AL-Bahir
Journal for Engineering and Pure Sciences
BJEPS

Volume 5 | Issue 1 Article 4

A Meta-Ensemble Predictive Model For The Risk Of Lung Cancer

Sideeqoh Oluwaseun Olawale-Shosanya
Department of Computer Science, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria

Olayinka Olufunmilayo Olusanya
Department of Computer Science, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria

Adeyemi Omotayo Joseph
Department of Computer Science, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria

Kabir Oluwatobi Idowu
Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA

Oyelade Babatunde Eriwa
Department of Mathematics and Statistics, College of Arts and Sciences, Bowling Green State University Ohio, United States

See next page for additional authors

Follow this and additional works at: https://bjeps.alkafeel.edu.iq/journal

 Part of the Applied Mathematics Commons, and the Artificial Intelligence and Robotics Commons

Recommended Citation

Olawale-Shosanya, Sideeqoh Oluwaseun; Olusanya, Olayinka Olufunmilayo; Joseph, Adeyemi Omotayo; Idowu, Kabir Oluwatobi; Eriwa, Oyelade Babatunde; Adebare, Adedeji Oladimeji; and Usman, Morufat Adebola (2024) "A Meta-Ensemble Predictive Model For The Risk Of Lung Cancer," *Al-Bahir*: Vol. 5: Iss. 1, Article 4.
Available at: https://doi.org/10.55810/2313-0083.1068

This Original Study is brought to you for free and open access by Al-Bahir. It has been accepted for inclusion in Al-Bahir by an authorized editor of Al-Bahir. For more information, please contact bjeps@alkafeel.edu.iq.

A Meta-Ensemble Predictive Model For The Risk Of Lung Cancer

Authors

Sideeqoh Oluwaseun Olawale-Shosanya, Olayinka Olufunmilayo Olusanya, Adeyemi Omotayo Joseph, Kabir Oluwatobi Idowu, Oyelade Babatunde Eriwa, Adedeji Oladimeji Adebare, and Morufat Adebola Usman

Source of Funding

No external Funding

Conflict of Interest

No Conflict of Interest

Data Availability

publicly available data

Author Contributions

All authors contributed

This original study is available in Al-Bahir: https://bjeps.alkafeel.edu.iq/journal/vol5/iss1/4

ORIGINAL STUDY

A Meta-ensemble Predictive Model for the Risk of Lung Cancer

Sideeqoh O. Olawale-Shosanya [a,*], Olayinka O. Olusanya [a], Adeyemi O. Joseph [a], Kabir O. Idowu [b], Oyelade B. Eriwa [c], Adedeji O. Adebare [a], Morufat A. Usman [a]

[a] Department of Computer Science, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria
[b] Department of Mathematics, Purdue University, West Lafayette, IN 47907, USA
[c] Department of Mathematics and Statistics, College of Arts and Sciences, Bowling Green State University Ohio, USA

Abstract

The lungs play a vital role in supplying oxygen to every cell, filtering air to prevent harmful substances, and supporting defense mechanisms. However, they remain susceptible to the risk of diseases such as infections, inflammation, and cancer that affect the lungs. Meta-ensemble techniques are prominent methods used in machine learning to enhance the accuracy of classifier learning systems in making predictions. This work proposes a robust predictive model using a meta-ensemble method to identify high-risk individuals with lung cancer, thereby taking early action to prevent long-term problems benchmarked upon the Kaggle Machine Learning practitioners' Lung Cancer Dataset. Three machine learning ensemble models—Random Forest, Adaptive Boosting (AdaBoost), and Gradient Boosting—were used to develop the meta-ensemble models proposed in this paper, whereby the three ensemble models were adopted as base classifiers while one of them was adopted as the meta-classifier. In addition, two of the ensemble models were used as base classifiers, while the third was used as a meta-classifier to evaluate lung cancer risk prediction. Different graphs were evaluated to show that people with these features are liable to develop lung cancer. The proposed model has immensely improved prediction performance. The meta-ensemble models were simulated using the Python simulation environment, and the 5-fold cross-validation technique was used. The model validation was carried out using several known performance evaluation methodologies. The results of the experiments showed that gradient boosting achieved a maximum accuracy of 100%, an area under the curve (AUC), and a precision of 100%. The proposed model was compared with novel machine learning methods and popular state-of-the-art (SOTA) deep learning techniques. It was confirmed from the results that the model in this study had the best accuracy at lung cancer risk prediction. This study's results can be utilized to enhance the performance of actual patient risk prediction systems in the future.

Keywords: Meta-ensemble model, Lung cancer, Machine learning, Ensemble models, Risk prediction

1. Introduction

Despite recent technological developments, the medical sciences are still not fully equipped to prevent and cure cancer. The main focus of the medical science community's technological advancement efforts is the containment and treatment of cancer diseases [1,2]. When it comes to death rates, cancer is recognized as the most serious complex disease, and lung cancer is the deadliest cancer in the world [3]. It has a greater impact on individuals and is now ranked seventh in the death rate index [4]. The World Health Organization (WHO) estimates that alcohol, tobacco, high body mass index (BMI), inadequate consumption of fruits and vegetables, and insufficient physical exercise may be responsible for around 33% of cancer-related fatalities. Your chances of acquiring cancer may be increased by specific risk factors, including the use of tobacco, a high intake of alcohol, being exposed to air pollution, radiation, sun exposure, or other unprotected UV light, the absence of physical

Received 17 March 2024; revised 7 May 2024; accepted 11 May 2024.
Available online 6 June 2024

* Corresponding author.
E-mail addresses: ssideeqoh@gmail.com (S.O. Olawale-Shosanya), olusanya_oo@tasued.edu.ng (O.O. Olusanya), tayo_009@hotmail.com (A.O. Joseph), kidowu@purdue.edu (K.O. Idowu), boyelad@bgsu.edu (O.B. Eriwa), adebaredj06@gmail.com (A.O. Adebare), morenikejiadebola@gmail.com (M.A. Usman).

https://doi.org/10.55810/2313-0083.1068
2313-0083/© 2024 University of AlKafeel. This is an open access article under the CC-BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

exercise, and starchy meals, sweets, processed grains, red and processed meat, sugar-filled beverages, and salty snacks, all of which are signs of an unhealthy diet [5,6].

According to the National Cancer Institute (2021), the likelihood of developing cancer increases with age. Generally, this risk continues to rise until individuals reach the age of 80, after which it begins to decline. The accumulation of risk factors over life, the length of time spent exposed to carcinogens, and aging's less efficient cell repair processes might all contribute to this. Your risk of cancer may also be increased by some pre-existing medical problems that induce inflammation [7,8]. It is a condition with a significant burden of symptoms, psychological anguish, and a low quality of life [9]. Lung cancer accounted for 1.8 million new instances of death in 2020, or 18% of all cancer-related deaths. Lung cancer had a far poorer 5-year survival rate (7%–25%) than other major malignancies [10]. Owing to lung cancer's high death rate, the illness's mortality distribution resembled that of its incidence, leading to a significant worldwide disease burden [11].

The yearly death toll of lung cancer is higher than that of colon, breast, and prostate cancer put together. Although some people believe that lung tumors are more deadly since they are frequently discovered at a later stage of the disease, treating lung cancer is one of the biggest obstacles to increasing patient survival. Diseases that are in their early stages can be operated on. However, systemic therapies are frequently the sole choice if cancer has spread throughout the lung and metastasized to other tissues [12]. There are several types of lung cancer, and it is important to treat each one differently. Lung cancers may be divided into approximately two main subgroups: small cell lung cancer (NSCLC), which makes up around 85% of instances and is further divided into lung squamous cell carcinoma (LSCC), large cell lung cancer (LCC), and lung adenocarcinoma (LUAD). Non-small cell lung cancer (SCLC) makes up the remaining 15% of occurrences of lung cancer [3].

Lung cancer is quite common in underdeveloped nations, where locals account for over half of all cases. Lung cancer ranks third among cancers in women and is the most common cancer in males [13]. The goal of treating any cancer is to eradicate or eliminate the malignant cells while sparing healthy ones. Radiation, chemotherapy, and surgery are the most commonly utilized conventional therapeutic modalities. These treatments can be used alone or in combination. Surgical resection is the most reliable and successful course of treatment for individuals with lung cancer [14].

Machine learning is crucial in the early phases of safe human existence for identifying and anticipating medical problems. The diagnosis procedure is streamlined and facilitated using machine learning (ML) [15]. For a time now, machine learning has dominated the medical industry. The health sector in every nation currently makes use of machine learning techniques. Actual illness detection may be explored using machine learning. Trait extraction is one of the major applications of machine learning. For instance, every characteristic of a disease has a genuine information container [16,17].

Machine learning improves data analysis by examining actual characteristics or information and determining the root of the issue. It enables medical professionals to identify the root of an illness. Image processing: Using a variety of machine learning algorithms, precise and practical photo analysis has been discovered. To save time and money and increase their value, this enables doctors to identify ailments earlier [18–20]. Drug production: Considering the rise in various diseases and the known amount, medications should have several functions. This issue has been solved by machine learning, which enables the pharmaceutical business to profit from its use. Improved illness forecasting, thanks to machine learning (ML) technology. ML controls the forecasting of early disease outbreaks so that necessary measures may be taken [4].

Nonetheless, it is more probable that a varied group of people would make wiser judgments than a single person. The same principle applies to machine learning, where a variety of models is always preferable to an individual model. Ensemble method is the term for the machine learning method that achieves diversity. It generates a model that may be used as a classifier or a repressor using training labels and training data as input. To reduce the mistakes caused by each algorithm and obtain a completely developed overall performance via a combined solo-ensemble model, ensemble-based learning chains many algorithms depending on the problem [21].

Accuracy and reliability are crucial prerequisites for machine learning techniques in medical diagnosis and cancer prediction [22]. The related works that were conducted previously in this field, emphasizing ensemble methods for improving performance, were presented [23]. Evaluated various ensemble learning techniques on the Surveillance, Epidemiology, and End Results (SEER) dataset to predict the 5-year survival rate for lung cancer. Five well-known ensemble techniques—Bagging, Dagging, AdaBoost, MultiBoosting, and Random SubSpace—as well as eight classification

algorithms—RIPPER, Decision Stump, Simple Cart, C4.5, SMO, Logistic Regression, Bayes Net, and Random Forest—as base classifiers were assessed for lung cancer survival prediction. Rapidminer Studio 7.1 and the Weka toolkit 3.8 were utilized to prepare the data and create the predictive models. The accuracy and area under the ROC curve (AUC) are used to assess the prediction's performance. Among the four ensemble methods, the AdaBoost algorithm demonstrated the highest efficiency in enhancing the performance of base classifiers.

[24] Employed 14 machine learning methods, including Naïve Bayes, Bayesian Network, SGD (stochastic gradient descent), K-nearest neighbors, support vector machine, artificial neural network, logistic regression, LMT (logistic model tree), random forest, random tree, rotation forest, J48, RepTree (reduced error pruning tree), and AdaBoostM1, to create effective models to identify high-risk people for developing lung cancer. The research used a publicly available dataset (Kaggle). With an accuracy, precision, recall, and F-measure of 97.1% and an AUC of 99.3%, the experiment results demonstrated that the suggested model, Rotation Forest, performed better than the other models [25]. We applied several classifiers as well as ensembles to a benchmark dataset from the UCI repository, which evaluated the discriminative power of multiple predictors to improve the effectiveness of lung cancer detection through symptoms. Support Vector Machine, Naïve Bayes, C4.5 Decision Tree, Neural Network, Multi-Layer Perceptron, and Gradient-Boosted Decision Tree are the classifiers. A comparison is made between the performance and popular ensembles like Majority Voting and Random Forest. Experiments were conducted using the Rapid Miner tool. Performance evaluations revealed that the gradient-boosted tree performed 90% accurately, outperforming both ensemble classifiers and all other individuals.

The researcher conducted diagnoses on patients suspected of having lung cancer and utilized data from public datasets such as “Cancer Patient,” “Survey Lung Cancer,” and “Cancer_Data” for an experiment. The research process encompassed exploratory data analysis (EDA), pre-processing, and classification, where EDA aimed to identify data types, missing values, attribute correlations, and outliers, while pre-processing involved data cleaning and discretization. Randomized oversampling was employed to address imbalanced data, followed by classification using the Gradient Boosted Decision Tree (GBDT), with experiments conducted on both imbalanced and balanced data scenarios. Testing involved varying learning rates and the number of trees through randomized search tuning, with training and testing data distribution utilizing 5-fold cross-validation. The dataset was also classified using k-nearest neighbor and support vector machine algorithms. Results indicated superior performance with balanced data compared to imbalanced data, with the GBDT achieving accuracies of 97% for “cancer patient” and 99% for “cancer_data” [26]. The study advances understanding by employing machine learning techniques to detect lung cancer early, aiming to enhance patient survival rates. The research comprises five stages, including data collection, preprocessing, partitioning for training and testing with 10-fold cross-validation, model training, and evaluation. Through experimentation with CatBoost and Random Forest classification methods, particularly with hyperparameter tuning via Bayesian optimization, the study demonstrates superior performance of the Random Forest model, achieving high accuracy (0.97106), precision (0.97339), recall (0.97185), f-measure (0.97011), and AUC (0.99974) in lung cancer detection [27].

The study examined various research articles on lung cancer prediction models employing machine learning and ensemble learning techniques. Additionally, they introduce novel ensemble learning methods developed using oversampling SMOTE on a survey dataset of 309 individuals with or without lung cancer. The ensemble techniques utilized, such as XGBoost, LightGBM, Bagging, and AdaBoost, undergo evaluation via a k-fold 10 cross-validation method, with predictive attributes encompassing age, smoking habits, physical symptoms, and lifestyle factors. The analysis reveals that XGBoost outperforms other ensemble techniques, achieving an accuracy of 94.42%, precision of 95.66%, recall of 94.46%, and AUC of 98.14% [28]. [29] Developed and created a multi-parameter artificial neural network to predict the risk of lung cancer. Utilizing individual health data, the study was able to accurately and precisely predict the risk of lung cancer. To train and validate the ANN model, adult data from the 1997–2015 National Health Interview Survey was utilized. 79.8% (95% CI: 75.9%–83.6%), 79.9% (79.8%–80.1%), and 0.86 (0.85–0.88) were the sensitivity, specificity, and AUC for the training set. Specificity was 80.6% (80.3%–80.8%), AUC was 0.86 (0.84–0.89), and sensitivity was 75.3% (68.9%–81.6%) for the validation set. The findings show that high specificity and moderate sensitivity are

achieved when lung cancer is detected through the use of an artificial neural network (ANN) based on personal health data.

[30] Pioneered an advanced learning algorithm rooted in AdaBoost for lung cancer detection via electronic nose (eNose) analysis. Breath signals were collected from volunteers, including both healthy individuals and those with lung cancer, and their features were optimized. The resulting improved AdaBoost classifier demonstrated exceptional precision of 98.47% in distinguishing between lung cancer patients and healthy subjects, with high sensitivity (98.33%) and specificity (97%), and remarkable stability across 100 randomized tests [31]. Introduced a novel classifier called Ada-GridRF, which utilized adaptive boost-based grid search optimization to fine-tune the hyperparameters of the base random forest model, effectively identifying malignant and non-malignant nodules in CT images. The proposed method offered enhanced performance speed and decreased computational complexity. Comparative analysis with other hyperparameter optimization techniques and traditional approaches demonstrated superior performance, surpassing even state-of-the-art deep learning methods like transfer learning and convolutional neural networks. Experimental results showcased the Ada-GridRF method achieving outstanding performance metrics, including 97.97% accuracy, 100% sensitivity, 96% specificity, 96.08% precision, 98% F1-score, 4% false positive rate, and 99.8% area under the ROC curve (AUC), with only 8 ms required for training.

[32] Introduced a two-phase approach to predict lung cancer survival, beginning with classification to estimate five-year survival probability, followed by regression to predict actual survival duration in months. Utilizing the SEER database, feature selection techniques such as RFE-RF and LASSO are employed to reduce dimensionality, while machine learning models trained with five-fold cross-validation show ensemble methods outperforming other algorithms, including Logistic Regression (LR), Random Forest (RF), Multilayer Perceptron (MLP), Adaboost, and Naïve Bayes (NB), in terms of performance metrics. Notably, the combination of the LGBM classifier with RFE-RF achieves the highest classification accuracy of 89.6% and an AUC score of 92.03 for survival durations up to 11 months. At the same time, in regression analysis, the LGBM regressor outperforms its counterparts with an MAE value of 7.53 and a RMSE value of 10.49 [4]. We applied various techniques, such as random forests, K-nearest neighbors, logistic regression, support vector classifiers, and radial basis function networks, to classify lung cancer data that was available in the UCI machine learning repository into benign and malignant categories. To classify the input data from cancer to non-cancerous, they were first preprocessed and converted into binary form. This was done using a well-known Weka classifier technique. The comparative approach showed that the recommended Radial Basis Function (RBF) classifier had an excellent accuracy rate of 81.25 percent.

[33] Used the AlexNet model for a deep-learning lung cancer prediction framework with the AlexNet model. The dataset was collected from several sources, including hospitals and open-source software. The study collected 100 images, of which 50 were of cancer and the remaining 50 were of normalcy. They employed data augmentation techniques for improved accuracy because the dataset was smaller. The Python Keras library acknowledges the picture growth processes. Convolutional neural networks (CNN) were used to train the model. To identify lung cancer, pre-trained ImageNet models such as LeNet, AlexNet, and VGG-16 were employed. The final fully connected layer of the model's features was applied individually as input to the softmax classifier. The accuracy obtained by combining the Softmax layer and AlexNet has reached 99.52% [21]. Gathered reports on lung biopsies, from which histological images are produced. To create a dataset, methods for image processing and data augmentation are used on the gathered histopathological images. This dataset is used to train three distinct models: ResNet50, variants-Inception-V3, and convolutional neural networks. Ensemble learning techniques were employed to help reduce the high variance of unseen data measured on training and validation datasets and give better accuracy than a single model.

[34] Proposed a KL divergence-based gene selection strategy for lung cancer prediction. The TCGA and ICGC portals provided the data that was used. The lung adenocarcinoma (LUAD) samples' RNA-seq gene expression data were taken from the TCGA and ICGC datasets. The data is processed using Python in the study so that TensorFlow and Sklearn can recognize the training format. The optimal model was determined and verified using the k-fold cross-validation technique. On the validation set, the deep learning model method based on KL divergence gene selection has an AUC of 0.99 [35]. We compared and analyzed several methods, like watershed segmentation, artificial neural networks, support vector machines, convolutional

neural networks, image enhancement, and image processing, to identify and diagnose lung cancer early on. The Super Bowl Dataset 2016, LIDC-IDRI, and LUNA16 were the datasets used for training. The techniques were applied to facilitate the process and increase accuracy.

This present study proposes a pioneering effort to adopt an ensemble model that combines the predictions of ensemble base classifiers as input to the meta-classifier to make the final prediction of lung cancer occurrence. Three ensemble models—Random Forest, Adaptive Boosting (AdaBoost), and Gradient Boosting—were used to develop the meta-ensemble models to identify those at high risk of lung cancer. These models utilize prevalent habits and symptoms or signs as input features to predict the likelihood of the disease. The performance was evaluated in terms of accuracy, precision, F1-score, recall, and area under the ROC (Receiver Operating Characteristic) curve (AUC). The experimental performance revealed that gradient boosting had superior accuracy when used as a meta-classifier. This assists in greatly improving the accuracy of lung cancer risk prediction. The proposed model had the best accuracy, which outperformed the other state-of-the-art (SOTA) work. This research work will provide great insight to assist doctors in enhancing the performance of actual patient risk prediction systems in the future. To verify the efficiency of the proposed work, a comparison of the performance of lung cancer prediction models is shown in Table 1.

2. Materials and methods

2.1. Technique of data collection and pre-processing

This study required the development of a predictive model for the risk of lung cancer. Data collected from an online resource was accessed by Kaggle machine learning practitioners on November 10, 2023, and can also be accessed from a data repository located online. The dataset contains information about lung cancer patients, which was downloaded from the repository as a.data file format and pre-processed into a comma-separated variable (.csv) file format.

2.1.1. Technique for collecting useful data

The dataset used in this study was obtained from an online repository from Kaggle machine learning practitioners and can be found at https://www.kaggle.com/datasets/ajisofyan/survey-lung-cancer/. Additionally, the dataset is available online at https://data.world/sta427ceyin/survey-lung-cancer. This can be accessed through the Data World repository.

The collected dataset contained 309 lung cancer patient records, which consisted of 16 attributes that were either nominal or numerically valued. The dataset, which was obtained in data file format from the repository, had the attributes used to characterize the data in the first row, after which the data about every lung cancer patient was defined. The collected dataset was used to identify the features considered to increase the risk of lung cancer. Among the 16 attributes, 15 were used as the input variables, while 1 was used as the target variable for the risk of lung cancer.

2.1.2. Technique for pre-processing collected data

Following the process of identifying and collecting the dataset required for constructing the meta-ensemble model aimed at predicting the risk of lung cancer. The collected data was pre-processed by checking the shape, data types, and missing data values of the dataset. The oversampling SMOTE method was used to handle the imbalanced data. We have to note that there was not much processing done on the dataset used, as there were no missing data values. Still, since the simulation environment only required numerical data, the nominal data was then converted.

2.2. The meta-ensemble model's description

The meta-ensemble model developed in this work incorporated the use of Random Forest (RF), AdaBoost, and Gradient Boosting (GB) classifiers as base learners, while Gradient Boosting (GB) was used as a meta-classifier, using the dataset collected for the risk of lung cancer. This research makes use of a framework that uses three ensemble models as base classifiers, while GB is used as the meta-classifier. The meta-classifier took as input the predictions of the three base learners' ensemble models to make the final prediction to develop the meta-ensemble model that was required for the prediction of the risk of lung cancer. To determine how the base learners' predictions were merged to produce the optimal prediction ensemble accuracy, the meta-learner was employed.

Below is the step-by-step representation of the meta-ensemble models:

- Base Ensemble Models:

Base Ensemble Model 1: Random Forest.
Base Ensemble Model 2: Gradient Boosting Machine (GBM).

Table 1. Comparison of the performance of lung cancer prediction models.

Author Name and Reference	Dataset Collected	Models	Performance of the Proposed Models
Safiyari et al. (2017) [23]	SEER dataset (643,924)	Bagging, Dagging, **AdaBoost (proposed),** MultiBoosting, Random SubSpace, RIPPER, Decision Stump, Simple Cart, C4.5, SMO, Logistic Regression, Bayes Net and Random Forest	Accuracy: 88.98% AUC: 94.9%
Dritsas et al. (2022) [24]	Kaggle dataset (309)	Naïve Bayes, Bayesian Network, Stochastic Gradient Descent, K-Nearest Neighbors, Support Vector Machine, Artificial Neural Network, Logistic Regression, Logistic Model Tree, Random Forest, Random Tree, **Rotation Forest (proposed),** J48, reduced error pruning tree, and AdaBoostM1	Accuracy: 97.1% Precision: 97.1% Recall: 97.1% F-Measure: 97.1% AUC: 99.3%,
Faisal et al. (2018) [25]	UCI repository (32)	MLP, Neural Network, Naïve Bayes, Support Vector Machine, Majority Voting, **Gradient Boosted Tree (proposed)**, and Random Forest	Accuracy: 90%, Precision: 87.82% Recall: 83.71% F1-score: 85.71%
Setiawan et al. (2023) [26]	Cancer Patient, Survey Lung Cancer, and Cancer_Data datasets	**Gradient Boosted Decision Tree (proposed),** k-nearest neighbor, and support vector machine	Accuracy: 97% (Cancer Patient) Accuracy: 99% (Cancer_Data)
Zamzam et al. (2024) [27]	Kaggle dataset (309)	CatBoost and **Random Forest (proposed)**	Accuracy: 97.11% precision: 97.34% recall: 97.19% F-measure: 97.01% AUC: 99.97%
Mamun et al. (2022) [28]	Kaggle dataset (309)	**XGBoost (proposed),** LightGBM, Bagging, and AdaBoost	Accuracy: 94.42% Precision: 95.66% Recall: 94.46% AUC: 98.14%
Gregory et al. (2018) [29]	National Health Adult dataset (1997–2015)	**Artificial Neural Network (proposed)**	Specificity: 80.6% (80.3%–80.8%) AUC: 0.86 (0.84–0.89) Sensitivity: 75.3% (68.9%–81.6%)
Hao et al. (2023) [30]	Electronic Nose (eNose) device dataset (142)	**AdaBoost (proposed)**	Precision: 98.47% sensitivity: 98.33% specificity: 97%
Subramanian et al. (2020) [31]	Hospitals and Open-Source Software Dataset (100 images)	**AlexNet, Softmax (Proposed),** LeNet, and VGG-16	Accuracy: 99.52%
Bhattacharjee et al. (2022) [32]	Computed Tomography (CT) Images	**Ada-GridRF Classifier (proposed)**	Accuracy: 97.97% Sensitivity: 100% specificity: 96% precision: 96.08% F1-Score: 98% AUC: 99.8%
Aggarwal et al. (2024) [33]	SEER Dataset	**LGBM with RFE-RF classifier (proposed),** Logistic Regression, Random Forest, Multilayer Perceptron, Adaboost, and Naïve Bayes,	Accuracy: 89.6% AUC score: 92.03%
Suvarchala et al. (2021) [4]	UCI repository (32)	Random Forests, **Radial Basis Function Networks (proposed),** K-Nearest Neighbors, Logistic Regression, Support Vector, and Classifiers	Accuracy: 81.25%
Liu et al. (2022) [34]	TCGA and ICGC dataset	**KL divergence-based gene selection (proposed)**	AUC: 99%
The Proposed	**Kaggle dataset (309)**	**Gradient Boosting (proposed), Random Forest and AdaBoost**	**Accuracy: 100% Precision: 100% AUC: 100%**

Base Ensemble Model 3: AdaBoost.

- Base Ensemble Model Predictions:

Base Ensemble Model 1 predicts: y1.
Base Ensemble Model 2 predicts: y2.
Base Ensemble Model 3 predicts: y3.

- Meta-Learner:

The predictions from the base ensemble models (y1, y2, y3) are used as input to a meta-learner, which is a higher-level model that combines these predictions.

- Final Prediction:

The trained meta-learner generates the final prediction using the base ensemble models' predictions as input. Fig. 1 displays the framework that was utilized to create the meta-ensemble model that was employed in this work.

2.2.1. Random forest (RF) classifier

Random forest is a popular machine-learning technique for classification and regression applications. It is part of the ensemble learning methods that combine several decision trees to provide more accurate predictions [36,37]. The following elements are included in the Random Forest classifier's mathematical expression:

- Decision Trees: This is a diagram that looks like a flowchart, in which every internal node stands for a feature, every branch corresponds to a decision based on that feature, and every leaf node is for the final class prediction.
- Decision Tree Ensemble: Random Forest creates a decision tree ensemble. A Random Forest classifier generates a collection of decision trees T_1, T_2, …, T_N, where N is the number of trees in the forest, given a training dataset containing features X and matching labels Y.
- Random Feature Selection: For every tree, Random Forest selects a random subset of features from the original feature set to introduce randomness. By doing this, over-fitting is decreased and generalization is enhanced since each tree is guaranteed to learn distinct parts of the data.

Mathematically, the Random Forest classifier's prediction for a given input data point x looks like this [36]:

For classification tasks:

Let H(x) represent the Random Forest classifier's output for input x.

$$H(x) \text{ is equal to the mode of } T\_1(x), T\_2(x), \ldots, T\_N(x) \quad (1)$$

where the most common class prediction among each decision tree for input x is indicated by the mode,

For regression tasks:

For each input x, let H (x) be the output of the random forest regression.

$$H(x) \text{ is equal to the mean of } T\_1(x), T\_2(x), \ldots, T\_N(x) \quad (2)$$

where mean is the average prediction of each decision tree for input x.

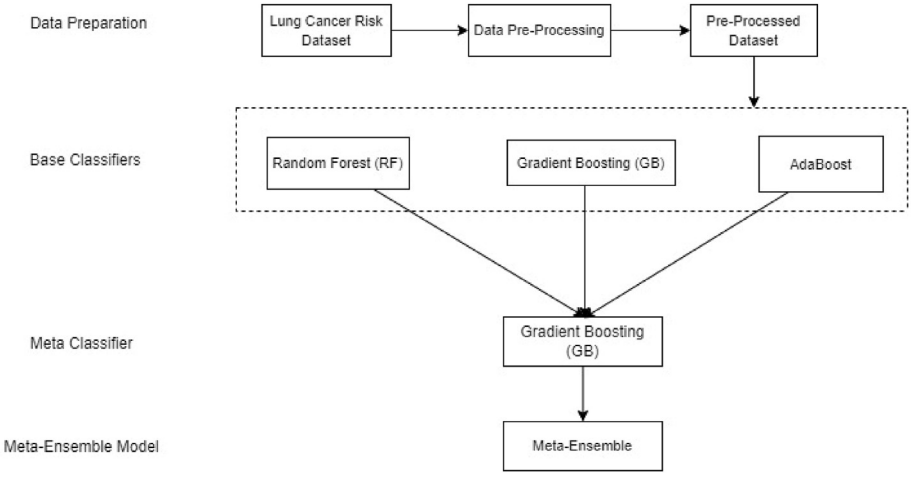


Fig. 1. Framework of meta-ensemble model using machine learning ensembles.

The following stages are involved in building a random forest classifier:

- Bootstrapped Sampling: The forest has N decision trees. From the original training dataset, N subsets are randomly picked (with replacement). Each subset is called a bootstrap sample.
- Tree Construction: Create a decision tree for each bootstrap sample using the chosen features (sometimes a random selection of all features) and the corresponding labels.
- Ensemble Creation: The Random Forest ensemble is created by combining the separate decision trees.
- Prediction: To generate a prediction for a new set of data, run the input through each decision tree in the ensemble. The final prediction may then be obtained by either an average (regression) or a majority vote (classification).

The Random Forest classifier is a powerful and extensively utilized machine learning algorithm because of its ability to handle complex datasets, reduce over-fitting, and provide accurate predictions in various applications [38,39].

2.2.2. Gradient boosting (GB) classifier

Applying the Gradient Boosting (GB) Classifier is a prominent technique in ensemble learning for classification tasks. By combining many weak learners—usually decision trees—gradient boosting is a potent boosting method that produces a powerful prediction model [40,41]. Several components are included in the mathematical representation of the Gradient Boosting Classifier:

- Decision Trees: Like the Random Forest, the Gradient Boosting Classifier employs decision trees as its base learners. But to avoid overfitting, the trees in gradient boosting are often shallow (sometimes called "weak" learners).
- Boosting Process: Gradient boosting revolves around the concept of boosting. It is an iterative process of consecutive learning that develops the ensemble. The goal of every succeeding model (tree) is to fix the errors that the earlier models committed.
- Objective Function: The loss (error) between the predicted values and the actual labels is represented by the objective function that the Gradient Boosting Classifier optimizes. The training procedure aims to reduce the objective function that measures the model's performance.

Gradient Boosting Classifier's mathematical forecast for a given input data point x looks like this [40]:

Let H(x) represent the Gradient Boosting Classifier's final prediction for input x, and let H_0(x) represent the beginning prediction (often a constant) for input x.

$$H(0) + \eta * \Sigma\_i = 1\hat{}N\ \varphi\_i * h\_i(x) = H(x) \quad (3)$$

where:

N is the number of boosting iterations (the number of trees in the ensemble).

η (eta) is the learning rate, which controls the step size at each iteration.

φ_i is the weight (also known as the "shrinkage") applied to the i-th tree's predictions.

The prediction of the i-th decision tree for input x is h_i(x).

New decision trees are added iteratively to the ensemble, known as the boosting process. The new tree concentrates on the data points where the earlier models were incorrect at every iteration. The contribution of every new tree to the final prediction is determined by the learning rate η. Although the training process takes longer with smaller values of η, improved generalization is frequently the result.

By generating decision trees and modifying weights to reduce prediction errors, the Gradient Boosting Classifier builds the ensemble while optimizing the objective function.

Because it can handle complicated datasets, handle missing values well, and perform classification tasks with high accuracy, the Gradient Boosting Classifier is widely utilized in many different machine learning applications. However, it necessitates meticulous adjustment of hyperparameters, including the learning rate and the number of boosting iterations, to achieve optimal performance [42].

2.2.3. AdaBoost (Adaptive boosting) classifier

Another well-liked ensemble learning technique for binary classification problems is the AdaBoost (adaptive boosting) classifier. A powerful predictive model is produced by merging several weak learners, usually known as decision stumps (shallow decision trees with a single split) [43,44]. The following elements are included in the mathematical statement of the AdaBoost classifier:

- Weak Learners (Decision Stumps): In AdaBoost, the weak learners are often straightforward classifiers. A decision stump is a one-level decision tree that uses a threshold value and a single feature to predict a result.
- Weighted Data Points: Based on how well the data points were classified in earlier iterations, weights are added to them in each AdaBoost

algorithm iteration. Data points that are incorrectly identified are given larger weights, making them more significant for later iterations.
- Alpha Values: Each weak learner in AdaBoost is given an alpha value that indicates how much of a contribution it made to the final prediction. The precision of the weak learners affects the alpha values. An improved weak learner's alpha rating indicates how important they are to the group.

In terms of mathematics, the AdaBoost classifier's prediction for a given input data point x is expressed as follows [45]:

Let H(x) be the AdaBoost classifier's final prediction for input x.

$$\text{Sign}(\Sigma\_i = 1\hat{}N\ \alpha\_i * h\_i(x)) = H(x) \quad (4)$$

where:

N is the total number of indecisive learners (weak learners) in the group.

The i-th weak learner's alpha value is represented by α_i.

The prediction made for input x by the i-th weak learner (decision stump) is denoted by h_i(x). Either +1 or −1 is returned based on the classification decision.

The following steps are included in the AdaBoost algorithm:

- Initialize Data Weights: Set all data points in the training dataset to the same weights.
- Iterative Training: Carry out several iterations (usually N iterations), each of which entails the following steps:

1. Using the existing data weights, train a weak learner (decision stump) on the training set.
2. Using the training set, determine the weak learner's weighted error rate.
3. Determine the weak learner's alpha value by calculating the weighted error rate.
4. Adjust the data weights by assigning incorrectly categorized data points a greater weight and correctly classified data points a lower weight.

- Ensemble Creation: Create the final ensemble by combining the weak learners based on their alpha values.
- Prediction: The last prediction is obtained by weighted voting of the predictions of each weak learner in the ensemble, which is used to produce predictions for new data.

When it comes to processing complicated datasets and attaining high accuracy in binary classification tasks, the AdaBoost classifier is efficient. By assigning greater weight to samples that were incorrectly categorized, it adjusts to the properties of the data, enabling it to concentrate on cases that are challenging to classify and enhancing performance. Its performance, however, might be adversely affected by outliers and noisy data because of its sensitivity [46,47].

2.3. Environment for model simulation

Since the supervised machine learning ensembles necessary for building the predictive model for lung cancer risk were identified, data obtained from Kaggle was utilized to simulate the model. The simulation was conducted using the Python programming language and a set of machine learning ensembles.

Python is a high-level programming language that has a reputation for being easy to understand and straightforward. Guido van Rossum invented it, and it was originally made available in 1991. Programmers may express ideas in fewer lines of code using Python's clear syntax and emphasis on code readability as compared to other languages [48,49]. Among Python's principal attributes are:

- Easy to Learn: Simple and clear grammar makes Python an easy language for novices to learn. It lowers software maintenance costs while emphasizing readability.
- Interpreted: Python is an interpreted language, meaning that each line of code is carried out individually. As a result, testing and development may go more quickly and don't require an additional compilation stage.
- Cross-platform: Python works with a variety of operating systems, such as Windows, Linux, macOS, and many more. This facilitates the writing of code that can run unchanged across several operating systems.
- Huge Standard Library: Python comes with a huge library of modules and functions that can be used for a wide range of applications, including web development, networking, file I/O, and more. For many typical jobs, this removes the need to develop code from the start.
- Third-Party Libraries: A wide range of third-party libraries and frameworks are available for Python, thanks to its thriving ecosystem. NumPy, Pandas, Django, Flask, and TensorFlow are a few examples of libraries that expand Python's capabilities for certain fields

encompassing scientific computing, web development, machine learning, and data analysis.
- Object-Oriented Programming (OOP): Python has support for OOP, which enables programmers to write modular and reusable code. It is appropriate for developing large-scale systems since it has features like classes, inheritance, and polymorphism.
- Dynamic Typing: Python has dynamic typing, which implies that variables' types are decided upon at runtime. While this offers flexibility, type handling, and variable assignments must be carefully considered.

Python is a versatile language with many uses, such as scientific computing, web development, automation, machine learning, data analysis, and artificial intelligence. It is a well-liked option among developers due to its adaptability, ease of use, and robust community support [50,51].

2.3.1. Results evaluation extraction using confusing matrix

It was necessary to overlay the classification results on a confusion matrix, as seen in Fig. 2, to assess the effectiveness of the meta-ensemble models employed to predict the risk of lung cancer. A confusion matrix is a table that has four different combinations of the actual and expected values. In other words, it serves to elucidate the performance of a classification model, sometimes referred to as a "classifier," when it is applied to a set of test data whose true values are known [52]. This contributes to providing information regarding the accuracy of the data classification. By counting the instances in which the model was properly and wrongly categorized, it provides a visual representation of performance. The figures for false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) are displayed in two rows and two columns [53].

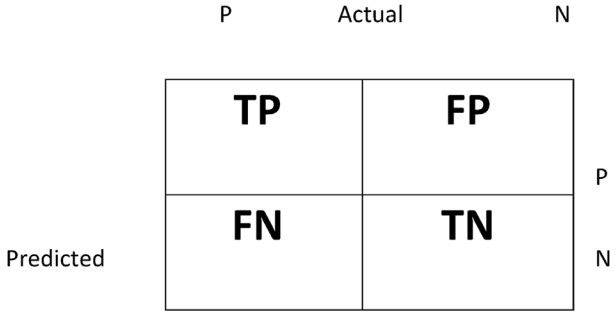


Fig. 2. Confusion matrix for model performance evaluation.

2.3.2. Explanation of performance evaluation metrics used to validate models

The prediction model's performance can be evaluated using the true positive/negative and false positive/negative values that were obtained from the confusion matrix. The following are explanations of the metrics' definitions and expressions [54]:

a. Precision: Precision is the ratio of accurate affirmative predictions to all positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

b. Recall (Sensitivity): Recall is the ratio of accurate positive predictions to all predicted outcomes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

c. F1 Score: The precision and recall harmonic means are combined to get the F1 score. The F1 score achieves a maximum of 1 and a minimum of 0. It offers an equilibrium between recall and precision.

$$\text{F1Score} = \frac{2*(\text{precision} * \text{recall})}{\text{Precision} + \text{recall}} \tag{7}$$

d. One of the most important performance evaluation metrics is accuracy, which is just the percentage of properly predicted observations to all observations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

e. ROC Curve: A graphical depiction of the model's performance at the lowest possible classification threshold is the Receiver Operating Characteristic Curve. It is a statistic for evaluating issues with binary categorization. Plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values is a probability curve that mainly distinguishes between "signal" and "noise." The curve displays two parameters: TPR and FPR. A True Positive Rate and a False Positive Rate are obtained, respectively, using equations (9) and (10).

True Positive Rate (Sensitivity/Recall): The ratio of correctly predicted positive observations to the total number of real positives.

$$\text{True Positive Rate} = \frac{TP}{TP + FN} \quad (9)$$

False Positive Rate (1-specificity/false alarms): This is the proportion of all genuine negative observations to all falsely predicted positive observations.

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \quad (10)$$

f. Area under the Curve (AUC): AUC calculates the area under the ROC curve and expresses how well the classifier can differentiate between positive and negative instances. It provides a single value that represents the model performance across different threshold values. A random model has an AUC of 0.5 while the model with perfect discrimination has an AUC of 1. Higher AUC values generally indicate better overall performance.

3. Results and discussion

3.1. Results of pre-processing and data collection

Data for this study was gathered from Kaggle Machine Learning Practitioners, an online repository. The target class, which determined the risk level of lung cancer, was one of the 16 attributes that made up the 309 records for the data collected for this study. The 16 attributes found in the gathered dataset were labeled using both categorical and numeric values. The patient's current age was indicated by a numerical label. The categorical or nominal attributes included the patient's gender, smoking status, yellow fingers, anxiety, peer pressure, chronic illness, fatigue, allergy, wheezing, alcohol intake, coughing, difficulty swallowing, shortness of breath, chest pain, and the patient's risk of lung cancer.

The male gender has a total number of 162, which accounts for 52%, while the female gender has a total number of 147, which accounts for 48% of the dataset. The youngest patient's age was 21, and the oldest was 87 years old. The description of the dataset according to the identified attributes, as presented in Table 2, was provided using a frequency distribution table. The shape of the datasets was checked, which was (309, 16). The dataset contains 33 duplicate instances and was removed, which reduced the shape of the dataset to (276, 16). The dataset has a target distribution imbalance issue. The imbalanced issue was handled using the random oversampling SMOTE method, which made the dataset balanced. The data types of the dataset were all integers except gender and lung cancer, which were object values. The object values were converted to integer values as required by the simulation environment. The object values are converted to integer values: gender: M−1, F-1, and lung cancer: Yes-1, No-0. The dataset was split into target and feature sets. The dataset does not have much pre-processing, as there was no missing value in the dataset used.

Table 2. Frequency distribution of the data collected based on the attributes.

Attribute	Distinct	Most Common	Next Most Common
Gender	2	M (162)	F (147)
Age	39	64 (20)	56 (19)
Smoking	2	2 (174)	1 (135)
Yellow Finger	2	2 (176)	1 (133)
Anxiety	2	1 (155)	2 (154)
Peer Pressure	2	2 (155)	1 (154)
Chronic Disease	2	2 (156)	1 (153)
Fatigue	2	2 (208)	1 (101)
Allergy	2	2 (172)	1 (137)
Wheezing	2	2 (172)	1 (137)
Alcohol Consumption	2	2 (172)	1 (137)
Coughing	2	2 (179)	1 (130)
Shortness of Breath	2	2 (198)	1 (111)
Swallowing Difficulty	2	1 (164)	2 (145)
Chest Pain	2	2 (172)	1 (137)
Lung Cancer	2	**True** 270 (87.38%)	**False** 39 (12.62%)

3.2. Discussion of identification and collection of data

After the necessary datasets had been identified and gathered for this research, a pair plot imported from the Seaborn Library was used to present the visualization of relationships between pairs of features against the target variable. Fig. 3 shows the pair plot for lung cancer risk prediction, which describes the dataset relationships based on the attributes identified. The results of the pair plot showed that, based on the relationship between the gender of patients and other features, there are high occurrences of lung cancer in both males and females. The results of the relationship between the patients' ages and genders showed that the majority of both genders have the occurrence of lung cancer between the ages of 40 and 79 years of age. The relationship between age indicated and smoking, yellow fingers, anxiety, chronic disease, fatigue, wheezing, allergy, alcohol consumption, coughing, peer pressure, swallowing difficulties, shortness of breath, and chest pain showed that patients who have the features have a higher occurrence of lung cancer compared to those who are without the features.

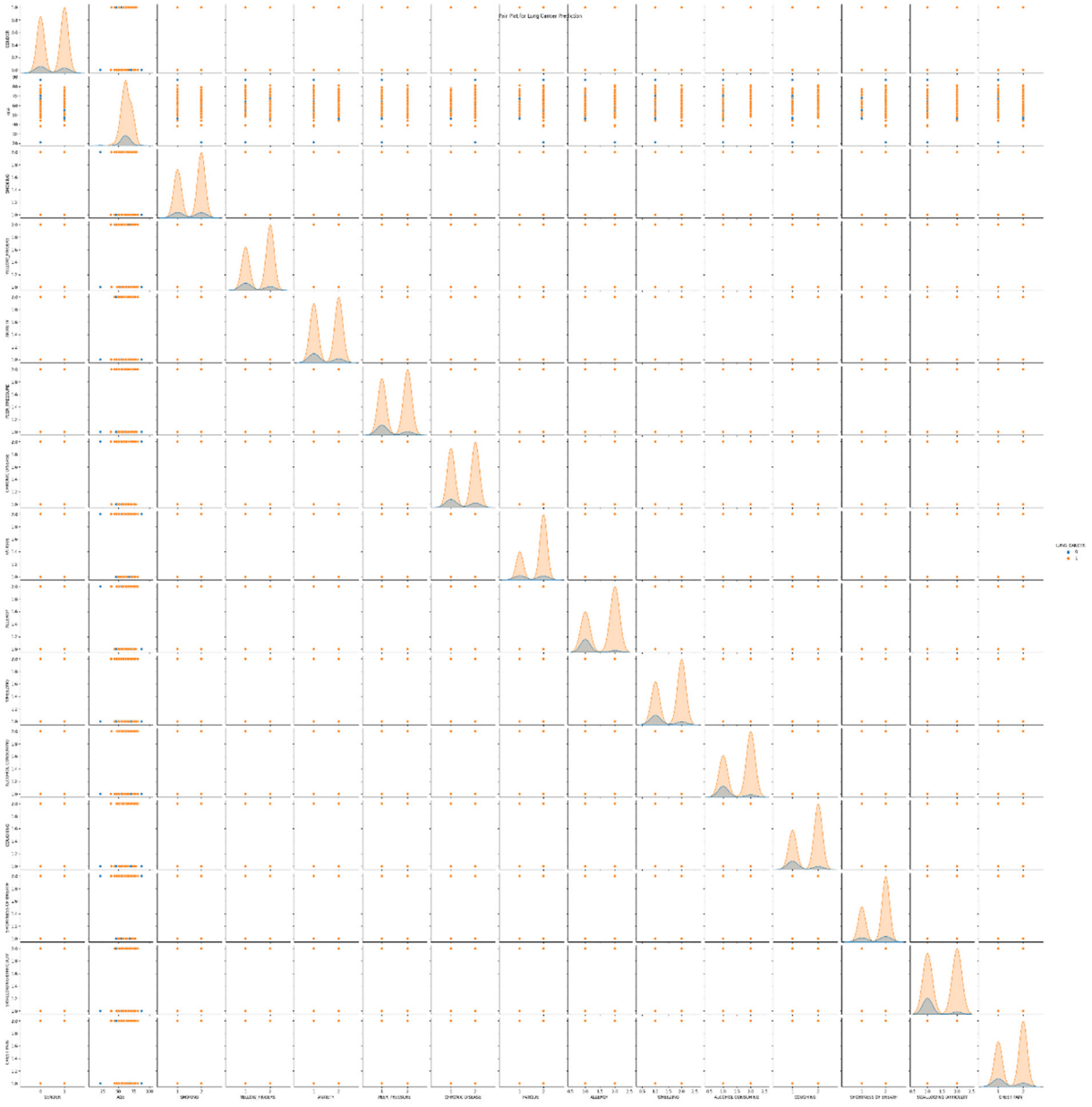


Fig. 3. The pair plot of lung cancer risk prediction.

The results of the relationships among all the remaining features showed that the majority of the features had the occurrence of lung cancer.

A correlation matrix heatmap was used to show the correlation coefficients between the attributes, as shown in Fig. 4. Each cell in the heatmap represents the correlation between two attributes. A correlation matrix is a table that shows the correlation coefficients between many variables. The numbers fall between −1 and 0. Perfectly positive correlations are denoted by a 1, perfectly negative correlations by a −1, and no correlations are denoted by a 0. A heatmap is a type of graphical data representation that is often used to visualize the correlation matrix. The colors in the matrix indicate the direction and degree of the connection between the variables; darker colors usually indicate stronger correlations. The results of the correlation matrix heatmap indicate correlation coefficients of 1, which showed a perfectly positive and stronger correlation, such that as one attribute increases, the other also increases. The diagonal values correlate with 1 because each attribute perfectly correlates with itself.

A count plot was used to show the visualization of the distribution of binary features concerning the

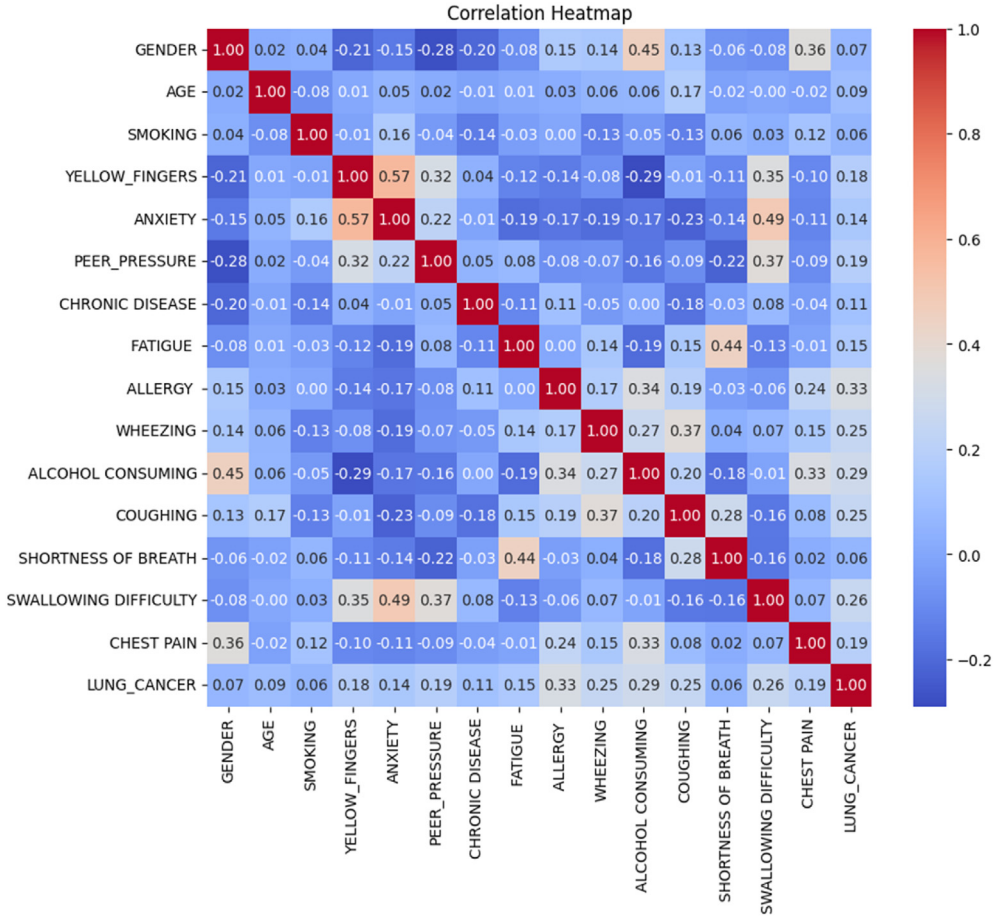
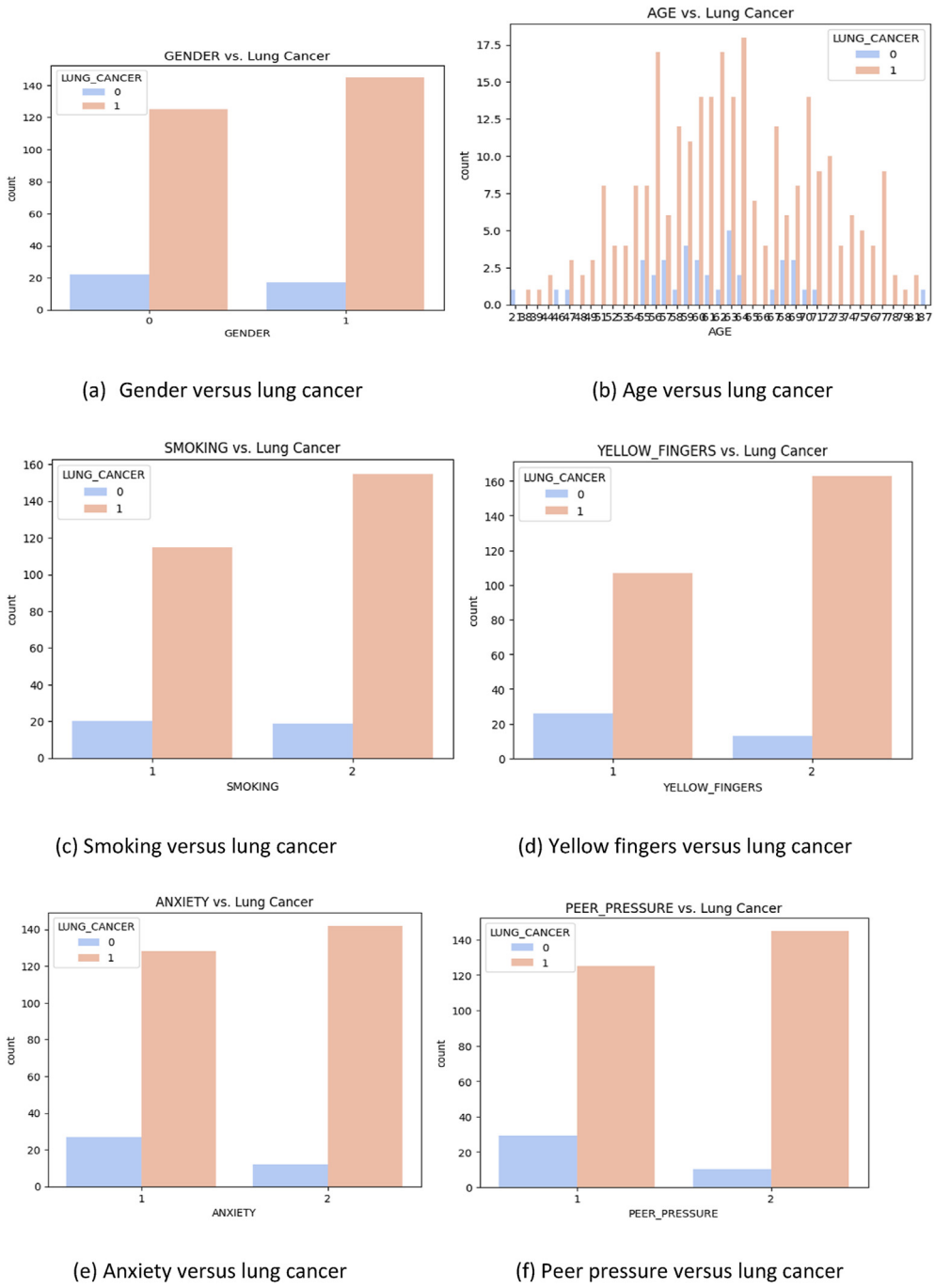


Fig. 4. The correlation heatmap of the dataset.

target variable, as presented in Fig. 5. The features indicate a blue box, while the target variable, which is lung cancer, indicates an orange box. The results of gender versus lung cancer showed that there are more occurrences of lung cancer in males than in females. The results of age versus lung cancer showed that the risk of having lung cancer starts to increase at the age of 40 and decreases at 80 years of age. Smoking versus lung cancer indicates that there are more occurrences of lung cancer in those who are smoking than in those who are not. Yellow fingers versus lung cancer indicates that those with yellow fingers are at a higher risk of lung cancer than those without yellow fingers. The results of anxiety versus lung cancer, peer pressure versus lung cancer, chronic disease versus lung cancer, and swallowing difficulty versus lung cancer showed that there is little difference between those who possess the mentioned features versus lung cancer, allergy versus lung cancer, wheezing versus lung cancer, alcohol consumption versus lung cancer, coughing versus lung cancer, shortness of breath versus lung cancer, and chest pain versus lung cancer, revealing that there is a higher occurrence of lung cancer in those who have the mentioned features than those who do not have the features.

This work also presents a boxplot to show the relationship between each feature and lung cancer, as shown in Fig. 6. The results of the boxplot of gender by lung cancer status showed that lung cancer affects both genders. The results of age by lung cancer status revealed that there is a very high occurrence of lung cancer within the age range of 55–68 years. The results of smoking, anxiety, yellow fingers, fatigue, chronic disease, chest pain, and shortness of breath based on lung cancer status showed that the higher the mentioned features, the higher the risk of having lung cancer. That is, they contribute to the risk factor of having lung cancer. The results of peer pressure and coughing by lung cancer status indicate that they have an impact on the risk of having lung cancer. The results of the boxplot of allergy, wheezing, alcohol consumption, and swallowing difficulty showed that they have little or no impact on the risk of lung cancer.



(a) Gender versus lung cancer

(b) Age versus lung cancer

(c) Smoking versus lung cancer

(d) Yellow fingers versus lung cancer

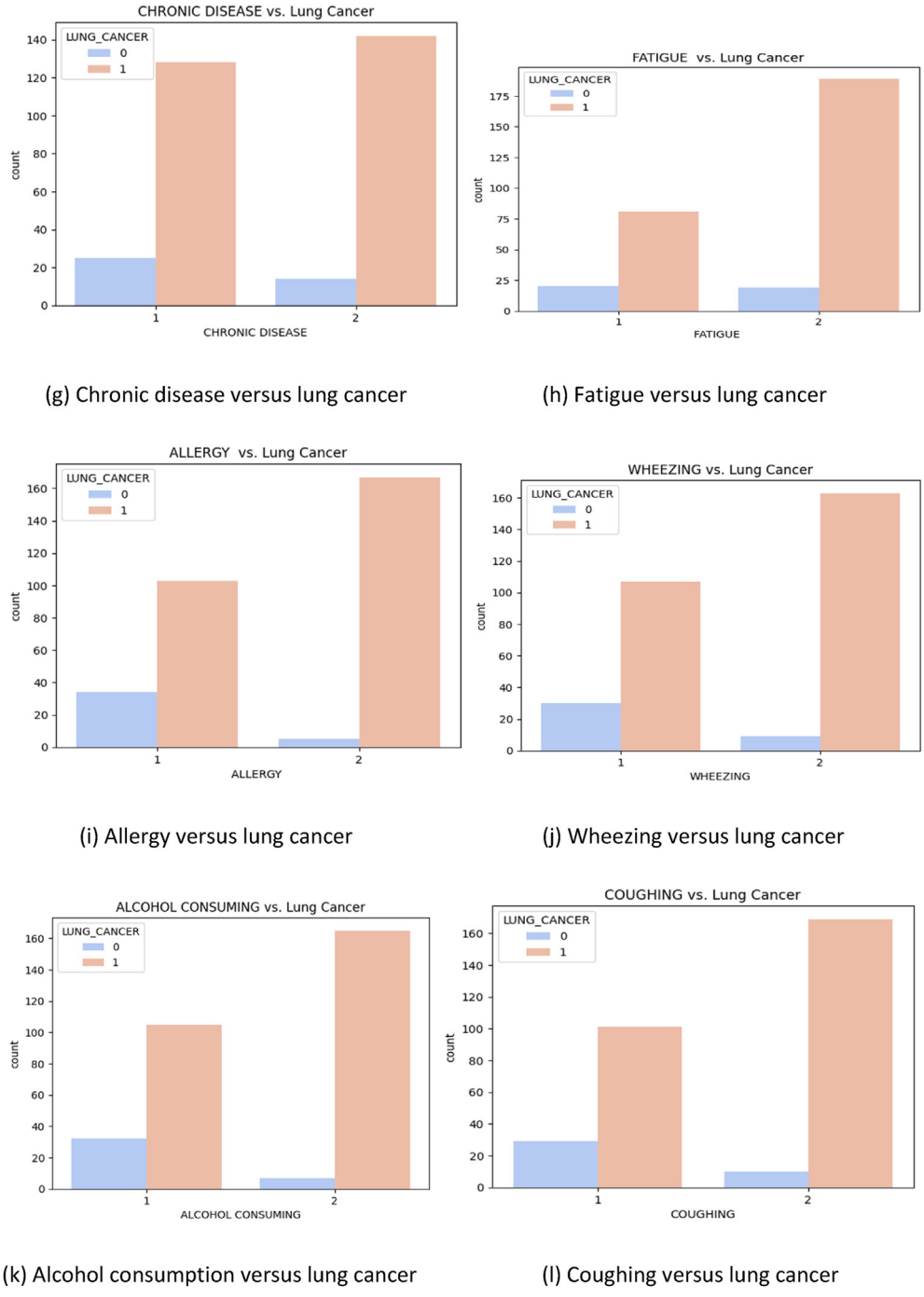
(e) Anxiety versus lung cancer

(f) Peer pressure versus lung cancer

Fig. 5. Visualization of the distribution of binary features against lung cancer.

3.3. Results of the meta-ensemble model development and simulation

This section presents the meta-ensemble model that is needed to predict the risk of lung cancer according to the framework after the data have been identified and described. The Python simulation environment was used to develop the meta-ensemble using Google Colaboratory (collab), from which the model was developed. The dataset was loaded from upload to session storage into the simulation environment, and all the necessary libraries were imported from Ski Kit-Learn (Sklearn). The base classifiers and the meta-classifiers were



(g) Chronic disease versus lung cancer

(h) Fatigue versus lung cancer

(i) Allergy versus lung cancer

(j) Wheezing versus lung cancer

(k) Alcohol consumption versus lung cancer

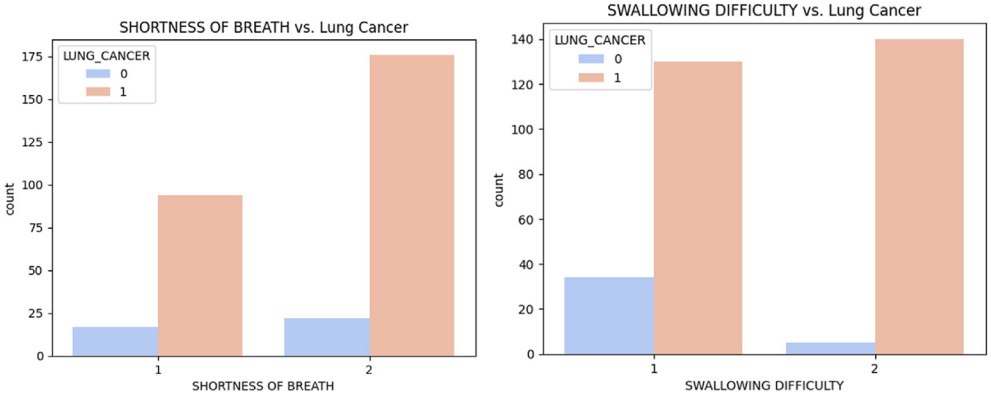
(l) Coughing versus lung cancer

Fig. 5. (Continued).

also imported by using the three identified ML ensembles as base classifiers while using one of them as a meta-classifier.

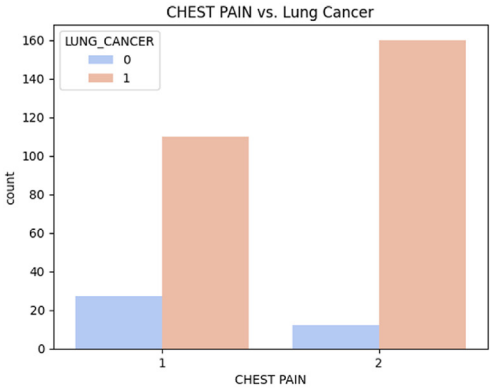
3.3.1. The meta-ensemble model Simulation's results

The meta-ensemble model was simulated using the model that the Python simulation environment offered by using Random Forest, AdaBoost, and Gradient Boosting as the base learners, while Gradient Boosting was used as the meta-classifier using the 5-fold cross-validation training process. As a result, the final meta-ensemble model for lung cancer risk prediction was created using the predictions supplied by the base learners as the input for the gradient boosting algorithm. The model operated for 3 s, with a meta-ensemble accuracy of



(m) Shortness of breath versus lung cancer

(n) Swallowing difficulty versus lung cancer



(o) Chest pain versus lung cancer

Fig. 5. (Continued).

100%. The confusion matrix of the ensemble models is presented in Fig. 7. The confusion matrix of the models was also plotted using Seaborn, as shown in Fig. 8.

Similarly, when Random Forest, AdaBoost, and Gradient Boosting were used as the base learner while Random Forest was used as a meta-learner, and when Random Forest, AdaBoost, and Gradient Boosting were used as the base classifiers while AdaBoost was used as a meta-classifier, they gave the meta-ensemble accuracy of 99.1% and 98.1%, respectively. Also, when Gradient Boosting and AdaBoost were adopted as base learners while Random Forest was adopted as the meta-classifiers, Random Forest and AdaBoost were adopted as base classifiers while Gradient Boosting was a meta-classifier, and when Random Forest and Gradient Boosting were adopted as base classifiers while AdaBoost was a meta-classifier, they had a meta-ensemble accuracy of 96.3%, 100%, and 97.2%, respectively. The meta-ensemble accuracy of the above models revealed that gradient boosting achieved the best meta-ensemble accuracy using the 3 ensembles as base learners, after which was Random Forest followed by AdaBoost, and also that gradient boosting achieved the best meta-ensemble accuracy using 2 ensembles as base learners, after which was Random Forest followed by AdaBoost, as shown in Table 3.

3.3.2. Model validation results using performance evaluation measures

Several performance evaluation measures, which were generated from the confusion matrices of the simulation results, were used to assess the validation outcomes of the three ensemble models. The number of accurate predictions, accuracy (expressed as a percentage), precision, recall, f1_score, and AUC ROC curves were used to access the model validation findings. The ensemble with the highest

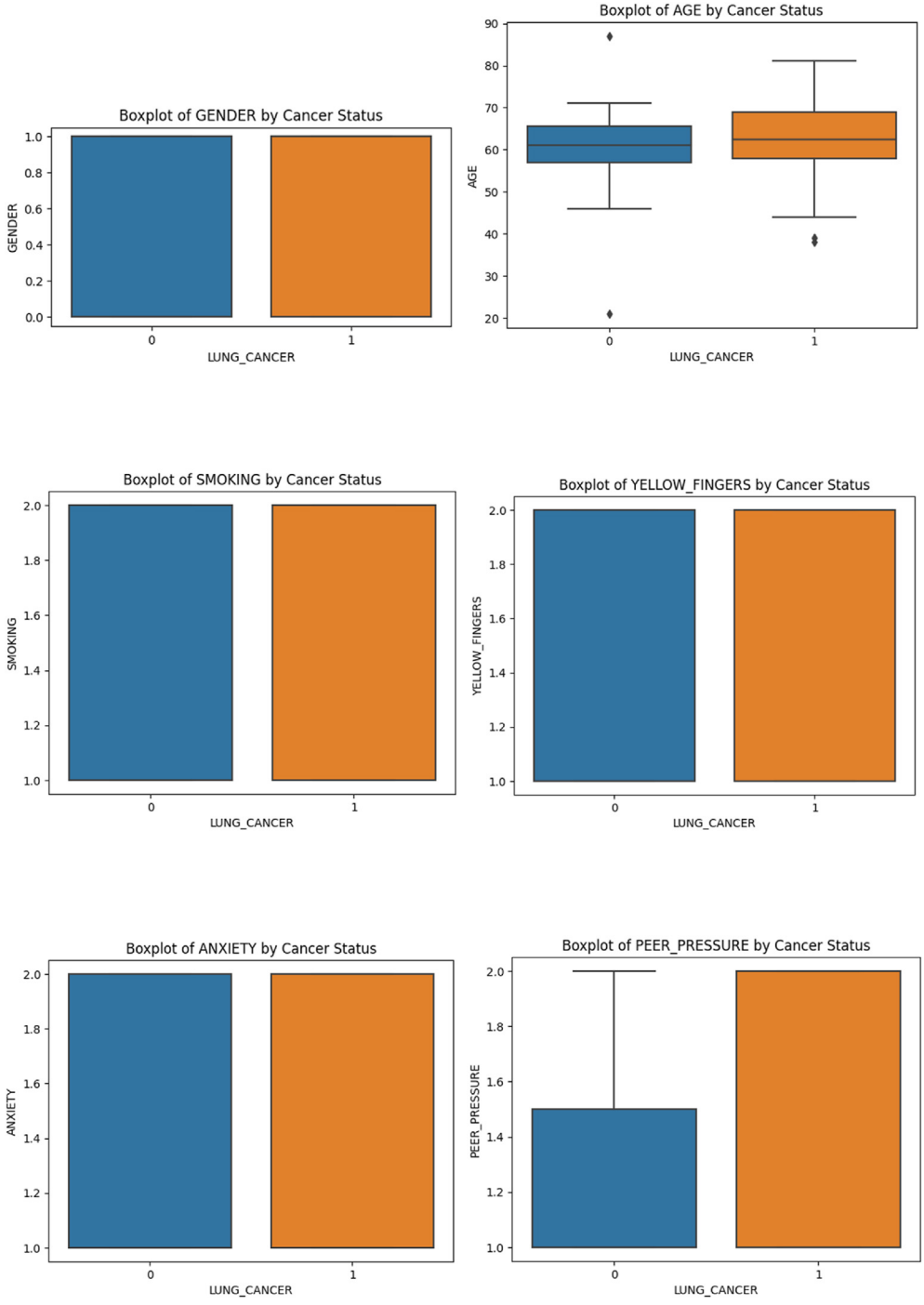


Fig. 6. Relationship between features and lung cancer status.

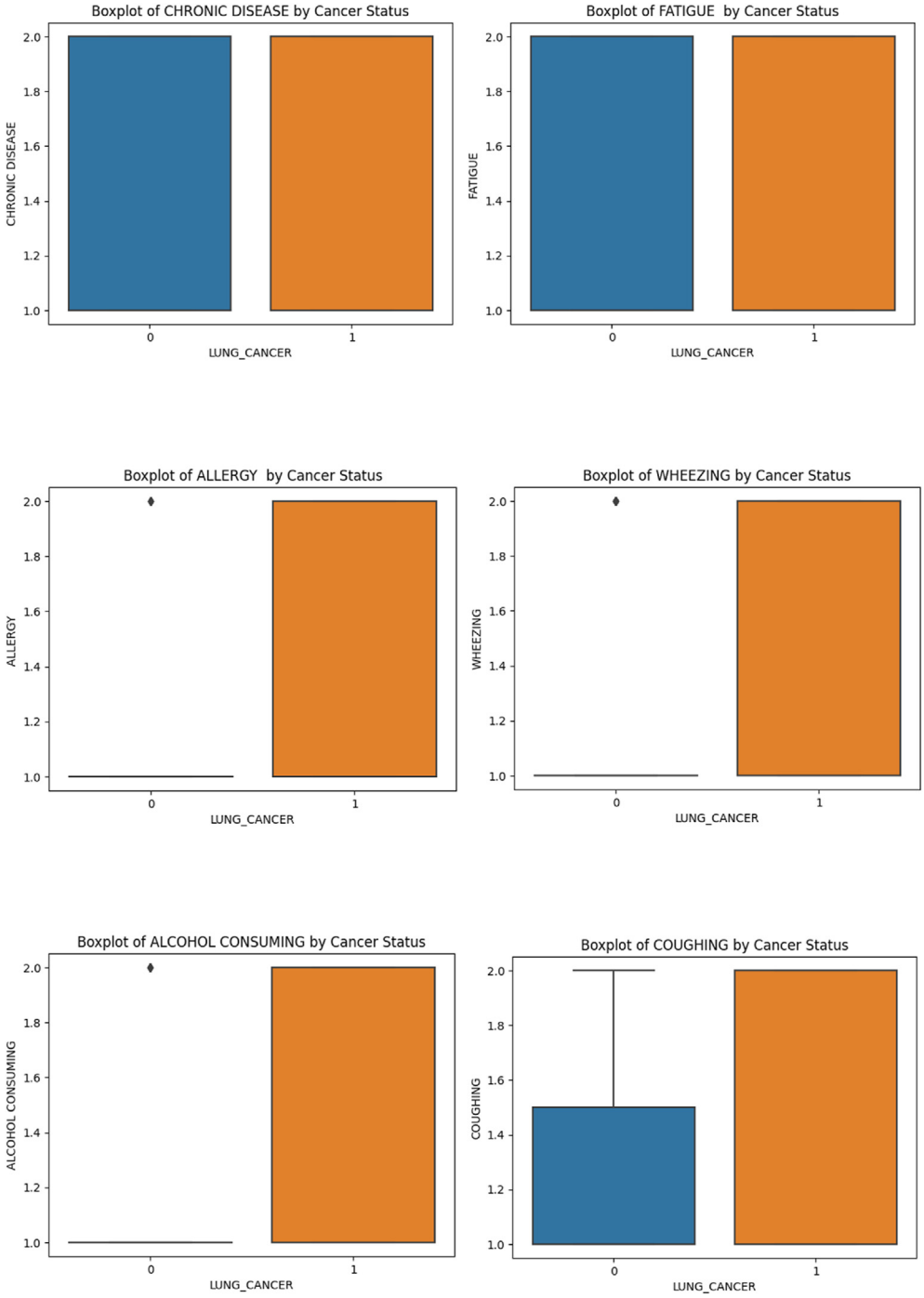


Fig. 6. (Continued).

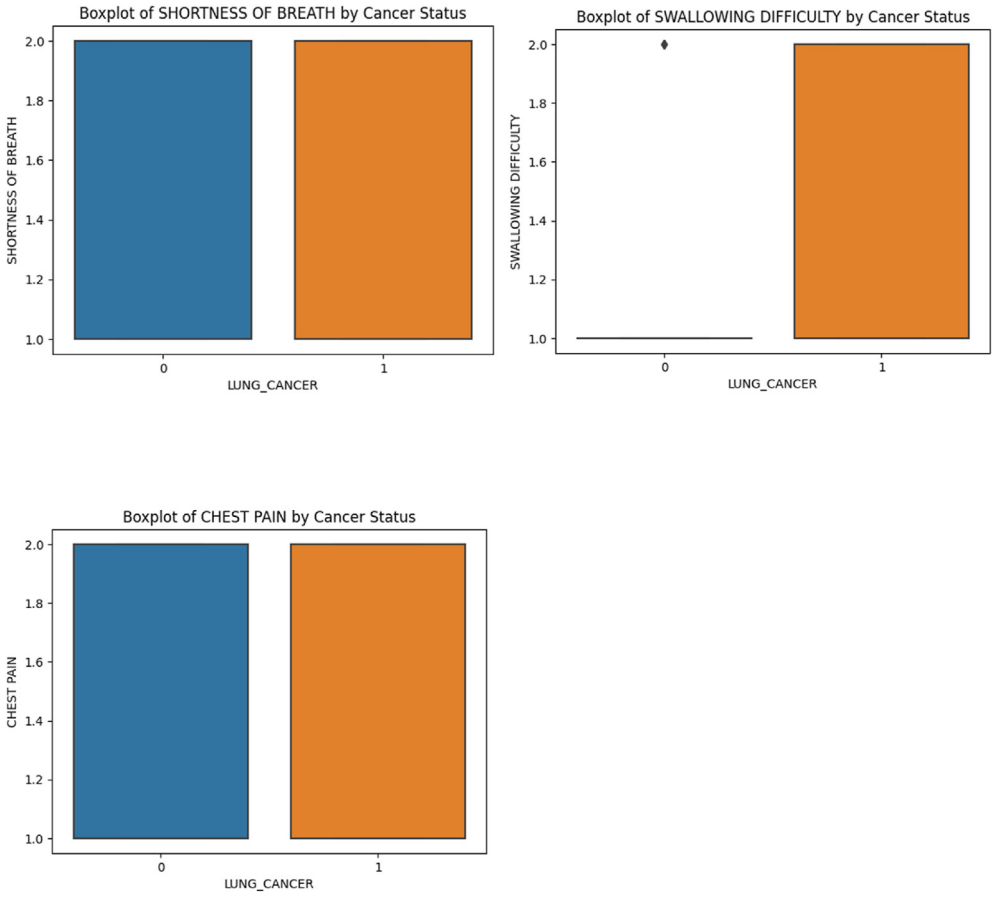


Fig. 6. (Continued).

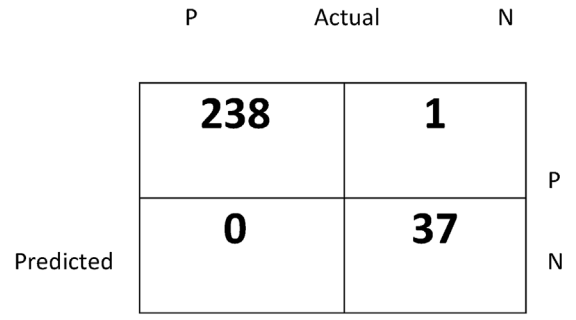
values for f1_score, AUC ROC curves, accuracy, precision, recall, and correct predictions is the best model. Table 4 presents an overview of the findings from the model validation process for the three ensemble models created using the performance evaluation metrics.

Out of the 276 dataset records utilized in this study, the Random Forest model produced 275 accurate classifications and 1 wrong one. The accurate classification consisted of 238 positive and 37 negative cases, while the inaccurate classification consisted of 1 negative case as positive and 0 positive cases as negative cases. The model accuracy, precision, recall, f1_score, and AUC were 99.1%, 100%, 98.2%, 99.1%, and 100%, respectively.

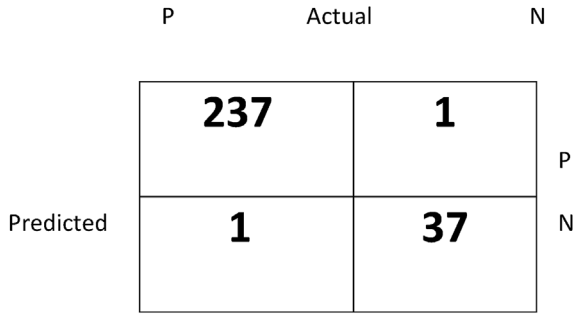
The results of the gradient boosting model had 274 accurate and 2 inaccurate classifications out of the 276 dataset records utilized in this study. The accurate classification consisted of 237 positive and 37 negative cases, whereas the inaccurate classification consisted of 1 negative case as positive and 1 positive case as negative. The model accuracy, precision, recall, f1_score, and AUC were 97.2%, 100%, 95%, 97.3%, and 100%, respectively.

Out of the 276 dataset records utilized in this study, the AdaBoost model produced 257 accurate classifications and 19 wrong ones. The accurate classification consisted of 232 positive and 25 negative cases, whereas the inaccurate classification consisted of 13 negative cases as positive and 6 positive cases as negative cases. The model accuracy, precision, recall, f1_score, and AUC were 94%, 93.1%, 95%, 94%, and 98%, respectively. The results of the area under the ROC curves showed that random forest and gradient boosting had the best AUC and outperformed the other third ensemble model, as presented in Fig. 9.

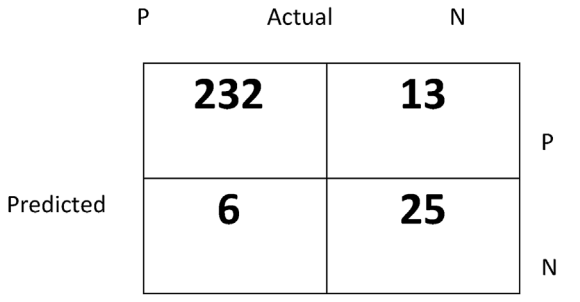
This research focuses on developing a meta-ensemble model for predicting the occurrence of lung cancer. The conclusion of the experimental results after applying SMOTE with 5-fold cross-validation showed that the gradient boosting model achieved a maximum performance of 100% when the three ensembles were used as base classifiers and



(a) *Random Forest Confusion Matrix*



(b) *Gradient Boosting Confusion Matrix*



(c) *AdaBoost Confusion Matrix*

Fig. 7. Results of confusion matrix for ensemble models performance evaluation.

Table 3. Results of meta-ensemble accuracy.

Classifiers		Accuracy (%)
Base	Meta	
RF GB AdaBoost	RF	99.1
RF GB AdaBoost	GB	100
RF GB AdaBoost	AdaBoost	98.1
GB AdaBoost	RF	96.3
RF AdaBoost	GB	100
RF GB	Adaboost	97.2

Table 4. Results of validation of ensemble models.

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1_score (%)	AUC (%)
Random Forest	99.1	100	98.2	99.1	100
Gradient Boosting	97.2	100	95	97.3	100
AdaBoost	94	93.1	95	94	98

the gradient boosting was used as a meta-classifier. The main contribution of this proposed work is the use of the meta-ensemble method to improve the prediction performance of lung cancer occurrence in individuals. The meta-ensemble method presented in this study can be integrated into the existing health information system to improve the decision-making process of medical experts regarding the risk of lung cancer among patients.

The weakness of this research paper is also pointed out. This study used a publicly accessible dataset rather than one sourced from a hospital unit or institute, which could have provided more diverse and detailed data. Moreover, obtaining access to sensitive medical information is challenging due to privacy concerns. Nevertheless, the dataset we utilized had valuable features that enabled us to produce dependable and accurate research outcomes.

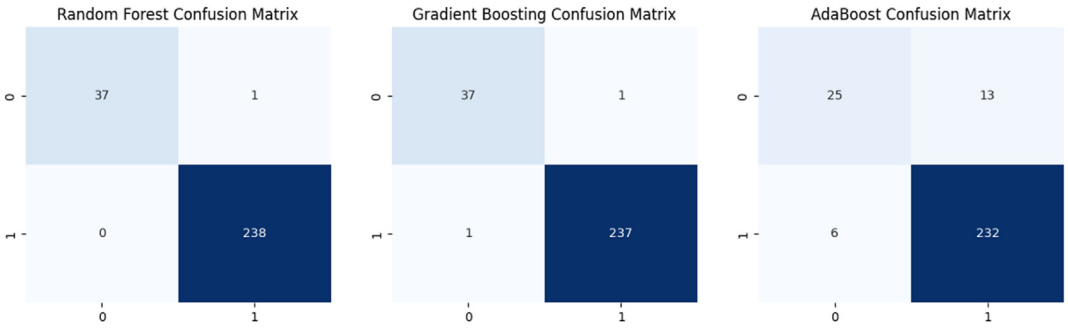


Fig. 8. Results of confusion matrix using seaborn.

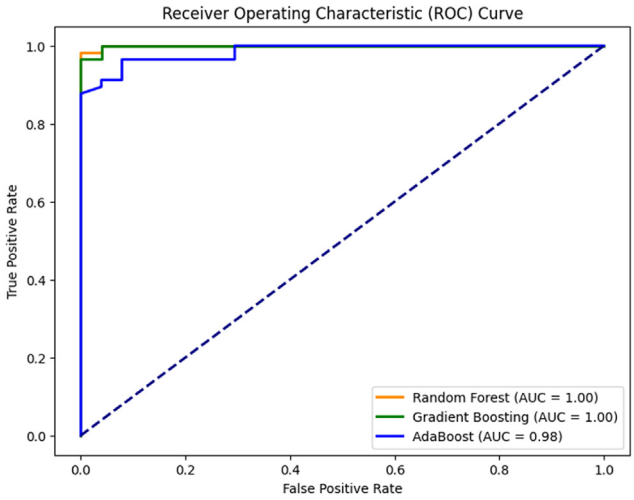


Fig. 9. Models evaluation based on AU (ROC) curves.

4. Conclusion

This research paper exploits ensemble learning models to develop a meta-ensemble method that identifies individuals showing signs of lung cancer by considering various features, such as symptoms. Three ensemble models, including RF, GB, and AdaBoost, were evaluated in terms of accuracy, precision, F1 score, recall, and AUC in the Python simulation environment. It was deduced from this study that according to the description of the dataset, the majority of patients who have lung cancer risk were part of the patients within the age range of 40–79 years of age but patients within the age range of 55–68 years have the highest occurrence of lung cancer; males have a higher occurrence of lung cancer than females; the higher the patients who are smoking, have yellow fingers, chronic disease, anxiety, fatigue, chest pain, and shortness of breath the higher the risk of having lung cancer; patients with peer pressure and coughing affect lung cancer risk; patients who have wheezing, allergy, alcohol consuming, and swallowing difficulty have little or no effect on the risk of lung cancer.

It was also concluded from the results that adopting Gradient Boosting as a meta-ensemble model achieved the overall best meta-ensemble accuracy, which is shown in Table 3, followed by Random Forest and then AdaBoost. Additionally, the results of the validation metrics were also compared, as seen in Table 4. Random Forest outperformed the other two ensembles in terms of accuracy, recall, and F1 score, followed by Gradient Boosting and AdaBoost. In terms of precision and AUC, both Random Forest and Gradient Boosting have the highest performance, followed by AdaBoost. Therefore, this research's results outperformed better compared to the models of references shown in Table 1.

Future work can focus on other diseases such as diabetes prediction, respiratory diseases through lung sound analysis employing deep neural networks, heart failure prediction, and tackling other ailments using machine learning algorithms for the betterment of humanity.

References

[1] Rahman A, Muniyandi R, Albashish D. Artificial neural network with Taguchi method for robust classification model to improve classification accuracy of breast cancer. 2021. p. 1–27. https://doi.org/10.7717/peerj-cs.344.
[2] Jain S, Nehra M, Kumar R, Dilbaghi N, Hu TY, Kumar S, et al. Internet of Medical Things (IoMT)-integrated biosensors for point-of-care testing of infectious diseases. Biosens Bioelectron 2021;179:113074. https://doi.org/10.1016/j.bios.2021.113074. ISSN 0956-5663.
[3] Ferone G, Lee MC, Sage J, Berns A. Cells of origin of lung cancers: lessons from mouse studies. 2020. p. 1017–32. https://doi.org/10.1101/gad.338228.120.
[4] Suvarchala V, Subbareddy PV, Madala SR. Lung cancer prediction using machine learning methodologiesvol. 8; 2021. p. 1265–72.
[5] Marino P, Mininni M, Deiana G, Marino G, Divella R, Bochicchio I, et al. Healthy Lifestyle and Cancer Risk: Modifiable Risk Factors to Prevent Cancer. Nutrients 2024; 16(6):800.
[6] Li W, Dong S, Wang H, Wu R, Wu H, Tang ZR, et al. Risk analysis of pulmonary metastasis of chondrosarcoma by establishing and validating a new clinical prediction model: a clinical study based on SEER database. BMC Muscoskel Disord 2021;22(1):529.
[7] Danlos FX, Voisin AL, Dyevre V, Michot JM, Routier E, Taillade L, et al. Safety and efficacy of anti-programmed death 1 antibodies in patients with cancer and pre-existing autoimmune or inflammatory disease. Eur J Cancer 2018;91: 21–9.
[8] Zhang JJ, Dong X, Liu GH, Gao YD. Risk and protective factors for COVID-19 morbidity, severity, and mortality. Clin Rev Allergy Immunol 2023;64(1):90–107.
[9] Scobie H. Understanding lung cancer screening participation hannah scobie BSc (Hons), MSc submitted in fulfillment of the requirements for the degree of doctor of philosophy institute of health and wellbeing, college of medical and veterinary life sciences, Univ. 2021.
[10] Lu T, Yang X, Huang Y, Zhao M, Li M, Ma K, et al. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. Cancer Manag Res 2019: 943–53. Jan 21.
[11] Huang J, Mphil YD, Tin MS, Mph VL, Ngai CH, Zhang L, et al. Distribution, Risk Factors, and Temporal Trends for Lung Cancer Incidence and Mortality. Chest 2022;161: 1101–11. https://doi.org/10.1016/j.chest.2021.12.655.
[12] Klein CA. Cancer progression and the invisible phase of metastatic colonization. Nat Rev Cancer 2020;20(11):681–94.
[13] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin 2018;68(6):394–424.
[14] Nageswaran S, Arunkumar G, Bisht AK, Mewada S, Kumar JNVRS, Jawarneh M, et al. Lung cancer classification and prediction using machine learning and image processing. 2022.

[15] Li W, Liu Y, Liu W, Tang ZR, Dong S, Li W, et al. Machine learning-based prediction of lymph node metastasis among osteosarcoma patients. Front Oncol 2022;12:797103.
[16] Zaman N, Gaur L, Humayun M, editors. Approaches and applications of deep learning in virtual medical care. IGI Global; 2022.
[17] Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. Int J Intell Netw 2022;3:58–73.
[18] Mokari A, Guo S, Bocklitz T. Exploring the steps of infrared (IR) spectral analysis: Pre-processing,(classical) data modeling, and deep learning. Molecules 2023;28(19):6886.
[19] Jayatilake SM, Ganegoda GU. Involvement of machine learning tools in healthcare decision-making. J Healthc Eng 2021;2021:6679512. 20 pages, https://doi.org/10.1155/2021/6679512.
[20] Shehab M, Abualigah L, Shambour Q, Abu-Hashem MA, Shambour MK, Alsalibi AI, et al. Machine learning in medical applications: A review of state-of-the-art methods. Comput Biol Med 2022;145:105458.
[21] Bokefode J, Rao MVP, Komarasamy G. ScienceDirect ScienceDirect Ensemble Deep Learning Models for Lung Cancer Diagnosis in Histopathological Application Images Ensemble Deep Learning Models for Lung Cancer Diagnosis in Histopathological Images. Procedia Comput Sci 2022;215:471–82. https://doi.org/10.1016/j.procs.2022.12.049.
[22] Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. J Biomed Inf 2022;127:103996.
[23] Safiyari A, Javidan R. Predicting lung cancer survivability using ensemble learning methods. 2017. p. 684–8.
[24] Dritsas E, Trigka M. Lung cancer risk prediction with machine learning models. 2022.
[25] Faisal MI, Bashir S, Khan ZS, Khan FH. Trends Eng Sci Technol 2018:1–4.
[26] Setiawan W, Pramudita YD. Mulaab, lung cancer classification using random oversampling and gradient boosted decision tree, vol. 16; 2023. p. 273–9.
[27] Zamzam YF, Saragih TH, Herteno R, Turianto D. Comparison of CatBoost and random forest methods for lung cancer classification using hyperparameter tuning bayesian optimization- based, vol. 6; 2024. p. 125–36. https://doi.org/10.35882/jeeemi.v6i2.382.
[28] Mamun M, Farjana A, Al Mamun M, Ahammed MS. Lung cancer prediction model using ensemble learning techniques and systematic review analysis. 2022. p. 187–93. https://doi.org/10.1109/AIIOT54504.2022.9817326.
[29] Gregory RH, Roffman DA, Decker R, Deng J. A multiparameterized artificial neural network for lung cancer risk prediction. 2018. p. 1–13.
[30] Hao L, Huang G. An improved AdaBoost algorithm for identification of lung cancer based on electronic nose. Heliyon 2023;9:e13633. https://doi.org/10.1016/j.heliyon.2023.e13633.
[31] Subramanian RR, Mourya RN, Reddy VPT, Reddy BN. Lung cancer prediction using deep learning framework, vol. 13; 2020. p. 154–60.
[32] Bhattacharjee A, Murugan R, Soni B. Ada-GridRF: A Fast and Automated Adaptive Boost Based Grid Search Optimized Random Forest Ensemble model for Lung Cancer Detection. Phys Eng Sci Med 2022;45:981–94. https://doi.org/10.1007/s13246-022-01150-2.
[33] Aggarwal P, Marwah N, Kaur R, Mittal A. Lung cancer survival prognosis using a two-stage modeling approach. Multimed Tool Appl 2024:1–28. https://doi.org/10.1007/s11042-024-18280-2.
[34] Liu S, Yao W. Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. BMC Bioinf 2022:1–11. https://doi.org/10.1186/s12859-022-04689-9.
[35] Cardis E, Richardson D. Prediction and classification of lung cancer using machine learning techniques prediction and classification of lung cancer using machine learning techniques. 2021. https://doi.org/10.1088/1757-899X/1099/1/012059.
[36] Shaik AB, Srinivasan S. A brief survey on random forest ensembles in the classification model. In: International Conference on Innovative Computing and Communications. 2; 2019. p. 253–60.
[37] Mienye ID, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects10; 2022. p. 129–49. 99.
[38] Ahmad I, Basheri M, Iqbal MJ, Rahim A. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection30; 2018. p. 33789–95. 6.
[39] Akhiat Y, Manzali Y, Chahhou M, Zinedine A. A new noisy random forest-based method for feature selection. Cybern Inf Technol 2021;21(2):10–28.
[40] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67.
[41] Zhang Y, Liu J, Shen W. A review of ensemble learning algorithms used in remote sensing applications. Appl Sci 2022;12(17):8654.
[42] Ribeiro MH, dos Santos Coelho L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput 2020;86:105837.
[43] Devi TJ, Gopi A. An Efficient Novel Approach for Early Detection of Mental Health Disorders through Distributed Machine Learning Paradigms from Public Societal Communication. Int J Intell Syst Appl Eng 2024;12(2):767–78.
[44] Sathishkumar R, Karthikeyan T, Shamsundar SM. Ensemble Text Classification with TF-IDF Vectorization for Hate Speech Detection in Social Media. In: 2023 International Conference on System, Computation, Automation and Networking (ICSCAN); 2023. p. 1–7.
[45] Wang F, Li Z, He F, Wang R, Yu W, Nie F. Feature learning viewpoint of AdaBoost and a new algorithm, vol. 7; 2019. p. 149890–9.
[46] Mehmood Z, Asghar S. Customizing SVM as a base learner with AdaBoost ensemble to learn from multi-class problems: A hybrid approach AdaBoost-MSVM. Knowl Base Syst 2021;217:106845.
[47] Baig MM, Awais MM, El-Alfy ES. AdaBoost-bas+ed artificial neural network learning. Neurocomputing 2017;248:120–6.
[48] Javed A, Zaman M, Uddin MM, Nusrat T. An analysis on python programming language demand and its recent trend in Bangladesh. In: Proceedings of the 2019 8th international conference on computing and pattern recognition; 2019. p. 458–65.
[49] Martelli A, Ravenscroft AM, Holden S, McGuire P. Python in a nutshell. O'Reilly Media, Inc; 2023.
[50] He S, Guo F, Zou Q. MRMD2. 0: a Python tool for machine learning with feature ranking and reduction. Curr Bioinf 2020;15(10):1213–21.
[51] Raschka S, Patterson J. Nolet, Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. Information 2020;11(4):193.
[52] Heydarian M, Doyle TE, Samavi R. MLCM: multi-label confusion matrix, vol. 10; 2022. p. 19083–95.
[53] AlSlaiman M, Salman MI, Saleh MM, Wang B. Enhancing false negative and positive rates for efficient insider threat detection. Comput Secur 2023;126:103066.
[54] Tharwat A. Classification assessment methods. Appl Comput Inform 2020;17(1):168–92.