Journal of Theoretical and Applied Information Technology

15th November 2023. Vol.101. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTEROPERABILITY AND EXPLAINABILITY OF MACHINE LEARNING CLASSIFIERS TO DETECT LUNG CANCER

JYOTIRMAY DEVNATH [1], MD. NAHID SULTAN[2], MD. FERDOUS WAHID[3], AHSAN HABIB[4]

[1, 2]Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.

[3]Department of Electrical and Electronic Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.

[4]School of Information Technology, Deakin University, Geelong, VIC 3225, Australia.

E-mail: [1]jyotirmay.pulak@gmail.com , [2]nahid.sultan@hstu.ac.bd, [3]mfwahid26@gmail.com,
[4]m.habib@deakin.edu.au

ABSTRACT

The prominence of lung disease as the leading cause of death in cancer necessitates utmost significance on early detection, prediction, and diagnosis of lung cancer, owing to time limitations and the intricacies of the ensuing clinical examination. Hence, the use of machine learning (ML) models may enable the early stage diagnosis of cancer as well as the characterization, stratification, and consequences of the disease. Therefore, several machine learning algorithms have been used in this paper to predict lung cancer, including Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGB), Multinomial Nave Bayes (MNB), Gradient Boosting Classifier (GBC), k-Nearest Neighbor (KNN), and Adaptive Boosting classifier (ABC). To make the data more trainable for the ML models, we proposed a preparation pipeline that included data cleaning, normalization, and data balancing. Nevertheless, healthcare practitioners may exhibit reluctance in embracing artificial intelligence (AI) models if the reasoning behind the generated predictions remains inscrutable. As a consequence, explainable artificial intelligence (XAI) is gaining popularity to meet the needs of healthcare practitioners. Hence, we employ XAI tools (SHAP and Shapash) to rank features, find partial dependencies, and correlate top feature dependencies to find the inner pattern of the features. We investigate both the global and local explainability of the ML model. RF, LR, and XGB among all algorithms exhibit 95% accuracy. To demonstrate the reasoning behind the prediction, we use XAI tools on RF. When it comes to lung disease, Allergy, Coughing, and Swallowing difficulty are of utmost importance. The expertise of the domain expert might be mapped to the developing field of XAI using this research.

Keywords: *Lung Cancer Prediction, Machine Learning, Classification, Explainable AI, Model Interpretability*

1. INTRODUCTION

Cancer, a fatal condition resulting from a confluence of genetic abnormalities and various biological anomalies, is exemplified by lung cancer, which is currently one of the deadliest and most prevalent cancers [1]. Consequently, its impact on mortality rates is substantial, primarily due to the lack of early-stage cancer signs. Since lung cancer is the third most prevalent type of cancer that affects both men and women, early detection is crucial for protecting against its deadly effects. Cancer is difficult to treat because of high intra-tumor heterogeneity (ITH) and the complexity of cancer cells, which leads to therapeutic resistance [2]. As cancer research methods have improved over the past few decades, many big collaborative cancer initiatives have been launched, leading to the creation of various clinical, medical imaging, and sequencing databases [3]. These databases help scientists investigate the full picture of lung cancer, from diagnosis to treatment to clinical outcomes [4]. Nowadays, machine learning (ML) is receiving a lot of interest from researchers as a potential method for early disease identification. ML involves the application of various algorithms and techniques to analyze medical data, such as imaging scans, pathology reports, and patient information, in order to identify and diagnose cancer at an early stage. In 2019, the authors proposed an ensemble of Weight Optimized Neural Network with Maximum Likelihood Boosting (WONN-MLB) for lung cancer disease detection from big data [5]. The goal of this study is to improve the accuracy and efficiency of the diagnostic process, which can potentially lead to better treatment outcomes and higher survival rates for patients. Another authors initiated a new approach to learn the partially observable Markov decision process (POMDP) to optimize the lung cancer detection process through improving the specificity. Authors measured 'detection ratios' by showing a comparison of the frequency of lung cancer detection in National Lung Screening Trial (NLST) vs Modified Logistic Hyper-Chaotic System (MLHCs) [6].

Explainable Artificial Intelligence (XAI) is crucial in lung cancer detection as it provides transparent insights into AI-driven diagnostic decisions. Understanding the reasoning behind predictions enhances trust, aids medical professionals in validating results, and ensures the ethical deployment of AI, fostering more effective and accountable healthcare interventions.

However, recently medical data analysis is getting interesting with the new emerging Explainable AI (XAI) field to find the internal story of the prediction instead of blind prediction of ML that are popular in different fields. To predict Lung cancer and explain the black-box ML model is the major contributions of this research. The overall contributions of the research are as follows:

1. We use ML algorithms to predict Lung Cancer from the secondary data with a preprocessing pipeline to make the data more trainable for the ML models.
2. We show the interoperability and Explainability of the black-box ML models to find the hidden story of the prediction.
3. We performed a comparative study among the ML algorithms to find the best algorithm to classify lung cancer.

The paper is organized as in section 2 the literature review, in section 3 the methodology, in section 4 the result and analysis and at last the conclusion and future work in section 5.

2. LITERATURE REVIEW

As one of the most common cancers, detecting Lung Cancer has gained extensive attention from researchers. Most of their detecting approaches are done by ML algorithms, deep learning, and ensemble. Moreover, some hybrid algorithms are used to classify lung disease from open-source datasets that are available in online data repositories.

To efficiently and effectively detect lung and colon cancer, Talukder et al., (2022) develop a hybrid ensemble feature extraction model. The model incorporates ensemble learning and deep feature extraction with high-performance filtering. The method is assessed on histopathological (LC25000) colon and lung datasets. The findings of the study determine that the proposed hybrid model can identify colon, lung, and (lung and colon) cancer with the accuracy of 100%, 99.05%, and 99.30%, respectively [7]. To improve the accuracy of prediction and lessen the time of prediction of lung cancer in big data handling, Chandrasekar et al., (2022) introduce the Multivariate Ruzicka Regressed eXtreme Gradient Boosting Data Classification (MRRXGBDC) technique and service-oriented architecture (SOA). The result reveals that compared to existing models, the MRRXGBDC model performs 10% better in prediction accuracy, reduces 50% false positives, and requires 11% lesser time to detect lung cancer [8].

Utilizing Artificial Intelligence (AI) and cloud platform approaches in the medical industry 4.0, Gu et al., (2022) develop an intelligent detection system for lung cancer. For similarity comparison, a cloud-based deep learning (DL) model is employed into this system and a content-based image recovery system is used. The results reveal that the proposed method performs better than some other baseline approaches [9]. Binson et al., (2021) introduced an e-nose system, a lung cancer detection method. They tested this approach on 199 participants including 93 controls, 55 chronic obstructive pulmonary disease (COPD) patients, and 51 lung cancer patients. It turned out that the ensemble learning method XGBoost performed far better than the other two models [10]. Radhika et al., (2019) intended to predict lung cancer employing classification algorithms such as SVM, Naive Bayes (NB), Logistic Regression (LR) and Decision tree (DT). The prime objective of this study was to ensure early diagnosis of lung cancer through assessing the performance of classifiers [11]. Zo et al., (2022) compared the screening detected lung cancer (SDLC) and clinical features and prognosis of incidentally detected lung cancers (IDLC). They involved the subjects with pulmonary nodules (<3cm) at the baseline CT scans, which were clinically defined as primary lung cancer in 2015.

Out of 553 samples, 344 (62.2%) SDLCs and 209 (37.8%) IDLCs were recognized [12].

Lung cancer detection in ML is challenging due to diverse data patterns, subtle anomalies, and the need for high accuracy. Complexities in feature extraction, limited labeled datasets, and interpretability issues pose research challenges, necessitating innovative ML approaches to enhance sensitivity, specificity, and overall diagnostic performance.

3. METHODOLOGY

This This research focuses on predicting lung cancer from medical data using ML algorithms with explainability. As the process of traditional ML algorithm training and testing, we separate the dataset into 80:20 ratio of training and testing data. Before fitting the data into ML models, we perform data preprocessing on the whole dataset that makes the dataset more trainable for ML models. We check the performance of the ML classifiers using some performance measure metrics. The best classifier is then used in the next step of analysis. We use several XAI tools on the top classifier to find the explainability of the model in terms of predicting lung cancer. The result is shown as the order of the objectives we claim in the introduction section. The overall process is shown in Figure 1.

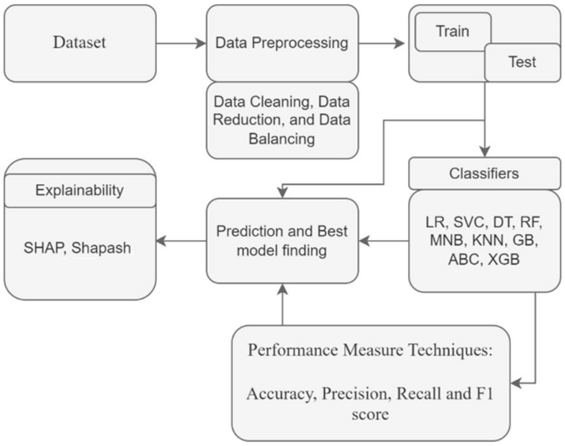
Dataset
Data Preprocessing
Train
Test
Data Cleaning, Data Reduction, and Data Balancing
Classifiers
LR, SVC, DT, RF, MNB, KNN, GB, ABC, XGB
Explainability
SHAP, Shapash
Prediction and Best model finding
Performance Measure Techniques:
Accuracy, Precision, Recall and F1 score

Figure 1: Overview of proposed methodology

3.1 Description of the Dataset

An affordable and effective cancer prediction system enables individuals to assess their cancer risk and make informed decisions accordingly. To train the proposed algorithm, we have gathered the data from a well-known online data source. The dataset comprises 309 instances, each consisting of 1 numeric, 14 nominal, and 1 target feature [13]. Table 1 shows dataset statistics.

3.2 Preprocessing Techniques

3.2.1 Data Cleaning

In data science, data cleaning is required to identify and repair mistakes, inconsistencies, and missing information for accurate and trustworthy analyses and models. In this field, missing value handling, wrong format correction and wrong data handling are all related topics [14]. The nominal values of the dataset were converted to numerical values in order to use in the training process. The target class of the dataset is converted into numerical value using label encoder.

3.2.2 Data Reduction

The performance of a model is negatively impacted when it contains data that is both unnecessary and duplicate. A total of 33 duplicate instances have been identified within the dataset. The duplicated data was eliminated and a dataset suitable for prediction was generated in order to create a machine learning trainable dataset.

Table 1: Statistical data on dataset features

Feature	Type of Feature	Mean (standard deviation) for numerical features /No. of values for categorical features.
Gender	Nominal	162/147 [Male/Female]
Age	Numerical	62.67 (8.21)
Smoking	Nominal	174/135 [Yes/No]
Yellow_Fingers	Nominal	176/133 [Yes/No]
Anxiety	Nominal	154/155 [Yes/No]
Peer_Pressure	Nominal	155/154 [Yes/No]
Chronic Disease	Nominal	156/153 [Yes/No]
Fatigue	Nominal	208/101 [Yes/No]
Allergy	Nominal	172/137 [Yes/No]
Wheezing	Nominal	172/137 [Yes/No]
Alcohol Consuming	Nominal	172/137 [Yes/No]
Coughing	Nominal	179/130 [Yes/No]
Shortness of Breath	Nominal	198/111 [Yes/No]
Swallowing difficulty	Nominal	145/164 [Yes/No]
Chest Pain	Nominal	172/137

		[Yes/No]
Lung Cancer	Nominal	270/39 [Yes/No]

3.2.3 Data Balancing

Imbalanced data results in fewer instances from the minority class for the model to learn from, making it difficult for the model to identify and generalize patterns, resulting in poor performance in that class. Therefore, conventional evaluation criteria such as accuracy can be deceiving. Models that are biased in favor of the majority class and have bad performance on the minority class can be the result of this class imbalance. We use Synthetic Minority Over-sampling Technique (SMOTE) data balancing technique that is an oversampling technique. Oversampling is a typical technique for this problem; it involves replicating or artificially generating instances of the minority class to enhance their representation in the sample [15]. The dataset contains 238 instances of lung cancer patients and 38 instances of healthy people before data balancing. Following the use of the SMOTE oversampling approach, each class has 238 instances.

3.3 Description of ML Algorithms

In In this analysis, we employ nine ML algorithms such as Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbor (KNN), Decision Tree (DT), Extreme Gradient Boosting (XGB), Multinomial Nave Bayes (MNB), Gradient Boosting Classifier (GBC), and Adaptive Boosting classifier (ABC). Short description of each algorithm is described below:

3.3.1 Support Vector Classifier (SVC)

SVC is a very effective supervised learning technique that can be utilized for pattern recognition and classification tasks. It divides data into various categories by making use of decision planes [16]. The robustness of the model may be ensured because to SVC's ability to maximize the margin between support vectors. This model works well in high-dimensional spaces and makes efficient use of memory, but it has difficulties when dealing with datasets that contain noise. Despite the fact that it has several drawbacks, SVC continues to be a popular option for complicated datasets and activities that need for accurate decision limits. When working with noisy data, users should exercise caution and explore preprocessing or alternate models, depending on the circumstances.

3.3.2 Random Forest (RF)

RF is a classifier that uses the average to boost the predicting accuracy of a particular dataset by employing a series of decision trees on various subsets of that dataset. Instead of depending on one decision tree, RF forecasts from each tree and projects the ultimate outputs depending on the overwhelming votes [17]. The variables for the classification problem are rated according to their significance. The accuracy improves as the volume of trees in the forest increases, reducing the detrimental effects of overfitting.

3.3.3 Logistic Regression (LR)

For forecasting the likelihood of a binary result, supervised machine learning approaches like LR work effectively [18]. Regularization is essential in logistic regression to reduce overfitting, especially when there are few training instances or many parameters to learn. Multiclass problems can be classified by employing LR. Since LR has a linear decision surface, it cannot address non-linear issues.

3.3.4 K-Nearest Neighbors (k-NN)

A popular classification technique used in a variety of applications is k-NN. The idea behind k-NN is that the examples predicted value may be similar to that of its neighbors [19]. The k-NN approach maps each applicant to a place in the predictor's vector space, specifies a metric there, and explains how many excellent risks there are among the k-nearest points in the training set. This method determines the likelihood of the predicted outcome.

3.3.5 Decision Tree (DT)

A decision tree is a flowchart-like paradigm that is used for classification and regression [20]. It makes decisions based on conditional statements and input data attributes. Beginning with a root node, the method recursively separates the data depending on attribute values until a stopping requirement is met. Each internal node represents a feature-based judgment, whereas leaf nodes indicate outcomes or forecasts. Decision trees are interpretable, can handle a variety of data formats, and are resistant to outliers. They can, however, overfit, and approaches such as pruning, and ensemble methods are employed to counteract this. Because of their simplicity and interpretability, they are frequently utilized in a variety of disciplines.

3.3.6 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting is a highly effective machine learning technique noted for its high performance. It is an ensemble method that combines several weak decision tree models to produce a reliable and accurate predictor [21]. XGB uses gradient boosting, a technique that iteratively trains new models to rectify prior models' mistakes. It uses gradient descent to optimize a given loss function and includes regularization to prevent overfitting. Because of its speed, scalability, and capacity to handle complicated datasets, XGB is a popular choice for a variety of applications in industry and academics.

3.3.7 Multinomial Naïve Bayes (MNB)

A probabilistic classifier frequently used for text classification problems is Multinomial Naive Bayes. It is based on the Naive Bayes algorithm, which takes the premise of feature independence. With MNB, the feature variables are assumed to have a multinomial distribution, making it appropriate for discrete feature data like word counts or term frequencies [22]. Given the feature values, the method determines the conditional probability of each class, and it chooses the class with the highest probability to serve as the predicted class. MNB has demonstrated strong performance in a variety of text classification tasks, particularly in spam filtering and sentiment analysis, despite its naive assumptions.

3.3.8 Gradient Boosting (GB)

Gradient Boosting is a powerful machine learning technique that combines several weak prediction models, often decision trees, to produce a strong and accurate predictive model [23]. The approach iteratively fits a base model to the data first, and then successively adds new models to fix the faults created by the prior ones. Using gradient descent optimization, each succeeding model is trained to minimize the residuals of the preceding models. This helps the ensemble model to improve its predictions progressively over iterations. Gradient Boosting is a powerful technique for addressing regression and classification problems, frequently beating other algorithms in terms of predicted accuracy.

3.3.9 Adaptive Boosting Classifier (ABC)

AdaBoost, or adaptive boosting, is a prominent machine learning ensemble method for classification tasks [24]. It combines several weak classifiers, usually decision trees or stumps, to form a strong and accurate classifier. AdaBoost works by training weak classifiers iteratively on the same dataset and modifying the weights of misclassified examples. The system lends more weight to misclassified examples in each iteration, driving succeeding weak classifiers to focus on those instances and increase overall accuracy. The final prediction is formed by averaging all weak classifier predictions based on their individual performance. AdaBoost is efficient at handling large datasets and is frequently utilized in a variety of applications.

3.4 Description of XAI Tools

3.4.1 SHapley Additive exPlanations (SHAP)

SHAP is an effective method for understanding predictions made by machine learning algorithms. SHAP let Users to comprehend how each characteristic contributed to the model output. It offers a local explanation for each prediction. The framework is founded on the Shapley values idea, a tried-and-true technique from cooperative game theory and has been proven to be both theoretically sound and experimentally successful.

3.4.2 Shapash

Shapash is a python library that intends to make machine learning models understandable and interpretable developed by Marketing in Australia of Infant Formulas (MAIF) Data scientists. This library provides visualization tools like model performance metrics, partial dependency plots and feature importance.

3.5 Description of Performance Measure Techniques

These metrics—Accuracy, Precision, Recall, F-1 Score, and ROC—are used to assess the effectiveness of the employed ML algorithms.

Accuracy is a performance measure technique such as the proportion of data instances that were properly categorized to all of the data instances. Although accuracy is one of the most fundamental performance indicators, it occasionally yields inaccurate results, particularly for imbalanced data sets. Mathematically,

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

Where, TP for True Positives, TN for True Negatives, FP for False Positives, and FN for False Negatives were used.

In terms of binary classification, precision is calculated by dividing the total number of TP by the total number of TP and FP. When FP reduction is the aim, precision operates accurately on unbalanced data. Precision is a useful statistic to utilize even when the rate of FP is large. Mathematically,

$$Precision = \frac{TP}{TP + FP}$$

Recall is also referred to as Sensitivity or True Positive Rate (TPR). Recall is often determined by dividing the total number of TP by the total number of TP and FN. Recall is appropriate while lowering FN from an unbalanced dataset. Mathematically,

$$Recall = \frac{TP}{TP + FN}$$

The F-1 score is the harmonic mean of Precision (P) and Recall (R). Only accuracy won't do to determine whether or not the model is relevant.

Only when Precision and Recall are both high will the model make sense. To compare the effectiveness of two classifiers, the F1-Score is computed. Its range is [0, 1], and as the F1-scores increase, a more logical model emerges. Mathematically,

$$F - 1\,Score = \frac{2\,PR}{P + R}$$

Using the Receiver Operating Characteristics (ROC) curve, one may visualize, group, and choose classifiers depending on how well they function. The False Positive Rate (FPR) is plotted on the X-axis and the TPR is plotted on the Y-axis at different threshold settings.

4. RESULT AND DISCUSSION

To predict lung cancer, we use several ML algorithms and tabulate the performance of the classifiers. Accuracy, precision, recall, and F1 score are the four performance metrics used to assess the effectiveness of the deployed ML algorithms. Table 2 shows the performance of the employed models. For effective data preprocessing at an early stage, the majority of algorithms exhibit good performance. LR, RF, GB, XGB achieve maximum 95% accuracy with a good precision, recall and f1 score. The performance of the other machine learning classifiers is similarly satisfactory. We use RF, the best performing model, to provide a more thorough analysis of the dataset. Figure 2 also uses a bar chart to display the performance of the classifiers.

Table 2: Performance of the ML classifiers

Algorithms	Precision (%)	Recall (%)	F1-Score (%)	Avg. Acc (%)
LR	0.95	0.95	0.95	0.95
SVC	0.94	0.95	0.94	0.94
DT	0.94	0.94	0.94	0.94
RF	0.95	0.95	0.95	0.95
MNB	0.87	0.87	0.87	0.87
KNN	0.94	0.94	0.94	0.94
GB	0.95	0.95	0.95	0.95
ABC	0.94	0.94	0.94	0.94
XGB	0.95	0.95	0.95	0.95

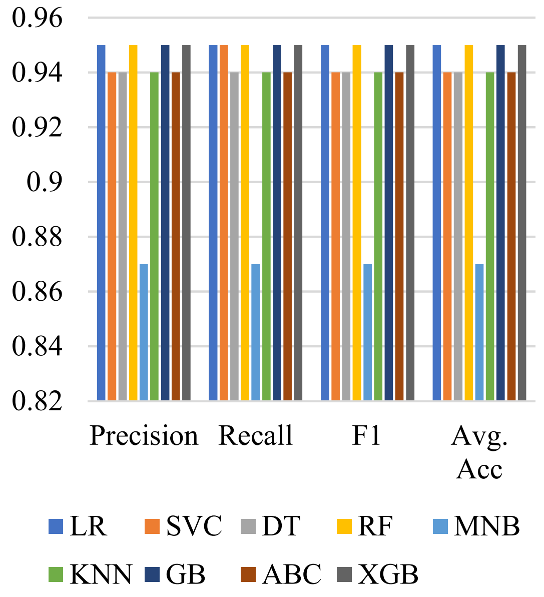

Figure 2: Performance of the models

Figure 3 depicts the contributions of numerous factors to lung cancer prediction, with ALLERGY being the most influential, while also demonstrating the impact of COUGHING and other features.

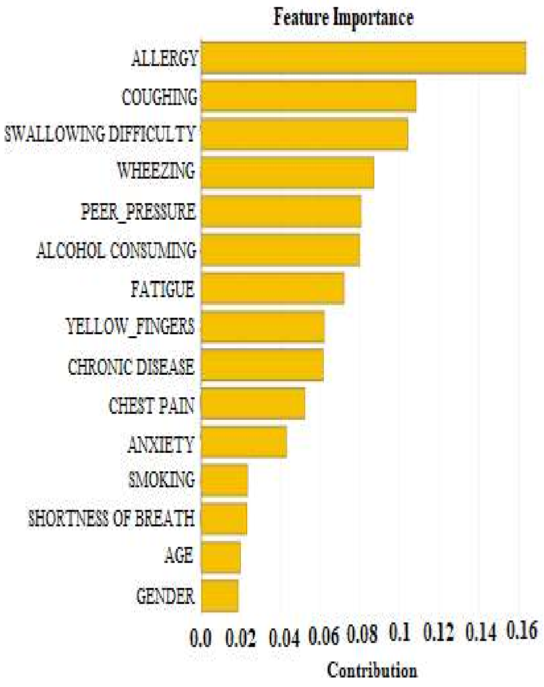

Figure 3: Features importance of the best model

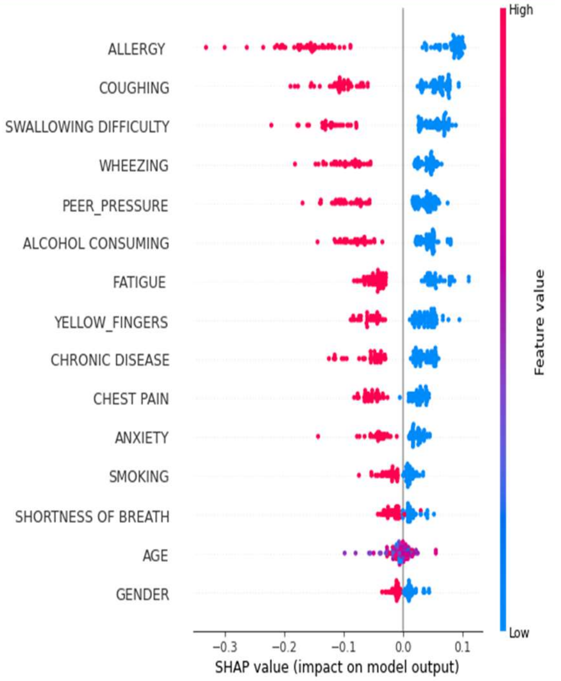

Figure 4: Plot for SHAP value 0

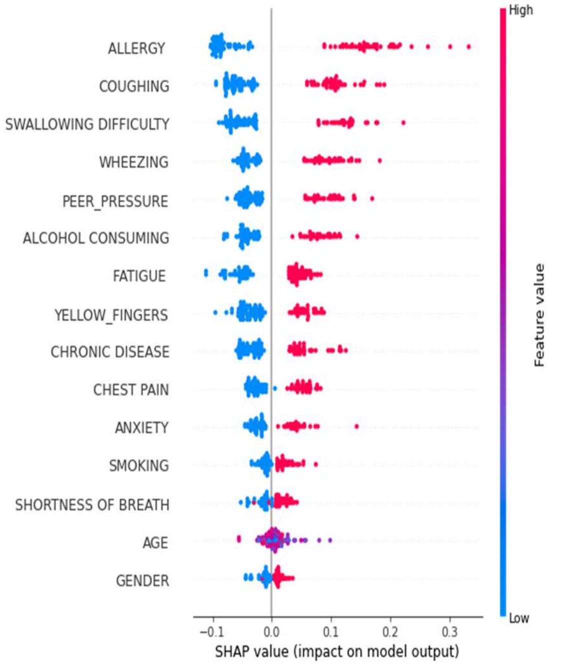

Figure 5: Plot for SHAP value 1

The global explanation of the model for various SHAP values is shown in Figure 4 and 5. Figure 4 display the features based on their significance in predicting lunger cancer. The top features, such as ALLERGY, COUGHING, and SWALLOWING DIFFICULTY, have a greater impact on the prediction model for lung cancer. The bottom features in the figures such as GENDER, AGE, and SHORTNESS of BREATH have the less influence to predict the lung cancer. Moreover, Figure 6 depicts the influence of different features values of the ALLERGY feature on the lung cancer prediction model.

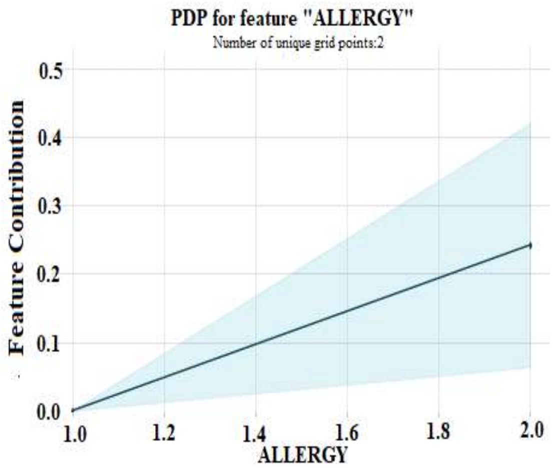

Figure 6: PDP plot for ALLERGY feature

The For different features values it's contributing differently to the lung cancer prediction model. The feature contribution to the mode will grow linearly from 0 to 0.24 if the value of the ALLERGY feature is increased from 1 to 2 along the x-axis of Figure 6.

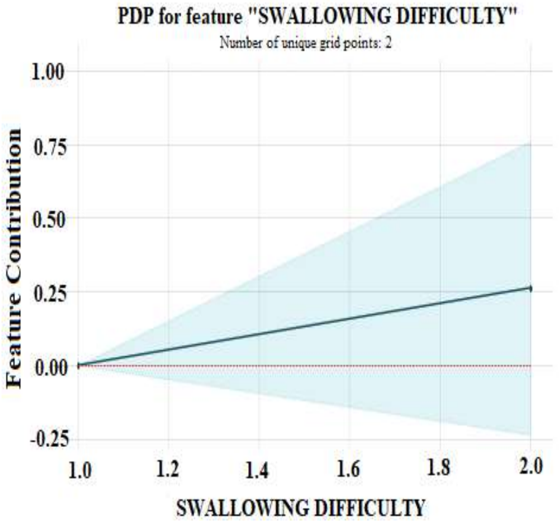

Figure 7: PDP plot for SWALLOWING DIFFICULTY feature

Figure 7 demonstrates that the contributions of feature to the model prediction rise slightly linearly from 0 to 0.25 when the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SWALLOWING DIFFICULTY value is increased from 1 to 2. It shows a linear relationship.

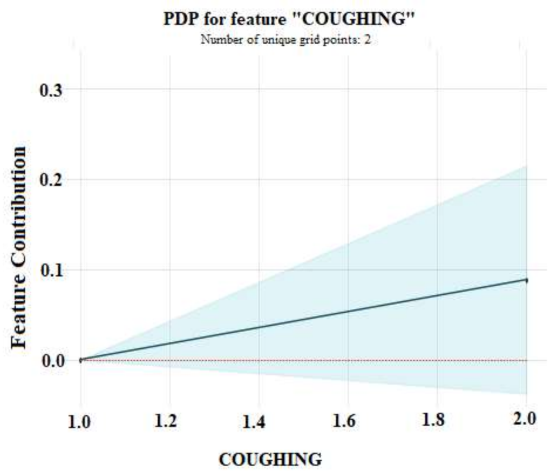

Figure 8: PDP plot for COUGHING feature

Figure 8 illustrates how the different feature values for the COUGHING feature contribute to the overall accuracy of the lung cancer prediction model. Figure 8 demonstrates that the features contribution to the model prediction will increase slightly linearly from 0 to around 0. if the COUGHING value is increased from 1 to 2.

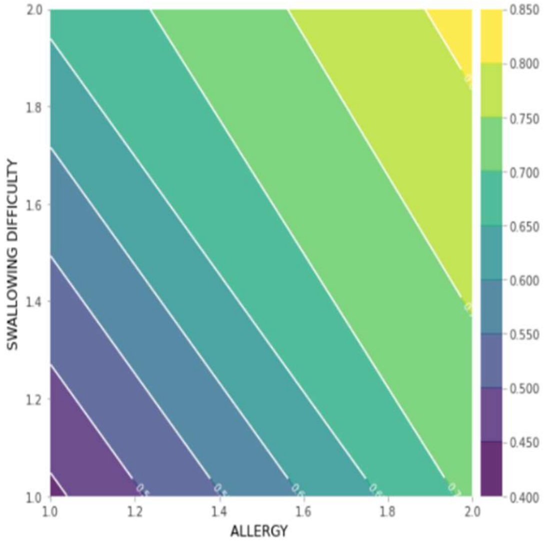

Figure 9: 2D PDP plot for SWALLOWING DIFFICULTY and ALLERGY feature

Figure 9 illustrates how the combined values of the COUGHING and SWALLOWING DIFFICULTY features have an effect on lung cancer prediction. There is a disparity in the contribution that various feature values make to the lung cancer prognosis. The relationship between these two factors in the development of lung cancer is seen in Figure 10.

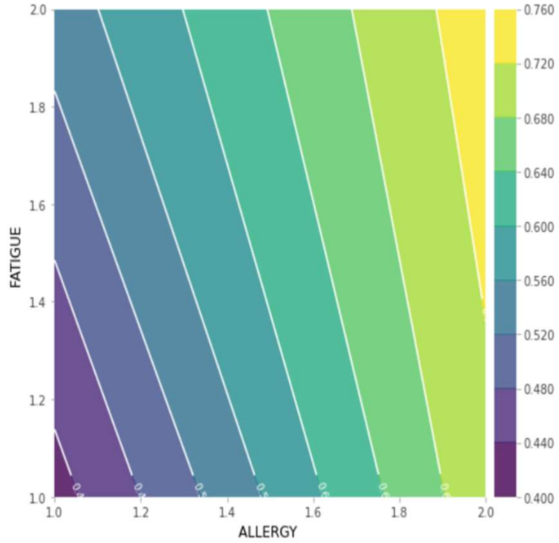

Figure 10: 2D PDP plot for FATIGUE and ALLERGY feature

Local model explainability for two individual instances is depicted in Figures 11 and 12. For those two instances the plots show which features are leading to predict the particular class of lung cancer. Coughing, wheezing, and alcohol use are the elements that are highlighted in blue in Figure 11, and these are the features that are positively contributing to the ability to predict the presence of lung cancer. In Figure 12, the features that are highlighted in blue are COUGHING, WHEEZING, and ALCOHOL_CONSUMING. These are the features that have a negative impact on the ability to forecast the absence of lung cancer.

5. CONCLUSION AND FUTURE WORK

The main focus of this research is to utilize machine learning techniques to predict the occurrence of lung cancer based on secondary data. Additionally, we aim to uncover the underlying factors contributing to this prediction by employing explainable artificial intelligence (XAI) tools. These tools facilitate the identification of feature importance, partial dependencies of the features, as well as global and local explainability. In order to enhance the trainability of the data, we perform preprocessing on the dataset and implementing data balancing techniques to mitigate potential biases in the decision-making process of the classifiers. Among all the algorithms LR, RF and XGB show maximum performance and we employ the XAI tools on RF to show the explainability. In global explainability, we rank the features and show it SHAP 0 and SHAP 1 explanation. Then show the partial dependencies of the top features using 1d and 2d PDP plot. This outcome of the proposed system can be a pathway to connecting the domain expert to the world of Artificial Intelligence. the XAI tools on RF to show the explainability. In global explainability, we rank the features and show it SHAP 0 and SHAP 1 explanation. Then show the partial dependencies of the top features using 1d and 2d PDP plot. This proposed system can be a pathway to connecting the domain expert to the world of Artificial Intelligence.

In future we will map domain knowledge with the XAI tools explanation and includes existing datasets for analysis. More dataset of this domain will add in further research.

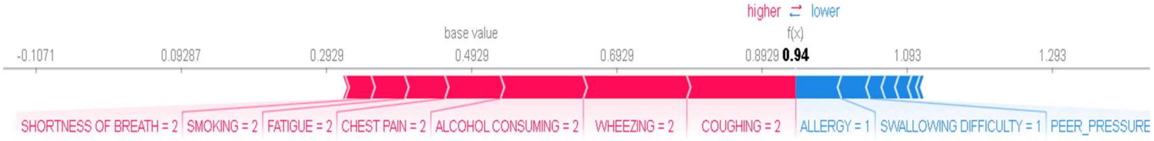

Figure 11: Individual instance explainability

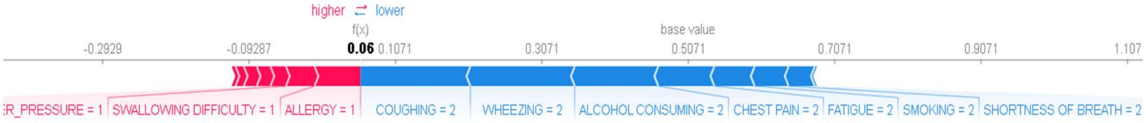

Figure 12: Individual instance explainability

REFERENCES:

[1] Ozcan, G., Singh, M., & Vredenburgh, J. J. Leptomeningeal Metastasis from Non–Small Cell Lung Cancer and Current Landscape of Treatments. Clinical Cancer Research, 2023, 29(1), 11-29.

[2] Ling S., Hu Z., Yang Z., Yang F., Li Y., Lin P., et al. Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. Proc Natl Acad Sci U S A. 2015; 112:E6496–E6505.

[3] International Cancer Genome Consortium, Hudson T.J., Anderson W., Artez A., Barker A.D., Bell C., et al. International network of cancer genome projects. Nature. 2010; 464:993–998.

[4] Pavlopoulou A., Spandidos D.A., Michalopoulos I. Human cancer databases (review) Oncol Rep. 2015; 33:3–18.

[5] ALzubi, J. A., Bharathikannan, B., Tanwar, S., Manikandan, R., Khanna, A., & Thaventhiran, C. (2019). Boosted neural network ensemble classification for lung cancer disease diagnosis. Applied Soft Computing, 80, 579-591.

[6] Petousis, P., Winter, A., Speier, W., Aberle, D. R., Hsu, W., & Bui, A. A. (2019). Using

sequential decision making to improve lung cancer screening performance. Ieee Access, 7, 119403-119419.

[7] Talukder, M. A., Islam, M. M., Uddin, M. A., Akhter, A., Hasan, K. F., and Moni, M. A., Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. Expert Systems with Applications, 2022, 205, 117695.

[8] Chandrashekar, N., & Pandi, A., Baicalein: A review on its anti-cancer effects and mechanisms in lung carcinoma. Journal of Food Biochemistry, 2022, 46(9), e14230.

[9] Gu, C., Dai, C., Shi, X., Wu, Z., & Chen, C., A cloud-based deep learning model in heterogeneous data integration system for lung cancer detection in medical industry 4.0. Journal of Industrial Information Integration, 2022, 30, 100386.

[10] Binson, V. A., Subramoniam, M., & Mathew, L., Detection of COPD and Lung Cancer with electronic nose using ensemble learning methods. Clinica Chimica Acta, 2021, 523, 231-238.

[11] P. R. Radhika, R A. Nair and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms", 2019 IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT), pp. 1-4, 2019, February.

[12] Zo, S., Shin, S. H., Jeong, B. H., Lee, K., Kim, H., Kwon, O. J., & Um, S. W.,102P Comparison of clinical characteristics and prognosis of incidentally detected and screening detected lung cancers. Annals of Oncology, 2022, 33, S79.

[13] "Data World," 18 9 2017. [Online]. Available: https://data.world/sta427ceyin/survey-lungcancer/workspace/file?filename=survey+lung+cancer.csv.

[14] Enders, C. K., Applied missing data analysis. Guilford Publications. 2022.

[15] Hasan, M., Islam, M. M., Sajid, S. W., & Hassan, M. M., The Impact of Data Balancing on the Classifier's Performance in Predicting Cesarean Childbirth. In 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), December 2022. pp. 1-4.

[16] Rabbi, M. F., Moon, M. H., Dhonno, F. T., Sultana, A., & Abedin, M. Z., Foreign Currency Exchange Rate Prediction Using Long Short-Term Memory, Support Vector Regression and Random Forest Regression. In Financial Data Analytics: Theory and Application, 2022, pp. 251-267. Cham: Springer International Publishing.

[17] Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P., Deep learning-based exchange rate prediction during the COVID-19 pandemic. Annals of Operations Research, 2021. 1-52.

[18] Soni, M., & Varma, S., Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (Ijert), 2020, Volume, 9.

[19] Datta, R. K., Sajid, S. W., Moon, M. H., & Abedin, M. Z., Foreign currency exchange rate prediction using bidirectional long short term memory. In The big data-driven digital economy: Artificial and computational intelligence, 2021, pp. 213-227. Cham: Springer International Publishing.

[20] Yu, L., Zhou, R., Chen, R., & Lai, K. K., Missing data preprocessing in credit classification: One-hot encoding or imputation?. Emerging Markets Finance and Trade, 2022, 58(2), 472-482.

[21] Kavzoglu, T., & Teke, A., Predictive Performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). Arabian Journal for Science and Engineering, 2022, 47(6), 7367-7385.

[22] Jiang, L., Wang, S., Li, C., & Zhang, L., Structure extended multinomial naive Bayes. Information Sciences, 2016, 329, 346-356.

[23] Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z., & Li, Z., Prediction of home energy consumption based on gradient boosting regression tree. Energy Reports, 2021, 7, 1246-1255.

[24] Walker, K. W., & Jiang, Z., Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. The Journal of Academic Librarianship, 2019, 45(3), 203-212.