Non-Invasive Detection of Lung Cancer: Integrating Genomic Data with Ensemble Learning for Improved Early Diagnosis

Dr. Shyamal S. Virnodkar[1], Purvi Shah[2], Shreya Shinde[2], Dr. Sangita B. Nemade3, Varsha P. Gaikwad[4], Naman Dagade[5]

[1]*Associate Professor, Department of Computer Engineering, K. J. Somaiya Institute of Technology, India.*
[2]*Bachelor of Technology, Department of Computer Engineering, K.J. Somaiya Institute of Technology, India.*
[3]*Assistant Professor, Computer Engineering Department, Government College of Engineering and Research, Avasari, Pune, India.*
[4]*Assistant Professor, Information Technology Department, Government College of Engineering, Aurangabad, Maharashtra, India.*
[5]*Bachelor of Technology, Department of Computer Engineering, K.J. Somaiya Institute of Technology, India.*
Email: shyamal@somaiya.edu,

KEYWORDS
Deep Learning, Deoxyribonucleic Acid, Machine Learning (ML), Ribonucleic Acid, Lung Cancer Detection.

ABSTRACT
The early and accurate diagnosis of lung cancer, a significant contributor to global cancer-related mortality, remains a paramount challenge in healthcare. Conventional diagnostic methods often lack the sensitivity required for early-stage detection, prompting the exploration of non-invasive alternatives. Leveraging advancements in genomics and bioinformatics, this study investigates the potential of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) analysis for early-stage lung cancer diagnosis. Two distinct datasets are utilized: GSE4115, comprising gene expression data from bronchial airway epithelial cells of smokers, and GSE33356, focusing on genomic alterations in Taiwanese female non-smoking lung cancer patients. The research employs comprehensive data pre-processing and feature reduction techniques, including normalization and Kernel Principal Component Analysis (KPCA). Subsequently, an ensemble of diverse learners, including Random Forests, AdaBoost, Bagging, Support Vector Machines (SVMs), and Neural Networks, is trained on the original datasets. A novel ensemble stacking approach is proposed, wherein initial predictions from the base learners are combined through logistic regression - the meta-learner to enhance predictive performance. The study aims to contribute to advancements in lung cancer detection by providing a more precise and non-invasive diagnostic method. By integrating DNA and RNA analysis with ensemble learning techniques, the research endeavours to enhance medical outcomes and potentially save lives through early-stage lung cancer detection.

1. Introduction

The identification of cancer continues to be a continuous challenge in the field of healthcare, where early and precise diagnosis is essential for effective treatment and better patient results [1]. Lung cancer is a major concern for global health, since it is responsible for the second greatest number of cancer-related deaths globally, according to the World Health Organization (WHO). Conventional techniques for identifying lung cancer typically depend on invasive procedures such as biopsies or chest X-rays. Although these approaches offer valuable data, they may not be appropriate for detecting early-stage conditions due to their limited sensitivity or patients' hesitancy to undertake invasive treatments.

Advancements in genomics and bioinformatics have recently provided new opportunities for detecting cancer without intrusive procedures, by analyzing DNA and RNA information. These datasets can identify fundamental molecular changes linked to the growth and advancement of cancer. DNA, the fundamental genetic code, can contain mutations that interfere with regular cellular functions and result in unregulated cell proliferation, a characteristic feature of cancer [2]. RNA, which serves as a bridge between DNA and proteins, provides insight into the condition of cells and can uncover alterations in the patterns of gene expression that are linked to the progression of cancer [3]. Through the examination of DNA and RNA mutations, scientists can identify lung cancer in its early stages, enabling more efficient therapy and enhancing the outlook for patients [4]. This study investigates the possibility of using DNA and RNA information to diagnose lung cancer without the need for intrusive procedures. Our proposal involves utilizing an ensemble stacking technique to capitalize on the advantages of various ML techniques. Ensemble approaches combine the predictions of many learning models to attain higher accuracy and resilience in comparison to individual algorithms [5].

This study focuses on training a diverse ensemble of learners directly on the original DNA and RNA datasets. These learners will include Random Forests (RFs), AdaBoost, Bagging, Support Vector Machines (SVMs), and

Neural Networks [6]. These algorithms incorporate several learning paradigms, each possessing its strengths and weaknesses. Random Forests exhibit sensitivity to overfitting and effectively manage high-dimensional data. AdaBoost employs an iterative process to retrain the model, with a specific emphasis on examples that were previously categorized incorrectly [7]. Bagging is a technique that generates numerous models by using random portions of the data, which helps to decrease variance. Support Vector Machines (SVMs) are highly effective in identifying the best possible hyperplanes for categorization in spaces with a large number of dimensions. Neural networks, which draw inspiration from the human brain, can acquire knowledge about intricate non-linear connections between characteristics and results [8]. Our theory is that by combining the predictions of diverse algorithms using ensemble stacking with a radial basis function (RBF) kernel, we can obtain higher accuracy and robustness in detecting lung cancer compared to using a single learning model [9]. The RBF kernel is commonly used for non-linear data, enabling the model to effectively capture intricate connections between DNA and RNA characteristics and the stage of lung cancer [10], [11].

The objective of this research is to make a valuable contribution to the continuous endeavour of enhancing approaches for detecting lung cancer. Through the utilization of DNA and RNA datasets and the capabilities of ensemble learning, our objective is to provide a more precise and non-invasive method for diagnosing lung cancer at an early stage. Consequently, this could result in enhanced medical outcomes and potentially save lives.

2. Literature Survey

DNA

Lung cancer is still one of the most frequently encountered devastating kinds of cancer worldwide. Revolutionary progress has been made in the field of cancer diagnosis, specifically concerning lung cancer, by combining DNA-based techniques with machine learning (ML) algorithms. Through ensemble learning, multiple classifiers can be trained on bootstrap samples, mitigating overfitting. At the start of the 2000s, M. P. S. Brown et al. [12] made a big discovery in the field of cancer study. The authors used SVM to sort microarray data and were able to tell the difference between cancer cells and normal cells with over 90% accuracy. The author's SVM model did better than other methods, showing that the SVM is good at figuring out complicated patterns in high-dimensional microarray datasets. Researchers Rongjun Chen [6] developed an SVM model for classifying lung cancer patients based on ctDNA methylation profiles. In the 2021 paper by D. Mathios et al. [13] the authors employed an SVM classifier to analyze a dataset from 365 individuals. To distinguish between small cell and non-small cell lung cancer. The 2021 study by Ferid Ben Ali et al., [14] explored machine-learning algorithms for classifying lung cancer types using DNA microarray data. Notably, Support Vector Machines (SVM) and Deep Neural Networks (DNN) exhibited exceptional performance Building on this foundation in 2008, P. P. Anglim et al. [15] employed the MethyLight technique to investigate DNA methylation in squamous cell lung cancer (SCLC) and adjacent non-tumor lung tissues, identifying 22 significantly hypermethylated loci. Utilizing a random forest classifier, they developed a panel of eight highly significant hypermethylated markers (GDNF, MTHFR, OPCML, TNFRSF25, TCF21, PAX8, PTPRN2, PITX2) to discriminate SCLC from non-tumor tissues. In the 2013 paper by A. A. Ponomaryova et al. [16] the potential of circulating DNA methylation, particularly in genes RARB2 and RASSF1A, as a diagnostic biomarker for lung cancer is explored. The use of additional algorithms such as PCA, Monte Carlo test, and MANOVA supported these findings.

J. Cabrera [17] In 2015 the Lung Cancer Classification Tool (LCCT), was a machine learning system using Support Vector Machines (SVM) to classify lung cancer subtypes from gene expression data. Quantile Normalization was found to be the most effective pre-processing method, achieving a high accuracy of 93.389%, aiming to improve lung cancer diagnosis and treatment selection. S. Wu et al. [18] 2018 proposed a gene selection method for cancer classification using L1/2 regularization in sparse logistic regression. A. A. ABRO et al. [19] 2020 introduced a stacking-based ensemble learning method for outlier detection, incorporating Rotation Forest, Random Forest, Bagging, Boosting, and Logistic Regression as base learners, along with a Meta-learner logistic regression. This research suggests the potential for hybridizing ensemble learning methods for continued performance enhancement in outlier detection tasks.

In their 2020 study, the authors Q. Wang et al. [20] introduced a method that combines random forest ensemble learning with a self-paced bootstrap sampling strategy. F. Ben Ali et al. [14] 2022 compared machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), k-nearest Neighbours (kNN), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN), for lung cancer type classification

using DNA microarray data. The study utilizes Principal Component Analysis (PCA) for dimensionality reduction and addresses dataset imbalance with SMOTE. In their 2023 study, authors K. Mary Sudha Rani and [11] utilized gene expression data from 192 smoking subjects to develop a lung cancer identification model. Employing Data Pre-processing, Feature Reduction via Kernel Principal Component Analysis, and Ensemble Learning with Random Forest and AdaBoost classifiers. The approach, led by N. Yao et al. [21] 2023 used a hybrid feature selection method combining Pearson's Correlation with univariate filters or recursive feature elimination, along with ensemble classifiers like Stochastic Gradient Boosting, Random Forest, and Support Vector Machine showcased promise for early lung cancer diagnosis and further exploration in bioinformatics methodologies.

In 2015 Z. Cai et al. [22] employed methods like Multi-category Receiver Operating Characteristic (Multi-ROC), Random Forests (RFs), and Maximum Relevance and Minimum Redundancy (mRMR) for feature selection, enhancing the discovery of accurate biomarkers. By combining ensemble-based approaches with Incremental Feature Selection (IFS), the research achieves a robust classification model capable of accurately distinguishing between lung adenocarcinoma, squamous cell lung cancer, and small cell lung cancer. The paper, authored by H. Hijazi et al. [23] 2012 utilizes ensemble learning to classify cancer subtypes from gene expression data, employing a mix of classifiers like decision trees, neural networks, and Random Forests, which combines bagging with the random selection of features to form an ensemble of decision trees, thus improving accuracy and identifying potential biomarkers. Authored by Mai Abdulla and Mohammad T. Khasawneh in 2020 introduced G-Forest, a novel ensemble classifier combining Genetic Algorithm with Random Forest to select informative features with low cost, outperforming state-of-the-art algorithms in accuracy and cost-effectiveness. The paper by S. Wang et al. [24] 2023 presents a methodology employing machine learning algorithms including a Generalised Linear Model (GLM), Gradient Boosting Machine (GBM), Random Forest, Deep Learning, and eXtreme Gradient Boosting (XGBoost) to construct the base and stacked ensemble models for lung cancer detection using cell-free DNA (cfDNA) fragment omic features.

RNA

In their respective studies, Huang et al. [25], Sherafatian et al. [26], A. Rahaman et al. [27], and Fei Yuan et al. [28] delved into the intricate landscape of lung cancer using machine learning methodologies, thereby contributing significantly to oncological research. Huang et al. scrutinized single-cell RNA sequencing data to discern cell markers pertinent to non-small cell lung cancer (NSCLC). Employing meticulous feature selection processes, including incremental feature selection (IFS), they pinpointed genes crucial for distinguishing cell subtypes within NSCLC samples. Similarly, Sherafatian et al. harnessed machine learning algorithms to unearth biomarkers for lung cancer diagnosis and subtyping, leveraging miRNA expression data from The Cancer Genome Atlas (TCGA). Their application of decision tree learning algorithms unveiled four miRNAs with diagnostic and subtype-discriminatory potential, offering insights for enhanced lung cancer diagnosis and classification.

In a subsequent study, A. Rahaman et al. [27] ventured into the analysis of RNA-Seq data to predict novel biomarkers for both small cell lung cancer (SCLC) and NSCLC. By employing machine learning techniques, they successfully annotated genes associated with lung cancer pathways, addressing data imbalance issues along the way. Their Random Forest model exhibited notable accuracy in classifying genes. Concurrently, Fei Yuan, Lin Lu, and Quan Zou delved into the transcriptomic terrain of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), leveraging Monte Carlo feature selection and support vector machine classifiers.

Furthermore, in the realm of cancer research, innovative methodologies continue to emerge to enhance diagnostic accuracy and therapeutic efficacy. Charkiewicz et al. [5] embarked on a groundbreaking exploration into non-invasive biomarkers for non-small cell lung cancer (NSCLC) through serum miRNA profiling. Leveraging next-generation sequencing techniques, they identified a panel of 28 upregulated miRNAs in early-stage NSCLC patients, culminating in the development of a robust gradient-boosting decision tree classifier. Meanwhile, Mohammed et al. [29] introduced a novel stacking ensemble deep learning model tailored for cancer-type classification using RNA-Seq data. By employing multiple one-dimensional convolutional neural networks (1D-CNNs) as base models and fusing their predictions through a meta-model neural network, their approach showcased superior accuracy compared to traditional machine learning algorithms, underscoring its potential as a robust tool for cancer classification.

3. Datasets

The datasets used in this study provide comprehensive insights into the molecular landscape of lung cancer, spanning a variety of populations and experimental methodologies. The first dataset, denoted as GDS2771 and referred to as GSE4115, includes gene expression data from 192 smokers, including those who have been pre-diagnosed (90), have had a lung cancer diagnosis (97), and are at high risk of getting cancer (5). This dataset is available through the Gene Expression Omnibus (GEO) database of the National Cancer Institute (NCBI), and it offers a useful tool for examining the underlying genetic anomalies linked to the development of lung cancer in smokers. Large airway epithelial cells are accessible, providing a less invasive option to standard lung biopsy techniques. Regular airway screenings may increase patient assessment and enable earlier detection. When these cells are directly exposed to inhaled carcinogens, it makes them vulnerable to early DNA alterations. This makes it easier to detect lung cancer early on when treatment efficacy is highest. Furthermore, field cancerization of the entire airway epithelium caused by cigarette smoke offers a chance to find more comprehensive biomarkers for detection as well as early carcinogenic processes. The feasibility of early detection and extensive screening programs, especially for high-risk groups like smokers, is further highlighted by the low cost of analyzing airway epithelial cells.

The second dataset focuses on genomic alterations and transcriptional modulation in Taiwanese female non-smoking lung cancer patients. It is named "Genome-wide screening of genomic alterations and transcriptional modulation in non-smoking female lung cancer in Taiwan" (GSE33356). SNP genotyping by SNP array, expression profiling by array, and genome variation profiling by SNP array are all included in this dataset. Originating from frozen tissue samples of the main tumors and surrounding normal tissue, it provides important new information about the genetic anomalies and transcriptional alterations linked to lung cancer in female non-smokers. This population is less well-researched than cases related to smoking. The analysis of gene expression patterns made possible by the integration of expression profiling by array offers a deeper understanding of the molecular mechanisms underlying the development and progression of lung cancer in female non-smokers.

4. Methodology

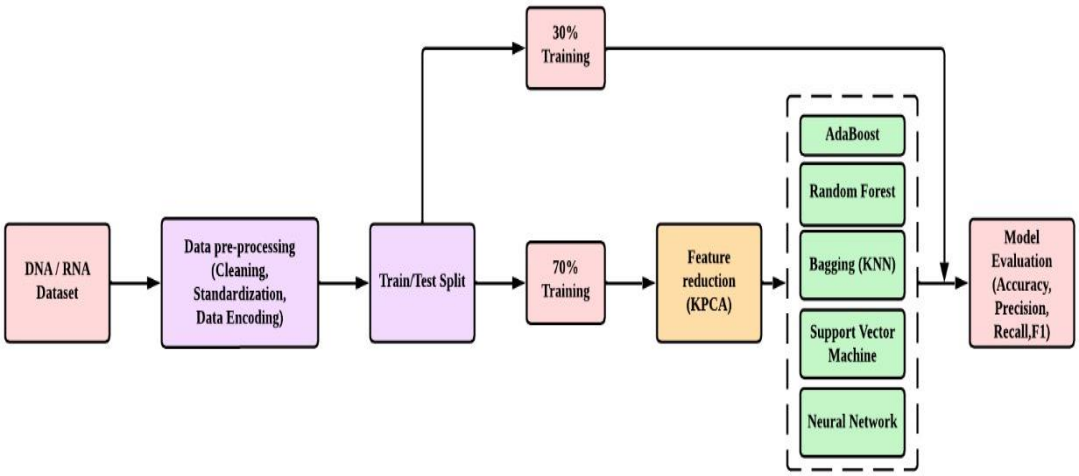

Fig 1. The proposed methodology for DNA-based cancer detection.

1. Data Pre-processing

The pre-processing of the GSE4115 dataset, which includes gene expression data from bronchial airway epithelial cells of smokers, was crucial in guaranteeing the quality and dependability of subsequent analyses aimed at DNA-based lung cancer diagnosis in Fig 1. The first steps involved normalizing the data to reduce technical biases that are inherent in microarray experiments. This was done to ensure that the gene expression values accurately represented the biological differences between the samples. The expression profile investigation was performed utilizing the Affymetrix Human Genome U133A Array platform. The probe ID was linked to the relevant gene symbol by utilizing the data stored on the platform (GPL96-15653.txt). Given the potential for many probes to be associated with the same gene sample, the domino effect was incorporated and

computed as a mean value. The Z-score was used to normalize all gene expression data. This was accomplished by computing the standard deviation (SD) and mean for each gene symbol, and then modifying the X value correspondingly. This was done to mitigate the influence of disparities in the levels of intrinsic expression reported in various genes. The modified equation provides the value X, which denotes the ratio of the average to the variability. The gene expression levels in each dataset were normalized using the approaches described below.

$$z = \frac{(x_{ij} - \mu_i)}{\sigma i}$$

Where x is the expression value of gene i in sample j, μ is the μ(i) and σ(i) respectively representing the mean and standard deviation of the expression vector for gene i across all samples.

2. Feature reduction

It is preferable to apply Kernel Principal Component Analysis (KPCA) rather than conventional Principal Component Analysis (PCA) while enhancing the GSE4115 dataset to make it simpler to comprehend and handle. Because KPCA can capture the nonlinear correlations found in high-dimensional genomic data, it is the recommended method. While KPCA works in a feature space specified by a kernel function, allowing the extraction of nonlinear features, PCA assumes linearity in the data distribution. Understanding intricate biological processes, including the patterns of gene expression linked to lung cancer, is made possible by this. In genomics research, nonlinearity is particularly important since the relationships between genes and phenotypes often exhibit intricate nonlinear patterns. By employing KPCA to efficiently reduce the dimensionality of the dataset, researchers can improve the interpretability of subsequent analyses. Important information contained in nonlinear relationships is preserved using this technique. Additionally, KPCA aids in addressing the difficulties brought on by the "curse of dimensionality," a prevalent issue in genomics datasets with a high feature count. It accomplishes this by removing unnecessary low-dimensional representations while preserving the data's fundamental structure. Consequently, a comprehensive analysis of the genomic landscape supporting the detection of lung cancer utilizing DNA in the GSE4115 dataset is ensured when employing KPCA in feature reduction attempts.

3. Ensemble Learning Methods

Ensemble learning is a method in machine learning that involves combining the predictions of numerous different models to generate a final prediction that is frequently more accurate and resilient than any individual model. Ensemble learning is based on the concept of combining the predictions of numerous models to benefit from their collective knowledge and reduce the limitations of individual models. Each model may exhibit distinct flaws or capture various facets of the data, and via intelligent aggregation of their predictions, these errors can be alleviated or nullified, leading to enhanced overall performance. Ensemble learning approaches generally entail generating a varied collection of base models, training them on the identical dataset, and subsequently merging their predictions by techniques such as averaging, voting, or more advanced methods like stacking. Ensemble learning utilizes the variety and combined expertise of numerous models to improve overall generalization and predictive accuracy in various machine learning tasks.

3.1. Averaging Methods (Bagging)

Bagging:

Bagging, also known as Bootstrap Aggregating, is an ensemble learning strategy that trains several instances of a base model on various subsets of the training data to increase the stability and accuracy of machine learning models. To generate several subsets, the dataset is sampled using replacement (bootstrap sampling), and a base model is subsequently trained on each subset. The final prediction is then generated by averaging or voting together the predictions from these models.

Random Forest Classifier:

Using decision trees as its foundation model, the well-known ensemble learning method Random Forest Classifier encapsulates the bagging concepts. The final prediction is obtained by combining the outputs of separate decision trees, which are trained on randomized subsets of the training data using the Random Forest technique. Every tree functions autonomously, reducing overfitting and adding to the collective insights of the ensemble. By adding a random component to the feature selection process, Random Forest increases the

resilience of the model and further diversifies the individual trees. In mathematical terms, the mode of the predictions made by each decision tree in a Random Forest consisting of "n" decision trees determines the final output class for a classification issue.

Mathematically, the classification prediction can be represented in Equation (3).

X= mode (y1, y2, y3…, yn)

Where:

x is the final output class.

(y1, y2, y3…, yn) are the forecasts from each of the decision trees.

3.2. Sequential Learning Methods (Boosting):

Boosting is an ensemble ML technique in which models are trained one after the other, to improve on the mistakes made by the preceding model. Boosting is the process of combining weak learners to produce a strong learner.

Assigning distinct weights to every training instance is known as weighted training data in boosting. Every instance is given the same weight at first, but as the boosting process goes on, the weights are changed in response to how well the earlier models performed. Accurately predicted instances are given lower weights in the subsequent iteration, while wrongly predicted instances receive larger weights from the prior model. This makes it possible for later models to concentrate more on the unpredictable cases.

AdaBoost

For classification tasks, AdaBoost, also known as Adaptive Boosting, is a well-liked ensemble learning technique. It trains a sequence of weak learners one after the other, giving misclassified cases more weight with each iteration. Based on their performances, these weak learners are joined to generate a strong learner, with each student contributing to the final projection. AdaBoost works well at increasing model accuracy by concentrating on examples that are difficult to categorize. It can adjust to various datasets, is less prone to overfitting, and is flexible. AdaBoost is widely utilized in many different sectors because of its effectiveness and versatility, despite its computational complexity.

3.3. Proposed Ensemble Stacking Method

In this section, we propose an ensemble stacking approach to enhance the predictive performance of our model in Fig 2. Ensemble stacking leverages the strengths of diverse models by combining their predictions to improve overall performance. In our scenario, we utilize a heterogeneous stacking approach, employing a mix of different types of models as base classifiers. These include decision tree ensemble (RF), boosting ensemble (Ada), KNN ensemble (Bagging), and artificial neural network and Logistic Regression as meta classifier.

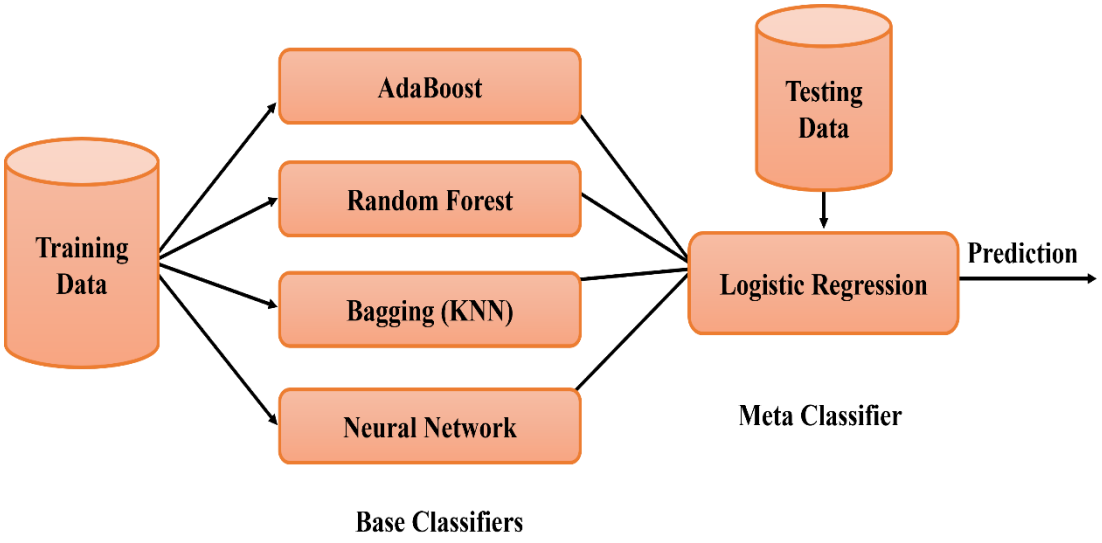

Fig. 2. The proposed ensemble stacking method for DNA-based cancer detection.

The ensemble stacking process involves the following steps:

1. Initial Predictions: Each base model is trained individually on the training data, and initial predictions are made on the validation set or through cross-validation.

2. Meta-learner Training: A separate meta-learner model is then trained on these initial predictions to learn how to effectively combine them for a final prediction. The meta-learner learns the relationships between the predictions of the base models and the actual target values, optimizing its parameters to minimize prediction error.

Logistic Regression:

The LR approach, also known as logistic regression, is typically used for binary classification tasks. Linear classification assigns a coefficient to each predictor, which quantifies the impact of that predictor on the outcome variable.

Contribution to the variance refers to the extent to which a variable affects the variability in the variable being studied.

In this particular scenario, the dependent variable, represented by the symbol 'Y,' takes on a value of 1 when the response is "Yes" and 0 when the response is "No." The model for Predicted Probabilities is represented by the natural logarithm (ln) of the odds ratio, which is mathematically described in Equations 7 to 9 as follows:

$$\mathrm{In}\left(\frac{P(Y)}{1-P(y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k \quad (7)$$

$$\frac{P(Y)}{1-P(y)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\beta_k X_k} \quad (8)$$

$$P(Y) = \frac{e^{\beta_0+\beta_1 X_1+\cdots\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots\beta_k X_k}} \quad (9)$$

$\mathrm{In}\left(\frac{P(Y)}{1-P(y)}\right)$= represent the log odds of the outcomes.

Y= binary result

$X_1, X_2, \ldots\ldots, X_k =$ predictor elements

$\beta_0, \beta_1, \ldots\ldots, \beta_k =$ coefficients of the predictors

The regression coefficients are represented as β0, β1, β2, and so on up to βk, with β0 explicitly functioning as the intercept. Equation (3.8) in logistic regression establishes a direct relationship between the probability of 'Y' and the predictor variables. The primary goal of logistic regression is to estimate the k + 1 unknown parameters (β) by maximum likelihood estimation, determining the values that most accurately correspond to the observed data. The regression coefficients represent the degree of the association between each independent variable and the output. They indicate the anticipated variation in the response variable when the matching predictor is present.

Ensemble stacking presents several advantages for our research endeavor. Firstly, integrating a diverse array of models as base learners, enables us to capitalize on the unique strengths of each model, thus enhancing our ability to capture multifaceted aspects of the underlying patterns and relationships within the data. Moreover, ensemble stacking adeptly addresses the potential weaknesses inherent in individual models by amalgamating their predictions. Given that each model may excel in certain domains while faltering in others, ensemble stacking offers a means to aggregate these predictions, thereby yielding more resilient and dependable outcomes. Importantly, the ensemble stacking approach holds the promise of achieving heightened accuracy compared to any singular model. By leveraging a meta-learner to combine predictions from multiple models, the ensemble can effectively harness the collective knowledge embedded within the diverse base models, facilitating more discerning and precise predictions.

5. Results and Discussion

The authors present the performance evaluation of various machine learning models applied to the GSE4115 dataset. They begin by highlighting the achievements of the best-performing model, a Random Forest classifier, which demonstrated an impressive accuracy of 0.810 and an F-1 score of 0.777. These results affirm the efficacy

of employing machine learning techniques for dataset analysis.

The author's evaluation reveals that the AdaBoost model exhibited comparatively lower performance, with an accuracy of 0.789 and an F-1 score of 0.766. Despite this, it is essential to note that all models, including AdaBoost, demonstrated the capacity to reliably predict genuine positives while mitigating false pessimistic predictions. Notably, the Random Forest classifier achieved the highest recall score of 0.789, underscoring its robust capability in accurately identifying true instances within the dataset in Table 1.

Upon implementing pre-processing and feature reduction techniques on the same dataset, we observed notable enhancements in the performance of both Random Forest and AdaBoost models. For Random Forest, the accuracy surged to 0.946 and the F-1 score reached 0.9, showcasing substantial improvements compared to the base results. Similarly, the AdaBoost model exhibited significant enhancements, with accuracy rising to 0.893 and the F-1 score increasing to 0.833. These findings underscore the efficacy of pre-processing and feature reduction methods in refining the dataset and enhancing model performance. The considerable boost in accuracy and F-1 scores highlights the importance of incorporating such techniques as integral components of the machine learning pipeline, ultimately leading to more accurate and reliable predictive models.

In addition to the comparison with Random Forest and AdaBoost models, we propose a Stacking Ensemble approach incorporating Random Forest (RF), AdaBoost, Bagging, Support Vector Machine (SVM), and neural network models to further enhance the accuracy of lung cancer detection. Our proposed model achieves an accuracy of 0.9464, indicating the proportion of correctly predicted instances out of the total instances tested. The precision of 1 signifies that all instances predicted as positive are indeed true positives, minimizing false positives. With a recall of 0.90, our model successfully identifies 90% of all actual positive instances, ensuring fewer false negatives. The F1-Score, at 0.947, shows a balanced measure of model performance. Lastly, the ROC AUC Score of 0.9974 reflects the model's ability to distinguish between positive and negative classes, with higher values indicating better discrimination capability in Table 2. These metrics collectively demonstrate the superior predictive capabilities of our proposed Stacking Ensemble model, offering a promising avenue for more accurate and reliable lung cancer diagnosis.

To validate the effectiveness of our proposed model, we conducted experiments on an RNA dataset as well. The performance metrics obtained on this dataset further corroborate the efficacy of our approach, with an accuracy of 0.944, precision of 0.952, recall of 0.952, F1-Score of 0.952, and ROC AUC Score of 0.993. These results underscore the robustness and generalizability of our proposed Stacking Ensemble model across different datasets, reaffirming its potential as a valuable tool for accurate lung cancer detection in clinical settings.

Table 1: DNA Dataset

Models	Accuracy	Precision	Recall	F1-Score	ROC AUC Score
Random Forest	0.946428571428571	1	0.9	0.9473684211	0.95
Ad boost	0.8928571429	0.9615384615	0.8333333333	0.8928571429	0.8974358974
Bagging (KNN)	0.9285714286				
Boosting (Decision Tree Classifier)	0.8571428571	0.8055555556	0.9666666667	0.8787878788	0.8487179487
Proposed Model	0.9464285714	1	0.9	0.9473684211	0.9974358974

Table 2: RNA Dataset.

Models	Accuracy	Precision	Recall	F1-Score	ROC AUC Score
Random Forest	0.777777777777777	1	0.578947368421052	0.733333333333333	0.789473684210526
Adaboost	0.916666666666666	0.875	1	0.933333333333333	0.9
Bagging (KNN)	0.694444444444444	1	0.476190476190476	0.64516129032258	
Boosting (Decision Tree Classifier)	0.888888888888888	0.869565217391304	0.952380952380952	0.909090909090909	0.876190476190476
Proposed Model	0.944444444444444	0.952380952380952	0.952380952380952	0.952380952380952	0.993650793650793

In our study, we explored the efficacy of ensemble learning techniques Bagging, Boosting, and Stacking in enhancing predictive accuracy in Fig 3. Among these methods, our proposed Stacking approach achieved the highest accuracy of 94.64%, surpassing both Bagging (92.86%) and Boosting (89.29%) in Fig 4. This underscores Stacking's superior performance in leveraging diverse models to improve predictive outcomes for advanced machine learning applications.

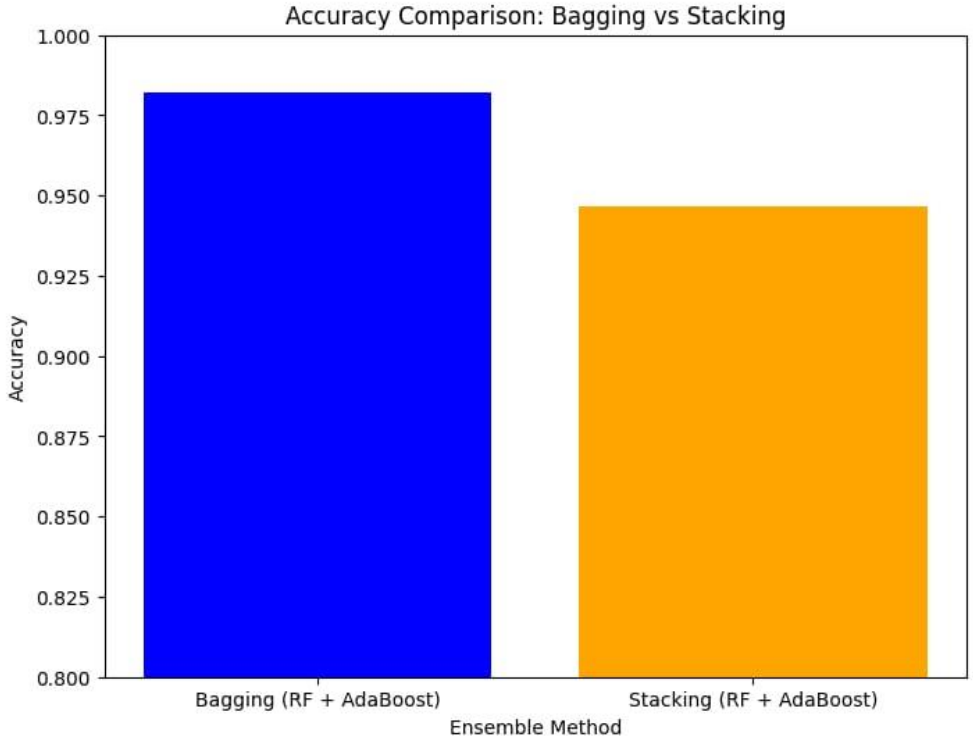

Fig. 3. Accuracy comparison: Bagging vs Stacking

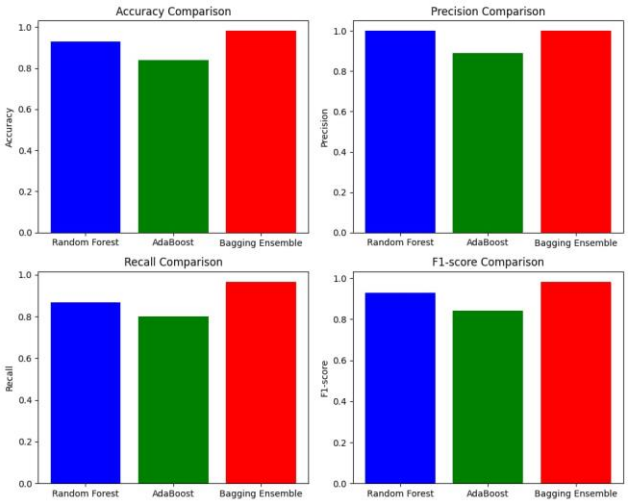

Fig. 4. Accuracy, Precision, Recall, F1-score comparison of RF, Ada-boost and Bagging Ensemble.

6. Conclusion

In conclusion, our study demonstrates the efficacy of machine learning methodologies in the analysis of the GSE4115 and GSE33356 datasets for lung cancer diagnosis. Through rigorous data pre-processing, including normalization and feature reduction via Kernel Principal Component Analysis (KPCA), we ensured data quality and reduced dimensionality to enhance interpretability. Leveraging ensemble learning methods such as Random Forest and AdaBoost, we achieved significant improvements in predictive accuracy compared to individual models. Ensemble stacking emerged as a promising approach, effectively combining the strengths of diverse models to yield robust and reliable predictions. Our findings underscore the potential of machine learning in genomics research, paving the way for the development of accurate diagnostic tools and personalized treatment strategies for lung cancer patients.

7. Future Scope

This study marks a substantial advancement in lung cancer detection by combining DNA and RNA datasets with

powerful ML approaches. While it has made progress in ensemble stacking and diagnostic accuracy, there is still potential for future investigation. Initially focused on establishing robust detection models, the research is motivated by a pressing need for new diagnostic tools. It is critical to identify potential bottlenecks that researchers may face, such as data privacy problems, scalability issues, and clinical validation challenges. Recognizing these obstacles allows researchers to develop more practical and feasible objectives for developing the area of lung cancer detection. By doing so, a path can be paved for new improvements to lung cancer diagnosis, ultimately improving the results for patients.

References

[1] W. H. Organization, "Cancer." Accessed: Nov. 13, 2024. [Online]. Available: https://www.who.int/health-topics/cancer#tab=tab_1

[2] Z.-H. Zhou, Ensemble Methods. Chapman and Hall/CRC, 2012. doi: 10.1201/b12207.

[3] et al. B. Alberts, "Molecular Biology of the Cell." Accessed: Nov. 13, 2024. [Online]. Available: https://www.sigmaaldrich.com/IN/en/product/sigma/m5940?utm_source=google&utm_medium=cpc&utm_campaign=15000381747&utm_content=129438265155&gclid=Cj0KCQiAlsy5BhDeARIsABRc6ZtcidJVOzDbfO_trcrZgdbDFrZ-f4VqRUXelEvshslp_tJzA4-mPJcaAjtEEALw_wcB

[4] M. Verma, P. Maruvada, and S. Srivastava, "Epigenetics and cancer," Critical Reviews in Clinical Laboratory Sciences. 2004. doi: 10.1080/10408360490516922.

[5] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.

[6] R. Chen and J. Lin, "Identification of feature risk pathways of smoking-induced lung cancer based on SVM," PLoS One, vol. 15, no. 6, p. e0233445, Jun. 2020, doi: 10.1371/journal.pone.0233445.

[7] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[8] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1023/A:1018054314350.

[9] L. D. Avendaño-Valencia and S. D. Fassois, "Natural vibration response based damage detection for an operating wind turbine via Random Coefficient Linear Parameter Varying AR modelling," J. Phys. Conf. Ser., vol. 628, no. 1, pp. 273–297, 2015, doi: 10.1088/1742-6596/628/1/012073.

[10] L. Vanneschi and M. Castelli, Multilayer perceptrons, vol. 1–3. 2018. doi: 10.1016/B978-0-12-809633-8.20339-7.

[11] R. Bahado-Singh, K. T. Vlachos, B. Aydas, J. Gordevicius, U. Radhakrishna, and S. Vishweswaraiah, "Precision Oncology: Artificial Intelligence and DNA Methylation Analysis of Circulating Cell-Free DNA for Lung Cancer Detection," Front. Oncol., vol. 12, May 2022, doi: 10.3389/fonc.2022.790645.

[12] M. P. S. Brown et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," Proc. Natl. Acad. Sci., vol. 97, no. 1, pp. 262–267, Jan. 2000, doi: 10.1073/pnas.97.1.262.

[13] D. Mathios et al., "Detection and characterization of lung cancer using cell-free DNA fragmentomes," Nat. Commun., vol. 12, no. 1, p. 5060, Aug. 2021, doi: 10.1038/s41467-021-24994-w.

[14] F. Ben Ali, D. Alrifai, M. Braoudaki, S. Adeleke, and I. Mporas, "Comparative Evaluation of Machine Learning Algorithms on Lung Cancer Type Classification from DNA Microarray Data," in 2021 International Conference on Biomedical Innovations and Applications (BIA), IEEE, Jun. 2022, pp. 33–36. doi: 10.1109/BIA52594.2022.9831234.

[15] P. P. Anglim et al., "Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer," Mol. Cancer, vol. 7, no. 1, p. 62, Dec. 2008, doi: 10.1186/1476-4598-7-62.

[16] A. A. Ponomaryova et al., "Potentialities of aberrantly methylated circulating DNA for diagnostics and post-treatment follow-up of lung cancer patients," Lung Cancer, vol. 81, no. 3, pp. 397–403, Sep. 2013, doi: 10.1016/j.lungcan.2013.05.016.

[17] J. Cabrera, A. Dionisio, and G. Solano, "Lung cancer classification tool using microarray data and support vector machines," in 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), IEEE, Jul. 2015, pp. 1–6. doi: 10.1109/IISA.2015.7387956.

[18] S. Wu, H. Jiang, H. Shen, and Z. Yang, "Gene Selection in Cancer Classification Using Sparse Logistic Regression with L1/2 Regularization," Appl. Sci., vol. 8, no. 9, p. 1569, Sep. 2018, doi: 10.3390/app8091569.

[19] A. A. ABRO, E. TAŞCI, and A. UGUR, "A Stacking-based Ensemble Learning Method for Outlier Detection," Balk. J. Electr. Comput. Eng., vol. 8, no. 2, pp. 181–185, Apr. 2020, doi: 10.17694/bajece.679662.

[20] Q. Wang, Y. Zhou, W. Ding, Z. Zhang, K. Muhammad, and Z. Cao, "Random Forest with Self-Paced Bootstrap

Learning in Lung Cancer Prognosis," ACM Trans. Multimed. Comput. Commun. Appl., vol. 16, no. 1s, pp. 1–12, Jan. 2020, doi: 10.1145/3345314.

[21] N. Yao et al., "Discovery of potential biomarkers for lung cancer classification based on human proteome microarrays using Stochastic Gradient Boosting approach," J. Cancer Res. Clin. Oncol., vol. 149, no. 10, pp. 6803–6812, Aug. 2023, doi: 10.1007/s00432-023-04643-z.

[22] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.-M. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods," Mol. Biosyst., vol. 11, no. 3, pp. 791–800, 2015, doi: 10.1039/C4MB00659C.

[23] H. Hijazi, M. Wu, A. Nath, and C. Chan, "Ensemble Classification of Cancer Types and Biomarker Identification," Drug Dev. Res., vol. 73, no. 7, pp. 414–419, Nov. 2012, doi: 10.1002/ddr.21032.

[24] S. Wang et al., "Multidimensional Cell-Free DNA Fragmentomic Assay for Detection of Early-Stage Lung Cancer," Am. J. Respir. Crit. Care Med., vol. 207, no. 9, pp. 1203–1213, May 2023, doi: 10.1164/rccm.202109-2019OC.

[25] G.-H. Huang, Y.-H. Zhang, L. Chen, Y. Li, T. Huang, and Y.-D. Cai, "Identifying Lung Cancer Cell Markers with Machine Learning Methods and Single-Cell RNA-Seq Data," Life, vol. 11, no. 9, p. 940, Sep. 2021, doi: 10.3390/life11090940.

[26] M. Sherafatian and F. Arjmand, "Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data," Oncol. Lett., Jun. 2019, doi: 10.3892/ol.2019.10462.

[27] L. C, P. S, A. H. Kashyap, A. Rahaman, S. Niranjan, and V. Niranjan, "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers," Cancer Inform., vol. 22, 2023, doi: 10.1177/11769351231167992.

[28] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," Biochim. Biophys. Acta - Mol. Basis Dis., vol. 1866, no. 8, p. 165822, Aug. 2020, doi: 10.1016/j.bbadis.2020.165822.

[29] M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir, and B. Omolo, "A stacking ensemble deep learning approach to cancer type classification based on TCGA data," Sci. Rep., vol. 11, no. 1, p. 15626, Aug. 2021, doi: 10.1038/s41598-021-95128-x.