KSÜ Mühendislik Bilimleri Dergisi, 27(3), 2024
Araştırma Makalesi

KSU J Eng Sci, 27(3), 2024
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 25.01.2024
Kabul Tarihi : 04.03.2024

Received Date : 25.01.2024
Accepted Date : 04.03.2024

PREDICTING LUNG CANCER USING EXPLAINABLE ARTIFICIAL INTELLIGENCE AND BORUTA-SHAP METHODS

AÇIKLANABİLİR YAPAY ZEKA VE BORUTA-SHAP YÖNTEMLERİYLE AKCİĞER KANSERİNİN ÖNGÖRÜLMESİ

Erkan AKKUR[1*] (ORCID: 0000- 0001-5573-5096)
Ahmet Cankat ÖZTÜRK[2] (ORCID: 0000- 0002-7082-6479)

[1] Turkish Medicine and Medical Agency, Ankara, Türkiye
[2] Presidency of The Republic of Turkey Secretariat of Defence Industries Agency, Ankara, Türkiye

*Sorumlu Yazar / Corresponding Author: Erkan AKKUR, eakkur@gmail.com

ABSTRACT

Machine learning algorithms, a popular approach for disease prediction in recent years, can also be used to predict lung cancer, which has fatal effects. A prediction model based on machine learning algorithms is proposed to predict lung cancer. Five decision tree-based algorithms were preferred as classifiers. The experiment was conducted on a publicly available data set that contained risk factors. The Boruta-SHAP approach was employed to reveal the most salient features in the dataset. The use of the feature selection method improved the performance of the classifiers in the prediction process. Experiments were conducted using all features and reduced features separately. When comparing all the classifiers' performances, the XGBoost algorithm produced the best prediction rate with an accuracy of 97.22% and an AUROC of 0.972. The proposed model has a good classification rate compared to similar studies in the literature. We used the SHAP (SHapley Additive exPlanation) approach to investigate the effect of risk factors in the dataset on the model output. As a result, allergy was found to be the most significant risk factor for this disease.

Keywords: Lung cancer, machine learning, explainable artificial intelligence, feature selection

ÖZET

Son yıllarda hastalık tahmini için popüler bir yaklaşım olan makine öğrenmesi algoritmaları, ölümcül etkileri olan akciğer kanserinin tahmininde de kullanılabilir. Bu çalışmada, akciğer kanserini tahmin etmek için makine öğrenmesi algoritmalarına dayalı bir tahmin modeli önerilmiştir. Sınıflandırıcı olarak beş karar ağacı tabanlı algoritma tercih edilmiştir. Deney, risk faktörlerini içeren kamuya açık bir veri seti üzerinde gerçekleştirilmiştir. Veri setindeki en belirgin özellikleri ortaya çıkarmak için Boruta-SHAP yaklaşımı kullanılmıştır. Öznitelik seçim yönteminin kullanılması sınıflandırıcılarının tahmin işleminde göstermiş oldukları performansları artırmıştır. Deneyler tüm özellikler ve indirgenmiş özellikler ayrı ayrı kullanılarak gerçekleştirilmiştir. Tüm sınıflandırıcıların performansları karşılaştırıldığında, 97.22% doğruluk ve 0.972 AUROC ile en iyi tahmin oranını üreten XGBoost algoritması olmuştur. Önerilen model, literatürdeki benzer çalışmalara kıyasla iyi bir sınıflandırma oranına sahiptir. Veri setindeki risk faktörlerinin model çıktısı üzerindeki etkisini araştırmak için SHAP (SHapley Additive exPlanation) yaklaşımını kullandık. Sonuç olarak, alerji bu hastalık için en önemli risk faktörü olarak bulunmuştur.

Anahtar Kelimeler: Akciğer kanseri, makine öğrenmesi, açıklanabilir yapay zeka, öznitelik seçimi

ToCite: AKKUR, E., & OZTURK A.C., (2024). PREDICTING LUNG CANCER USING EXPLAINABLE ARTIFICIAL INTELLIGENCE AND BORUTA-SHAP METHODSI. *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(3), 792-803.

INTRODUCTION

The unbalanced proliferation of cells in the lung can lead to malignant tumoral formations known as lung cancer (LC). LC is acknowledged globally as one of the most frequent and lethal forms of cancer (Sung et al., 2021). LC is regarded as the second most common cancer in men, following prostate cancer and breast cancer in women. According to global statistics, this disease was estimated to cause 2.2 million new diagnoses and 1.8 million deaths in 2020 (Li et al., 2023). The incidence of this disease is growing globally because of increased access to tobacco and industrialization. If this disease is detected early, the survival rate can be increased. Imaging techniques and examination of symptoms can help early diagnosis. Imaging modalities such as thoracic computed tomography and chest X-rays are used to diagnose LC. Despite providing reliable results, these techniques may be expensive. Shortness of breath, chest pain, fever, wheezing, and persistent cough are symptoms that can help predict this disease at an early stage. (Latimer & Mott, 2015). The onset of this disease can be predicted using early symptoms using machine learning (ML) algorithms. These algorithms use statistical analysis and mathematical optimizations and can provide predictive results (e.g. disease, no disease) based on input data, such as text or images. ML algorithms are widely used in the healthcare industry to predict diseases such as breast cancer, liver cancer, cardiovascular disease, kidney disease, COVID-19, and diabetes. (Kaplanoglu & Nasab, 2023; Turk & Kokver, 2022; Turk et al., 2022).

The prediction performance of the ML algorithm is directly affected by the number of features in the datasets. A limited number of features may result in classes not being properly separated in certain cases. A large number of features can also lead to problems such as increased training time and degraded classification performance for high-noise features. Therefore, it is important to select the relevant features to be given as input to ML algorithms. Feature selection (FS) is defined as selecting the most important attributes in the dataset. FS is one of the processes that enhance the classification performance of ML algorithms (Cai et al., 2018; Theng, 2023). Due to the black box feature of ML algorithms and their complex structure, it is difficult to understand and interpret the behaviour of the models in the prediction process. Despite providing satisfactory results for problems such as disease prediction, these algorithms also present problems like model interpretability. Explainable artificial intelligence (EAI) is considered a process package that can assist in the comprehension and explanation of model results. With these methods, the prediction results of the models can be made more understandable by investigating the relationships between inputs and outputs for each sample. Furthermore, these methods enable a predictive model designed for disease prediction to be more comprehensively explained in a clinical setting (Confalonieri et al., 2021; Arrieta, 2020).

This study aims:

- to select the best features to achieve maximum accuracy in lung cancer prediction using the Boruta-SHAP technique, a powerful feature selection method,
- to achieve the best classification rate in predicting lung cancer using different ML algorithms.
- to investigate the effect of the attribute associated with SHAP (SHapley Additive exPlanation) values, one of the EAI methods, on the prediction of the model that achieves the best classification rate.

RELATED WORKS

Recent related studies on predicting LC utilizing ML approaches are presented in this section. One study used different classifiers for early prediction of LC (Fasial et al., 2018). A dataset from the UCI data repository was used to implement the models designed in the study. The Gradient-boosted Tree (GBT) algorithm was able to achieve the highest classification rate with 90% accuracy based on the comparisons. Patra conducted a study to predict lung cancer by comparing different classifiers (Patra, 2020). The experiments were performed using Weka tools. The dataset was taken from UCI Respiratory. The Radial Basis Function (RBF) algorithm showed the best result with an accuracy of 81.25%. Abuya studied combining PCA with ensemble learning (EL) (Abuya, 2023). The experiments were applied to the ELVIRA Biomedical Data Set Repository (EBDSR) dataset. The proposed method achieved 98.25% accuracy. Agarwal et al. tried to predict lung cancer by comparing different classifiers with each other (Agarwal et al., 2022). The study was implemented at COLAB. The experiments were applied to a dataset from Kaggle. The Random Forest (RF) algorithm scored the best success rate with 92.3% accuracy. Dristas and Trigka carried out a study on LC risk prediction using various classifiers (Dristas & Trigka, 2022). The imbalance problem in the dataset was dealt with through the use of the SMOTE (Synthetic Minority Oversampling Technique). The Rotation Forest technique obtained the highest prediction rate with an accuracy of 97.1%. Dirik sought to predict LC with nine different classifiers (Dirik, 2023). As a result of the study, the Naive Bayes (NB) algorithm yielded a better prediction rate than the other classifiers with an accuracy of 91%. Nasser and Abu-Naser utilized an artificial neural

network (ANN) to predict LC (Nasser & Abu-Naser, 2019). In the study, the dataset was split 80:20 and the proposed method achieved a 96.67% accuracy rate. Omar and Nassif presented a comparative study using FS and ML algorithms (Omar & Nassif, 2023). The Principal Component Analysis (PCA) and Correlation-based feature selection (CFS) approach were used to identify the most discriminating features. Three classifiers were utilized to predict LC. Multiplayer Perceptron (MLP) algorithm outperformed with 90% accuracy. Ojha applied six different ML algorithms for disease prediction and compared their performance results (Ojha, 2023). The dataset used in the study was split 80:20 into training and test data. Logistic Regression (LR) gave the best result with 94.7% accuracy. Table 1 provides a summary of research on LC prediction using ML algorithms.

Table 1. Existing Studies LC Prediction

Study	Dataset	Feature Selection	The Best Classifier	Accuracy (%)
Fasial et al. (2018)	UCI repository	-	GBT	90.00%
Patra (2020)	UCI repository	-	RBF	81.25%
Abuya (2023)	EBDSR	PCA	EL	98.25%
Agarwal et al. (2022)	Kaggle	-	RF	92.30%
Dristas & Trigka (2022)	Kaggle	Gain Ratio	Rotation Forest	97.10%
Dirik (2023)	Kaggle	-	NB	91.00%
Nasser & Abu-Naser (2019)	Kaggle	-	ANN	96.67%
Omar & Nassif (2023)	Kaggle	CFS and PCA	MLP	90.00%
Ojha (2023)	Kaggle	-	LR	94.70%

The literature on LC prediction has utilized various methods of FS and classification. These studies have shown that ML algorithms can be used to successfully predict LC. The Boruta-SHAP method used in this study is the first time it has been used in the literature for LC prediction, as far as we know. Furthermore, there is no research on how to interpret lung cancer prediction models using EAI methods.

MATERIALS AND METHODS

This section explains the model built for the prediction of LC. Figure 1 illustrates the suggested predictive models. In the suggested approach, lung cancer data is first acquired and then the dataset is preprocessed. The features in the dataset are given as input to the ML algorithms before and after applying FS. At this stage, the effect of the FS method on the classification performance of ML algorithms was analyzed. Then, using different performance metrics, the classifier with the best prediction result was determined. Finally, SHAP analysis was used to interpret the model with the best prediction result.

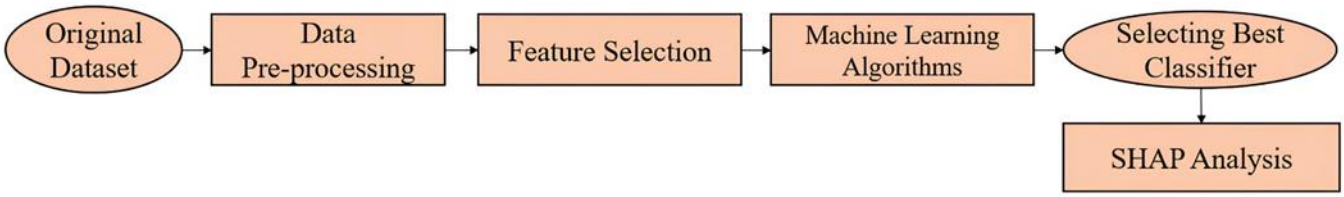

Figure 1. The Suggested Framework of The Suggested Model

Description of the Dataset

The "Lung Cancer" dataset acquired from the Kaggle online web page (Lung Cancer Prediction Dataset, 2013) was employed. It consists of 309 samples and 15 different features, including basic characteristics of the individuals as well as information about their harmful habits and the presence of different abnormalities that can be observed in their health. It also contains as an output attribute a label indicating whether the person has been diagnosed with lung cancer or not.

Data preprocessing facilitates data analysis and improves the performance and speed of models, which is an important element of ML algorithms. The process starts with detecting and handling missing or duplicate values in the data set.

Data Pre-processing

When the dataset is analyzed, no missing data is found. However, 33 duplicate samples are observed in the dataset. Hence, the duplicate sample is removed to complete the dataset. As a result, a total of 276 samples are acquired in the dataset, 238 with lung disease and 38 without lung disease. There is an imbalance problem in the class distribution

in the dataset. The Adaptive Synthetic Sampling (ADASYN) approach was used to solve this problem. The characteristic of this synthetic data generation technique is that small samples are not repeated and more data is generated for hard-to-learn samples (He et al., 2008). Thus, the balance between the number of people diagnosed with lung cancer and those who were not diagnosed was achieved. The distribution of the output feature before and after data balancing is shown in Figure 2.

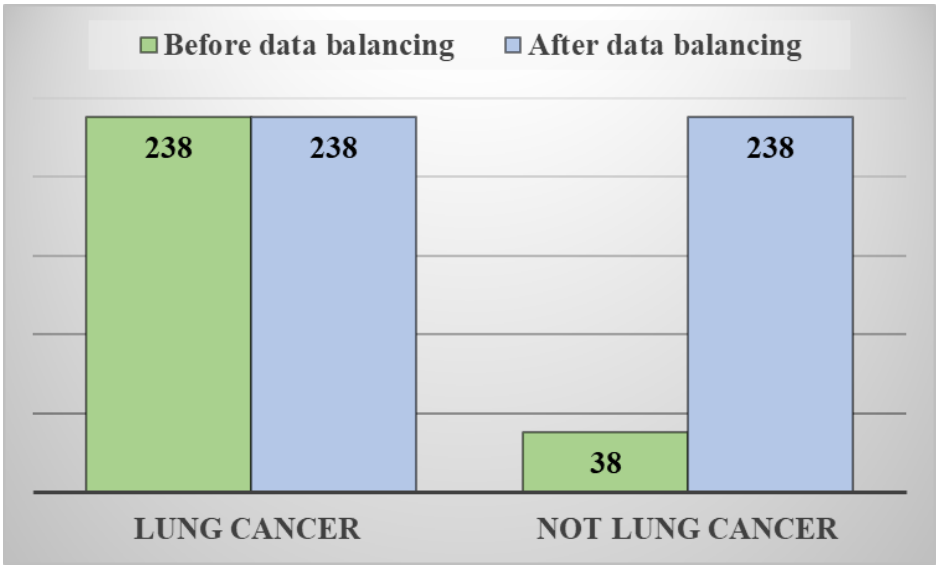

Figure 2. Data Balancing Process

Boruta-SHAP Feature Selection Approach

The Boruta-SHAP method, which combines the Boruta algorithm and SHAP values, was employed in this study to select significant features from the dataset. The Boruta method is an FS algorithm that is built around the RF technique's working principle. It attempts to select the entire set of features related to the dependent variable through a wrapper algorithm. It uses the RF algorithm as a classifier to filter out the features that are related to the target variable among all the features to form a new subset. It extracts the relative importance of all features in the dataset concerning the dependent variable, highlights the important features, and eliminates the unnecessary ones. By using a black box prediction model with the best prediction accuracy, features that are relatively closely connected to the target variable can be identified (Kursa MB, Rudnicki, 2010; Keany, 2020). The flowchart of the Boruta algorithm is shown in Figure 3.

The algorithm comprises the following steps:

1- Duplicates of the original dataset are created, and the variable values in these duplicate datasets are compared. The original and compared data are merged to train an RF model that measures variable importance.
2- A Z-score is calculated for each variable. The Z-score is derived from the Random Forest model which is the standardized version of the calculated variable significance values.
3- The one with the highest Z score among the shadow variables is determined.
4- When the resulting Z-score exceeds the highest Z-score, it is classified as significant; otherwise, it is classified as insignificant.
5- Repeat the above operations until all variables are labelled.

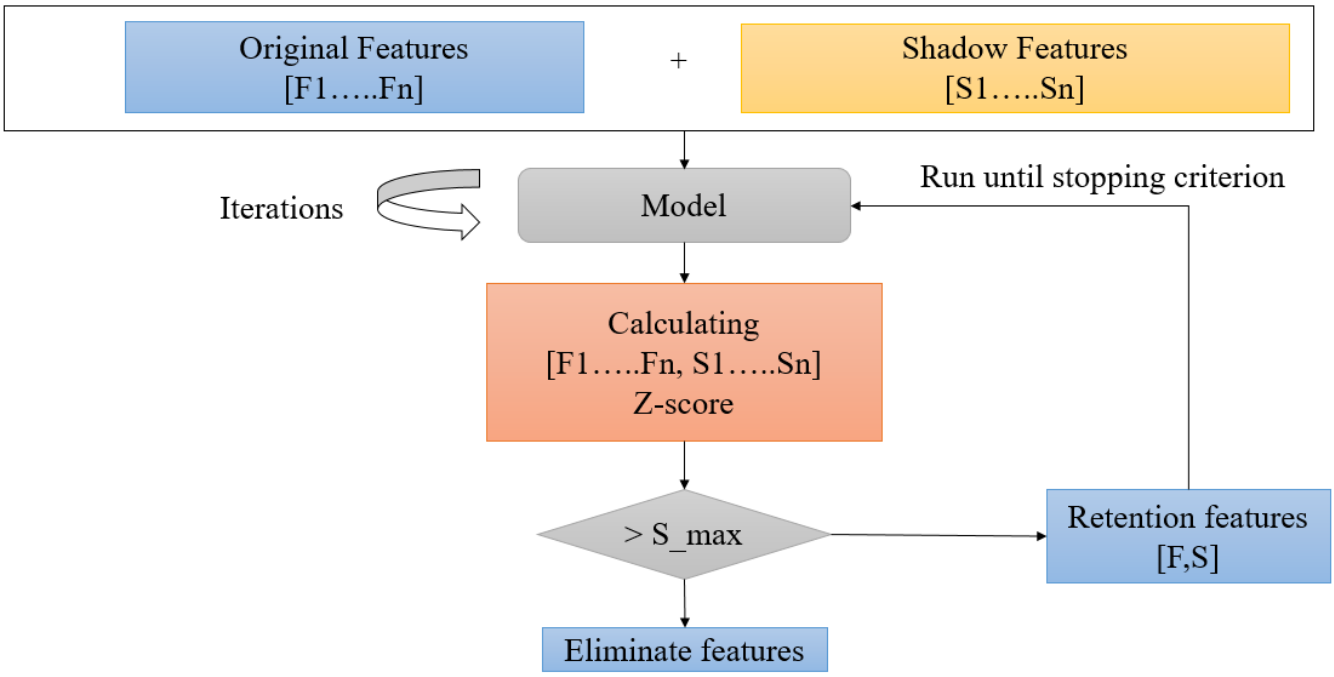

Figure 3. Boruta-SHAP Flowchart

The Boruta-SHAP method combines the Boruta method with SHAP values (Keany, 2020). The Boruta algorithm is an approach that seeks to capture all features of importance that can be used for prediction. The importance of a feature is measured by permutation importance, which is defined as the loss in model accuracy due to the random mixing of features in a dataset (Kursa and Rudnicki, 2010). However, this measure of importance makes the algorithm computationally expensive and is not considered a reliable measure of global feature importance, which led to the replacement of permutation importance with SHAP importance in the Boruta-SHAP algorithm (Keany, 2020). This method of variable selection Boruta-SHAP built in the Python programming language for its implementation library was used. Iteration, one of the parameters of the Boruta-SHAP method (iteration) number was set as 100.

Machine Learning Algorithms

This study investigates Decision Tree (DT) and decision tree ensemble models to predict lung cancer. DT is an effective ML method that is simple but powerful. In the prediction process, the algorithm exploits the recursive partitioning of different features according to a tree structure (Charbuty & Abdulazeez, 2021). The DT-based ensemble model is an ML approach that blends multiple trees to produce more accurate predictions. A multi-tree structure is preferable in ensemble models and is beneficial for error correction. In community models, the strengths of one tree can balance the weaknesses of another tree (Tsiligaridis, 2023). As tree-based ensemble models, Random Forest (RF), Extra Tree Classifier (ETC), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost) algorithms are employed in this study. The RF algorithm leverages trees as the base classifier and exploits an additional level of randomization. In this method, multiple trees are trained on a random subset of features. Merging the predictions of multiple trees through majority voting produces the final prediction (Palimkar, 2022). In the ETC algorithm, a random split point of each feature is selected to enhance the variety between the trees and minimize overfitting (Geurts et al., 2006). The XGBoost algorithm inserts a new model that can reduce errors in the existing model. The algorithm computes a step function by using the second derivative of the loss function. This step ensures that the output of the new model is adjusted to minimize the error of the existing model (Chen & Guestrin, 2016). The AdaBoost algorithm aims to create a stronger classifier by combining multiple weak classifiers. At each stage, it predicts by re-running the classifier by augmenting the weight of the inaccurate predictions made as a result of the preceding phase (Wang, 2012).

SHAP (SHapley Additive ExPlanations) Analysis

The SHAP is an approach that uses coalitional game theory to explain the outcome of any ML algorithm (Lundberg & Lee, 2017). The SHAP values consider each feature value of the data sample as the “player and the prediction as the “payout” and investigate the distribution of the “payout” across different features. In a sense, SHAP values can

be thought of as the average marginal contribution of a feature given all possible coalitions. The contribution of values may be either positive or negative. The positive values strengthen the prediction process, while negative ones weaken it. The absolute values of the feature correspond to its impact on the model. Larger values have a greater overall impact. The objective of the method is to present the prediction for any instance x_i by taking into account contributions from individual features (Yao et al., 2022). Using Equation 1 can provide the SHAP values in a model with a prediction function f(x) and m features.

$$\emptyset_i = \sum_{p \subseteq N \setminus \{i\}} \frac{|s|!\,(m-|s|-1)!}{m!} [f_x(s \cup \{i\}) - f_x(s)] \quad (1)$$

The formula expressed in Equation 1 is the sum of all possible subsets (p) of all feature values except feature value i. |s|! is the number of permutations of features that precede feature value i. (m-|s|-1)! is the number of permutations of features that follow feature value i. $\emptyset_i$ represents the SHAP value. The difference operation in the equation expresses the marginal contribution of adding the first feature value to s (Lundberg & Lee, 2017). The SHAP values can be classified into Kernel SHAP, Tree SHAP, and Deep SHAP (Yao et al., 2022). Since tree-based methods are used in this study, the Tree SHAP method was used. All of these analyses were carried out in Python through the "SHAP" package.

EXPERIMENTAL RESULTS

Experiment Setups

In this study, we used Jupyter Notebook 3.8.16 in Python, which provides libraries for data pre-processing, visualization, prediction, and classification to implement the prediction model for predicting lung cancer. The dataset was randomly subdivided 7:3 into training and test sets. A 5-fold cross-validation technique was used to validate the prediction of the classifiers during the training process. The performance of the classifiers in the test set was evaluated by accuracy, precision, recall, and F1-score, respectively.

Evaluation

In this research, five ML algorithms were employed for LC prediction. Different perspectives are used to evaluate the classifications' prediction performance. The models' functionality using all features in the dataset is first examined. The impact of the FS method on the performance of the classifier is also analyzed.

Classification Results of Using the Full Feature Set

Using the entire feature set, all ML models in this section are evaluated as datasets to predict the outcome of a binary disease. Table 2 presents the results produced by the ML models for predicting LC using the full feature set. In lung cancer prediction, the DT algorithm had the lowest prediction rate (92.36% accuracy, 91.78% precision, 93.0% recall, and 92.41 % F1-score) and the XGBoost algorithm had the highest classification performance, with an accuracy of 95.83%, precision 95.83%, recall of 95.83%. and F1-score of 95.83%.

Table 2. Classification Results of ML Algorithms Using the Full Feature Set

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	92.36	91.78	93.06	92.41
AdaBoost	93.06	93.06	93.06	93.06
ETC	93.75	92.00	95.83	93.88
RF	93.75	90.91	97.22	93.96
XGBoost	95.83	95.83	95.83	95.83

Boruta-SHAP Feature Selection Results

The Boruta-SHAP method was utilized to identify the significance of features in the dataset, resulting in model selection flexibility. It allows for a visual representation of the selected features (Kim et al., 2022). Figure 4 depicts the impact of each feature on prediction via Boruta-SHAP. Green bars mark accepted features, yellow bars mark tentative features, and red bars mark rejected features. The blue box represents the minimum, maximum, median, and mean features. The FS process led to the selection of nine significant risk factors. The selected features were allergy, swallowing difficulty, age, gender, smoking, alcohol consumption, peer pressure, fatigue, and yellow fingers, respectively.

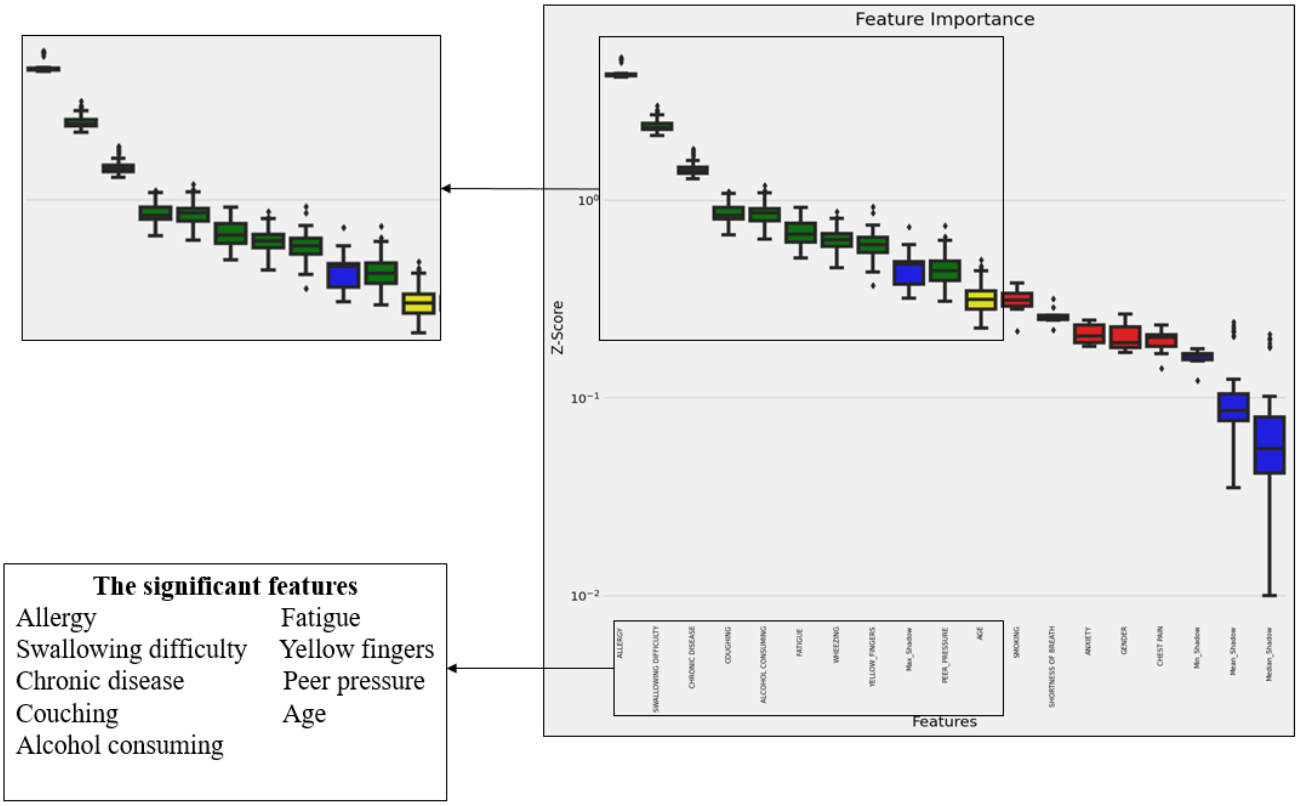

Figure 4. Selection of Features by Boruta-SHAP

Classification Results Using Reduced Feature Set

After feature selection, the selected features were given as input to the classifiers. The performance of the classifiers with the reduced feature set is presented in Table 3. In lung cancer prediction, in line with the results obtained before using the feature selection method, the DT algorithm had the lowest prediction rate (94.44% accuracy, 95.71% precision, 93.06% recall, and 94.37% F1-score), while the XGBoost algorithm had the highest classification performance with 97.22% accuracy, 95.95% precision, 98.61% recall, and 97.26% F1-score.

Table 3. Classification Results of ML Algorithms Using the Reduced Feature Set

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DT	94.44	95.71	93.06	94.37
AdaBoost	95.14	92.21	98.61	95.30
ETC	95.83	94.59	97.22	95.89
RF	96.53	94.67	98.61	96.60
XGBoost	97.22	95.95	98.61	97.26

The Effect of The Boruta-SHAP Method on the Performance of ML Algorithms

The performance of the classifiers in terms of accuracy using all features and selected features is shown in Figure 5. The use of the feature selection algorithm improved the accuracy of the DT, AdaBoost, and ETC algorithms by 2.08%, the RF algorithm by 2.78%, and the XGBoost algorithm by 1.39%. It indicates that by reducing the number of features using a feature selection technique, classifiers can achieve a higher prediction rate.

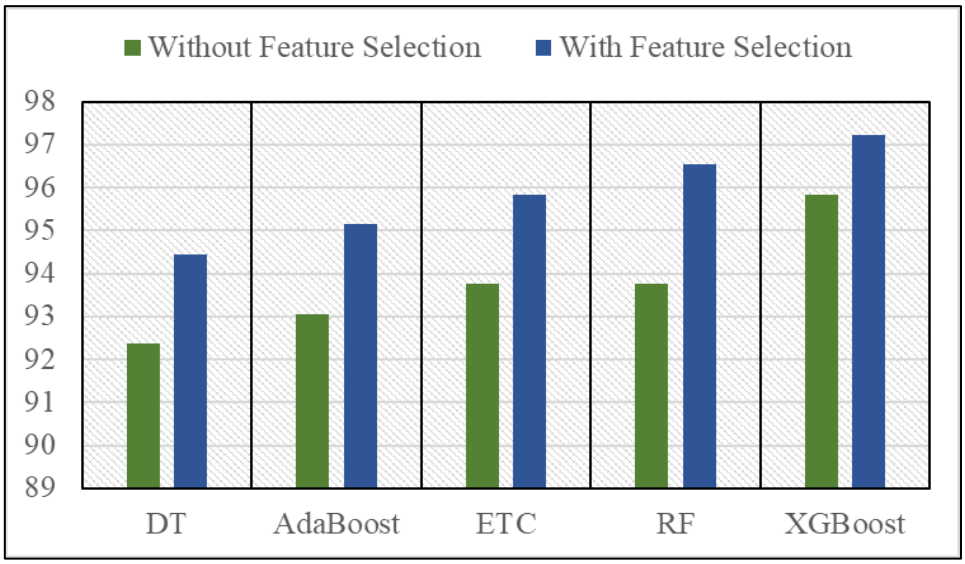

Figure 5. The Effect of the Boruta-SHAP Method on the Performance of ML Algorithms

Determining the ML Algorithm with the Best High Classification Performance for Predicting Lung Cancer

In this study, we utilized ROC (Receiver Operating Characteristic Curve) to identify the ML algorithm with the best prediction rate in the process of predicting LC. This metric can visually show the performance of classification models at all classification thresholds. At different classification thresholds, it plots the false positive rate against the true positive rate. The area under the curve (AUC) is expressed as the area under the ROC curve. It indicates if the data is sufficient to enable precise differentiation between predicted values (Bradley, 1997). The AUROC values are illustrated in Figure 6. With an AUROC of 0.972, the XGBoost algorithm is the most accurate predictor, while the DT algorithm has the worst performance among the classifiers with an AUROC of 0.944.

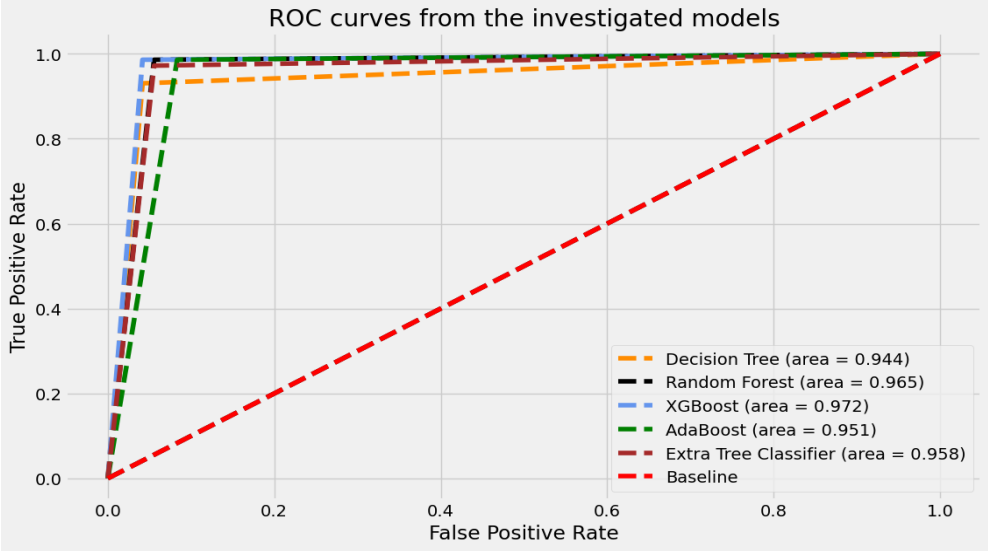

Figure 6. AUROC Curves of Classifiers

Comparison of the Suggested Predictive Model with Similar Studies in the Literature

There have been considerable studies in the literature on LC prediction using the ML algorithm. Table 4 presents a benchmark of the suggested model against models from other earlier studies using the same dataset. During analysis of the comparison table, it is observed that the suggested model has a good classification rate compared to other existing models. The FS method was used in certain studies on this subject in the literature, but it was not favoured in others. Our study utilized the Boruta-SHAP method, which offers flexibility in model selection and a visual

representation of selected key variables. At the same time, to our knowledge, the Boruta-SHAP method has not been used before in the diagnosis of lung cancer.

Table 4. Comparison of Suggested Model with Similar Studies

Studies	The Best Classifier	Accuracy (%)
Agarwal et al. (2022)	RF	92.30%
Dristas & Trigka (2022)	Rotation Forest	97.10%
Dirik (2023)	NB	91.00%
Nasser & Abu-Naser (2019)	ANN	96.67%
Omar & Nassif (2023)	MLP	90.00%
Ojha (2023)	LR	94.70%
Suggested model	XGBoost	97.22%

Model Interpretability

This study presents the SHAP approach for modelling interpretability. The significance of the features in the XGBoost model with the highest prediction rate was analyzed. Figure 7 demonstrates the feature prominence of the XGBoost model using the SHAP method. In analyzing SHAP values, it appears that allergy, fatigue, alcohol consumption, and yellow fingers are significant variables in predicting LC. As can be seen in Figure 7, the variable that contributes the most to the prediction model is allergy (+1.46), followed by fatigue (+1.18) and alcohol consumption (+1.15).

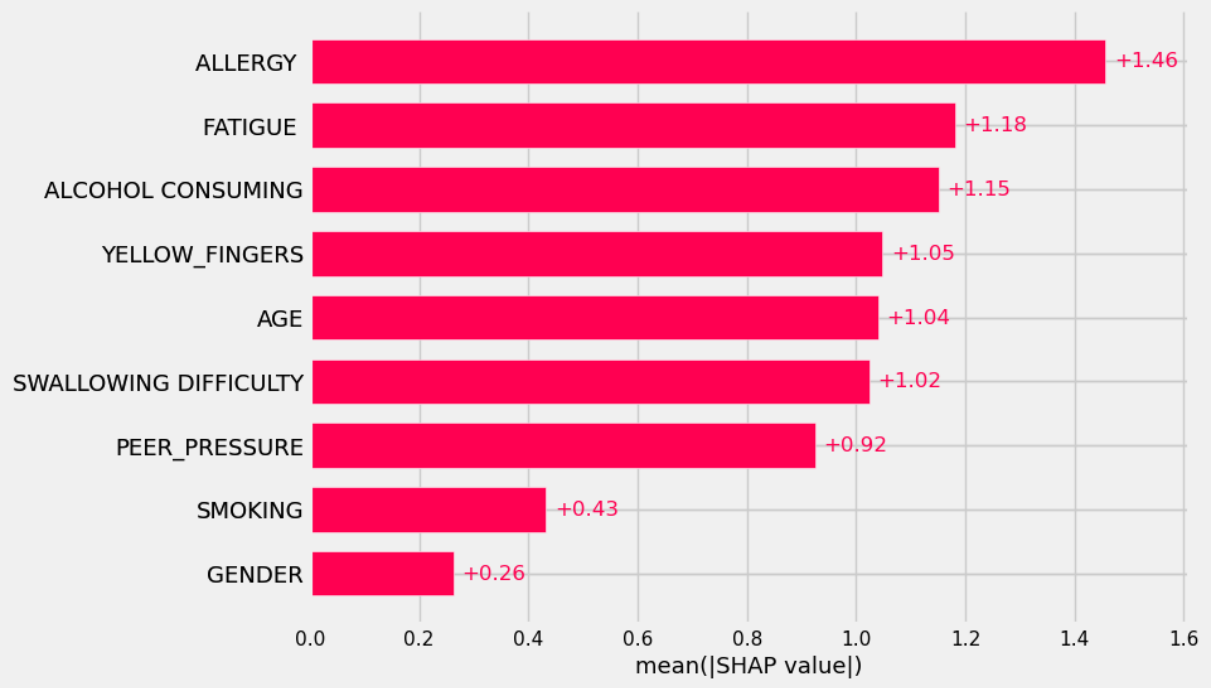

Figure 7. Feature Importance Based on SHAP Values

Another visual representation of SHAP analysis is the Beeswarm, which is displayed in Figure 8. It aims to display the distribution of multiple variables in a way that minimizes the friction between points. In the graph, the x-axis represents the data points and the y-axis represents the density of the points. The higher the SHAP value of a feature, the higher the log-likelihood of lung disease in the lung disease prediction model. Each sample in the dataset is run through the model, and each patient in the dataset is run through the model. A dot is produced for each feature association value so that a patient gets a dot in the row of each attribute. The dots are coloured according to the value of the feature for that patient and stacked vertically to show density. In Figure 8, the allergy is the most significant risk factor for lung cancer patients.

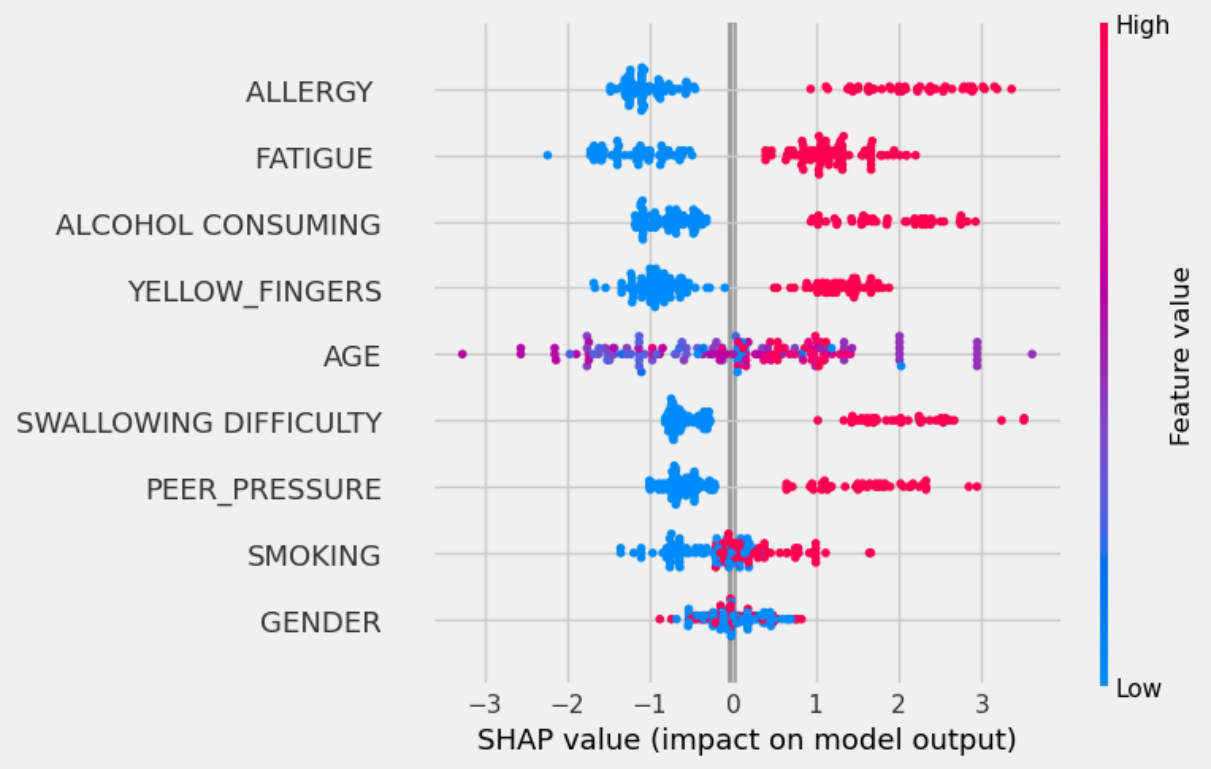

Figure 8. SHAP Summary Plot for The XGBoost Algorithm

CONCLUSION

LC causes many deaths worldwide. The earlier it can be diagnosed, the lower the mortality rate. For the diagnosis of this disease, it is important to identify and monitor the risk factors affecting the disease. In recent years, ML algorithms have become an important approach for the prognosis or prediction of various diseases. In this study, we attempt to predict LC using DT-based ML techniques. For this purpose, DT, AdaBoost, RF, ECT, and XGBoost algorithms were used for prediction. The suggested approach was applied to a publicly available dataset containing the risk factors of people with LC. The Boruta-SHAP approach was used to identify the most meaningful features in the dataset. To examine the effect of the FS algorithm, the performance of the classifiers was evaluated before and after the use of the FS algorithm. According to the analysis, the Boruta-SHAP approach had a positive impact on the performance of all classifiers. Experimental results reveal that with a lower number of features, LC disease can be correctly predicted at a higher rate. Comparing the performance of the classifiers, the XGBoost algorithm produced the best prediction rate with an accuracy of 97.22% and an AUROC of 0.972. In addition, this study seeks to bridge the gap in the interpretability of ML algorithms. Understanding the inner workings of a predicting model is crucial. For this purpose, we utilized the SHAP technique, that is one of the explainable artificial intelligence methods. It can help measure the contribution of features to the model. The XGBoost model with the best prediction rate was selected for SHAP analysis. As a result, the allergy feature was found to be the most prominent risk factor in predicting LC. Even though the model presented in this study delivers a reasonable prediction rate, it also has constraints. The dataset used in this study is publicly accessible. However, we consider it to be a reliable dataset as it includes risk factors for LC. A dataset from a hospital or an institute would have provided data with richer and more diverse features. However, access to medical data is restricted due to privacy reasons. In future studies, it is planned to construct deep learning-based prediction models on a dataset to be taken from the hospital environment.

REFERENCES

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3), 209-249.

Li, C., Lei, S., Ding, L., Xu, Y., Wu, X., Wang, H., Zhang, Z., Gao, T., Zhang, Y., Li, L. (2023). Global burden and trends of lung cancer incidence and mortality. *Chin Med J (Engl)*, 136(13):1583-1590

Latimer, K. M., & Mott, T. F. (2015). Lung cancer: diagnosis, treatment principles, and screening. *American family physician*, 91(4), 250-256.

Kaplanoglu, E., & Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1).

Turk, F. &. Kokver, Y. (2022). Application with deep learning models for COVID-19 diagnosis, *SAUCIS*, vol. 5, no. 2, pp. 169–180.

Turk, F., Luy, M., Barıscı, N. & Yalcınkaya, F., (2022), Kidney tumour segmentation using two-stage bottleneck block architecture, *Intelligent Automation and Soft Computing*, 33(1).

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70-79.

Theng, D., & Bhoyar, K. K. (2023). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 1-63.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1), e1391.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.

Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early-stage prediction of lung cancer. In 2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST) (pp. 1-4). IEEE.

Patra, R. (2020). Prediction of lung cancer using machine learning classifier. In: Chaubey, N., Parikh, S., Amin, K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore. DOI: 10.1007/978-981-15-6648-6_11.

Abuya, T.K. (2023). Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis. *Journal of Data Analysis and Information Processing*, 11, 175-199.

Agarwal S., Thakur S. and Chaudhary A. (2022, October). Prediction of lung cancer using machine learning techniques and their comparative analysis. 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India. DOI**:** 10.1109/ICRITO56286.2022.9965052.

Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, 6(4), 139.

Dirik, M. (2023). Machine learning-based lung cancer diagnosis. *Turkish Journal of Engineering*, 7(4), 322-330.

Nasser, I. M., & Abu-Naser, S. S. (2019). Lung cancer detection using artificial neural network. *International Journal of Engineering and Information Systems (IJEAIS)*, 3(3), 17-23.

Omar A. C. and Nassif A. B. (2023). Lung cancer prediction using machine learning based feature selection: A comparative Study, 2023 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, pp. 1-6. DOI: 10.1109/ASET56582.2023.10180436.

Ojha T. (2023), Machine learning based classification and detection of lung cancer, *Journal of Artificial Intelligence and Capsule Networks* 5(2):110-128.

Lung Cancer Prediction Dataset (2013). Available online: https://www.kaggle.com/datasets/mysarahmadbhat/lungcancer?fbclid=IwAR0uQ5K3mEbQZJcwQGYqlLJ5RydvsK2oU1Sa5vYvit0ECoqkx6vPR43JAM. / Accessed 02.01.2024.

He, H., Bai, Y., Garcia, E.A., Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

Kursa MB, Rudnicki WR. (2010). Feature selection with the Boruta package. *J. Stat. Softw*. 36(11): 1-13.

Keany E. (2020). Boruta-Shap: A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values. Zenodo: Geneva, Switzerland.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.

Tsiligaridis J., (2023). Tree-Based ensemble models and algorithms for classification, 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, pp. 103-106.

Palimkar, P., Shaw, R.N., Ghosh, (2022). A. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced Computing and Intelligent Technologies*; Springer: Berlin/Heidelberg, Germany, pp. 219–244.

Geurts P., Ernst D. & Wehenkel L. (2006). Extremely randomized trees, *Machine Learning*, vol.63, pp.3-42.

Chen T. & Guestrin C. (2016). XGBoost: A scalable tree boosting system. In Proc. of the 22Nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining pp. 785–94.

Wang, R. (2012). AdaBoost for feature selection, classification and its relation with SVM, a review. *Physics Procedia*, *25*, 800-807.

Lundberg S.M. & Lee S.I. (2017). A unified approach to interpreting model predictions." *Advances in neural information processing systems*, 30.

Yao L., Leng Z., Jiang J. & Ni F. (2022). Modelling of pavement performance evolution considering uncertainty and interpretability: a machine learning based framework, *International Journal og Pavement Engineering*, 23(14):5211-5226.

Kim, J. Lee, J. & Park, M. (2022). Identification of smartwatch-collected lifelog variables affecting body mass index in middle-aged people using regression machine learning algorithms and SHapley Additive Explanations. *Appl. Sci.* 12, 3819.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159.