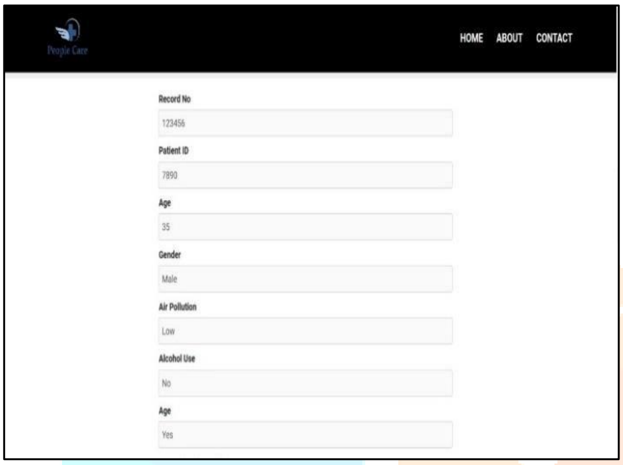
Fig.4 Registration page



Upload the Cancer Report:

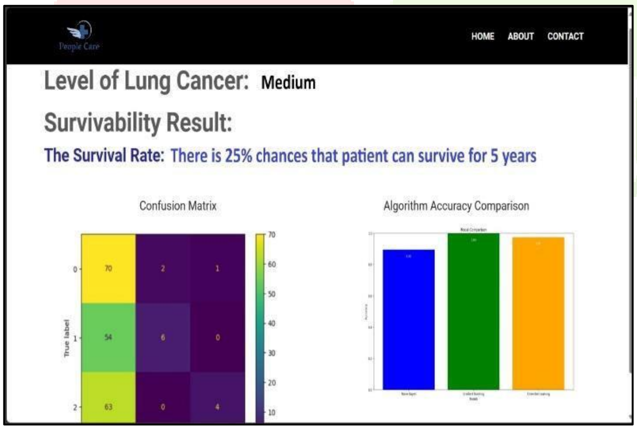
No file chosen

Fig.5 Upload medical report



Record No
123456
Patient ID
7890
Age
35
Gender
Male
Air Pollution
Low
Alcohol Use
No
Age
Yes

Fig.6 Feature extraction from report



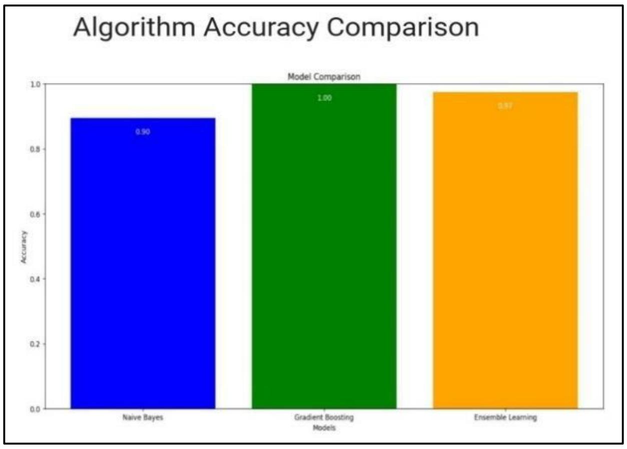
Level of Lung Cancer: Medium

Survivability Result:

The Survival Rate: There is 25% chances that patient can survive for 5 years

Fig.7 Survivability Result

Fig.9 Algorithm Comparison



X. CONCLUSION

In conclusion, the lung cancer survival prediction system represents a significant advancement in the field of oncology care, offering personalized insights into patient prognosis and treatment outcomes. Through the integration of machine learning algorithms and predictive analytics, the system has the potential to revolutionize how healthcare providers approach lung cancer management. By leveraging genomic data analysis, real-time monitoring capabilities, and telemedicine integration, the system can deliver tailored recommendations and support to patients, enhancing accessibility to care and improving treatment adherence. Ongoing optimization of machine learning models and collaborative research initiatives further underscore the system's commitment to advancing lung cancer research and clinical practice. Moreover, by prioritizing data privacy and security measures, the system ensures the confidentiality and integrity of patient information, fostering trust and compliance with regulatory standards. Overall, the lung cancer survival prediction system holds immense promise for improving patient outcomes, driving innovation in oncology care, and ultimately, contributing to the fight against lung cancer.

XI. FUTURE SCOPE

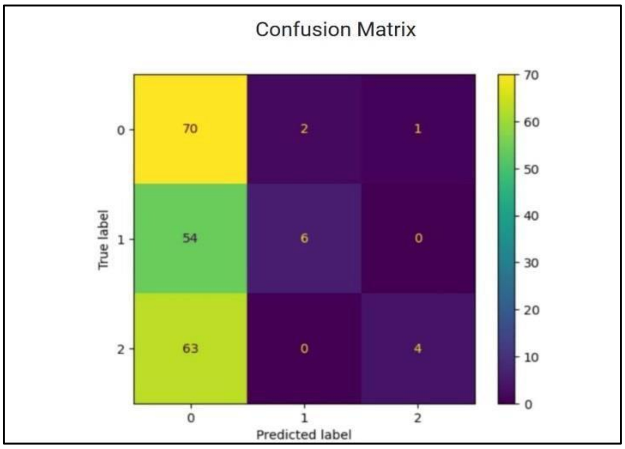


Fig.8 Confusion Matrix

In the domain of lung cancer survival prediction, future developments are poised to significantly enhance system effectiveness and utility. Integration of genomic data analysis offers a promising avenue for deeper insights into genetic influences on outcomes, refining predictions and providing personalized prognosis. Real-time monitoring and feedback mechanisms, coupled with ongoing model optimization, ensure adaptability and accuracy across diverse patient populations and evolving healthcare landscapes.

Additionally, the integration of telemedicine capabilities and emphasis on patient support and education represent critical strides toward accessible, proactive care and improved patient engagement. Collaborative research initiatives and robust data privacy measures further enhance the system's role in advancing lung cancer care, fostering research insights, and ensuring patient trust and compliance.

XII. REFERENCES

[1] Sonia kukreja, Munish Sabharwal, Mohd Asif Shah and D.S. Gill, "A Heuristic Machine Learning-Based Optimization Technique to Predict Lung Cancer Patient Survival", Volume 2023.

[2] Kun-Hsing Yu, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, and Daniel L. Rubin, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features", Nature Communications volume 7, Article number: 12474 (2016).

[3] R. K. Singh and M. Sivabalakrishnan, "Feature selection of gene expression data for cancer classification: a review," Procedia Computer Science, vol. 50, pp. 52–57, 2015.

[4] Tina M. St. John M.D. (2005).:" With Every Breath: A Lung Cancer Guidebook" .1(1):75-82. ISBN 0-9760450-2-8, www.lungcancerguidebook.org.

[5] Hamid KarimKhani Z and et.al. (2015).:" A comparative survey on data mining techniques for breast cancer diagnosis and prediction Survey". Indian Journal of Fundamental and Applied Life Sciences.5 (S1): 4330- 4339 ISSN: 2231– 6345.

[6] V. Krishnaiah et al. (2013).:" Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies. 4 (1):39 – 45.

[7] International Agency for Research on Cancer.
GLOBOCAN Lung Cancer Facts Sheet 2020.

[8] Kaggle, "Lung cancer prediction dataset,"2018, https://www.kaggle.com/datasets/thedevastator/cancerpatients-and-air-pollution-a-new-link

[9] Prof. Pritam Ahire,Akanksha Kale,Kajal Pasalkar,Sneha Gujar,Nikita Gadhave, ECG MONITORING SYSTEM&quot;, International Journal of Creative ResearchThoughts (IJCRT), ISSN:2320-2882, Volume.9, Issue 3, pp.407-412, March 2021,
Available
http://www.ijcrt.org/papers/IJCRT2103052.pdf.

[10] Xibin DONG, Zhiwen YU, Wenming CAO, Yifan SHI, Qianli MA:"A survey on ensemble learning". School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China,2019.

[11] Rohini Hanchate, Jayesh Ramani, Mayur Lalchandani, Ramendra Singh, A System based on Data Mining Techniques for Predicting Heart Diseases , https://www.irjet.net/archives/V7/i6/IRJET-V7I6351.pdf , Volume: 07 Issue: 06 | June

[13] Rajalaxmi R R, Kavithra S, Gothai E, Natesan P, Thamilselvan R,"A Systematic Review Of Lung Cancer Prediction Using Machine Learning Algorithm", 2022 International Conference on Computer Communication and Informatics (ICCCI).

[14] A. Malekloo, E. Ozer, M. AlHamaydeh, and M. Girolami, ''Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights,'' Struct. Health Monitor., vol. 21, no. 4, pp. 1906–1955, Jul. 2022.

[15] IBOMOIYE DOMOR MIENYE, AND YANXIA SUN,"A Survey of Ensemble Learning: Concepts,Algorithms, Applications, and Pros

[16] H. M. Abdul Fattah,K. M. Azharul Hasan, Sunanda Das ,"A Voting Classifier for the Treatment of Employees' Mental Health Disorder", IEEE, Volume:08 September 2021.

[17] Y. Sun, Z. Li, X. Li, and J. Zhang, ''Classifier selection and ensemble model for multi-class imbalance learning in education grants predic- tion,'' Appl. Artif. Intell., vol. 35, no. 4, pp. 290–303, Mar. 2021.

[18] Khushi Kumari Jha, Roshan Jha, Ankita Kumari Jha, Md Al Mahedi Hassan, Saurav Kumar Yadav,"A Brief Comparison On Machine Learning Algorithms Based On Various Applications: A Comprehensive Survey", 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS).

[19] S. V. Joshi and R. D. Kanphade, &quot;Deep Learning Based Person Authentication Using Hand Radiographs: A Forensic Approach, in IEEE Access, vol. 8, pp. 95424- 95434, 2020, 10.1109/ACCESS.2020.2995788.

[20] Joshi, S.V., Kanphade, R.D. (2020). Forensic Approach of Human Identification Using Dual Cross Pattern of Hand Radiographs. In: Abraham, A., Cherukuri, A.,

[21] Melin, P., Gandhi, N. (eds) Intelligent Systems Design and Applications. ISDA 2018, 2018. Advances in Intelligent Systems and Computing, vol 941. Springer, Cham. https://doi.org/10.1007/978-3-030-16660-1_105.

2020
www.irjet.net p-ISSN: 2395-0072 , 2020

[12] Pritam Ahire , Rohini Hanchate Predictive and Descriptive Analysis for Healthcare Data, A Hand book on Intelligent Health Care Analytics Knowledge Engineering with Big Data " https://www.wiley.com/en-us/Handbook+on+Intelligent+Healthcare+Analy
ti
cs%3A+Knowledge+Engineering+with+Big+D
ata
-p-9781119792536 Published by Scrivener Publishing, Wiley Group,2021.