Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 50-57, 10, 2023

Araştırma Makalesi

IJANSER
https://alls-academy.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 50-57, 10, 2023

Research Article

An Explainable Artificial Intelligence Based Early Lung Cancer Risk Prediction Using LightGBM

Ömer Miraç Kökçam[1*], Yunus Santur[2] and Muhammed Emre Çolak[2]

[1]*Department of Software Engineering, Fırat University, Türkiye*
[2]*Department of Artificial Intelligence and Data Engineering, Fırat University, Türkiye*

[*]*(omkokcam@firat.edu.tr) Email of the corresponding author*

(Received: 12 October 2023, Accepted: 23 October 2023)

(2nd International Conference on Recent Academic Studies ICRAS 2023, October 19-20, 2023)

ATIF/REFERENCE: Kökçam, Ö. M., Santur, Y. & Çolak, M. E. (2023). An Explainable Artificial Intelligence Based Early Lung Cancer Risk Prediction Using LightGBM. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(10), 50-57.

Abstract – Lung cancer is the most common cancer in the world and is also the deadliest. Early diagnosis can improve patients' life expectancy and reduce the cost of treatment. The aim of this study is to help physicians and patients by using explainable artificial intelligence methods to early diagnose risk of lung cancer. This study will create an opportunity for physicians to make early diagnosis and treatment strategies for patients. In this study, a machine learning model was developed to predict the risk of lung cancer by using the LightGBM algorithm. Furthermore, the SHAP method is used to explain why and how the model's predictions are made, thus making the AI model reliable. These explanations increase the acceptability and reliability of the predictions made by model, while helping physicians and patients understand the model's decisions. The results obtained show that the developed LightGBM model can predict the risk of lung cancer with a 100% accuracy rate. The model has achieved great results in terms of accuracy rate and sensitivity. In addition, the SHAP analyses explain which features each of the model's predictions is based on, which will increase the confidence of physicians and patients in the decisions of artificial intelligence.

Keywords – Cancer Risk Prediction, Machine Learning, Data Engineering, Explainable Artificial Intelligence

I. INTRODUCTION

Air pollution poses a significant worldwide health challenge, particularly within the backdrop of swift economic growth and the proliferation of urban areas. One of the most important effects of lung cancer is air pollution [1][2]. Lung cancer is the leading cause of cancer-related deaths worldwide. According to the report of the World Health Organization (WHO), lung cancer causes approximately 1.6 million deaths worldwide every year. Globally, it is estimated that the number of cancer cases will double by 2050, and lung cancer is expected to lead this list. The main reason behind these alarming statistics is that lung cancer is often diagnosed in advanced stages. Therefore, screening of high-risk groups, such as smokers, passive smokers, and those working in oil fields, is essential for early detection of lung cancer [3].

Figure 1 shows the distribution of cancer-related deaths worldwide in 2020 by cancer type [4]. The number of deaths by cancer type is an important indicator of public health and a critical data source in development and evaluation for cancer control strategies. Therefore, the visual representation of the number of cancer-related deaths reported in 2020 is extremely valuable for tracking progress in

the fight against lung cancer and shaping future health policies.

Figure 1. Cancer-releated deaths worldwide.

Many studies in the literature on lung cancer prediction focus on developing various decision support systems to support physicians and patients. Salman et al. [5] examined various methods to improve the performance of machine learning models which works on small datasets to predict cancer patients. In their study, they conducted a detailed analysis on four different machine learning models (Naive Bayes, Decision Tree, K-Nearest Neighbor, and Logistic Regression) using data augmentation techniques and feature ranking/reduction methods. The accuracy rates they obtained were 100%, 89%, 99.8%, and 99.9%, respectively. Cetin et al. [6] developed an artificial neural network model for early detection of lung cancer. This model aims to detect the presence of lung cancer based on data from symptoms of patients. They achieved an accuracy rate of 98.75%. Nasser has devised an Artificial Neural Network (ANN) for discerning the presence or absence of lung cancer within the human body. The outcomes of this endeavor have demonstrated the ANN model's remarkable ability to identify the presence or absence of lung cancer with a notable accuracy rate of 96.67% [7]. Abdullah et al. [8], on the other hand, used a publicly available dataset targeting a different lung cancer risk classification. It used various algorithms such as CNN, SVM and K-nearest neighbors on the dataset. The accuracy rates obtained from these are 95.56%, 92.11%, 88.40% respectively. Patra [9] used the same dataset used by Abdullah et al and applied J48, KNN, Naive Bayes and RBF via WEKA on the dataset. It achieved an accuracy rate of 78.12%, 75%, 78.12% and 81.25% respectively. Dritras et al. [10] used multiple machine learning models such as NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF, and AdaBoostM1 on the same dataset as Wan. They achieved accuracy rates of 95%, 95%, 96%, 95%, 96%, 94%, 96%, 94%, 95%, 95%, 93%, 93%, 97%, 95%, respectively. Purba et al. [11] utilized various machine learning algorithms to determine the risk of lung cancer in patients. The results of this study demonstrated exceptional abilities in identifying the risk, with accuracy rates of 99% in the logistic regression model, 98% in Linear SVM, 97% in kernel SVM, 62% in decision tree, and a remarkable 100% accuracy rate in both random forest and gradient boosting methods. Gupta et al. [12] collected approximately 15000 images from the Kaggle platform and selected 1200 photos from the image dataset they obtained and used machine learning algorithms for more accurate identification and early diagnosis of lung cancer. They converted the images to 64 by 64 pixels and applied various filters. Then, using KNN, Random Forest and SVM algorithms, Gupta et al. achieved an accuracy rate of 48.7%, 84.2% and 82.1%, respectively.

In this study, detection of lung cancer is based on risk factors. The performance of ensemble algorithms has been investigated and, in this context, LightGBM (LBGM) classifier algorithms are used. It is thought that the proposed system will help healthcare professionals and patients in the detection and classification of lung cancer.

The study is structured as follows: Section 2 provides comprehensive information on the dataset used and the methodology applied. In Chapter 3, the obtained results are meticulously presented, including a rigorous evaluation of the performance of the algorithm in the context of lung cancer diagnosis and the explainability of the algorithm used. Chapter 4 provides an in-depth discussion of the findings and summarizes the study with a well-founded conclusion.

II. Materials and Method

This section discusses the proposed algorithm for the detection and identification of lung cancer. The algorithm used is the LightGBM algorithm developed by Microsoft [13]. High accuracy rate was obtained with LightGBM algorithm.

A. *Dataset*

This dataset contains a database of 1000 records that examines the risk factors and symptoms of lung cancer. The records consist of 26 different features, such as the patient's age, gender, exposure to air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung

disease, balanced diet, obesity, smoking habit, passive smoking status, chest pain, coughing up blood, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of the nails, and snoring. Meaningless features such as index and patient id in the dataset were removed. The remaining 24 features are divided into three categories (Low, Medium, and High) that represent the patient's lung cancer risk levels. All features and their scopes in the dataset are shown in detail in Table 1.

This dataset is of great importance for medical research, health policies, and clinical studies. Data analysis and modeling can play a critical role in assessing the cancer risk of patients and developing preventive measures. This dataset can form the foundation for future work on lung cancer and help health professionals better assess the risk levels of patients.

Table 1. Features and their scopes of the dataset

#	Feature	Scope	Description
1	Age	Various ages from 14 to 73	The patient's age.
2	Gender	1, 2	The patient's gender.
3	Air Pollution	1 - 8	The patient's level of exposure to air pollution.
4	Alcohol Use	1 - 8	The patient's alcohol use level.
5	Dust Allergy	1 - 8	The patient's level of dust allergy.
6	Occupational Hazard	1 - 8	The patient's occupational hazard level.
7	Genetic Risk	1 - 7	The patient's genetic risk level.
8	Chronic Lung Disease	1 - 7	The patient's level of chronic lung disease.
9	Balanced Diet	1 - 7	The patient's balanced nutritional level.
10	Obesity	1 - 7	The patient's obesity level.
11	Smoking	1 - 7	The patient's smoking level.
12	Passive Smoker	1 - 8	The patient's passive smoking level.
13	Chest Pain	1 - 9	The patient's level of chest pain.
14	Coughing of Blood	1 - 9	The level at which the patient coughs up blood.
15	Fatigue	1 - 6 and 8, 9	The patient's fatigue level.
16	Weight Loss	1 - 8	The patient's level of weight loss.
17	Shortness of Breath	1 - 7 and 9	The patient's level of shortness of breath.
18	Wheezing	1 - 8	The patient's wheezing level.
19	Swallowing Difficult	1 - 8	The patient's level of swallowing difficulty.
20	Clubbing of Finger	1 - 9	The level of clubbing of the patient's fingernails.
21	Frequent Cold	1 - 7	Frequency of patient's cold
22	Dry Cough	1 - 7	The patient's dry cough level.
23	Snoring	1 - 7	The patient's snoring level
24	Level	1, 2 and 3	The patient's risk level (Low, Normal, High)

B. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a foundational step in data science, involving the systematic investigation and comprehension of a dataset's essential characteristics. EDA is instrumental in uncovering latent patterns, insights, and structure within data through a combination of visual and statistical techniques. The process entails scrutinizing the distribution, outliers, missing values, and correlations, with the primary goal of extracting meaningful insights about the dataset. EDA serves as a critical initial phase in data analysis projects, enabling data scientists to gain a profound understanding of the data's nature and attributes. Consequently, it facilitates more effective data preprocessing and prepares the ground for subsequent modeling and prediction endeavors [14][15].

Multiclass lung cancer classification was performed using the Lung Cancer dataset containing publicly available data from kaggle [16].

The Lung Cancer dataset contains 365 high, 332 medium and 303 low entries from a total of 1000 samples as shown in Figure 2, it is a fairly balanced dataset.

As seen in Figure 3, the correlations between the dataset features are visually presented in the form of a heatmap. This heatmap provides a clear overview of how various features relate to one another.

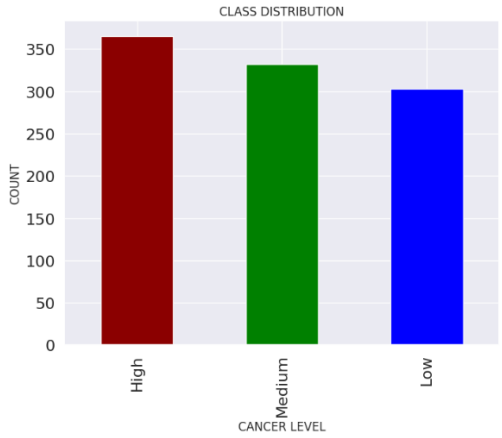

Figure 2. Cancer level distribution of the dataset.

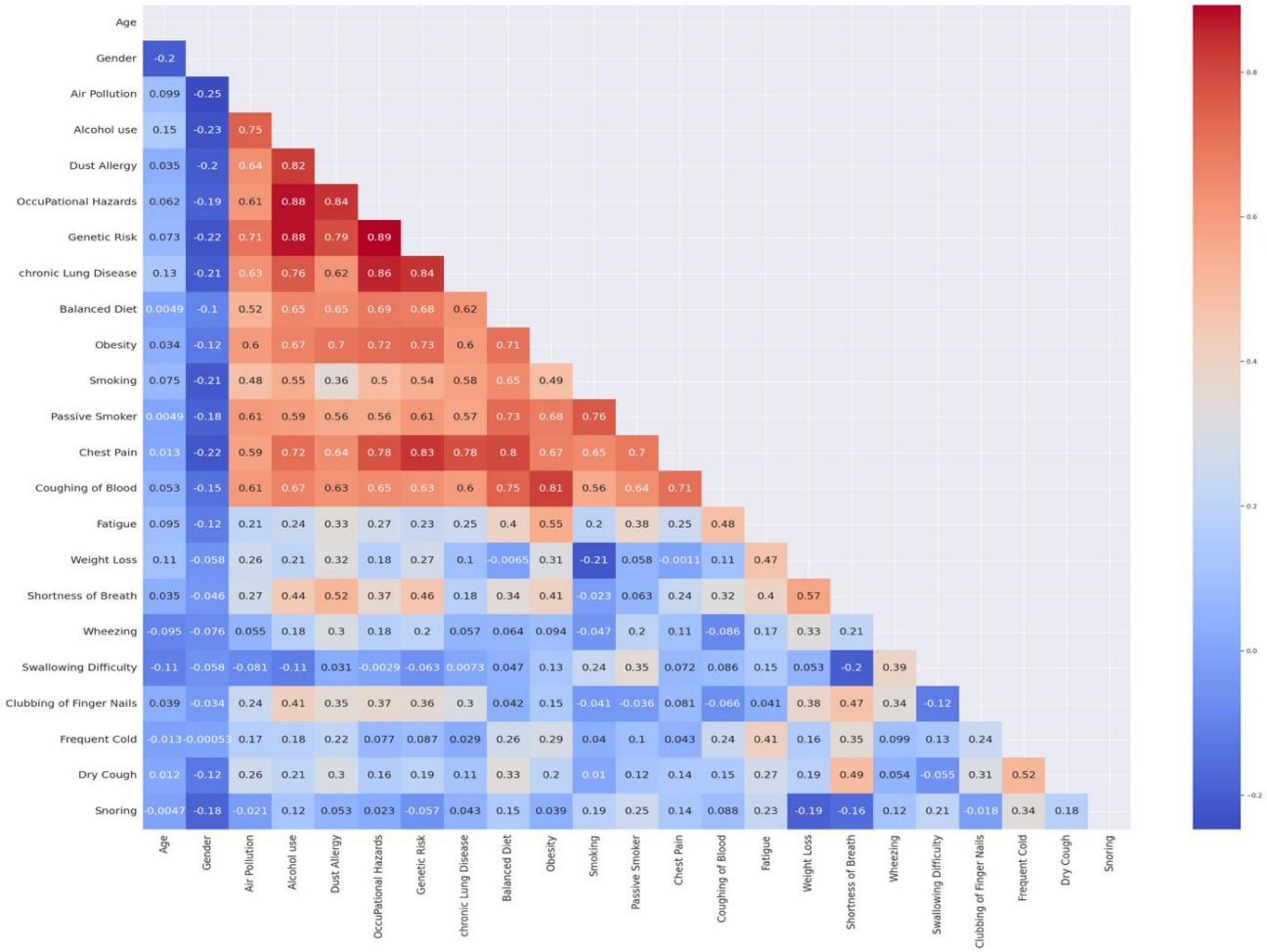

Figure 3. Heat map of correlations between features in the dataset.

The Kernel Density Estimation (KDE) plot provides valuable insights into the distribution of age within the dataset. As illustrated in Figure 4, the KDE graph showcases the probability density of age values. Notably, the peak density occurs around 35 years old, suggesting that individuals in their forties are more prevalent in the dataset. This visualization helps understand the distribution of age values and serves as a valuable tool for exploring age-related characteristics of the dataset.

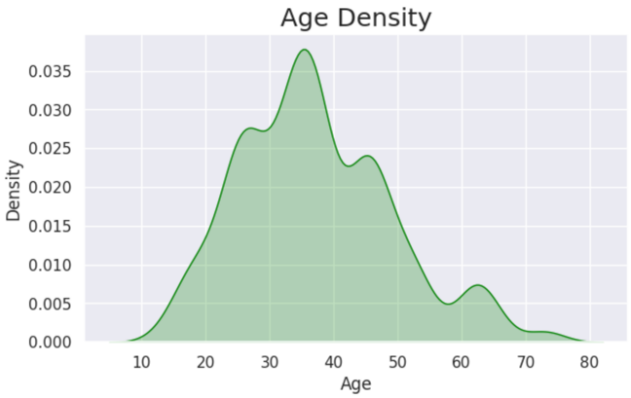

Figure 4. Age density prediction table of the dataset.

C. *Method*

The input to the system is clinical data containing various ordered categorical values from a patient suspected to be at risk of lung cancer. These categorical data were then subjected to standardization. The "Level" column in the data set represents the target of the model, and the values of the target are coded numerically. (0: Low, 1: Normal, 2: High). Then, the data set was divided into two as train and test and given to the LightGBM model.

The risk class of the patient at risk of lung cancer given to the model was classified as low, normal and high. The following diagram, presented in Figure 5, is used to show the general structure of the proposed model.

D. *Data Preprocessing*

In the data preprocessing phase, the numerical input values in the dataset underwent a critical transformation. This meticulous procedure entailed the normalization of these values to conform to a standardized range from -1 to 1. The primary objective behind this normalization process was to enhance the model's capacity to discern and utilize inherent correlations within the dataset, thereby elevating its overall performance and predictive capabilities.

Equation (1) below succinctly illustrates the precise methods used in the normalization and standardization process. In this equation, 'x' represents the original input values, and 'x'' denotes the transformed values, reflecting the enhanced, range-bound, and standardized data [17].

This comprehensive preprocessing strategy collectively contributed to the model's robustness, resilience, and its capacity to extract meaningful insights from the data. It thereby paved the way for superior predictive performance.

$$x' = x - \min(x) \,/\, \max(x) - \min(x) \quad (1)$$

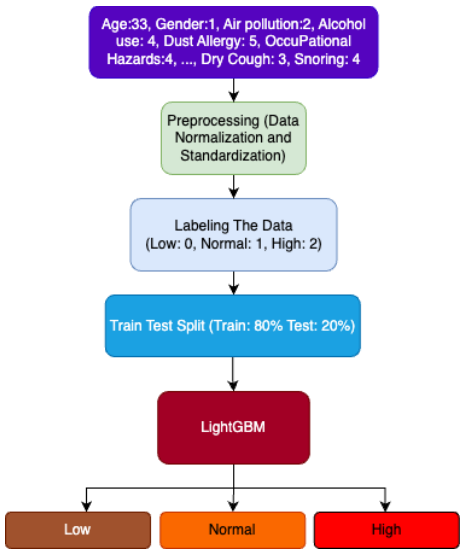

Figure 5. Flow of a LightGBM-based lung cancer risk prediction classification model.

E. *Light Gradient-boosting Machine (LightGBM)*

LightGBM is an open-source machine learning framework specifically designed for fast and effective model training on large datasets. It uses a tree-based modeling approach, and these trees work with a gradient boosting machine (GBM), which is an ensemble learning method that divides data into smaller parts and then combines them to make predictions on those parts. LightGBM offers some significant innovations to improve the processing speed and memory efficiency of this algorithm. One of its most distinctive features is its histogram-based learning approach, which summarizes data quickly with pre-computed histograms. This allows for fast training and predictions. Additionally, LightGBM can automatically handle categorical variables and can handle large datasets with low memory usage. Therefore, LightGBM has become a popular choice for performance-oriented machine learning applications on large datasets [18].

F. *Evaluation Metrics*

Evaluation metrics are important tools used to measure the performance of a classification model. Some of these metrics that are commonly used are F1 score, precision, recall and accuracy metrics. Confusion matrix is used as a key component to calculate these metrics. Figure 6 shows a binary class confusion matrix and Figure 7 shows a multiclass confusion matrix.

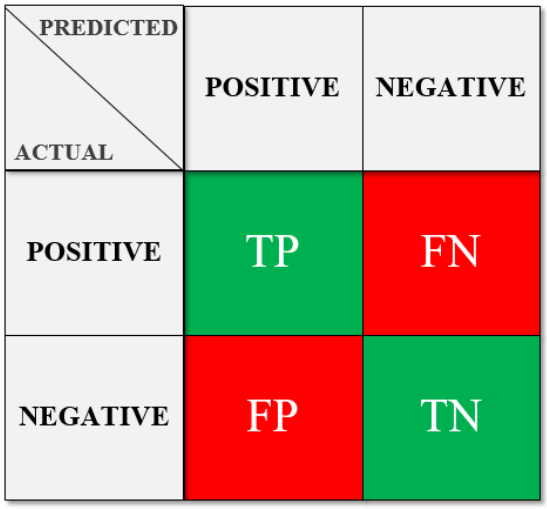

Figure 6. A confusion matrix for binary classification.

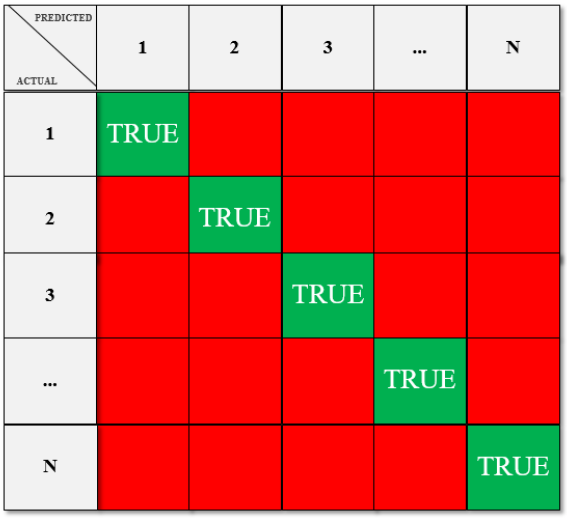

Figure 7. A confusion matrix for multiclass classification.

A confusion matrix is a table that summarizes the results of a classification model. The terms in the complexity matrix below are explained:

- True positive (TP): The number of samples that were truly positive and correctly classified as positive.
- True negative (TN): The number of samples that were truly negative and correctly classified as negative.
- False positive (FP): The number of samples that were actually negative but were incorrectly classified as positive.
- False negative (FN): The number of samples that were truly positive but were incorrectly classified as negative.

Accuracy is the most common performance measure. It is calculated as the ratio of the number of correctly classified samples to the total number of samples. Sensitivity is the ratio of the number of true positives to the number of false positives. It measures the proportion of samples that are truly positive and samples that are correctly classified as positive. Recall is the ratio of the number of true positives to the number of true positives plus the number of false negatives. It measures the proportion of samples that are truly positive and samples that are correctly classified as positive. The F1 score is a harmonic average of precision and recall.

Measuring performance based on accuracy alone is not a sufficient measure. Therefore, values such as precision, recall, and f1 score also need to be included. The equations for performance measurements appear in Equations (2) – (5).

$$precision = TP / (TP + FP) \quad (2)$$

$$recall = TP / (TP + FN) \quad (3)$$

$$F1 = (2 \times precision \times recall) / (precision + recall) \quad (4)$$

$$accuracy = TP + TN / (TP + FN + TN + FP) \quad (5)$$

III. RESULTS

In this section, the results of the LightGBM model trained for low, normal and high lung cancer risk classification are presented. In addition, various evaluation metric values of the experimental findings and the explanation of the model with SHAP are also shown in the following sections.

A. LightGBM Model Performance

LightGBM offers the best evaluation metrics. The accuracy, recall, precision and f1-score of the LightGBM method discussed here are shown in Table 2. The relative sizes of the training and testing datasets were chosen as 0.8 and 0.2, respectively. The data set was randomly mixed for the results obtained.

Table 2. LightGBM's evaluation results

	Precision	Recall	F1-Score
Low	1.00	1.00	1.00
Medium	1.00	1.00	1.00
High	1.00	1.00	1.00
Accuracy			1.00

The confusion matrix for the experimental results obtained from applying to the test dataset are shown in Figure 8.

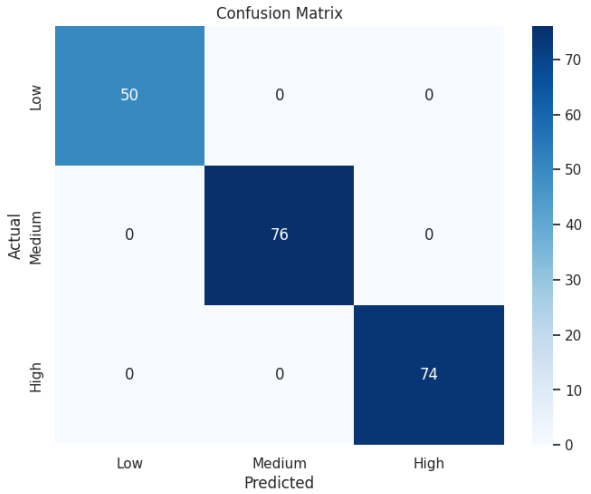

Figure 8. Confusion matrix of experimental results.

B. SHAP Results

To understand why a machine learning (ML) model makes the predictions it does, it is essential to interpret and analyze its results after evaluating its performance. This involves understanding the relationships between the features and the target variable, identifying any relevant patterns or trends in the data, and determining the features that are most important for the model's predictions. In this study, the SHAP (Shapley Additive Explanations) algorithm [19] was used for the explainability of artificial intelligence. Figure 9 shows the relationship between the features and the SHAP algorithm applied to the experimental results.

SHAP is a powerful explainability method used to explain the predictions of machine learning models and understand the contribution of features. This study aims to make model predictions more understandable by using the SHAP algorithm in lung cancer risk classification.

We analyzed the contribution of each feature using SHAP to explain our lung cancer risk estimates. As a result of our SHAP analysis, we observed that passive smoker is an important factor in risk estimates and the risk of cancer increases as passive smoker increases. We also found that risk factors such as wheezing and obesity have a significant impact on the predictions.

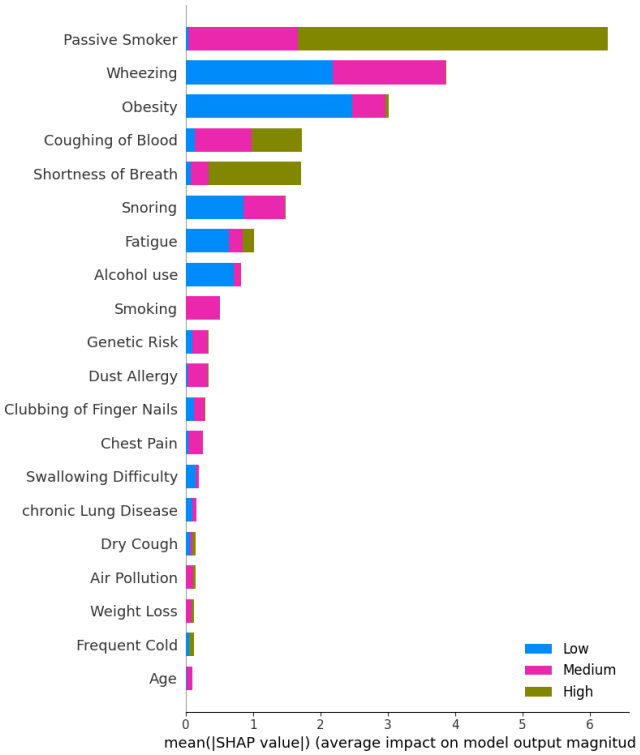

Figure 9. SHAP analysis results.

IV. Discussion and Conclusion

The presented study addresses the critical issue of lung cancer, which is not only the most common cancer worldwide but also one of the deadliest. Early diagnosis is crucial to improve patient outcomes and reduce the burden of treatment costs. The main aim of the study is to help both doctors and patients by using explainable artificial intelligence methods to predict the risk of lung cancer.

Using the LightGBM algorithm for risk prediction is a notable choice. Known for its efficiency and speed, this machine learning algorithm demonstrated the ability to achieve an impressive 100% accuracy rate in this study. Additionally, the inclusion of explainability through the SHAP method increases the reliability of the model and its acceptance by healthcare professionals and patients.

One of the strengths of the study lies in its commitment to the explainability of the model. The use of the SHAP algorithm to explain the model's predictions is commendable. SHAP provides information about the features that influence the model's decisions. In particular, passive smoking, wheezing and obesity are identified as important risk factors and increase the interpretability of the model.

In this study, the LightGBM algorithm was used with a small csv data set with 1000 data. This study opens several avenues for future research. In the future, using a larger dataset and incorporating different machine learning and deep learning methods and validating the performance of the model in different patient populations are important steps. Additionally, integration with real-world clinical applications and healthcare systems should be explored to bring this technology closer to practical use.

In conclusion, this study provides a robust framework for lung cancer risk prediction and demonstrates the potential of artificial intelligence, specifically the LightGBM algorithm, to achieve unprecedented accuracy in risk assessment. Incorporating model interpretability through SHAP benefits both medical practitioners and patients by adding transparency and confidence to predictions. This research marks a significant advance in the field of lung cancer prediction, with the potential to assist physicians in early detection, improve patient outcomes, and reduce healthcare costs.

ACKNOWLEDGMENT

This work was supported by the Republic of Türkiye Ministry of Industry and Technology Attratcion Centers Supporting Program Under Grant No: TRB1/22/CMDP-E1/0001.

REFERENCES

[1] Y. Xue, L. Wang, Y. Zhang, Y. Zhao, ve Y. Liu, "Air pollution: A culprit of lung cancer", *J. Hazard. Mater.*, c. 434, s. 128937, Tem. 2022, doi: 10.1016/j.jhazmat.2022.128937.

[2] I. A. J. D. Kusumawardani, P. Indraswari, ve N. L. G. Y. Komalasari, "Air Pollution and Lung Cancer", *J. Respirasi*, c. 9, ss. 150-158, May. 2023, doi: 10.20473/jr.v9-I.2.2023.150-158.

[3] Nooreldeen, R.; Bach, H. Current and Future Development in Lung Cancer Diagnosis. Int. J. Mol. Sci. 2021, 22, 8661. https://doi.org/10.3390/ijms22168661

[4] International Agency for Research on Cancer. 2023. Available online: https://gco.iarc.fr/today (accessed on 11 October 2023).

[5] N. A. Salman and S. T. Hasson, "A Prediction Approach for Small Healthcare Dataset," 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), Split/Bol, Croatia, 2023, pp. 1-5, doi: 10.23919/SpliTech58164.2023.10193552.K. Elissa, "Title of paper if known," unpublished.

[6] V. Cetin, H. H. Yumrukaya and C Bakir, "Detection of Lung Cancer with Enhanced Feed Forward Backpropagation Artificial Neural Networks", Proceedings of the International Conference on Industrial Engineering and Operations Management Manila, Philippines, March 7-9, 2023,

[7] Nasser, Ibrahim, "Lung Cancer Detection Using Artificial Neural Network" (2019). International Journal of Engineering and Information Systems (IJEAIS), Vol. 3 Issue 3, March – 2019, Pages: 17-23 , Available at SSRN: https://ssrn.com/abstract=3700556

[8] Abdullah, Dakhaz & Mohsin Abdulazeez, Adnan & Sallow, Amira. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. Qubahan Academic Journal. 1. 141-149. 10.48161/qaj.v1n2a58.

[9] R. Patra, "Prediction of Lung Cancer Using Machine Learning Classifier", *Computing Science, Communication and Security*, N. Chaubey, S. Parikh, ve K. Amin, Ed., Singapore: Springer Singapore, 2020, ss. 132-142.

[10] Dritsas, E.; Trigka, M. Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn. Comput. 2022, 6, 139. https://doi.org/10.3390/bdcc6040139

[11] W. Purba, S. Wardani, D. F. . Lumbantoruan, F. C. I. Silalahi, and T. L. Edison, "OPTIMIZATION OF LUNG CANCER CLASSIFICATION METHOD USING EDA-BASED MACHINE LEARNING", JUSIKOM PRIMA, vol. 6, no. 2, pp. 43-50, Feb. 2023.

[12] A. Gupta, Z. Zuha, I. Ahmad, ve Z. Ansari, "A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms", In Review, preprint, Agu. 2022. doi: 10.21203/rs.3.rs-1912967/v1.

[13] Hamed, E.A.-R.; Salem, M.A.-M.; Badr, N.L.; Tolba, M.F. An Efficient Combination of Convolutional Neural Network and LightGBM Algorithm for Lung Cancer Histopathology Classification. Diagnostics 2023, 13, 2469. https://doi.org/10.3390/diagnostics13152469

[14] A. G V S, A. Kethaan, ve G. V. N. Akshay Varma, "Fraud Detection System Employing Machine Learning Techniques for Credit Card Transactions", *International Journal of Advanced Research in Science Communication and Technology*, c. 3, ss. 2581-9429, Agu. 2023, doi: 10.48175/IJARSCT-12492.

[15] A. Shabbir, M. Shabbir, A. R. Javed, M. Rizwan, C. Iwendi, ve C. Chakraborty, "Exploratory data analysis, classification, comparative analysis, case severity detection, and internet of things in COVID-19 telemonitoring for smart hospitals", Journal of Experimental & Theoretical Artificial Intelligence, c. 35, sy 4, ss. 507-534, May. 2023, doi: 10.1080/0952813X.2021.1960634.

[16] Air Pollution, Alcohol, Smoking & Risk of Lung Cancer Dataset. https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link Access Date : 09 Sep 2023

[17] Kuzu A, Santur Y. Early Diagnosis and Classification of Fetal Health Status from a Fetal Cardiotocography Dataset Using Ensemble Learning. *Diagnostics*. 2023; 13(15):2471. https://doi.org/10.3390/diagnostics13152471

[18] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. Adv. Neural Inf. Process Syst. 2017, 30, 3149–3157.

[19] S. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', ArXiv170507874 Cs Stat, Nov. 2017, http://arxiv.org/abs/1705.07874