Predictive Analytics for Early Lung Cancer Risk Using Machine Learning

Sumukh Acharya
Dept. of Computer Science Engineering
PES University
Bangalore, India
sumukh.acharya@gmail.com

Atharva Amol Kinage
Dept. of Computer Science Engineering
PES University
Bangalore, India
kinageatharv@gmail.com

Devi Krishikesh Reddy
Dept. of Computer Science Engineering
PES University
Bangalore, India
krishi0106@gmail.com

Desetti Shashank
Dept. of Computer Science Engineering
PES University
Bangalore, India
shashankdesetti@gmail.com

Divya Ebenezer Nathaniel
Assistant Professor
Dept. of Computer Science Engineering
PES University
Bangalore, India
divya.en@pes.edu

Umme Haani
Assistant Professor
Dept. of Computer Science Engineering
PES University
Bangalore, India
umme.haani@pes.edu

***Abstract*—This research utilizes machine learning methods for enhancing the early detection of lung cancer. A detailed synthetic medical dataset that includes records for 22,811 patients, each with 788 health-related parameters, consisting of both numeric and categorical forms. The target variable stands for lung cancer presence or absence. Extensive data filtering was and imputation was done using the Random Forest approach. The dataset was condensed to 89 key features. Feature selection and reduction was performed using Principal Component Analysis (PCA), Brain Storm Optimization (BSO), Recursive Feature Elimination (RFE) and SelectKBest (SelectK). Four machine learning models—XGBoost, Support Vector Machine (SVM), CatBoost, and K-Nearest Neighbors (KNN) were trained on each of transformed datasets after feature selection and reduction. The model's performance was tested on the transformed datasets through 5-fold cross-validation, focusing on accuracy and recall. An ensemble model was built to combine individual model outputs, in an effort to improve overall predictive accuracy and reliability. This study is aimed at determining the best feature selection technique to improve early detection for lung cancer. The results showed that Recursive Feature Elimination (RFE) was the best feature selection algorithm, achieving the highest accuracy of 98.746%, recall of 96.245%, precision of 98.582%, and F1 Score of 97.4% for the ensemble model. The study concluded that the application of using machine learning models in predicting an early risk in lung cancer increases the chances of survival with minimal survival time and minimal costs incurred.**

Keywords—Machine Learning, Early Detection, Lung Cancer, Feature Selection, Cross-Validation, Ensemble Model

I. Introduction

The threat of lung cancer is high all around the world because its early symptoms are not as observable as other types and it usually surfaces at the final stages. Thus, non-invasive, predictive diagnostic methods have been urgently in demand because of the expensive, invasive existing techniques. The following research can be used to show how synthetic datasets are used to overcome issues faced by ML Models while working on real-world data, which are not available due to privacy concerns.

This paper involves developing machine learning models for prediction of early risk of lung cancer by analyzing routine medical data. Different techniques for feature selection have been researched, and also how these methods are used in the ensemble model. Therefore this paper determines which approach improves model accuracy and dependability in medical research. And this will support doctors at early diagnosis of lung cancer and can help in increasing sruvival rates and decreasing costs.

This work strengthens the predictive capability and thus contributes to enhanced decision-making within medical departments especially Oncology department.

II. Related Works

Most recent research in the field of early risk prediction in various diseases like Lung Cancer and Gastric Cancer uses various machine learning models like Extreme Gradient Boosting (XGBoost), K-nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM) to make the predictions. The following studies make substantial contributions to advancing our knowledge in this domain.

M. S. Bhuiya et al. [1] XGBoost, LightGBM, AdaBoost, Logistic Regression, and Support Vector Machine were compared and XGBoost gave the best results. The study involved the following stages: data collection, data preprocessing and training using various machine Learning Algorithms. The dataset used contained 5000 patients with 25 features. This dataset only contained patients with lung related issues, no background patients. Randomized filter was used to to replace the missing data, No feature selection models were used, the existing 25 features were used to train the data.

C.-H. Liu et al. [2] Discusses the effect of feature selection on missing value imputation in medical datasets, where missing data often affects the performance of machine learning algorithms. It compares three feature selection methods information gain (IG), genetic algorithm (GA), and decision tree (DT) combined with imputation techniques like k-nearest neighbor (KNN), multilayer perceptron (MLP), and support vector machine (SVM). Tests on five datasets with different numbers of dimensions, was tested in MCAR conditions, showed that feature selection accompanied by imputation improves the classification performance over just using imputation. GA

and IG favor lower-dimensional datasets, while DT prefers higher dimensions.

N. Jothi et al. [3] Discusses the use of a Genetic Algorithm for feature selection to enhance the classification accuracy of the disease using a Support Vector Machine classifier. It follows the Knowledge Discovery in Databases methodology and stresses the role of feature selection in reducing the complexity of high-dimensional medical data and improving classification performance. For testing, five datasets have been employed - Breast Cancer, Parkinson's, Heart Disease, Statlog (Heart), and Hepatitis-the GA-based approach showed an increase in accuracy when compared to those models with no feature selection at all. The authors also compare other techniques of feature selection such as Particle Swarm Optimization (PSO) and Ant Bee Colony (ABC) along with the classifiers K-Nearest Neighbor (KNN), Naïve Bayes, and Random Forest. The GA-SVM method promises to deliver some good results for aiding medical diagnosis, but this study also recognizes its computational intensity and susceptibility to local optima.

P. Saha et al. [4] Some work focused on improving the accuracy of the classification models through efficient methods for feature selection. Some adopted the hybrid filter-wrapper approach on document clustering while others used hybrids on supervised classifications based on different approaches, and is ranked based on the Laplacian Score. Some have divided their process into two phases: they rank attributes through some criteria followed by the determination of the best subset of these ranked attributes. Another model uses mutual information to rank features and classes, then applies the Shapley value to measure the contribution of features. For the diagnosis of the genetic variant of oligodendroglioma, a hybrid filter-wrapper method was used. Hybrid models have been developed to enhance accuracy and performance in prediction.

P. Theerthagiri et al. [5] Many studies have used machine learning for cardiovascular disease (CVD) prediction. One study proposed a recursive feature elimination-based gradient boosting (RFE-GB) algorithm for accurate heart disease prediction. This algorithm achieved high accuracy (89.7%) and an AUC value of 0.84, outperforming other techniques. RFE-GB makes use of recursive feature elimination. Here, it ranks the features and removes the lowest ranked one in each round. These are then used by the technique of Gradient Boosting to classify the CVD patients. Other studies have used techniques such as extreme gradient boosting.

A. Chen et al. [6] Uses 5 datasets, each of them include 30k, 60k, 90k, 120k, 150k patients. XGBoost is trained on this data using 10 fold cross-validation. Repeated training with more patient data gradually improves the model accuracy. The dataset used is a synthetic dataset, generated by Synthea, an open-source, synthetic patient generator which models the medical history of synthetic patients. It was not comapared with various available and well-known feature selection algorithms.

E. Dritsas et al. [7] Through Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology and the RapidMiner software, different models and sampling methods were built. The dataset contained 309 patients and 9 features, most of which were lifestyle patterns.Since there was no missing values or any outliers, no preprocessing was done. Feature ranking was done using Gain Ratio and Random Forest. Those features that have little or no importance were still used to train the data. Synthetic Minority Over-sampling Technique (SMOTE) was used to tackle highly skewed class distribution of the participants among the Lung Cancer (87.4%) and Non-Lung Cancer classes, after which the classes' distribution was 50-50, which is worse than real life patient data for whom lung cancer is less likely. Rotation Forest (RotF) emerged as the most accurate model. XGBoost wasn't used in comparison.

M. R. Afrash et al. [8] Uses a dataset of 2029 patients. Relief feature selection algorithm was used to select 11 features, all of which were lifestyle patterns to train the ML models. XGBoost emerged as the most accurate model with the best metrics.

A. Pfob et al. [9] Pfob et al. 2022 proposed A step-by-step guide for developing machine learning models in medical applications using open-source tools, such as a mammography dataset that is used to classify breast masses. Five algorithms had statistically equivalent performance: logistic regression, Extreme Gradient Boosting (XGBoost), Multivariate Adaptive Regression Splines (MARS), Support Vector Machine (SVM), and a neural network with Area Under the Receiver Operating Characteristic (AUROC) values between 0.88 and 0.89. The authors stress that these techniques are crucial for generalizability and reproducibility in medical ML research, emphasizing the need for data preparation and statistical testing in model evaluation.

A. B. H. Prof. D. Sharma et al. [10] The work presents a method for lung cancer detection using an ensemble model with Decision Tree, Random Forest, and Artificial Neural Network classifiers. It is believed this work will enhance the accuracy of the diagnosis and will help in early detection of lung cancer by combining the strengths of each model. The ensemble model was trained with a dataset consisting of 309 patient's symptoms and lifestyle characteristics. The accuracy of the ensemble model stood at 98.37 % in comparison with the individual classifiers. The study states that this ensemble learning potential can improve the accuracy of the machine.

M. K. Gould et al. [11] Developed an XGBoost-based model to predict non-small cell lung cancer (NSCLC) 9–12 months before diagnosis using routine clinical and lab data. The dataset included 6,505 NSCLC cases and 189,597 controls. The model outperformed the modified Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Model 2012 (PLCOm2012 model), achieving an AUC of 0.86 vs. 0.79. This approach promises earlier detection of lung cancer and better screening strategies.

S. P. Maurya et al. [12] Compared 12 machine learning algorithms on a dataset of 310 instances with 16 features, including symptoms and habits. K-Nearest Neighborwas the best model with an accuracy of 92.86% and Bernoulli Naïve Bayes ranked second with an accuracy of 91.07%. Correlation and classification metrics were used for analysis of the dataset. The study depicts the strength of machine learning for early lung cancer detection and individual treatment strategies.

P.-L. Benveniste et al. [13] This study presents an XGBoost machine learning model to predict the risk of lung cancer within five years based on the Prostate, Lung,

Colorectal, and Ovarian (PLCO) dataset, validated on the NLST dataset. The model was able to achieve a Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) of 82% on the Prostate, Lung, Colorectal, and Ovarian (PLCO) dataset and 70% on the National Lung Screening Trial (NLST) dataset. The comparison with United States Preventive Services Task Force (USPSTF) guidelines finds the same recall but higher precision so the model demonstrates enhanced screening efficacy. The study also emphasizes that Area Under the Curve - Precision Recall (AUC-PR) is more important as a metric in an imbalanced dataset, such as lung cancer screening, where correct identification of positives is the most important thing.

N. Pudjihartono et al. [14] This review covers feature selection methods for machine learning-based disease risk prediction using genetic data. Model creation is complicated because of "curse of dimensionality," where features are more than the samples. Methods are categorized into filters, wrappers, and embedded methods. Filter methods are based on the statistical tests in ranking features. Wrapper methods depend on the classifier performance, and embedded methods, on the integration of selection within the algorithm. Hybrid approaches combine several methods, like using filters for feature reduction followed by the wrapper approach. Whereas ensemble methods make use of combining outputs from various algorithms. According to the authors, there is no one-size-fits-all optimum method, and the two-stage or hybrid approach is best practice.

Savannah L. Bergquist et al. [15] This paper addresses one of the essential challenges which is predicting the severity level of lung cancer based on healthcare claimed data and those that are critical in oncology-related research. An ensemble machine-learning tool is constructed to classify cancer patients receiving chemotherapy into early- versus late-stage categories, supported by augmented administration claims data classification rules. The authors realized that their ensemble machine learning algorithm, which uses a median based classification rule, performed better than decision tree, producing 93% accuracy, 92% specificity and 93% sensitivity. The study utilizes data from the Surveillance, Epidemiology and End Results (SEER) cancer registry program and Medicare claims. The results shows the possibility of using administrative data in measuring the quality and outcomes of cancer care and improving risk adjustment methods. The super learner prediction function pooled five algorithms, and the most substantial increments in performance were achieved from the addition of extra claims variables.

III. Methodology Of The Proposed System

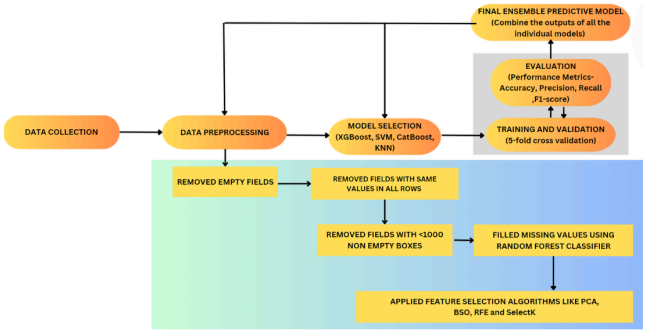

Fig. 1. Design Diagram

Fig.1, the design diagram outlines the process of working on the Synthetic Lung Cancer dataset, including data preprocessing, model selection, training and validation, evaluation and ensemble model building. Here is a step-by-step description of the flow:

A. Data Collection

The data collection process was accomplished by obtaining a dataset from Harvard Dataverse that contains Synthea synthetic patient data. It includes records of 22,811 patients along with 788 related heatlh parameters. The health parameter codes were mapped with their respective name and the preprocessing was carried out.

B. Data Preprocessing

The system starts with a strong data preprocessing pipeline that guarantees high-quality input for the predictive models. It involves an advanced filtering mechanism that dynamically assesses each column, eliminating columns that are either empty, uniform values that cause redundancy, or have sparse data which means entries of less than 1000 to maintain statistical significance. To prevent potential bias and ensure patient privacy, the "ptnum" column, typically used for patient identification, is excluded. Random Forest algorithm is used instead of the simple mean or median method to impute missing values. This captures all relationships between variables with better accuracy.

C. Feature Selection

The feature selection module employs a multi-faceted approach to identify the most relevant predictors:

Principal Component Analysis (PCA): Principal Componet Analysis (PCA) is a dimensionality reduction technique that was used to transform the original features into a new set of orthogonal features called principal components. These components capture the maximum variance in the data.

Brain Storm Optimization (BSO): Brain Storm Optimization (BSO) is a metaheuristic optimization algorithm inspired by the brainstorming process. A population of candidate solutions is generated (feature subsets) and then performance is evaluated using a classifier. The algorithm iteratively generates new candidate solutions by combining and modifying existing ones, then selects the best-performing candidates to form the next population. This process of brainstorming and selection continues until convergence or a maximum number of iterations is reached.

Recursive Feature Elimination (RFE): Recursive Feature Elimination (RFE) is a feature selection technique that we used to recursively remove the least important feature based on the model's performance. It starts with all features and iteratively removes the least significant feature until the desired number of features is reached. This process is fine-tuned to balance model complexity and performance.

SelectKBest: SelectKBest is a univariate feature selection method that selects the top K features based on univariate statistical tests. This method uses the 'f_classif' scoring function, which is typically used for classification tasks, and calculates the Analysis of Variance (ANOVA) F-Value between the feature and target variables providing a measure of the linear dependency between the feature

variable and the target variable. This is used to rank the features by their relevance to the target variable.

D. Model Training

The system leverages a diverse set of machine learning algorithms, each chosen for its unique capabilities

1. XGBoost (Extreme Gradient Boosting): XG Boost is known for it's speed and performance. It's implementation is fine-tuned to handle the complexities of medical data, with custom regularization to prevent overfitting.

2. Support Vector Machine (SVM): A kernelized SVM is used to capture nonlinear relationships in the data, with careful parameter tuning to balance the bias-variance tradeoff.

3. CatBoost: This algorithm excels in handling categorical variables which are common in medical datasets. The implementation includes automatic feature combination generation to capture complex interactions.

4. K-Nearest-Neighbours: The KNN algorithm is optimized with a dynamic weighing scheme that adjusts based on the density of data points in the feature space.

E. Cross-Validation and Ensemble Learning

To ensure robust performance estimates and leverage the strengths of multiple models:

Stratified 5-Fold Cross-Validation: Stratified sampling is used to maintain class distribution across folds, crucial for imbalanced medical datasets. This allowed the model to generalize new data more accurately and makes the evaluation unbiased.

Ensemble Model Integration: The ensemble method used more complex techniques than just relying on majority voting. To implement this, a bagging approach is used where each base model (XGBoost, CatBoost, SVM and KNN) is wrapped in a BaggingClassifier with 10 estimators. This technique involves training multiple instances of each base model on different subsets of the training data to reduce variance and improve stability. The predictions from these bagging classifiers are then combined to produce the final ensemble prediction.

F. Evaluation and Reporting

The evaluation module provides comprehensive insights into model performance.

Metric Suite: For the individual models, the metrics Accuracy and Recall were calculated. The plotted line graphs for accuracy and recall compare the performance of four machine learning models—XGBoost, SVM, CatBoost, and KNN—for different feature selection algorithms, namely PCA, BSO, RFE, and SelectK. Other than calculating the metrics, ROC curve was plotted which illustrates the performance of the ensemble model by plotting True Positive Rate (TPR) values against False Positive (FPR) values at different threshold levels. The area under the curve (AUC) is calculated and displayed.

IV. Results And Discussion

In this study, 4 different models were trained on each of the transformed datasets and evaluated the performance of each model using 5-fold cross-validation, recording accuracy, and recall metrics. The following are the results:

TABLE I. Comparison Of Different Models For Different Feature Selection Algorithms

Feature Selection Algorithms	Models							
	XgBoost		SVM		CatBoost		KNN	
PCA	98.1	94.4	97.8	94.1	97.9	94.1	94.7	83.29
BSO	95.9	90.7	93.7	84.6	96.0	90.4	91.6	78.7
RFE	98.7	96.3	98.4	94.9	98.7	96.1	97.7	93.87
SelectK	98.6	95.8	98.2	94.2	98.7	96.0	97.1	92.7

Graphs were plotted to visualize and compare the accuracy and recall values. This was done to understand the difference between accuracy and recall values obtained for the various feature selection algorithms. The following are the graphs:

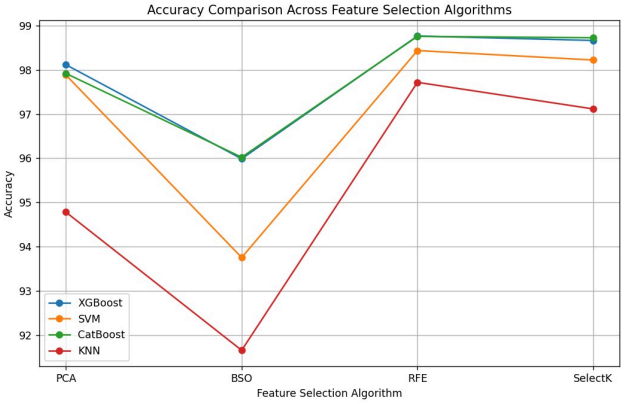

Fig. 2. Accuracy comparison across feature selection algorithms

Fig.2 illustrates the comparison of the accuracy across the different feature selection algorithms for all the various ML models.

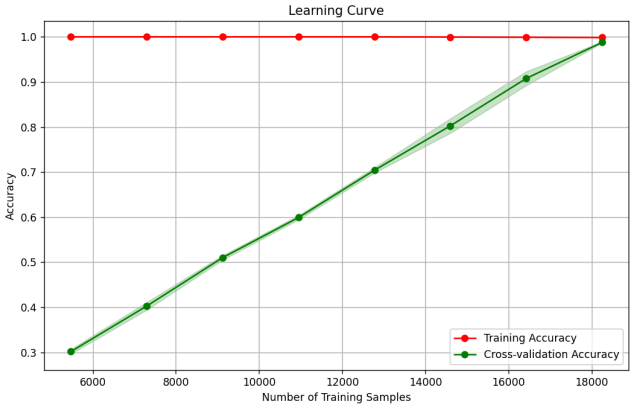

Fig. 2.1. Converging Learning Curve, reassures us of minimal overfitting

Fig.2.1 illustrates the convergence of the learning curve where it is seen that as the number of training samples increases the cross validation accuracy and the training accuracy converges, which indicated minimal overfitting.

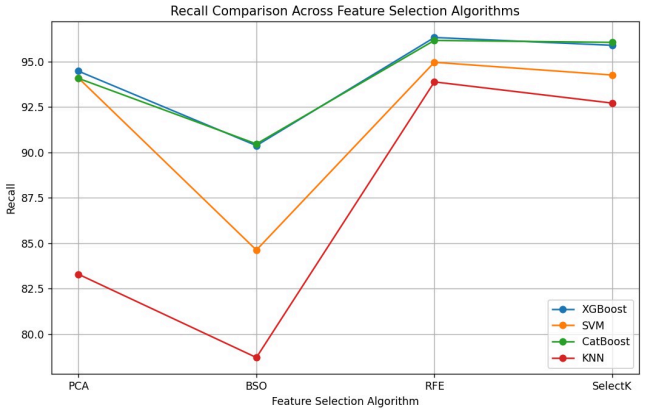

Fig. 3. Recall comparison across feature selection algorithms

Fig.3 illustrates the comparison of the accuracy across the different feature selection algorithms for all the various ML models

The output of the individual models were combined to form an ensemble model. The predictions from these bagging classifiers are then combined using majority voting to produce the final ensemble prediction. The following are the results:

TABLE II. ENSEMBLE MODEL ANALYSIS FOR FEATURE SELECTION ALGORITHMS

Feature Selection Algorithm	Ensemble Model			
	Accuracy	Precision	F1-Score	Recall
PCA	98.15	97.806	96.144	94.538
BSO	95.857	91.91	91.47	91.034
RFE	98.746	98.582	97.4	96.245
SelectK	98.654	98.309	97.211	96.137

V. CONCLUSION AND FUTURE WORK

The results depicted that RFE was the strongest feature selection algorithm, which yielded an accuracy of 98.746%, recall of 96.245%, precision of 98.582%, and F1-Score of 97.4% in the ensemble model. An ensemble model combines the individual models, improving the accuracy and reliability. The project concluded that machine learning models can significantly improve lung cancer survival rates while reducing medical costs.

Future work will involve exploring deep learning models, such as LSTM, Tab-R, and training the detected model for lung cancer detection. We shall also validate synthetic features against real-world Electronic Medical Records (EMR) data. The model will be further translated into a public-friendly access website, in order to easily detect early stages of lung cancer, with perfect error-free as well as usability.

REFERENCES

[1] M. S. Bhuiyan, I. K. Chowdhury, M. Haider, A. H. Jisan, R. M. Jewel, R. Shahid, and M. Z. Ferdus, "Advancements in Early Detection of Lung Cancer in Public Health: A Comprehensive Study Utilizing Machine Learning Algorithms and Predictive Models," Journal of Computer Science and Technology Studies, vol. 6, no. 1, pp. 113-121, Jan 2024. doi: 10.32996/jcsts

[2] C.-H. Liu, C.-F. Tsai, K.-L. Sue, and M.-W. Huang, "The Feature Selection Effect on Missing Value Imputation of Medical Datasets," Applied Sciences, vol. 10, no. 7, pp. 1-15, March 2020. doi: 10.3390/app10072344

[3] N. Jothi, W. Husain, N. A. Abdul Rashid, and S. M. Syed-Mohamad, "Feature Selection Method using Genetic Algorithm for Medical Dataset," International Journal on Advanced Science Engineering Information Technology, vol. 9, no. 6, Dec 2019. ISSN: 2088-5334

[4] P. Saha, S. Patikar, and S. Neogy, "A Correlation - Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), pp. 69-72, Dec 2020. doi:10.1109/GUCON48875.2020.9231179

[5] P. Theerthagiri, and V. J, "Cardiovascular Disease Prediction using Recursive Feature Elimination and Gradient Boosting Classification Techniques," Jun 2021. doi:10.48550/arXiv.2106.08889

[6] A. Chen, and D. O. Chen, "Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data," Scientific Reports, vol. 12, no. 1, p. 17917, Oct 2022. doi: 10.1038/s41598-022-23011-4

[7] E. Dritsas, and M. Trigka, "Lung Cancer Risk Prediction with Machine Learning Models," Big Data Cogn. Comput., vol. 6, no. 139, Nov. 2022. doi: 10.3390/bdcc6040139

[8] M. R. Afrash, M. Shafiee, and H. Kazemi-Arpanahi, "Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors," BMC Gastroenterol., vol. 23, no. 6, 2023. doi: 10.1186/s12876-022-02626-x

[9] A. Pfob, S.-C. Lu, and C. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison," BMC Med. Res. Methodol., vol. 22, no. 282, 2022, doi: 10.1186/s12874-022-01758-8

[10] A. B. H. Prof. D. Sharma, C. H. N., S. Bhat, B. V. A. Suhana, A. Raj, and G. Ashok, "Enhancing Lung Cancer Early Detection: A Hybrid Ensemble Model," J. Electrical Systems, vol. 20-10s, 2024, pp. 2595-2602

[11] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, "Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data," Am. J. Respir. Crit. Care Med., vol. 204, no. 4, pp. 445-453, Aug. 2021, doi: 10.1164/rccm.202007-2791OC

[12] S. P. Maurya, P. S. Ghodia, R. Mishra, and D. P. Singh, "Performance of machine learning algorithms for lung cancer prediction: a comparative approach," Sci. Rep., 2023, doi: 10.1038/s41598-024-46514-5

[13] P.-L. Benveniste, J. Alberge, L. Xing, and J.-E. Bibault, "Development and external validation of a lung cancer risk estimation tool using gradient-boosting," arXiv:2308.12188v1 [cs.LG], Aug. 23, 2023

[14] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," Front. Bioinform., vol. 2, Article 927312, Jun. 2022, doi: 10.3389/fbinf.2022.927312

[15] Savannah L. Bergquist, Gabriel A. Brooks, Nancy L. Keating, Mary Beth Landrum, and Sherri Rose, "Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data," Proc. Mach. Learn. Res., vol. 68, pp. 25–38, Aug. 2017