Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 67-75, 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. 67-75, 10, 2023

Research Article

Predictive Modeling of Diseases with Explainable Artificial Intelligence Using LightGBM

Elif Bahar Özdoğru[1*], Yunus Santur[2] and Mustafa Ulaş[2]

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

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

(Received: 16 October 2023, Accepted: 23 October 2023)

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

ATIF/REFERENCE: Özdoğru, E. B., Santur, Y. & Santur, M. (2023). Predictive Modeling of Diseases with Explainable Artificial Intelligence Using LightGBM. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(10), 67-75.

Abstract – The continuous exploration of the intricate connections among symptoms, patient attributes, and diseases within the intricate landscape of human health represents an ongoing pursuit. Data-driven methodologies have ushered in novel opportunities for comprehending these intricate relationships. Especially with the COVID-19 pandemic, the paradigms of disease understanding, diagnosis, and treatment management have assumed unprecedented significance. This study, powered by LightGBM and SHAP, has the potential to provide invaluable support to experts in decision support systems, early diagnosis of diseases, personalized treatment plan applications, strengthening medical interventions with case-oriented treatment predictions by producing advanced diagnosis and treatment strategies at demographic scales and analyzing risk factors, developing evidence-based public health policies and proactive health services, researchers. Furthermore, this research can be effectively leveraged in epidemiological investigations to ascertain the correlations and emerging trends between various diseases and the influencing health determinants all with an impressive 81% accuracy.

Keywords – Disease Prediction, Machine Learning, AI-Driven Healthcare, SHAP, Decision Support Systems

I. INTRODUCTION

Predicting diseases accurately and at an early stage holds paramount significance in the realm of healthcare and medical research [1]. As the adage goes, "prevention is better than cure," and early disease detection epitomizes this principle. Timely identification of diseases not only enhances the chances of successful treatment but also has far-reaching implications in terms of healthcare cost reduction and improved patient outcomes. In an era marked by the exponential growth of medical data and the increasing prevalence of chronic and complex diseases, the ability to forecast and diagnose health conditions with precision has emerged as a pressing imperative [2].

Marios and his colleagues examined the symptoms of Parkinson's disease in their study. To provide more information about the disease's impact on patients' quality of life, they evaluated 265 consecutive Parkinson's patients by asking them to list their three most bothersome symptoms in the last 6 months. They made certain classifications and feature distinctions. They found further evidence of the diversity of experiences about Parkinson's disease and its symptoms. They stated that as the disease progresses, the most troubling problems perceived by patients are lack

of response to medication and non-motor aspects of the disease [3].

In their study, Edwards and colleagues used a large-scale biomedical literature database to create a symptom-based human disease network and investigate the connection between clinical manifestations of diseases and their underlying molecular interactions. They found that the symptom-based similarity of two diseases was strongly associated with the number of shared genetic relationships and the degree of interaction of the associated proteins. They associated the diversity of clinical manifestations of a disease with the connectivity patterns of the underlying protein interaction network. They have produced a comprehensive, high-quality map of disease-symptom relationships. As a result of the research, they presented it as a resource that helps to identify unexpected relationships between diseases, disease etiology research or drug design [4].

In their study of disease symptoms, Cohen and colleagues reviewed research on the role of stress in infectious diseases, as measured by illness behaviors (symptoms and use of health services) or confirmed pathology. Finding significant evidence for a relationship between stress and increased sickness behavior, they found less convincing but important evidence for a similar relationship between stress and infectious pathology. Introverts, isolates, and people who lack social skills also said they were at high risk for both illness behaviors and pathology. They have proposed psychobiological models of how stress may influence the onset and progression of infectious diseases and a psychological model of how stress may influence disease behaviour [5].

Arnold and colleagues used a semistructured interview administered to primary family caregivers to assess the prevalence and nature of psychiatric pathology in 175 community-dwelling, well-diagnosed Alzheimer's patients. They found that symptoms indicative of depression in the cognitively intact were almost ubiquitous in this demented population. They regularly reported various psychotic features. They discussed the significance of these findings for the recognition and treatment of reversible psychiatric disorder [6].

Zoabi and colleagues aimed to ensure rapid and effective diagnosis of COVID-19. Prediction models combining various features have been developed to predict infection risk. These are intended to assist medical personnel worldwide in prioritizing patients, especially in the context of limited healthcare resources. In this study, they created a machine learning approach that trained on recordings from 51,831 people tested (4,769 of whom were confirmed to have COVID-19). Their model predicted COVID-19 test results with high accuracy using eight binary features: gender, age ≥60, known contact with an infected person, and the appearance of five initial clinical symptoms. Overall, based on nationwide data made public by the Israeli Ministry of Health, they have developed a model that detects COVID-19 cases with simple features accessed by asking basic questions [7].

Artificial intelligence and machine learning have played instrumental roles in addressing the unprecedented challenges posed by the COVID-19 pandemic. These technologies have demonstrated their efficacy in several key areas. For instance, AI-driven predictive models have been employed for epidemiological forecasting, helping authorities anticipate disease spread and allocate resources efficiently. Machine learning algorithms have been pivotal in the rapid development of diagnostic tools, enabling quicker and more accurate detection of the virus. Additionally, AI-powered drug discovery has expedited the identification of potential treatments and vaccines [8-11].

In this study, we employ advanced machine learning techniques, specifically LightGBM and SHAP, to address the vital task of predicting diseases based on patient symptoms. Our primary objective revolves around the development and application of explainable artificial intelligence algorithms for this purpose. The utilization of machine learning approaches has demonstrated their significant utility across various domains within the medical field. [12-15].

II. Materials and Method

In the following section, the methodology for disease prediction based on symptoms is presented, utilizing the LightGBM algorithm developed by Microsoft. Exceptional levels of accuracy have been achieved in this study through the implementation of the LightGBM algorithm, a crucial factor for ensuring reliable disease prediction.

A. Dataset

This dataset contains a comprehensive database of 349 records of disease symptoms and patient profiles. Each record represents a data point that associates patients with a specific disease or medical condition and has different characteristics such as symptoms, age, gender, blood pressure, and cholesterol levels. The dataset provides valuable information for understanding how disease symptoms are related to the patient's personal characteristics and health status, for diagnosis or for studies on treatment management. Additionally, an output variable is available to use this data to relate it to outcomes such as diagnosis or prognosis of the disease. The output variable has two values (Positive, Negative). The features in the dataset and their descriptions are given in Table 1. This dataset can serve as a resource for various analyzes and research in healthcare and assist healthcare professionals in evaluating patients.

Table 1. Features and their descriptions and scopes of the dataset

#	Feature	Scope	Description
1	Disease	116 various diseases	This category denotes the specific medical condition.
2	Fever	Yes - No	This field signifies the status of fever in the patient, indicated by binary values.
3	Cough	Yes - No	This parameter indicates if the patient has a cough, delineated by binary responses.
4	Fatigue	Yes - No	This attribute reveals the status of fatigue as experienced by the patient, distinguished through binary choices.
5	Difficulty in Breathing	Yes - No	This category discerns whether the patient encounters difficulties in breathing, with responses represented in binary form.
6	Age	19 - 90	The age of the patient is documented in years, providing a measure of their chronological age.
7	Gender	Male - Female	The patient's gender is categorized as Male or Female.
8	Blood Pressure	Low – Normal - High	This variable captures the patient's blood pressure level, categorized as Low, Normal or High, offering info about cardiovascular health.
9	Cholesterol Level	Low – Normal - High	This attribute records the patient's cholesterol level, categorized as either Normal or High, providing information about metabolic health.
10	Outcome Variable	Positive - Negative	This pivotal component serves as the outcome variable. It is characterized as Positive or Negative, offering a definitive assessment of the disease status in each case.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) stands as a fundamental step in data mining and analytics, serving as a pivotal process for comprehending and assimilating a given dataset. EDA encompasses a blend of statistical and visual techniques employed to discern patterns, relationships, and outliers within the dataset. This process aids in understanding the characteristics, central tendencies, dispersions, and distributions of the data. Furthermore, it aids in identifying relationships among various variables and uncovering potential deficiencies or anomalies within the dataset. EDA plays a crucial role as the initial examination of the data, preceding more comprehensive analyses and the construction of data mining models. As such, the effective application of EDA forms the cornerstone of data-driven decision-making and discoveries [16].

The creation of patient profiles with binary classification attributes was conducted using a dataset that encompasses disease symptoms and publicly available data from Kaggle [17].

The disease symptoms and patient profile dataset is a balanced dataset containing 186 positive and 163 negative entries from a total of 349 samples, as shown in Figure 1.

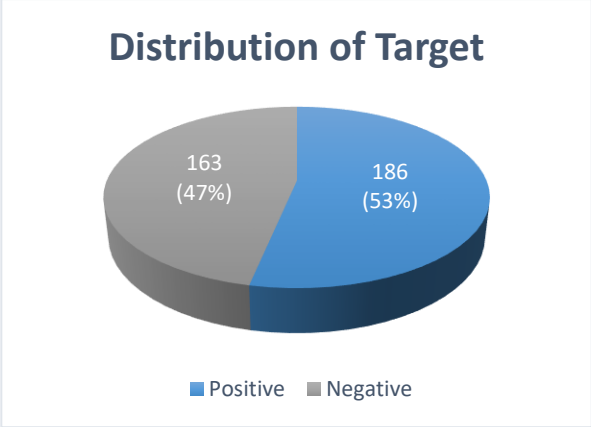

Figure 1. Outcome variable distribution of the dataset.

The correlations between the dataset features are shown in Figure 2 in the form of a heatmap. The outcome variables in the dataset are coded as negative (0), positive (1). In this empirical study, classification was performed using LightGBM.

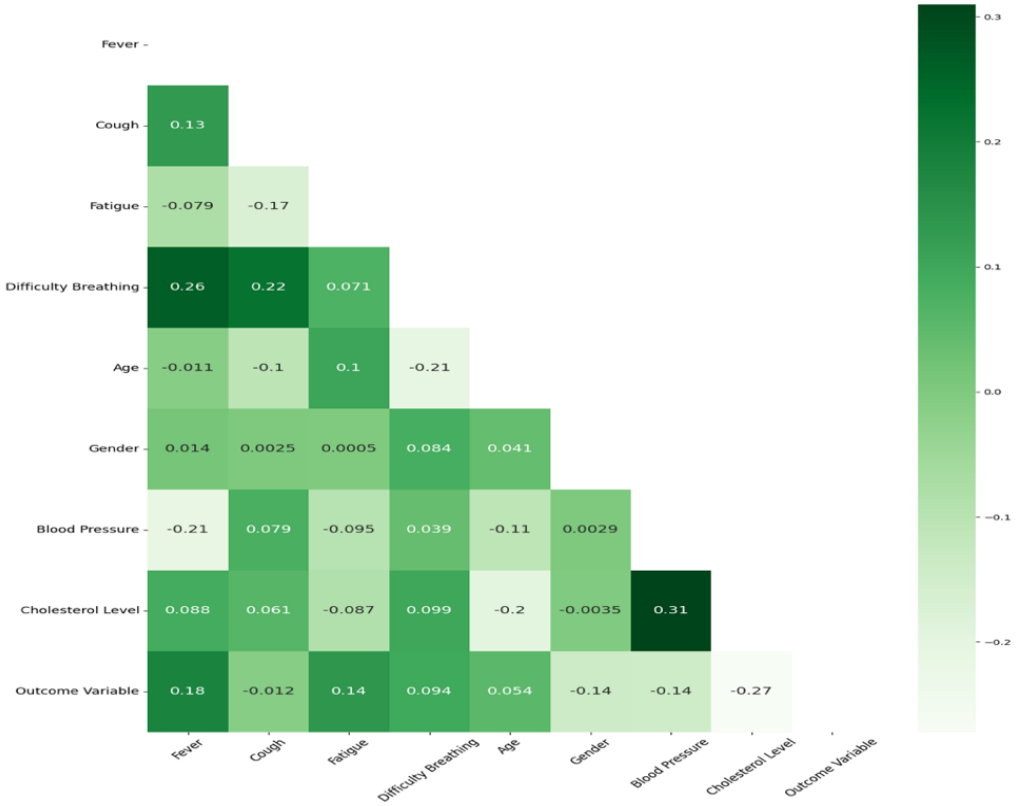

Figure 2. Heat map showing the correlation levels of features in the dataset with each other.

C. *Method*

Categorically coded data taken from the disease symptoms and patient profile dataset is entered into the system as input. These categorical data were then subjected to label encoding. The "Outcome Variable" column in the dataset represents the target of the model, and the values of the target are coded numerically. (0: Negative, 1: Positive). Then the dataset is divided into 85% and 15% train and test, respectively. Then, the LightGBM model was trained with the train dataset and the results were predicted as positive or negative with the test model.

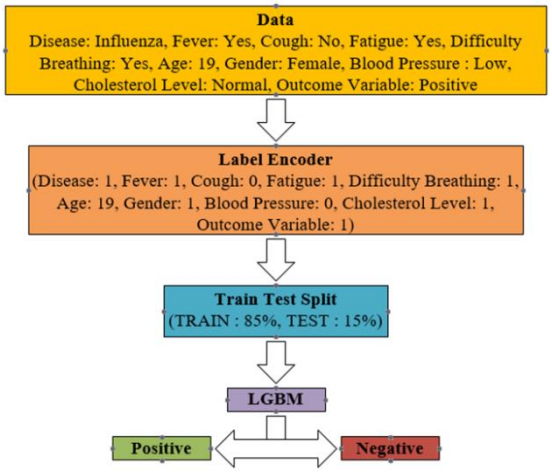

Figure 3. Flow chart of disease symptoms and patient profile classification.

D. *Evaluation Metrics*

Evaluation metrics are measurements used to measure and compare model performance in fields such as data analytics, machine learning, and statistical modelling. These metrics help us understand how well or poorly a model is performing. They fall into two basic categories: classification and regression metrics. Classification metrics evaluate how well a model predicts categorical results and include measurements such as accuracy, sensitivity, specificity, and F1 score. Regression metrics evaluate how well the model predicts continuous values and include measurements such as mean square error (MSE), mean absolute error (MAE), and R-squared. These metrics play a critical role in model selection, hyperparameter tuning, and interpretation of results because they allow us to objectively evaluate the success of a model. Since classification is made in this study, only classification metrics are specified. The basic classification metrics are:

- Accuracy: Represents the ratio of correctly predicted samples to the total number of samples. That is, it shows the rate at which all classes are classified correctly. However, it can be misleading in unbalanced datasets.
- Precision: It refers to the ratio of samples predicted to be positive to samples that are actually positive. That is, it aims to reduce the number of false positives.
- Recall or Sensitivity: Indicates how much of the samples that are true positives are correctly predicted as positive. It aims to reduce the number of false negatives.
- F1 Score: Calculated by taking the harmonic average of sensitivity and specificity. It is a useful metric in unbalanced classification problems.

These classification metrics are used to evaluate a model's classification ability, and which metric should be used may vary depending on the problem context.

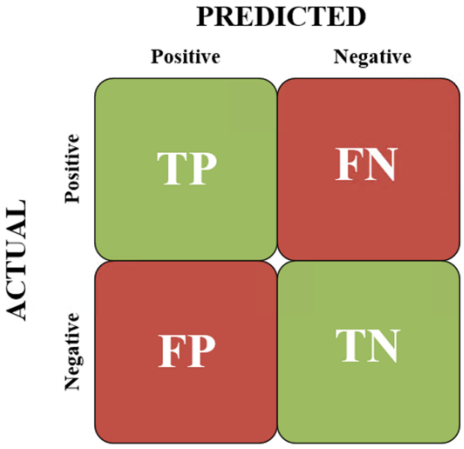

Figure 4. A confusion matrix structure for binary classification.

Figure 4 gives the confusion matrix structure for binary classification. For classification metrics, TP, FN, FP and TN given in Figure 4 are explained below.

- TP (True Positive): TP represents the number of samples that a classification model correctly predicts as positive. That is, it is the number of true positive examples that the model correctly classifies as positive.
- FN (False Negative): FN refers to the number of samples that a classification model incorrectly predicts as negative. That is, the number of examples that the model classifies as negative, but are actually positive.
- FP (False Positive): FP represents the number of samples that a classification model incorrectly predicts as positive. That is, the number of examples that the model classifies as positive, but are actually negative.
- TN (True Negative): TN refers to the number of samples that a classification model correctly predicts as negative. That is, it is the number of true negative examples that the model correctly classifies as negative.

The formulas for accuracy, precision, recall and f1 score are specified in equations 1, 2, 3 and 4 respectively.

$$accuracy = TP + TN / (TP + FN + TN + FP) \quad (1)$$

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

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

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

E. Light Gradient-boosting Machine (LightGBM)

LightGBM, is a highly influential and powerful machine learning algorithm that has gained prominence in recent years. Its importance in the field of machine learning and data science cannot be overstated, primarily due to its exceptional efficiency and effectiveness in handling large-scale datasets and complex tasks. LightGBM stands out for its unique gradient boosting framework, which employs a histogram-based learning approach to optimize decision trees [18]. This innovative technique significantly accelerates the training process, making LightGBM one of the fastest gradient boosting frameworks available. Furthermore, its ability to manage high-dimensional data, categorical features, and imbalanced datasets with ease, while delivering state-of-the-art predictive performance, has made it a fundamental tool for practitioners in various domains. In this era of big data and complex modelling tasks, LightGBM's speed, scalability, and accuracy render it an indispensable resource for researchers and practitioners alike, enabling them to tackle challenging problems with unprecedented efficiency and precision. [19].

III. RESULTS

In this section, we will meticulously dissect the results stemming from the application of the LightGBM model to the task of classifying patient profiles into positive and negative categories. Not only will we assess various evaluation metrics to gauge the model's performance, but we will also embark on a comprehensive exploration of how the model's predictions are elucidated, thanks to the SHAP algorithm's explanatory power. This multifaceted analysis promises to provide valuable insights into the effectiveness and interpretability of our classification model.

A. LightGBM Model Performance

The LightGBM model has proven its capability by delivering outstanding evaluation metrics. Our evaluation includes metrics such as accuracy, recall, precision, and F1 score, which specifically highlight the effectiveness of the LightGBM method. To ensure a robust evaluation, we carefully chose the proportions of the training and testing datasets, with 85% dedicated to training and 15% to testing. These measurements from the test results are carefully presented in Figure 5 to shed light on the adequacy of the model in patient profile classification.

Table 2. LightGBM's evaluation results

	Precision	Recall	F1-Score
0	0.73	0.86	0.79
1	0.89	0.77	0.83
Accuracy			0.81
Macro Avg.	0.81	0.82	0.81
Weighted Avg.	0.82	0.81	0.81

We also scrutinized the experimental results by constructing a confusion matrix upon applying the model to the test dataset. This confusion matrix is meticulously depicted in Figure 5, serving as a visual representation of the model's performance in distinguishing between positive and negative patient profiles.

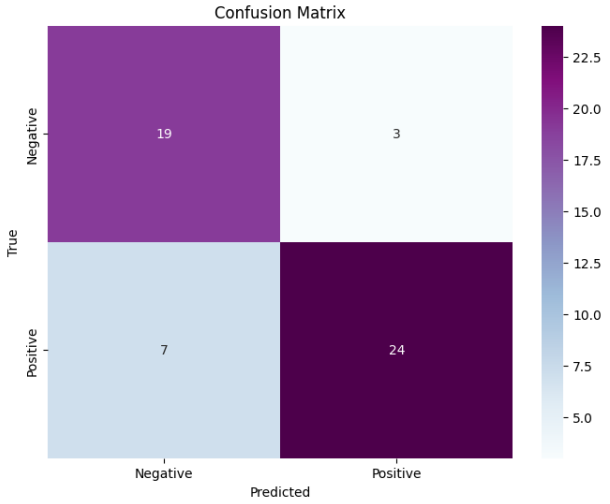

Figure 5. Confusion matrix of test results.

The ROC curve, as illustrated in Figure 6, provides a clear and succinct visualization of our model's performance. It enables us to discern how well the model distinguishes between positive and negative patient profiles, with higher areas under the curve indicating superior classification accuracy.

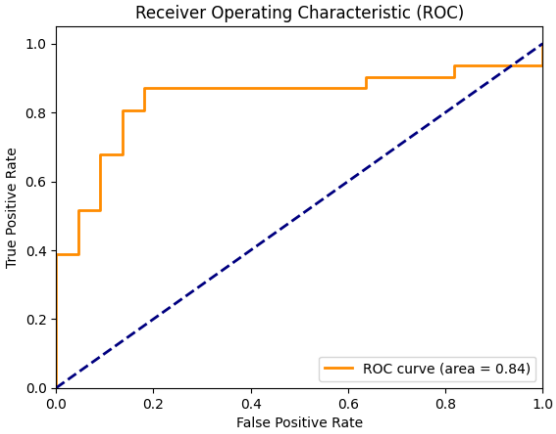

Figure 6. ROC curve of experimental results.

B. XAI Results

The ability to interpret how factors such as cholesterol level, age, gender, fever, cough, fatigue, difficulty breathing and blood pressure affect the LightGBM algorithm can lead to more informed medical decisions and better patient care.

To achieve this understanding, we turn to a powerful tool known as SHAP (Shapley Contribution Annotations). In 1953, Shapley introduced SHAP, a framework rooted in game theory [20]. SHAP provides a transparent and interpretable way to understand the impact of different features on a patient's health [21]. It is particularly valuable for understanding why a patient exhibits certain health profile values or symptoms.

The figure presented as Figure 7 summarizes the results of this analysis and shows the impact of each feature on the patient's health profile. Here, cholesterol level and age factors appear to be more effective in the decision making of the LightGBM algorithm.

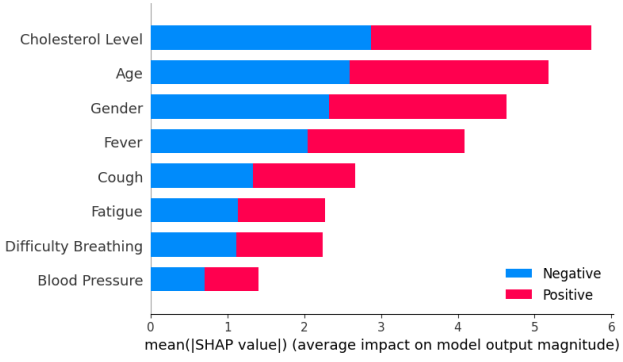

Figure 7. SHAP analysis results – Visualizing the impact of key features on patient profiles.

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.

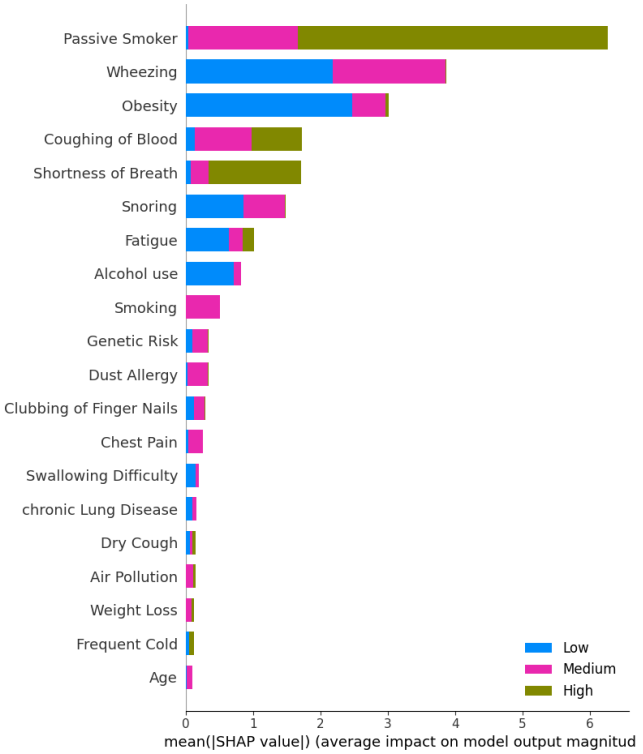

Figure 9. SHAP analysis results.

IV. Discussion and Conclusion

This study addresses a pressing problem, diagnosis of diseases. The significance of early and precise diagnosis cannot be overstated, as it not only contributes to improved patient outcomes but also plays a pivotal role in mitigating the burgeoning healthcare expenditures. Primary objective of presented study is to provide a

substantive contribution to the area of medical diagnostics, with a specific focus on the development and deployment of explainable artificial intelligence methodologies for disease prediction. These methods are envisioned as a means to empower both medical experts and patients by offering a transparent and interpretable approach to disease diagnosis, thereby fostering a more informed decision-making process and ultimately enhancing the quality of healthcare delivery.

The choice of employing the LightGBM algorithm in this study holds a paramount importance. LightGBM's unique gradient boosting framework, has been critical in achieving the high level of predictive accuracy that underpins the success of this research. This algorithm, renowned for its efficiency and scalability, not only facilitates the handling of complex, high-dimensional medical data but also enhances the speed and precision of disease prediction.

Furthermore, the integration of SHAP (SHapley Additive exPlanations) interpretable artificial intelligence techniques has greatly contributed to the transparency and comprehensibility of the predictive models. SHAP values have enabled us to understand the impact of various symptoms and features on disease prediction outcomes, providing invaluable insights for medical practitioners and patients. SHAP values have enabled us to understand the impact of various symptoms and features on disease prediction outcomes, providing invaluable insights for medical practitioners and patients.

The combined utilization of LightGBM and SHAP within this research not only underscores the significance of employing advanced machine learning methods but also ensures that the resulting disease prediction models are not only highly accurate but also interpretable and actionable, thereby enhancing their practical utility in real-world healthcare scenarios.

It is imperative to acknowledge the limitations of this study, primarily pertaining to the dataset utilized. The dataset comprised a relatively modest sample size, consisting of 349 data in CSV format. This limitation has potential implications for the robustness and generalizability of the predictive models developed. A larger dataset with a more extensive pool of patient cases and symptom profiles could have provided a more comprehensive and robust foundation for our disease prediction models. It is, therefore, a critical point of critique that a more expansive and comprehensive dataset would have been preferable, offering a broader foundation for the disease prediction models and potentially enhancing their overall efficacy and reliability.

An additional limitation of this study pertains to the allocation of data for model training. In this study, a conservative approach was adopted, with only 15% of the available dataset utilized for training the predictive models. This conservative approach, while motivated by a desire for methodological caution, raises legitimate concerns about the utilization of the dataset's full potential. By restricting the training dataset to a relatively small fraction, there is a possibility that the predictive models may not have been optimally fine-tuned, and their capacity to capture the intricate relationships between symptoms and diseases might have been compromised. A more equitable allocation of data for training, validation, and testing would have been ideal, enabling a more comprehensive model calibration. It is vital to acknowledge that the chosen allocation strategy, despite its good intentions, may have had implications for the models' performance, thus necessitating further consideration in subsequent research endeavors.

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] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339.

[2] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[3] Politis, M., Wu, K., Molloy, S., G. Bain, P., Chaudhuri, K. R., & Piccini, P. (2010). Parkinson's disease symptoms: the patient's perspective. Movement Disorders, 25(11), 1646-1651.

[4] Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2014). Human symptoms–disease network. Nature communications, 5(1), 4212.

[5] Cohen, S., & Williamson, G. M. (1991). Stress and infectious disease in humans. Psychological bulletin, 109(1), 5.

[6] Merriam, A. E., Aronson, M. K., Gaston, P., Wey, S. L., & Katz, I. (1988). The psychiatric symptoms of Alzheimer's disease. Journal of the American Geriatrics Society, 36(1), 7-22.

[7] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), 3.

[8] Huang, S., Yang, J., Fong, S., & Zhao, Q. (2021). Artificial intelligence in the diagnosis of COVID-19: challenges and perspectives. International journal of biological sciences, 17(6), 1581.

[9] Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., & Cheng, X. (2020). Artificial intelligence and machine learning to fight COVID-19. Physiological genomics, 52(4), 200-202.

[10] Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N., ... & Yuan, J. S. (2020). Artificial intelligence for COVID-19 drug discovery and vaccine development. Frontiers in Artificial Intelligence, 65.

[11] Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. Drug discovery today, 24(3), 773-780.

[12] Waldstein, S. M., Vogl, W. D., Bogunovic, H., Sadeghipour, A., Riedl, S., & Schmidt-Erfurth, U. (2020). Characterization of drusen and hyperreflective foci as biomarkers for disease progression in age-related macular degeneration using artificial intelligence in optical coherence tomography. JAMA ophthalmology, 138(7), 740-747.

[13] Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., ... & Mohyuddin, W. (2020). Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. Ieee Access, 8, 109581-109595.

[14] Schork, N. J. (2019). Artificial intelligence and personalized medicine. Precision medicine in Cancer therapy, 265-283.

[15] Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić-Jovičić, M., & Kostić, V. S. (2019). Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—A review. Clinical neurology and neurosurgery, 184, 105442.

[16] Dey, S. K., Rahman, M. M., Siddiqi, U. R., & Howlader, A. (2020). Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. Journal of medical virology, 92(6), 632-638.

[17] Disease Symptoms and Patient Profile Dataset. https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset
Access Date : 12 Sep 2023.

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[19] Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, 13(1), 6-10.

[20] Shapley, L. S. (1953). A value for n-person games.

[21] Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis & Prevention, 136, 105405.