Journal of Pathology Informatics 14 (2023) 100307

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer

Sarreha Tasmin Rikta [a], Khandaker Mohammad Mohi Uddin [a,*], Nitish Biswas [a], Rafid Mostafiz [b], Fateha Sharmin [c], Samrat Kumar Dey [d]

[a] Department of Computer Science and Engineering, Dhaka International University, Dhaka 1205, Bangladesh
[b] Institute of Information Technology, Noakhali Science and Technology University, Noakhali, Bangladesh
[c] Department of chemistry, University of Chittagong, Chittagong, Bangladesh
[d] School of Science and Technology, Bangladesh Open University, Gazipur 1705, Bangladesh

ARTICLE INFO

Keywords:
Lung cancer
Explainable machine learning
ROS
SHAP
GBM
Mobile app

ABSTRACT

Lung cancer has been the leading cause of cancer-related deaths worldwide. Early detection and diagnosis of lung cancer can greatly improve the chances of survival for patients. Machine learning has been increasingly used in the medical sector for the detection of lung cancer, but the lack of interpretability of these models remains a significant challenge. Explainable machine learning (XML) is a new approach that aims to provide transparency and interpretability for machine learning models. The entire experiment has been performed in the lung cancer dataset obtained from Kaggle. The outcome of the predictive model with ROS (Random Oversampling) class balancing technique is used to comprehend the most relevant clinical features that contributed to the prediction of lung cancer using a machine learning explainable technique termed SHAP (SHapley Additive exPlanation). The results show the robustness of GBM's capacity to detect lung cancer, with 98.76% accuracy, 98.79% precision, 98.76% recall, 98.76% F-Measure, and 0.16% error rate, respectively. Finally, a mobile app is developed incorporating the best model to show the efficacy of our approach.

Introduction

The most lethal kind of cancer, globally, is lung cancer. It is one of the main causes of cancer fatalities in both women and men.[1,2] One type of cancer that begins in the lungs is lung cancer. When the body's cells start to proliferate out of control, cancer develops. Lung cancer frequently develops over a long period of time and primarily affects persons between the ages of 55 and 65.[3] Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the 2 main kinds of lung cancer. NSCLC accounts for about 80%–85% of lung cancer cases. Most often, smokers or former smokers develop this type of lung cancer. More than 85% of lung cancer cases are caused by current or previous cigarette smokers.[4] Compared to other forms of lung cancer, it is more prevalent in women than men and is more likely to affect younger people. Contrarily, SCLC, also known as oat cell carcinoma, accounts for 10%–15% of all cases of lung cancer. The growth rate of SCLC and the formation of big tumors that have the potential to spread far throughout the body are virtually directly correlated with cigarette smoking. They frequently begin in the bronchi in the center of the chest. The overall number of cigarettes smoked has an impact on the death rate from lung cancer.[5] According to the World Health Organization (WHO), lung cancer was the leading cause of cancer-related death in 2020, taking 1.80 million lives.

Both the mortality rate and the number of people affected by this disease are predicted to rise along with the increase in the world's population. The 5-year survival rate for lung cancer is only 18%, which highlights the importance of early detection and diagnosis. Medical imaging, such as computed tomography (CT) scans, is commonly used in the diagnosis of lung cancer. However, the interpretation of medical images can be challenging and time-consuming for radiologists. This fatality rate can be decreased with early detection and treatment. Machine learning algorithms can be quite beneficial in that circumstance to correctly forecast the malignancy.[6–8] However, due to the complexity and black-box nature of many machine learning algorithms, it is difficult to understand how the model is making predictions or decisions and to identify potential errors or biases. This can be a major concern in fields such as healthcare, finance, and criminal justice, where the consequences of model errors can be severe. Another limitation of black-box models is that they are often sensitive to the choice of hyperparameters, which can make them difficult to optimize and generalize to new data. Additionally, black-box models can be prone to overfitting, which can lead to poor performance on unseen data. Overall,

* Corresponding author at: Department of Computer Science and Engineering, Dhaka International University, Dhaka 1205, Bangladesh.
E-mail addresses: jilanicsejnu@gmail.com (K.M.M. Uddin), samrat.sst@bou.ac.bd (S.K. Dey).

http://dx.doi.org/10.1016/j.jpi.2023.100307
Received 30 January 2023; Accepted 20 March 2023
Available online 24 March 2023

black-box models have been very successful in various applications, but the lack of interpretability and transparency is a major concern. Explainable machine learning (XML) is a new approach that aims to provide transparency and interpretability for black-box models, which can help overcome these limitations.

Due to this, post hoc techniques have recently become increasingly popular as a solution to the problem of presenting black-box models in a way that is understandable by humans. Such explanations are frequently used to assist domain experts in finding discriminatory biases in black-box models.[9,10]

Local, model-agnostic methods that concentrate on explaining specific predictions of a given black-box classifier, such as LIME[11] and SHAP,[12] are among the most well-known of these techniques. These techniques produce perturbations of a particular instance in the data and track the impact of these perturbations on the black-box classifier's output in order to quantify the contribution of individual features to a given prediction. These approaches have been employed in a variety of fields, including law, medicine, finance, and science[13–15] due to their generality, to explain a variety of classifiers, including neural networks and sophisticated ensemble models.

The 4 explainable principles,[16] however, are used in this experiment to forecast lung cancer. Transparency is the first major principle that explains models in a clear and understandable way, for example, by highlighting the most important features that led to a certain diagnosis. This can be achieved by using techniques such as feature importance, feature selection, and model interpretability methods such as LIME, SHAP, and others. According to the second principle named Fairness, the model should not discriminate against certain groups of patients based on factors such as age, race, or gender. This can be achieved by using fairness-aware algorithms, such as those that explicitly optimize for group fairness metrics, or by using bias correction methods, such as re-sampling or adversarial training. And the third principle is Robustness. According to this principle, the model should be robust to small changes in the input data and produce consistent predictions even when presented with new or unseen data. This can be achieved by using techniques such as cross-validation, regularization, and ensembling to improve model generalization. In accordance with the final tenet, accountability, the model should be able to provide an explanation for its predictions in case of errors or mistakes, and it should be possible to understand the causes of these errors and take corrective measures. This can be achieved by using techniques such as model monitoring, model auditing, and model governance to ensure that the model is behaving as expected and that any issues or biases are identified and addressed in a timely manner. The above explainable machine learning principles will help to build a trustable and understandable model which can help to improve the accuracy of predictions and ensure that the model is fair, robust, and accountable.

However, we proposed SHAP (SHapley Additive exPlanation) values in this experiment to increase trust, responsibility, debugging, and many other tasks. Game theory concepts[17] and local explanations are used to form the foundation of SHAP. Explainable strategies have been highlighted in renowned journals and are also gaining popularity in medical applications. Cosgriff and Celi[18] show how to analyze deep neural network models using explanatory methodologies using high-frequency electronic patient records. Explanatory models were described in Lundberg et al.,[19] as a way to supplement ML models in forecasting mortality for patients with kidney failure. Nowadays, image data analysis, X-rays, CT scans, ultrasounds, and other imaging techniques use explicable models. Lundberg et al.[20] explain how models forecast hypoxemia during surgery work. The list of techniques used in medicine is more thoroughly discussed in Singh et al.[21] In addition, research into lung conditions is ongoing. Numerous articles have been written about using artificial intelligence (AI) to treat lung conditions. Xi et al.[22] employed exhaled aerosols to detect lung structural illness using machine learning algorithms like Random Forests (RF) and Support Vector Machines. In the extensive investigation, RF models were also employed.[23] To find lung cancer, scientists suggest a handmade e-nose device. Even though there has been a lot of research on this subject, we are still driven to work with lung cancer. This article is focused on the explainability of models used in lung cancer so that people may comprehend how a model works internally and have faith in the experiment's prediction. The proposal's summary is shown in Fig. 1.

Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LightGBM) are 3 ensemble-based classifiers that have been used in this study. Among all the classifiers, GBM has the best accuracy, coming up at 98.76%. Finally, a smartphone app is developed to integrate the best model. The method for applying the proposal is shown in Fig. 2.

The following are the contributions to the proposed research efforts for lung cancer:

1. GBM achieves a higher accuracy of 98.76% in this investigation.
2. For the best outcomes, 3 classifiers—XGBoost, LightGBM, and GBM—are used in this scenario.
3. Data balance techniques, feature scaling, PCA (Principal Component Analysis), and hyperparameter tuning have been used to attain the highest level of accuracy.
4. SHAP is utilized hereto make the gradient-boosting output understandable, meaningful, and trustworthy to humans.
5. Finally, a user-friendly smart phone application that can calculate the result based on real-time inputs has been developed.

There are 4 sections for the remaining portions of this research project. Section 2 demonstrates the related work and Section 3 provides the materials and methods. The 4 subsections in Section 3 are dataset description and data pre-processing, PCA and Hyperparameter tuning, machine learning models, andSHAP. The analysis and discussion of the results are covered in Section 4 which contains 5 subsections. These subsections are environmental setup, classification accuracy, model evaluation, SHAP result analysis, and the creation of the mobile app. The task is finally finished in Section 5.

Related work

There has been a growing interest in the use of eXplainable Machine Learning (XML) techniques for lung cancer prediction in recent years. Some notable studies in this field include:

In one study, Masrur Sobhan and Ananda Mohan Mondal[24] proposed a pathway to identify significant lung cancer class- and patient-specific genes that could support the development of effective medicines for lung cancer patients. They used the 2 SHAP variants known as "tree explainer" and "gradient explainer," for which the classification algorithms "tree-based classifier," XGBoost, and "deep learning-based classifier," convolutional neural network, respectively, were applied. The class-specific top 100 genes and the differentially expressed genes, both of which are population-based biomarkers. Few genes were found to be shared by the patients, indicating that each individual with lung cancer is represented by a different set of patient-specific genes that were found. This test demonstrates that XGBoost achieves 96.3% accuracy.

In this study,[25] machine learning (ML) models are used to predict the length of stay (LOS) for patients with lung cancer. The methodology is forth to address imbalanced datasets for classification-based methods employing electronic medical records (EHR). They forecasted the average length of stay (LOS) for ICU patients with lung cancer using the MIMIC-III dataset and supervised ML algorithms. The Random Forest (RF) Model performed better than other models during the 3 stages of the framework and delivered the expected results. They described the predictive model's (RF) outcome using the SMOTE class balance technique to comprehend the most significant clinical factors that contributed to predicting lung cancer LOS using the RF model utilizing SHAP.

Another study by Jamie et al.[26] used 3 XAI techniques—SHAP, LIME, and Scoped Rules—to show how usable it is to add an explainable tertiary appendix to ML models and to give data interpretability for large-scale EHR datasets. The Simulacrum, a synthetic dataset produced by Health Data

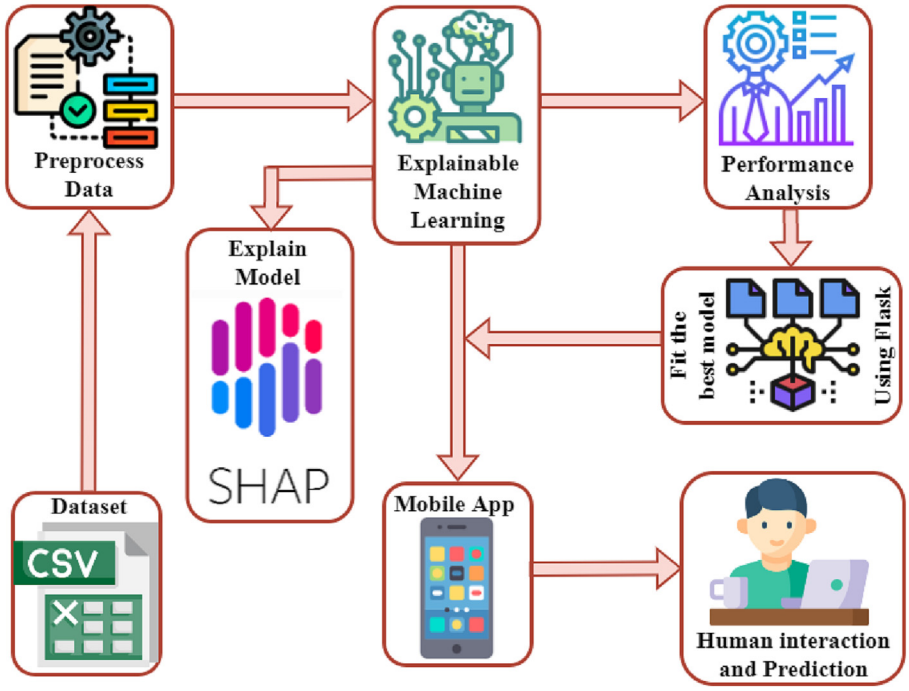


Fig. 1. Overview of the proposed work.

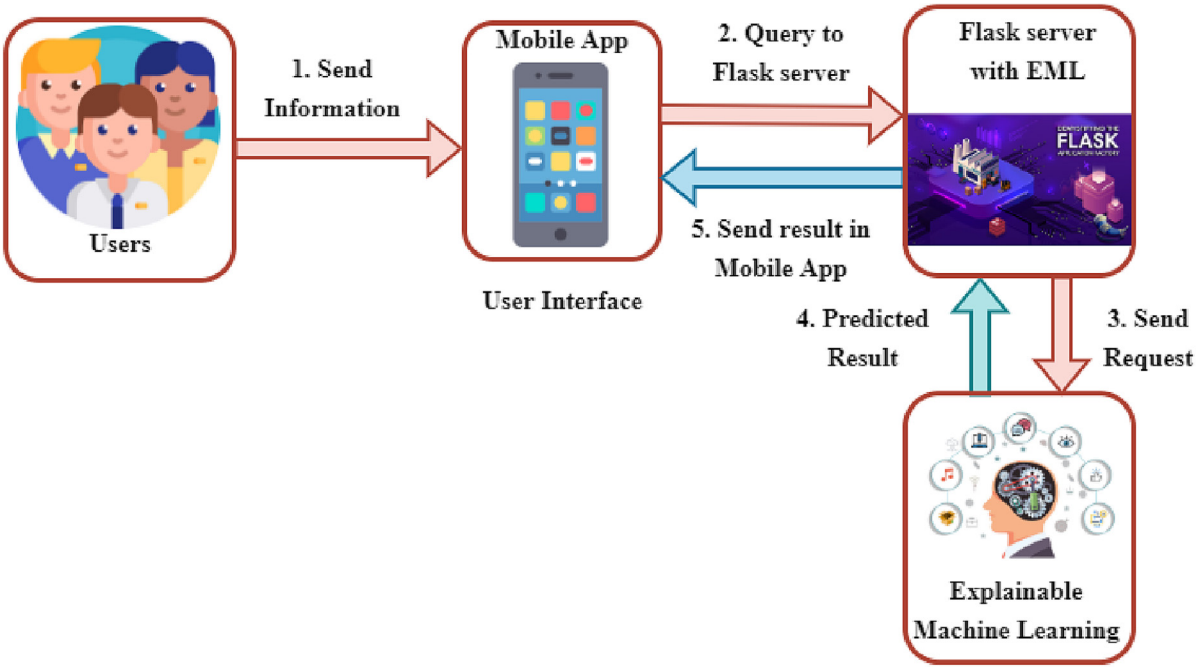


Fig. 2. Working flow of the mobile application.

Insight CiC using anonymized cancer data supplied by Public Health England's National Cancer Registration and Analysis Service (NCRAS), served as the source of the data for this experiment.They contrasted EHR features based on the weighted prediction relevance calculated by XAI models. However, in this study, 3 classifiers—Logistic Regression, XGBoost, and EBM—were utilized, with XGBoost displaying the greatest performance in terms of classification accuracy.

By using the example of models used to assess lung cancer risk in lung cancer screening by low-dose computed tomography, Katarzyna et al.[27] proposed selected approaches from the XAI field in another work. This method aids in a better understanding of the comparison of the 3 lung cancer risk prediction models used in lung cancer screening, namely the BACH model, PLCOm2012 model, and LCRAT model. The study's model performance and accuracy are not discussed by the authors. They only concentrate on comprehending how the models act for various patients. For this investigation, they employed the domestic lung cancer database.

Elias Dritsas and Maria Trigka[28] employed a variety of machine learning classifiers, including NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF, and AdaBoostM1, to identify people who are at a high risk of developing lung cancer. In order to identify the model with

the highest predictive performance, these classifiers are assessed in terms of accuracy, precision, recall, F-Measure, and AUC. Their source for the dataset was obtained from Kaggle. The RotF model from this experiment performs at the highest level.

Another study on lung cancer research was carried out by Muntasir et al.[29] in which they examined a number of earlier studies that were concerned with lung cancer prediction models and contrasted the results with their models. They created the XGBoost, LightGBM, AdaBoost, and bagging ensemble learning approaches, among others, to forecast lung cancer. The model validation approach was carried out using K-fold 10 cross-validation. The best accuracy in this experiment is 94.42%, which is achieved with XGBoost.

Patra[30] examined various machine learning classifiers, including Radial Basis Function Network (RBF), K-Nearest Neighbors (KNN), J48, Support Vector Machine (SVM), Logistic Regression, Artificial Neural Network (ANN), Nave Bayes, and Random Forest, for predicting lung cancer. The dataset, which includes 32 occurrences and 57 attributes, was gathered from the "UCI repository." RBF achieved an accuracy of 81.25%, which the authors deemed to be higher than all the other algorithms.

Another study by Sim et al.[31] suggested a study of health-related quality of life (HRQOL) in 5-year lung cancer survival prediction using a variety of machine learning models, including Decision Tree, Logistic Regression, Bagging, Random Forest, and AdaBoost. To assess model performance, 2 different feature sets were utilized with K-fold 5 cross-validations. Data from 809 lung cancer surgery survivors who underwent surgery were compared to the model performances. This experiment's findings indicated that AdaBoost had the highest accuracy of 94.8%.

Overall, these studies mainly focus on demonstrating the potential of using explainable machine learning for lung cancer research, as it can improve the performance and interpretability of machine learning models. They provide interpretable insights into the underlying mechanisms of the disease and the factors that contribute to its development.

Materials and methods

This section explains our recommended approach and the techniques utilized to detect lung cancer. Fig. 3 depicts the proposed approach for predicting lung cancer. The suggested approach looked into label encoding, feature selection methods with standard scalars, where each feature's values in the data have a zero mean and unit variance, Principal Component analysis (PCA), which compresses the data, hyperparameter tuning with grid search, which is used to determine the best model, and performance

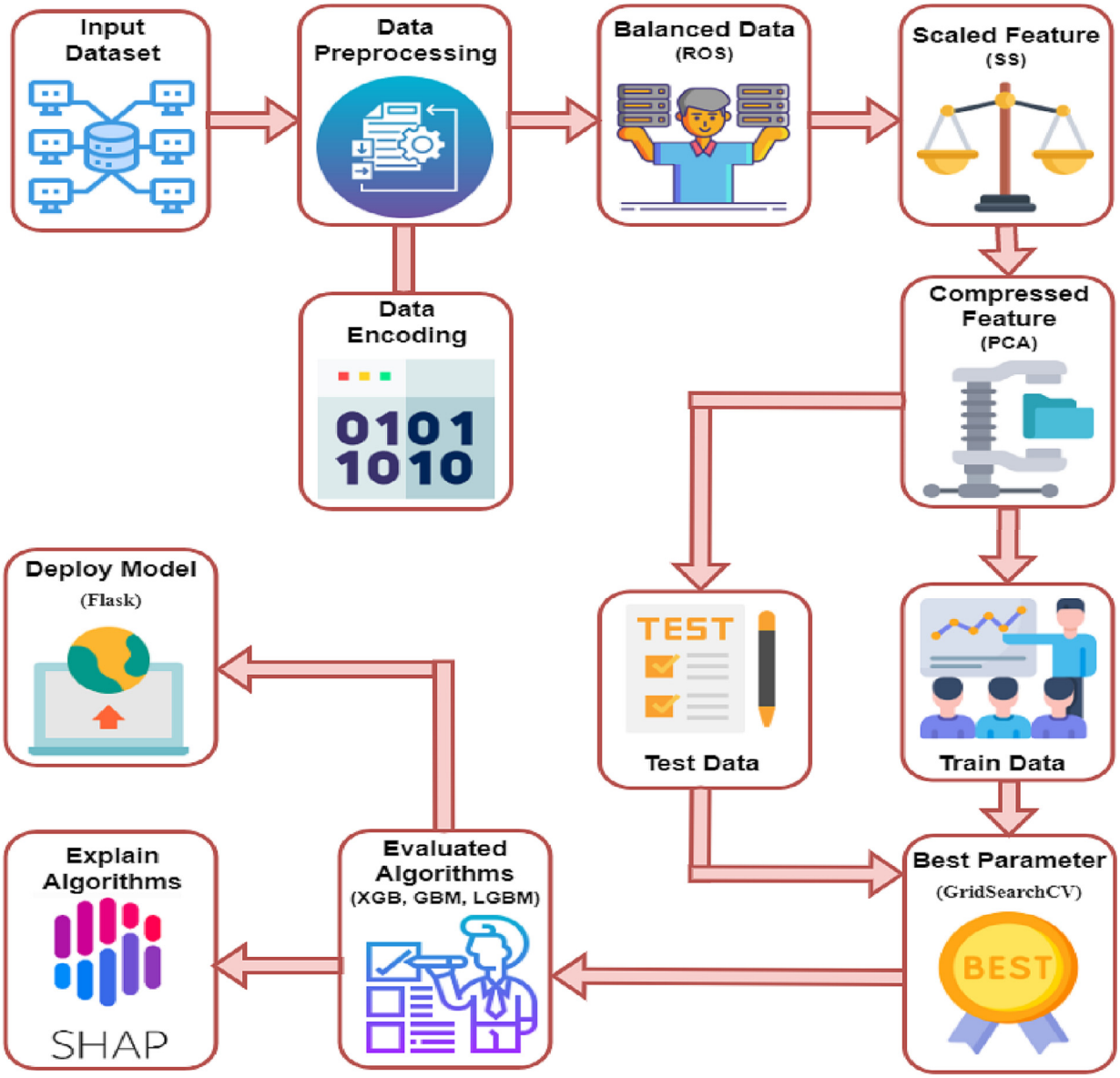


Fig. 3. Working procedure of proposed methodology.

matrices, which are used to improve the efficacy of a classification model. This experiment's results showed that using all machine learning classifiers to predict lung cancer is a promising direction to go in. Below is a depiction of the algorithm used in this study for simplicity of understanding.

Algorithm: Working Procedure of XML Lung Cancer Prediction
Input: Kaggle Lung Cancer Dataset
Output: Predicted value of XML Lung Cancer (Yes or No)

1. ***Begin***
2. data ← load dataset;
3. shap ← load shap;
4. **Procedure DO_EXPLAINABLE(** model, x **)**
5. explainer ← shap.Explainer(model, x);
6. shap_values ← explainer(x);
7. ***if*** plots ***is equal*** bar
8. shap.bar(shap_values);
9. ***else if*** plots ***is equal*** beeswarm
10. shap.beeswarm(shap_values);
11. ***else***
12. shap.waterfall(shap_values);
13. **end procedure**
14. ***pre-processing:***
15. ***if***data.dtypes ***is*** equal object or string
16. encoding the data;
17. x ← data.drop[lung];
18. y ← data.lung;
19. ***balancing_data:***
20. x_os, y_os = RandomOverSampling(x,y);
21. ***feature_scaling:***
22. scaled_x ← scaling_the_feature(x_os);
23. ***feature_optimizing:***
24. pca_x = PCA(n_components = 9).fit(scaled_x);
25. x1, x2, y1, y2 ← split_data of pca_x and y_os;
26. ***for*** i in ***range***(***len***(models)):
27. checking for HTP;
28. model ← train_model using x1and y1;
29. predict ← testing_model using x2 and y2;
30. computes performance evaluation metrics;
31. **DO_EXPLAINABLE(**model, x**);**
32. ***End***

Dataset description and data pre-processing

A dataset called "Lung Cancer" was obtained from the Kaggle[32] and contains 309 occurrences and 16 attributes, of which 15 attributes are predictive and 1 is the class attribute. Lung cancer is the class attribute, and the predictive attributes are, in order, gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness of breath, swallowing trouble, and chest pain. Table 1 describes each feature of the dataset.

One of these characteristics is that Gender and Lung Cancer both contain categorical values that have been transformed into numerical values (0,1) during the data pre-processing stage via label encoding. The dataset's noise, missing values or information, and unbalanced data[33] may reduce the accuracy of the result. That is why these undesired items from the dataset were eliminated prior to running the machine-learning model. The best output for the dataset is achieved by data pre-processing. However, this dataset does not contain any missing values, but this dataset was completely imbalanced. Random oversampling (ROS)[34] is used here to address the issue of imbalanced datasets. It involves duplicating examples from the minority class in order to balance the class distribution. This is done by randomly selecting examples from the minority class and adding them to the dataset until the class distribution is balanced. In order to conduct this experiment, the minority class has been increased by 70%. After oversampling, a total of 216 rows in the dataset have a value of 1, indicating that malignancies were discovered, while 270 rows in the dataset have a value of 0, indicating that no cancers were discovered. Fig. 4 shows the imbalanced data before imbalanced and after imbalanced.

Table 1
Descriptions of each characteristic in the dataset.

Attributes name	Description
Gender	This characteristic indicates whether a person is male or female.
Age	The person's age is recorded using this feature.
Smoking	This characteristic indicates whether or not the participant smokes.
Yellow fingers	This characteristic shows if the user has yellow fingers or not.
Anxiety	The presence or absence of anxiety is indicated by this feature.
Peer pressure	This characteristic lets the user know whether they are susceptible to peer pressure or not.
Chronic disease	This feature indicates if the participant has a chronic disease or not
Fatigue	The participant's level of fatigue is indicated by this attribute.
Allergy	This characteristic shows whether or not the person has allergies.
Wheezing	This feature reveals if the individual has wheezing or not.
Alcohol	This feature shows whether or not the person drinks.
Coughing	This feature shows whether the participant has a cough or not.
Shortness of breath	This feature shows whether the participant suffers from shortness of breath or not
Swallowing difficulty	This feature indicates whether the user has swallowing problems or not.
Chest pain	This feature shows whether or not the individual is experiencing chest pain.
Lung cancer	This feature specifies whether or not the individual has been diagnosed with lung cancer or not.

PCA and hyperparameter tuning

The dataset of this experiment consists of 16 features that are highly dimensional. Due to overfitting, these numerous features make it difficult to achieve the optimal outcome. In order to improve the performance of the outcome, Principal Component Analysis (PCA) is applied to the dataset, which reduces the 16 characteristics to 9. PCA is a dimensionality-reduction technique that is frequently used to reduce the dimensionality of big datasets.[35] It is done by condensing a large collection of variables into a smaller one while retaining the majority of the data in the larger set.[36]

However, machine learning employs hyperparameter tuning,[37] where the value of the parameter is chosen before the algorithm is taught. That particular set of hyperparameters maximizes the performance of the model and produces better results with fewer errors by minimizing a preset loss function. GridSearchCV and hyperparameter tuning are consequently merged in this study. GridSearchCV is a method in the scikit-learn library for Python that is used to perform an exhaustive search over a specified

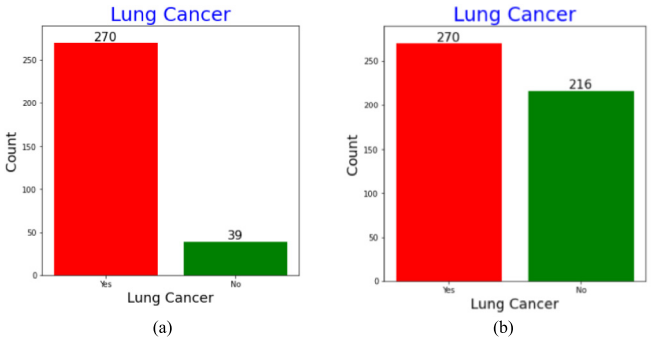


Fig. 4. Data balanced: (a) Before data balancing and (b) after data balancing.

parameter space for an estimator. This technique for hyperparameter tuning does a comprehensive cross-validation search to find the ideal values for the desired hyperparameters.[38] The model evaluates and verifies each unique set of dictionary values.[39] But in order to get the promising accuracy from GridSearchCV, we specified the parameters for GBM, XGBoost, and LightGBM. The best GBM parameters n_estimators = 100, learning_rate = 1, and max_depth = 1 was chosen to yield an accuracy of 98.76%.

GridsearchCV was then used to build XGBoost, with the best parameters being objective = "binary: logistic", random_state = 45, eval_metric = "auc," and n_estimators = 100, producing an accuracy of 98.27%. Following the development of the LightGBM model with default parameters, predictions were made on an unseen test set. But the accuracy was poor. Consequently, the model was built using GridsearchCV, which achieved an accuracy of 98.89% when the ideal values for num_leaves = 31, learning_rate = 1, and n_estimators = 100 were used. The 10-fold cross-validation was employed for conducting this experiment.Thus, the best model with the highest accuracy is selected for each set of hyperparameters.

Machine learning classifiers

Data modeling is carried out here using 3 ensemble-based machine learning algorithms: Gradient GBM, LightGBM, and XGBoost.GBM is the learning process that incrementally fits new models to provide a more precise estimate of the response variable. To create the final forecasts, it aggregates the predictions from many decision trees. Every decision tree's nodes use a distinct subset of information to decide which split is the best. This indicates that no 2 trees are exactly alike, and as a result, various signals can be extracted from the data by each tree. Each subsequent tree also takes into account any blunders or errors produced by the preceding trees. So, each decision tree that comes after it is constructed using the flaws of the prior trees. A gradient-boosting machine algorithm builds the trees in a sequential manner in this way.

Another well-known boosting method is XGBoost. XGBoost is actually just a modified version of the GBM algorithm. The goal of this classifier is to accurately classify data by calculating weak classifiers iteratively.[40] Apply the accuracy and logistic loss criterion to select the best model in the hypothesis space and make the best prediction of the test data under the evaluation criterion using the sample data that are provided that are independent of one another. In fact, it comprises a number of regularization methods that lessen overfitting and enhance performance in general.

Large volumes of data can be handled with ease with LightGBM. The optimal split is chosen by LightGBM using a histogram-based strategy to expedite the training process. Any continuous variable is separated into bins or buckets rather than using individual values. This shortens the training period and uses less memory. However, 3 machine-learning algorithms were examined in this study for their capacity to forecast the emergence of lung cancer, with GBM showing the highest level of accuracy.

SHAP (SHapley Additive explanation)

SHAP is a comprehensive method for analyzing the results of any machine learning model developed by Lundberg et al.[41] The SHAP provides a way to calculate the contribution of each feature and is based on game theory and local explanations. The model generates a prediction value for each prediction sample, and the SHAP value is the score given to each feature in the dataset.[42] In order to support iML, SHAP was created and made available as a set of python tools. For each feature, SHAP provides a list of Shapley values for a particular datum. This is based on the notion that predictions can be described by supposing that each feature is a "player" in a game where the prediction is the payout.[43] The Shapley value, a strategy from coalitional game theory, explains how to equally distribute the "payout" across the characteristics. Numerous distinctive factors will be present in our existing dataset. Each distinctive variable can be viewed as a player in the game theory sense. The benefits of several participants working together to complete a project may be seen in the prediction results generated by utilizing this dataset to train the model. Shapley's value distributes the advantages of cooperation evenly by taking into account each player's contributions. The standard measure of feature relevance merely indicates which features are significant, and its effects on prediction outcomes are unknown. The key benefit of the SHAP value is that it may demonstrate both the positive and negative effects of the impact of the attributes in each sample.

Table 2
Environment setup of the proposed system.

Resource	Details
CPU	Intel® Core™ i-3-1005G1 CPU @ 1.20GHz
RAM	12GB
GPU	Intel® UHD Graphics
Software	Anaconda
Language	Python

Result and discussion

Firstly, the data is split into training (65%) and testing (35%). The optimal model with the highest accuracy is examined using a variety of machine learning methods, including feature scaling, PCA, ROS, and hyperparameter tuning.The best model was selected using all these machine learning techniques.

Environmental setup

This experiment involves some resources. Table 2 presents the materials used for this study's model development.

Classification accuracy

The efficiency of the classification systems is assessed using a number of well-known matrices, such as accuracy, recall (also known as sensitivity), precision, and F1-score.[44] Table 3 displays how well GBM, XGBoost, and LightGBM performed in terms of machine learning techniques. However,

Table 3
Evaluation of explainable machine learning methods.

Methods	Precision	Recall	F_Measure	Accuracy	Error
GBM	98.79%	98.76%	98.76%	98.76%	0.012%
XGB	96.41%	96.27%	96.28%	96.27%	0.037%
LGBM	96.97%	96.89%	96.89%	96.89%	0.031%

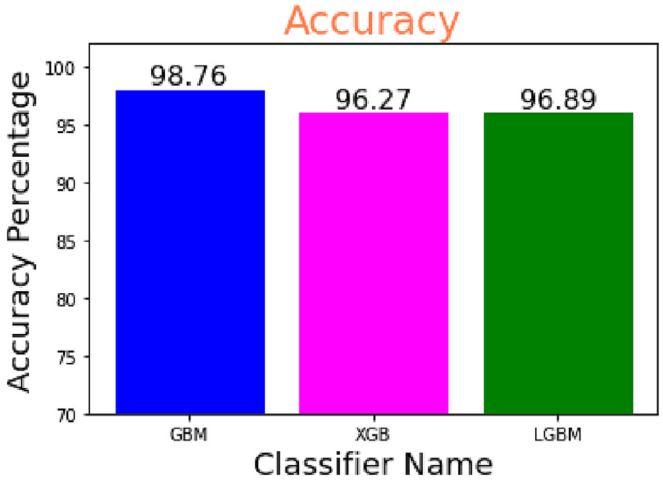


Fig. 5. Accuracy for the GBM, XGB, and LGBM.

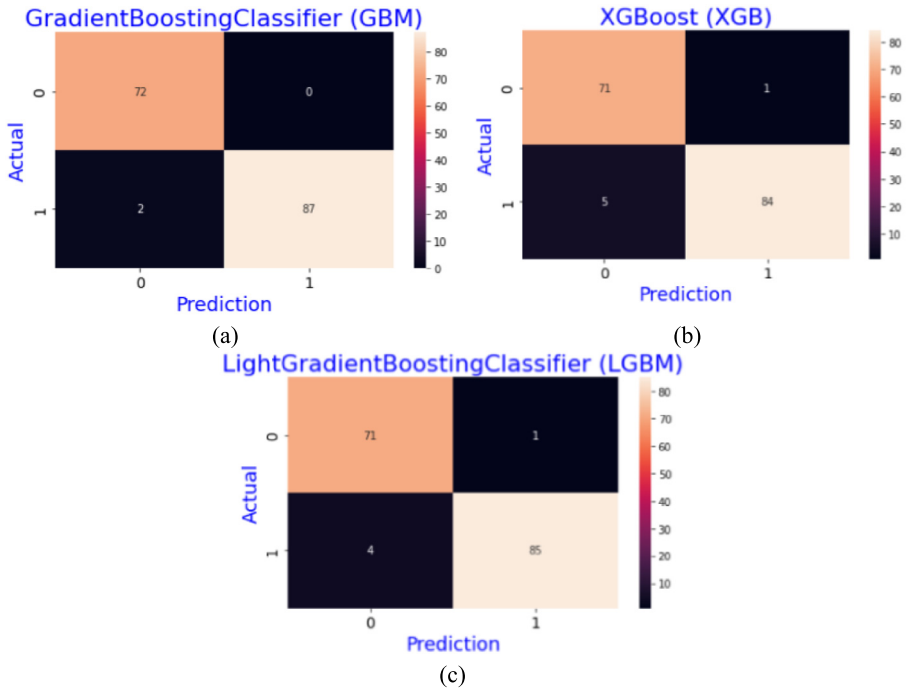


Fig. 6. Confusion matrix for: (a) GBM and (b) XGB, and (c) LGBM.

we have included the Accuracy, Precision, Recall, and F1-Score results for the purpose of observing the models' performance.

The comparison research showed that the GBM classifier outperformed all others with an accuracy of 98.76%. All metrics showed low performance for the XGB classifier. And among these classifiers, LGBM ranked second highest. Since accuracy is a reliable indicator of balanced data, it is considered as the experiment's key performance metric. However, Gradient Boosting Machine achieved the best-balanced accuracy in this investigation, as shown in Fig. 5.

Model evaluation

A key component of creating a powerful machine learning model ismodel evaluation. In this experiment, various evaluation metrics, such as the confusion matrix (accuracy, precision, recall, F measure, Error), and the AUC-ROC curve, are utilized to judge the performance or caliber of the model. The number of true-positive, true-negative, false-positive, and false-negative predictions made by the algorithm is determined by the confusion matrix. True positives are the number of instances where the algorithm successfully identified the positive class, whereas true negatives are the number of instances where the method correctly anticipated the negative class. False positives are the number of occasions when the algorithm predicted a positive class when the actual class was negative, and false negatives are the number of occasions when the system predicted a negative class when the actual class was positive. A variety of performance indicators, including accuracy, precision, recall, and F1 score, can be calculated using the matrix.[45] Fig. 6 shows the confusion matrix of each classifier.

These performance metrics allow us to assess how well our model processed the given data. These evaluation matrices are defined in Eqs (1)–(5).

$$\text{Accuracy}\ (\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{1}$$

$$\text{Precision}\ (\%) = \frac{TP}{TP + FP} \times 100 \tag{2}$$

$$\text{Recall}\ (\%) = \frac{TP}{TP + FN} \times 100 \tag{3}$$

$$\text{F\_Measure}\ (\%) = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \times 100 \tag{4}$$

$$\text{Error}\ (\%) = \frac{FP + FN}{TP + TN + FP + FN} \times 100 \tag{5}$$

Where TP, FP, TN, and FN stand for True Positives, False Positives, and True Negatives, respectively. The AUC-ROC curve is shown in Fig. 7. Here, the AUC-ROC curve is used to show how well the classification model performs on graphs. It is a favored and important statistic for evaluating how well the categorization model is working.

Binary classification issues can be evaluated using the ROC curve as a statistic. This probability curve, which basically distinguishes the "signal" from the "noise," displays the TPR (True positive rate) versus the FPR (False Positive Rate) at different threshold values. The ROC curve is summarized using the Area Under the Curve (AUC), which measures a classifier's capacity to differentiate between classes. The performance of the model in separating the positive and negative classes is inversely correlated with the AUC.

SHAP result analysis

Finally, the values of their explanatory factors are utilized to determine the Shapley value explanations of lung cancer in the test set. In the discipline of machine learning, the more explainable a model is, the simpler it is to understand and comprehend the predictions that have been made. In order to explain the model outputs and determine the extent to which a certain characteristic contributes to the outcomes of a particular event, SHAP is used here. It is a powerful tool for feature importance that allows to understand which features are most important in driving a specific prediction,

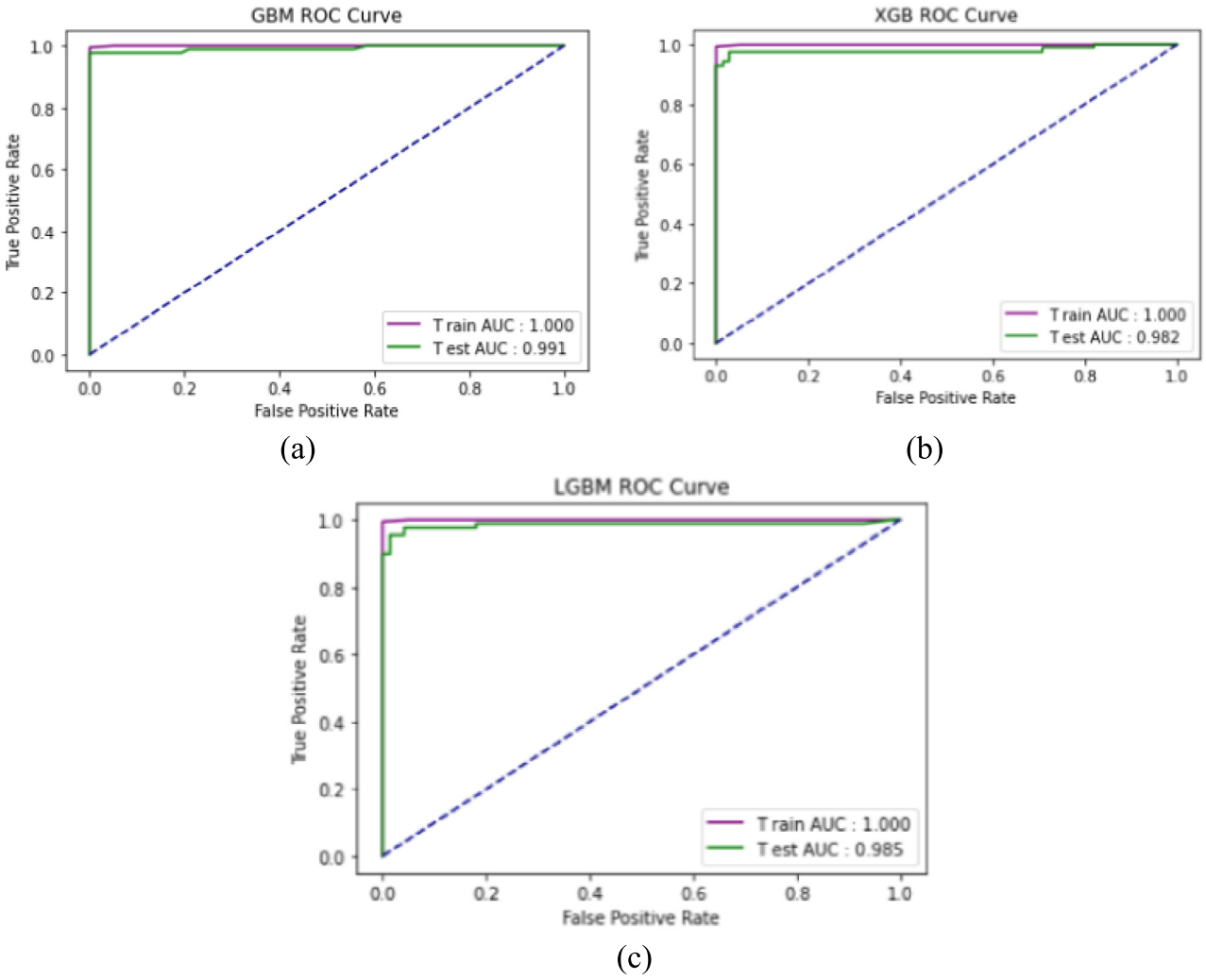


Fig. 7. ROC curve: (a) GBM, (b) XGB, and (c) LGBM.

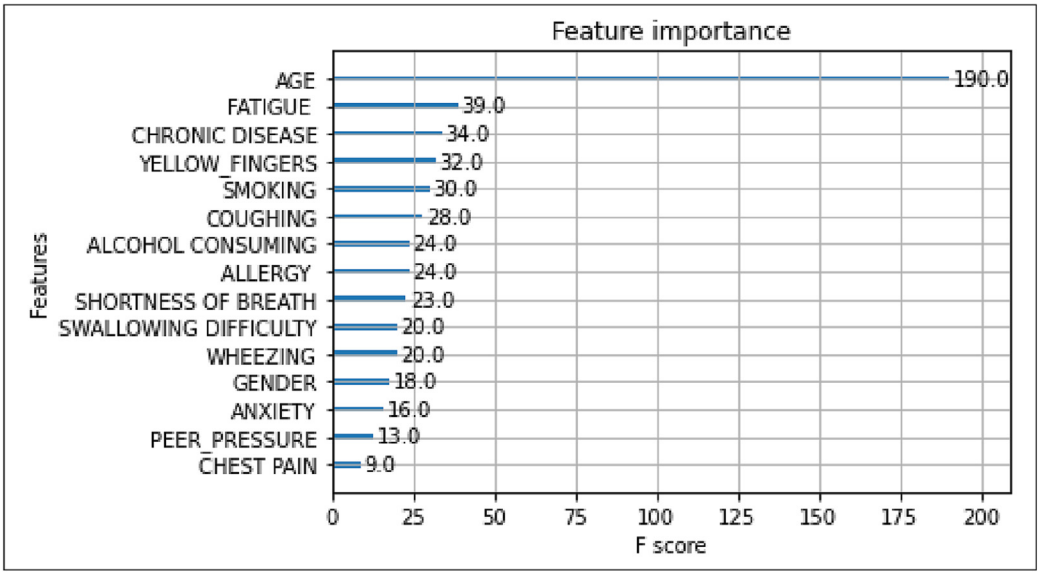


Fig. 8. Features importance for lung cancer prediction.

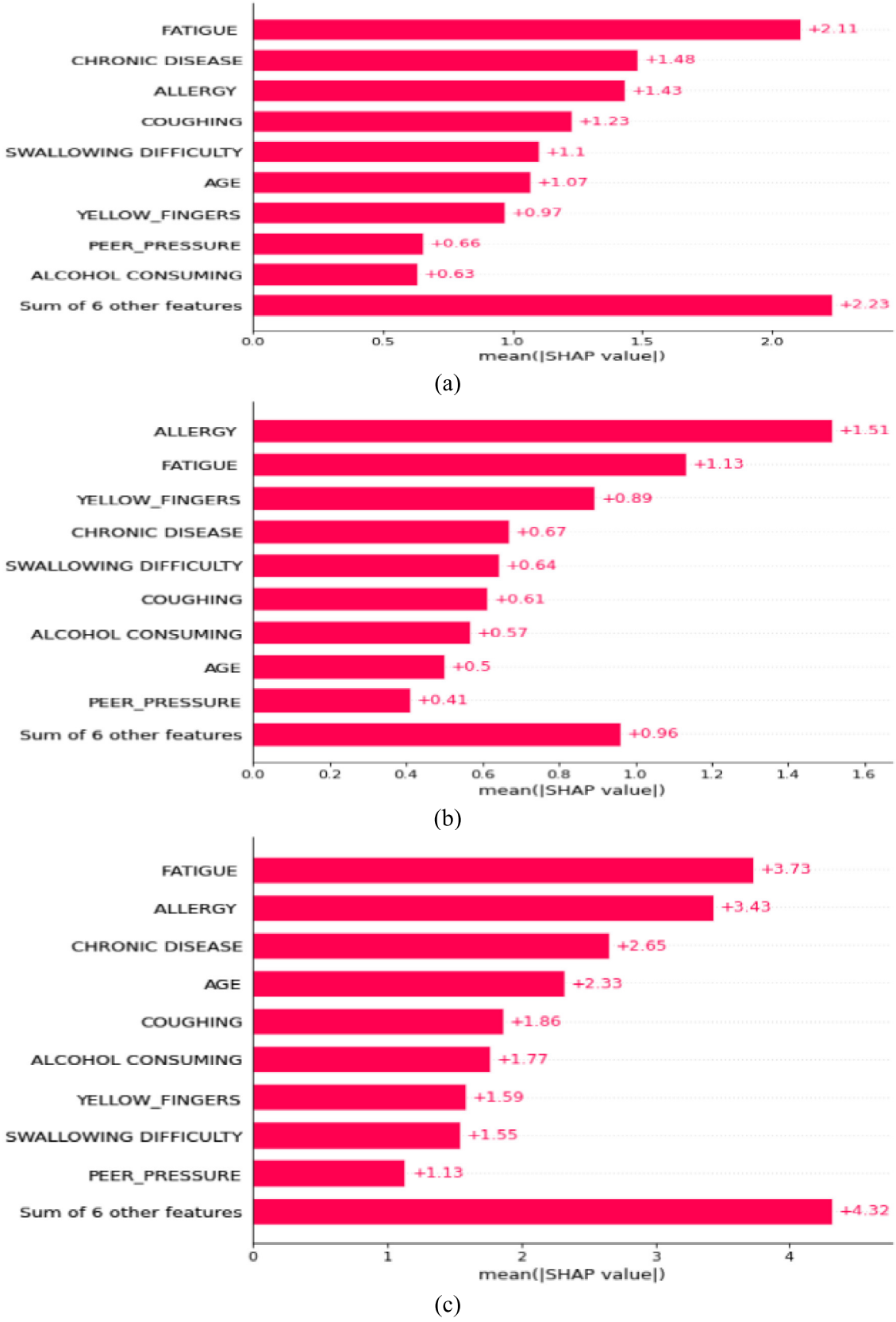
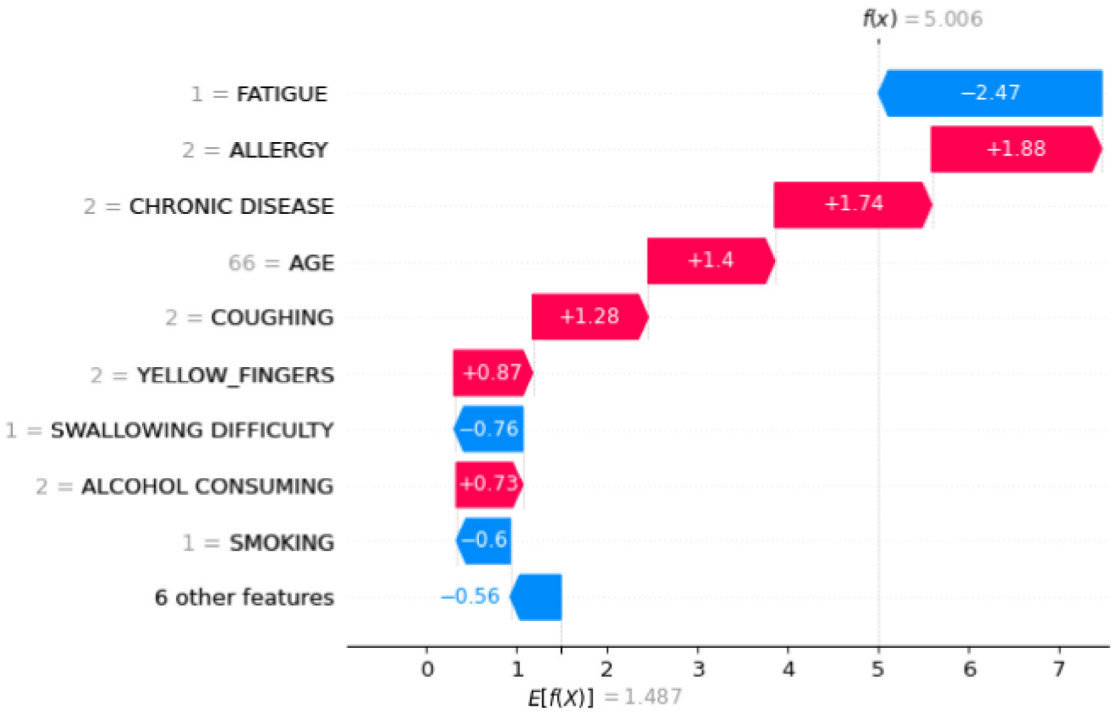
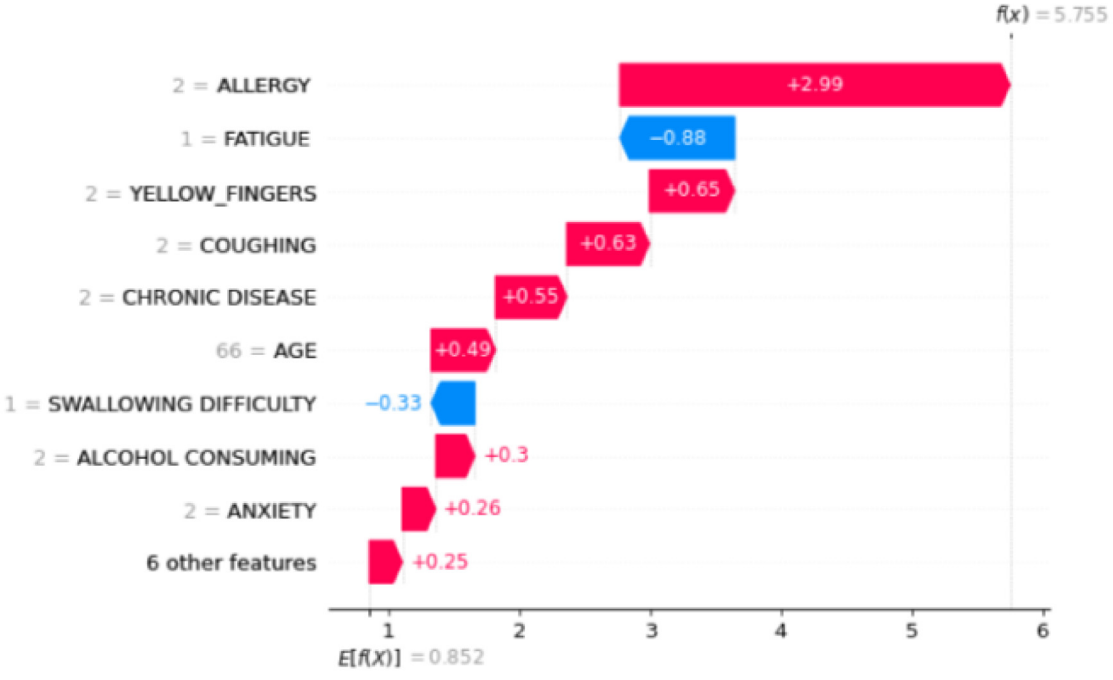


Fig. 9. SHAP Bar plot for: (a) GBM, (b) XGBoost, and (c) LGBM.

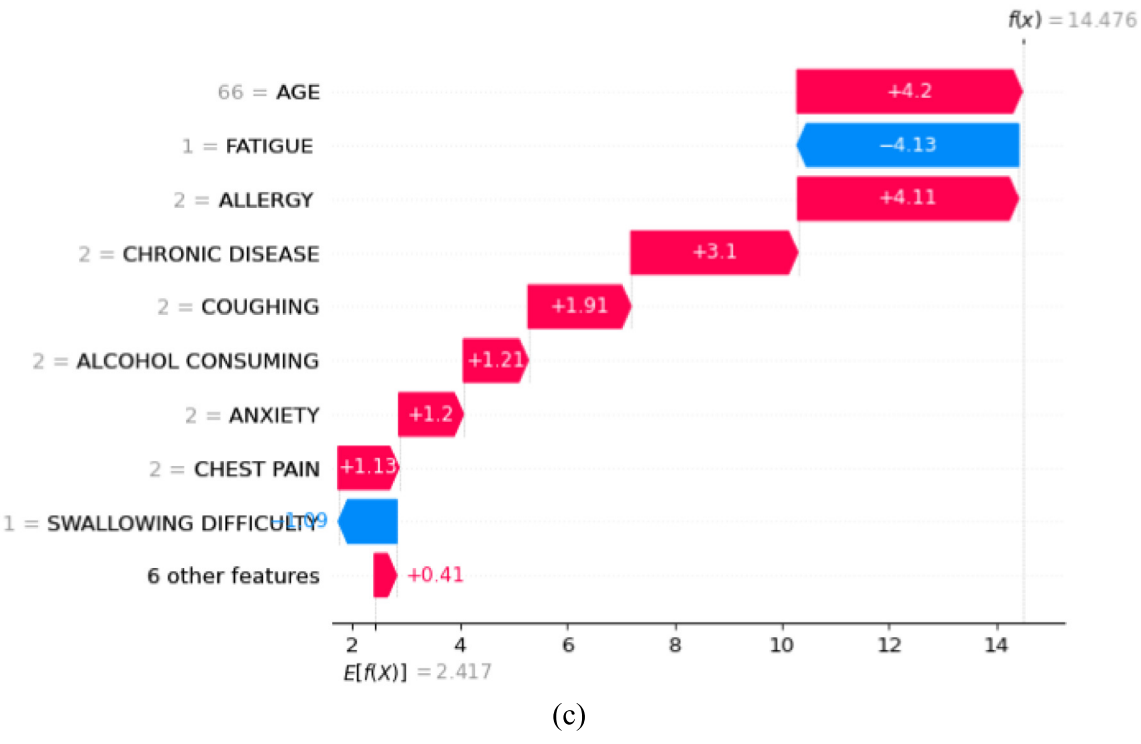


(a)



(b)

Fig. 10. SHAP waterfall plot for: (a) GBM, (b) XGB, and (c) LGBM.



(c)

Fig. 10 (*continued*).

and how different feature interactions contribute to the overall prediction of the model. Feature importance strategy is developed as a result of this analysis. Fig. 8 illustrates the importance of variables.[46]

The graph makes it evident that AGE, which accounts for 190.0 of the totals, is the main contributor for lung cancer prediction. FATIGUE is the next most important feature, followed by CHRONIC DISEASE, YELLOW FINGERS, and so on, which are organized according to their significance. The least-contributing factor, by a factor of 9.0, is CEST PAIN.

More graphing SHAP plots are available to aid with human comprehension of the expected outcome. The bar plot of GBM, XGBoost, and LGBM is shown in Fig. 9.

Each model's associated importance is different. The most important component of GBM, as shown in Fig. 9(a), is FATIGUE, which must imply absolute SHAP values that are significantly higher than those of any other characteristic. CHRONIC DISEASE contributed the second-highest amount (+1.48). FATIGUE is the secondary contributor to XGBoost, according to 9(b), whereas ALLERGY is the main contributor, accounting for +1.51. Similar to GBM, FATIGUE plays a prominent role for 9(c), followed by ALLERGY, CHRONIC DISEASE, and other conditions.

Another, visualization technique is the Waterfall Plot, which visualizes the contribution of each feature to the prediction. In a SHAP Waterfall Plot, the features are listed along the x-axis and the SHAP values are represented by bars that extend from the baseline (usually zero) to the final prediction. Positive SHAP values indicate that the feature had a positive impact on the prediction, while negative SHAP values indicate that the feature had a negative impact. The height of each bar represents the magnitude of the feature's contribution to the prediction. It provides a clear and intuitive way to understand how each feature contributes to the prediction and how they interact with each other. It can help identify which features are the most important, which have the strongest positive or negative effects, and how they relate to the final prediction.Fig. 10 represents the waterfall plot for GBM, XGBoost, and LGBM.

Fig. 10(a) demonstrates that FATIGUE has a SHAP value of -2.47, which has a negative influence on prediction, while ALLERGY has a positive impact on prediction with a SHAP value of +1.88, and so on. All SHAP values added together will equal E[f(x)]—f (x). ALLERGY has a +2.99 positive effect on prediction for Fig. 10(b). To predict lung cancer, FATIGUE has a negative impact of -0.88 and YELLOW_ FINGERS has a favorable impact of +0.65 for XGBoost. Comparatively, AGE has the most positive contribution to the prediction of 10(c), whereas the combined contributions of the other 6 variables are the least.

Another plotting technique of SHAP is Beeswarm which is shown in Fig. 11. It is a type of scatter plot that is used to display the distribution of a large number of individual observations in a way that minimizes the overlap between points. In this plot, the data points are represented by small dots that are placed along the x-axis, with the y-axis showing the density of the points. The dots are arranged in a way that they are as close as possible to their x-value without overlapping.

The FATIGUE is typically the most significant component for GBM, as shown in Fig. 11(a). Then, CHRONIC_DISEASE and ALLERGY are the second and third most crucial factors for prognosis, respectively.On the other hand, ALLERGY contributes most to the prediction of stroke for 11(b). The likelihood of experiencing a forecast will also increase as ALLERGY levels rise. For XGBoost, FATIGUE and YELLOW_FINGERS are the second and third-highest risk factors for lung cancer, respectively.Fig. 11(c) shows that FATIGUE has the biggest influence on prediction. ALLERGY and CHRONIC_DISEASE are the next 2 most crucial characteristics.

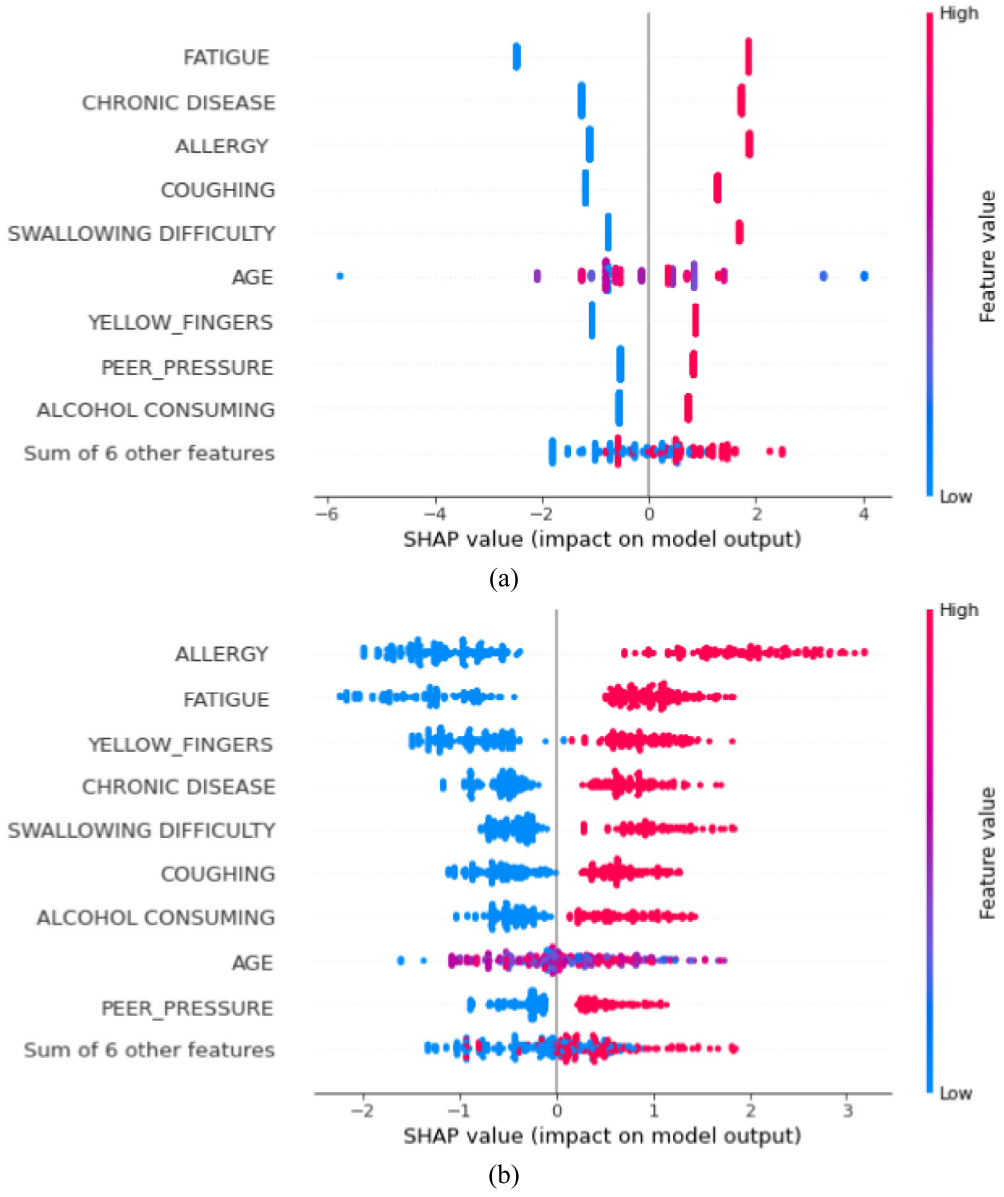


Fig. 11. SHAP Beeswarm plot for: (a) GBM, (b) XGB, and (c) LGBM.

As we can see from this section, GBM obtains the maximum accuracy of 98.76% with a precision of 98.79%, recall of 98.76%, and F-measure 98.76% with a 0.012% error rate. These signs collectively indicate that lung cancer detection can be modeled using GBM and are highly significant overall. Finally, the 35% testing set is used to test GBM once it has been retrained on the entire 65% training set. Data for lung cancer have shown to benefit from using SHAP values for model explainability. It is interesting to note that this work demonstrates how the idea of importance given to features by the absolute SHAP values may be stretched to be utilized as a feature selection method. Explainable properties, which are used in this approach, could be beneficial for feature selection, a common pre-processing step in machine learning. We anticipate that feature selection based on SHAP values will become a popular strategy among machine learning practitioners.

However, a lot of research has been done on this subject by numerous researchers using diverse methods and producing a range of findings. Accurate cancer forecasting is crucial since lung cancer affects people all around the world. The significance of lung cancer inspired us to pursue this topic. This study focuses on the use of XML to forecast lung cancer and demonstrated a useful implementation (mobile app) that can predict cancer-based on given inputs. A comparison of this work with earlier research is provided in Table 4 in order to understand the existing knowledge on a topic and identify gaps in the literature that their own study can address.

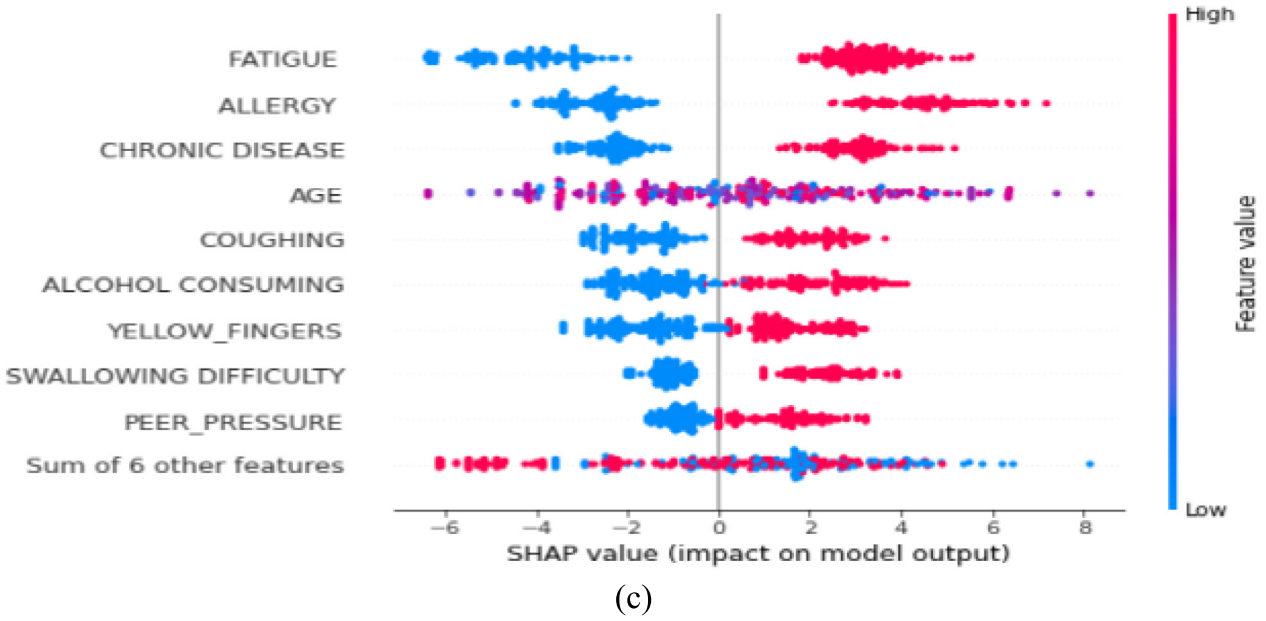


Fig. 11 (*continued*).

Table 4
Comparison of our work with the most related works.

Author	Dataset	Proposed models	Performance
Lung cancer prediction (XML)			
Sobhan et al.,[1] (2022)	UCSC Xena database, 1415 instances	XGBoost	Accuracy: 96.3 %
Alsinglawi et al.,[2] (2022)	MIMIC-III data, 53 423 instances	RF	AUC: 98% (95.3%-100%), Recall: 98% (95.3%-100%).
Katarzyna et al.,[3] (2022)	Domestic Lung Cancer dataset, 34 393 individuals	Used models BACH, PLCOm2012, and LCART.	Not mentioned the performance and accuracy. Only focus on comprehending how the models act for various patients.
Jamie et al.,[4] 2021	Simulacrum dataset, 1 322 100 instances	XGBoost	Precision: 78 %, Recall:78% Accuracy: 78%
Lung cancer prediction (ML)			
Elias Dritsas and Maria Trigka,[5] 2022	Kaggle dataset, 309 instances	Rotation forest (RotF)	AUC: 99.3%, F-Measure, precision, recall, and accuracy: 97.1%.
Muntasir et al.,[6] 2022	Kaggle dataset, 309 instances	XGBoost	AUC: 98.14%, Precision:95.66% Accuracy:94.42% Recall: 94.46%
Patra,[7] 2020	UCI repository, instances 32	Radial Basis Function Network	Accuracy:81.25% F-score: 81.3% AUC: 74.9% Precision:81.3% Recall: 81.3%
Sim et al.,[8] 2020	HRQOL data, 809 individuals	AdaBoost	AUC: 94.9% Accuracy: 94.8%
Proposed	Kaggle, 309 instances and total 16 features	GBM (XML)	Accuracy: 98.76% F-score: 98.76% AUC: (train-1.0, test-0.991) Precision: 98.79% Recall: 98.76%

Implementation of mobile app

The practical component of this study is shown in this section. The experiment's application app, which was created using the best model, is depicted in Fig. 12. React Native was used to build this application. This program has a user feedback form with input fields that forecasts breast cancer and collects user comments. The model is initially constructed as a pkl file in the Jupyter notebook after which a flask application is used to create an api. This api is used to add a machine learning model to an Android app, and the results are then shown on the screen. Fig 12(a) and (b) depict the mobile app where anyone can submit input for forecasting the outcome. Fig 12(c) and (d) show the output after entering the inputs.

Conclusion and future work

In this study, we have made an effort to close the gap regarding the interpretability of these risk models. In various sectors, explainable machine learning algorithms have gained significant traction. Lung cancer issues have already been solved using XML techniques. However, we have also shown how to use explainable machine learning techniques to predict lung cancer. XML approach offers research insights into the characteristics that are most crucial for cancer prognosis. For this, GBM, XGBoost, and LGBM are explained using SHAP. The ability to satisfy multiple desirable qualities, such as consistency, locality, and missingness, makes SHAP a popular option for model interpretability. It is crucial to demonstrate how the inner workings of a medical system.

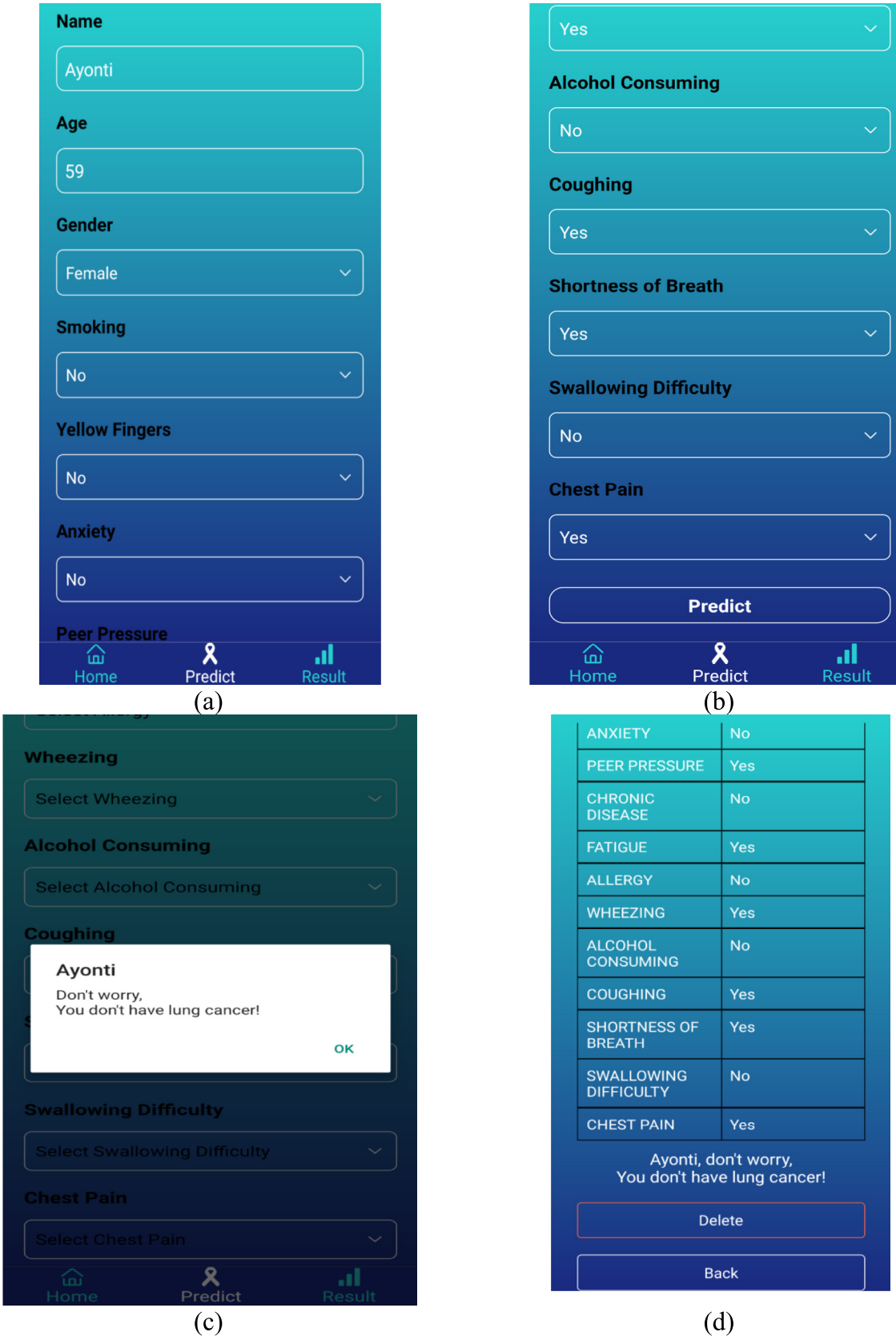


Fig. 12. Android application (a,b) input field (c,d) result field.

In SHAP, the relevance of features is determined by their contribution to the model's output, regardless of the model being used. These contributions are employed in this case as a feature selection approach and to rank features in terms of relevance. In this experiment, SHAP proved to be superior to other popular feature selection methods. This finding suggests that using SHAP as a feature selection mechanism can be a good strategy for

machine learning solutions that need to be interpretable. We intend to further examine SHAP using deep learning in further works. Additionally, an analysis of the image with explainability will be conducted.

Author Contribution

STR, NB, and KMMU were responsible for the conceptualization and design of the study. They also had full access to all the study's data and accepted responsibility for the accuracy of the model generation and the study's data. All of the contributors worked together to write the article. The report was critically revised with input from STR, SKD, and others. All of the results and data presentation techniques were produced by NB and KMMU. The final version has been reviewed and approved by all authors, who also contributed to the data collection and analysis.

Funding

None.

Ethical Approval

Not required.

Consent to participate

Not required.

Data availability

On reasonable request, the corresponding author will provide the data that support the study's findings.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: A tool for early detection. International Journal of Cancer Jan 1, 2007;120(1):1–6. https://doi.org/10.1002/ijc.22331.
2. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. [Online]. Available: www.ijcsit.com.
3. Qiang Y, Guo Y, Li X, Wang Q, Chen H, Cuic D. The Diagnostic Rules of Peripheral Lung Cancer Preliminary Study Based on Data Mining Technique. [Online]. Available: www.elsevier.com/locate/jnmu.
4. Shopland DR, Eyre HJ, Pechacek TF. ARTICLES Smoking-Attributable Cancer Mortality in 1991: Is Lung Cancer Now the Leading Cause of Death Among Smokers in the United States? [Online]. Available: http://jnci.oxfordjournals.org/.
5. Karabatak M, Ince MC. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl 2009;36(2 PART 2):3465–3469. https://doi.org/10.1016/j.eswa.2008.02.064.
6. Stokowy T, Wojtaś B, Krajewska J, Stobiecka E, Dralle H, Musholt T, et al. A two miRNA classifier differentiates follicular thyroid carcinomas from follicular thyroid adenomas. Mol Cell Endocrinol Jan. 2015;399:43–49. https://doi.org/10.1016/j.mce.2014.09.017.
7. Zhang R, bin Huang G, Sundararajan N, Saratchandran P. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. IEEE/ACM Trans Comput Biol Bioinform Jul. 2007;4(3):485–494. https://doi.org/10.1109/TCBB.2007.1012.
8. Wang Y, et al. Gene selection from microarray data for cancer classification - a machine learning approach. Comput Biol Chem Feb. 2005;29(1):37–46. https://doi.org/10.1016/j.compbiolchem.2004.11.001.
9. Kim B, et al. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. 2018.
10. Tan S, Caruana R, Hooker G, Lou Y. Distill-and-compare: auditing black-box models using transparent model distillation. AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; Dec. 2018. p. 303–310. https://doi.org/10.1145/3278721.3278725.
11. Ribeiro MT, Singh S, Guestrin C. 'Why should i trust you?' Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016. Aug. 2016. p. 1135–1144. https://doi.org/10.1145/2939672.2939778.
12. Lundberg SM, Allen PG, Lee S-I. A Unified Approach to Interpreting Model Predictions. [Online]. Available: https://github.com/slundberg/shap.
13. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak Jul. 2019;19(1). https://doi.org/10.1186/s12911-019-0874-0.
14. Ibrahim M, Louie M, Paisley J. *Global Explanations of Neural Network Mapping the Landscape of Predictions Ceena Modarres Center for Machine Learning, Capital One*. 2019. https://doi.org/10.1145/aies062.
15. Whitmore LS, George A, Hudson CM. *Mapping chemical performance on molecular structures using locally interpretable explanations*. Nov. 2016.[Online]. Available: http://arxiv.org/abs/1611.07443.
16. Phillips PJ, et al. Four Principles of Explainable Artificial Intelligence Gaithersburg, MD. Sep. 2021. https://doi.org/10.6028/NIST.IR.8312.
17. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. KnowlInfSyst Nov. 2014;41(3):647–665. https://doi.org/10.1007/s10115-013-0679-x.
18. Cosgriff Cv, Celi LA. Exploiting temporal relationships in the prediction of mortality. Lancet Digital Health Apr. 1, 2020;2(4):e152–e153. https://doi.org/10.1016/S2589-7500(20)30056-X.Elsevier Ltd.
19. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell Jan. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
20. Lundberg SM, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng Oct. 2018;2(10):749–760. https://doi.org/10.1038/s41551-018-0304-0.
21. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. J Imaging Jun. 20, 2020;6(6). https://doi.org/10.3390/JIMAGING6060052.MDPI AG.
22. Xi J, Zhao W, Yuan JE, Cao B, Zhao L. Multi-resolution classification of exhaled aerosol images to detect obstructive lung diseases in small airways. Comput Biol Med Aug. 2017;87:57–69. https://doi.org/10.1016/j.compbiomed.2017.05.019.
23. Li W, Jia Z, Xie D, Chen K, Cui J, Liu H. Recognizing lung cancer using a homemade e-nose: A comprehensive study. Comput Biol Med May 2020;120. https://doi.org/10.1016/j.compbiomed.2020.103706.
24. M. Sobhan and A. M. Mondal, "Explainable Machine Learning to Identify Patient-specific Biomarkers for Lung Cancer", https://doi.org/10.1101/2022.10.13.512119.
25. Alsinglawi B, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. Sci Rep Dec. 2022;12(1). https://doi.org/10.1038/s41598-021-04608-7.
26. Kobylińska K, Orłowski T, Adamek M, Biecek P. Explainable machine learning for lung cancer screening models. Appl Sci (Switzerland) Feb. 2022;12(4). https://doi.org/10.3390/app12041926.
27. Duell J, Fan X, Burnett B, Aarts G, Zhou S-M. A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records. [Online]. Available: http://www.ncin.org.uk/about.
28. Dritsas E, Trigka M. Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing Nov. 2022;6(4):139. https://doi.org/10.3390/bdcc6040139.
29. Mamun M, Farjana A, al Mamun M, Ahammed MS. Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. 2022 IEEE World AI IoT Congress, AIIoT 2022; 2022. p. 187–193. https://doi.org/10.1109/AIIoT54504.2022.9817326.
30. Patra R. Prediction of lung cancer using machine learning classifier. Communications in Computer and Information Science. CCIS; 2020. p. 132–142. https://doi.org/10.1007/978-981-15-6648-6_11.
31. ah Sim J, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. Sci Rep Dec. 2020;10(1). https://doi.org/10.1038/s41598-020-67604-3.
32. Lung Cancer Prediction Dataset. Available online: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer?fbclid=IwAR0uQ5K3mEbQZJcwQGYqlLJ5RydvsK2oU1Sa5vYvit0ECoqkx6-vPR43JAM.
33. Ahmed N, et al. Machine learning based diabetes prediction and development of smart web application. Int J Cognit Comput Eng Jun. 2021;2:229–241. https://doi.org/10.1016/j.ijcce.2021.12.001.
34. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. 2020 11th International Conference on Information and Communication Systems, ICICS 2020; Apr. 2020. p. 243–248. https://doi.org/10.1109/ICICS49469.2020.239556.
35. Tharwat A. Principal component analysis - a tutorial. Int J Appl Pattern Recognit 2016;3(3):197. https://doi.org/10.1504/ijapr.2016.079733.
36. Kumar S. Effective Hedging Strategy For Us Treasury Bond Portfolio Using Principal Component Analysis. [Online]. Available: https://ssrn.com/abstract=4007786.
37. Biswas N, Uddin KMM, Rikta ST, Dey SK. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Anal Nov. 2022;2:100116. https://doi.org/10.1016/j.health.2022.100116.
38. Mir Ishrak A, Dhruba M, Haider N, et al. *Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking*. 2018.
39. Saleh H, et al. Stroke prediction using distributed machine learning based on apache spark. Int J Adv Sci Technol 2019;28(15):89–97. https://doi.org/10.13140/RG.2.2.13478.68162.
40. Wang D, Zhang Y, Zhao Y. LightGBM: an effective miRNA classification method in breast cancer patients. ACM International Conference Proceeding Series; Oct. 2017. p. 7-11. https://doi.org/10.1145/3155077.3155079.

41. Lundberg SM, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018;2(10):749–760. https://doi.org/10.1038/s41551-018-0304-0.
42. Li R, et al. Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. JCO Clin Cancer Informatics 2020;4:637–646. https://doi.org/10.1200/cci.20.00002.
43. Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Sci Rep Dec. 2022;12(1). https://doi.org/10.1038/s41598-022-05112-2.
44. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag Jul. 2009;45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
45. Luque A, Carrasco A, Martín A, DE LAS Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit Jul. 2019;91:216–231. https://doi.org/10.1016/j.patcog.2019.02.023.
46. Fisher A, Rudin C, Dominici F. *All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously*. 2019.