International Journal of
Digital Health & Patient Care
e-ISSN : 3023-851X
https://ndpapublishing.com/index.php/

Evaluation of the Performance of Machine Learning Algorithms in Disease Prediction

Alparslan Göktürk Güneş*[1] and Volkan Altuntaş[2]

[1]*Bursa Technical University, Faculty of Engineering and Natural Sciences, Bursa, Turkey*
[2]*Bursa Technical University, Faculty of Engineering and Natural Sciences, Computer Engineering Department, Bursa, Turkey*

Article Info
Received: 19 February 2025
Revised: 17 May 2025
Accepted: 25 May 2025
Published: 30 June 2025

Keywords
Machine Learning
Disease Prediction
Naive Bayes
AdaBoost

ABSTRACT

Today, machine learning is widely applied in various disciplines such as technology, healthcare, law, cybersecurity, and image recognition. When examining the research, it is evident that the scope of machine learning applications is expanding day by day. In this study, the goal was to develop a classifier model using machine learning algorithms for disease diagnosis in the healthcare field. In the scope of the study, the performance of various machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), Decision Trees (CART), Random Forest, Gradient Boosting, and AdaBoost was compared for disease prediction. The dataset used in the study was obtained from the Kaggle platform and includes records where diseases are predicted based on various symptoms. The dataset is organized into two different CSV formats for training and testing. The training dataset was used for the model's learning process, while the testing dataset was used to evaluate the accuracy and performance of the model. The dataset contains a total of 4,962 records and consists of 133 columns, with 132 independent variables (symptoms) and 1 dependent variable (disease) for classification. The dataset includes 41 different diseases, and there are 120 examples for each disease. When comparing the accuracy performance of the algorithms used in the study, the highest success rates were achieved with Naive Bayes, Support Vector Machines (SVM), and Gradient Boosting algorithms. Jupyter Notebook was used in the processes of data preparation and model development.

1. INTRODUCTION

Modern computers, with significant increases in storage and processing capacities, are gaining the ability to perform a wide range of different tasks in various fields. These increases allow computers to use their processing power more efficiently and perform previously impossible or time-consuming tasks in a shorter period of time. Additionally, the growing capacity levels of computers enable them to work independently on specific tasks, and the number of tasks a computer can perform within a given time frame is a key indicator for evaluating its performance [7]. This, in turn, allows for a more accurate measurement of computer performance, making it possible to shorten processing times and, consequently, use computer resources more efficiently. With the increased use of computers in many fields, the amount of processable data is also constantly increasing. The amount of processable data has reached 44 trillion gigabytes (GB) with an approximately 22-fold increase from 2010 to 2022 [8]. When examining the amount of data produced, it is seen that the vast majority of this data comes from digital transactions, sensors, and social media sources. In light of the increase in data, data-driven methods have started to be used and developed more. Among these methods, machine learning and artificial intelligence are effectively used in many different fields such as medicine, industry, weather forecasting, risk analysis, and law. Today, healthcare providers and data scientists are collaborating to develop machine learning techniques and medical diagnostic systems [9]. In light of the aging population and the growing use of personalized gene therapies, machine learning models are expected to play a critical role in the future delivery of healthcare [10]. In particular, the application of motion analysis in the healthcare field stands out as another important area that accelerates advancements in this domain. For example, studies predicting POMA-G scores based on spatiotemporal analyses of gait parameters demonstrate the potential applications of motion

*Corresponding author

*e-mail: gokturk.gunes@btu.edu.tr
ORCID ID: 0000-0002-6200-554X

How to cite this article

Güneş, A. G., and Altuntaş, V. (2025). Evaluation of the Performance of Machine Learning Algorithms in Disease Prediction. *Int. J. Digital Health & Patient Care*, 2(1), 17-23

Research Article/ DOI: 10.5281/zenodo.15754286

analysis and machine learning methods in healthcare [17]. Additionally, reliability and validity studies of innovative ROM measurement methods using Microsoft Kinect V2 also serve as an important reference [18]. Based on the information mentioned above, in this study, a disease detection model based on symptoms was developed using a dataset containing symptoms for each disease and six different machine learning algorithms. The success of the model was evaluated separately for each algorithm, and the model showing the highest performance was selected. It is predicted that the developed model could be used by doctors in clinical decision support processes and serve as an effective method during the decision-making phase.

In the study, a performance evaluation was conducted on 20 different machine learning algorithms to detect DDOS attacks. Ensemble learning algorithms, such as Random Forest and XGB, demonstrated better results compared to simpler algorithms like Logistic Regression and Naive Bayes [4]. In the study, various machine learning algorithms were used to predict taxi departure times at Istanbul Airport, and among these algorithms, ANN (Artificial Neural Networks) showed the best performance with the lowest error rate. To improve the model's performance, the data size was reduced using the PCA (Principal Component Analysis) method. It is believed that the findings of the study could contribute positively to reducing flight delays [5].

In the study, sentiment analysis was conducted on data obtained from Facebook using different machine learning algorithms to evaluate corporate performance. The study, which used SVM (Support Vector Machines), Naive Bayes, and Logistic Regression, concluded that SVM produced the most successful results [1]. In their study, they developed a model for detecting threats in the field of cybersecurity by using algorithms such as KNN (K-Nearest Neighbors), Gradient Boosting, SVM (Support Vector Machines), Random Forest, and Logistic Regression. Among the algorithms used, RF (Random Forest) showed the best performance in threat prediction. It is expected that the findings of the study will make a positive contribution to the formulation of cybersecurity strategies [2]. In their study, they attempted to predict malicious nodes in IoT (Internet of Things) networks using classification algorithms. The study used a dataset consisting of 10,000 records with 21 attributes [3].

2. MATERIALS AND METHODS

2.1. Dataset

The dataset used in the study was obtained from the Kaggle platform. The dataset is provided in two different CSV formats: one for training and the other for testing. The training dataset was used for model training, while the test dataset was used for performance evaluation of the model. The dataset contains a total of 4962 records and consists of 133 columns, of which 132 are independent variables (symptoms) and 1 is the dependent variable (disease). The dataset includes 41 different diseases, with 120 examples for each disease. The diseases and their frequencies in the dataset are shown in Table 1.

Table 1. Diseases in the dataset

No	Disease Name	Count
1	Fungal infection	120
2	Hepatitis C	120
3	Hepatitis E	120
4	Alcoholic hepatitis	120
5	Tuberculosis	120
6	Common Cold	120
7	Pneumonia	120
8	Dimorphic	120
9	Heart attack	120
10	Varicose veins	120
11	Hypothyroidism	120
12	Hyperthyroidism	120
13	Hypoglycemia	120
14	Osteoarthristis	120
15	Arthritis	120
16	Vertigo	120
17	Acne	120
18	Urinary tract infection	120
19	Psoriasis	120
20	Hepatitis D	120
21	Hepatitis B	120
22	Allergy	120
23	Hepatitis A	120
24	GERD	120
25	Chronic cholestasis	120
26	Drug Reaction	120
27	Peptic ulcer diseae	120
28	AIDS	120
29	Diabetes	120
30	Gastroenteritis	120
31	Bronchial Asthma	120
32	Hypertension	120
33	Migraine	120
34	Cervical spondylosis	120
35	Paralysis	120
36	Jaundice	120
37	Malaria	120
38	Chicken pox	120
39	Dengue	120
40	Typhoid	120
41	Impetigo	120

2.2. Data Preprocessing

To prepare the data for model creation, various data preprocessing steps such as converting categorical data into numerical values, filling in missing data, scaling, and normalization are applied using the Pandas library in the Python programming language. Scaling and normalization are considered crucial steps to ensure that the values with different metrics in the dataset contribute equally to the model's performance [6].

2.3. Data Preparation

The data used in machine learning can be divided into two main categories: categorical data and numerical data. Categorical data represents qualitative attributes such as a person's education level, marital status, and gender, while numerical data represents quantitative characteristics such as salary, height, and personal expenses. Since machine learning algorithms can only operate on numerical data, they cannot work with raw string data. Therefore, categorical data must be converted into numerical values. Additionally, standardizing different types of data into a numerical format is crucial for the model's performance [11]. In this study, string data from the dataset has been converted into numerical data, making it ready for use with the model.

2.4. *Tools and Method*

The models created in the study were developed using the Python programming language, along with the Sci-Kit Learn and XGBoost libraries. During the data preprocessing phase, the NumPy, Pandas, and Matplotlib libraries were utilized.

3. RESULTS and DISCUSSION

In the study, during the model creation phase, 85% of the data from the total training dataset was used for the training set, and 15% was set aside for the test set. To evaluate the model's performance, metrics such as accuracy, F1 score, precision, and recall were calculated. For assessing the performance of the model across different algorithms, the following algorithms were used in sequence: Decision Trees, Support Vector Machines (SVM), Random Forest, Gradient Boosting, AdaBoost, and Naive Bayes.

3.1. Decision Tree

Decision Trees are a machine learning algorithm that resembles a flowchart, allowing for a clearer understanding of the decision-making process and is commonly used in classification and regression problems [12]. In this study, the metrics of the decision tree algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 0.92. The confusion matrix for the Decision Tree algorithm is shown in Figure 1.

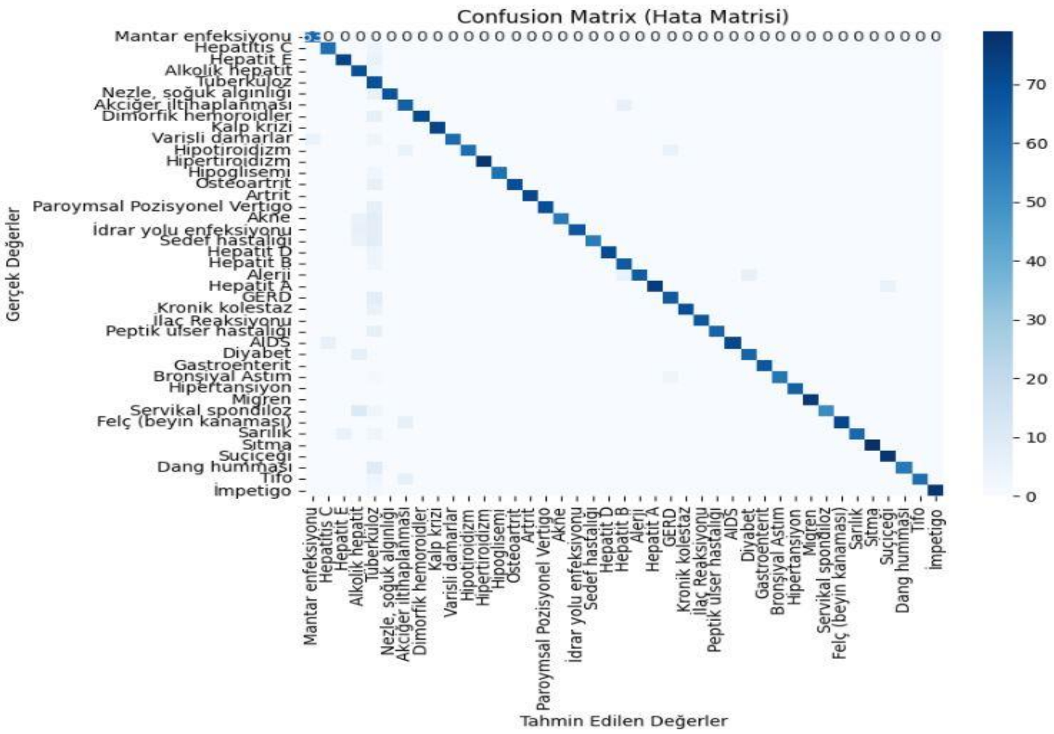

Figure 1. Decision tree confusion matrix

3.2. Support Vector Machine

Support Vector Machines is a machine learning algorithm used in classification and regression problems. It works by determining an optimal hyperplane that separates different classes and maximizes the margin between the nearest points of each class [13]. In this study, the metrics of the Support Vector Machines algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 1.0. The confusion matrix for the Support Vector Machines algorithm is shown in Figure 2.

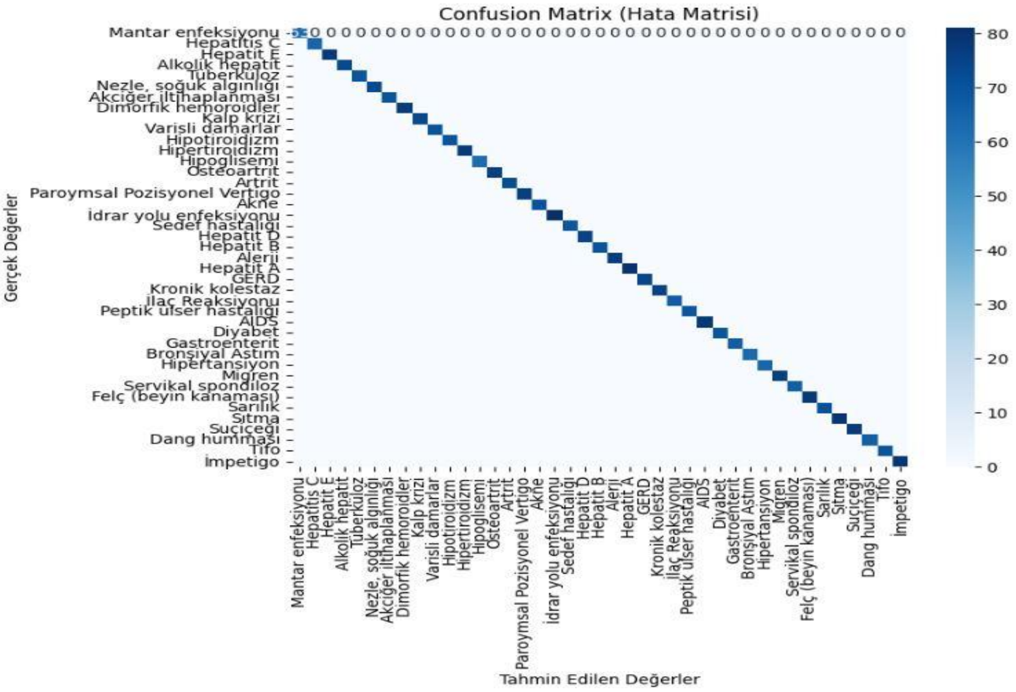

Figure 2. Support vector machine confusion matrix

3.3. Random Forest

Random Forest is a machine learning algorithm used in applications such as large datasets, hazard prediction, and performance analysis of electronic devices. The Random Forest algorithm demonstrates better performance compared to other machine learning algorithms [14]. In this study, the metrics of the Random Forest algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 0.85. The confusion matrix for the Random Forest algorithm is shown in Figure 3.

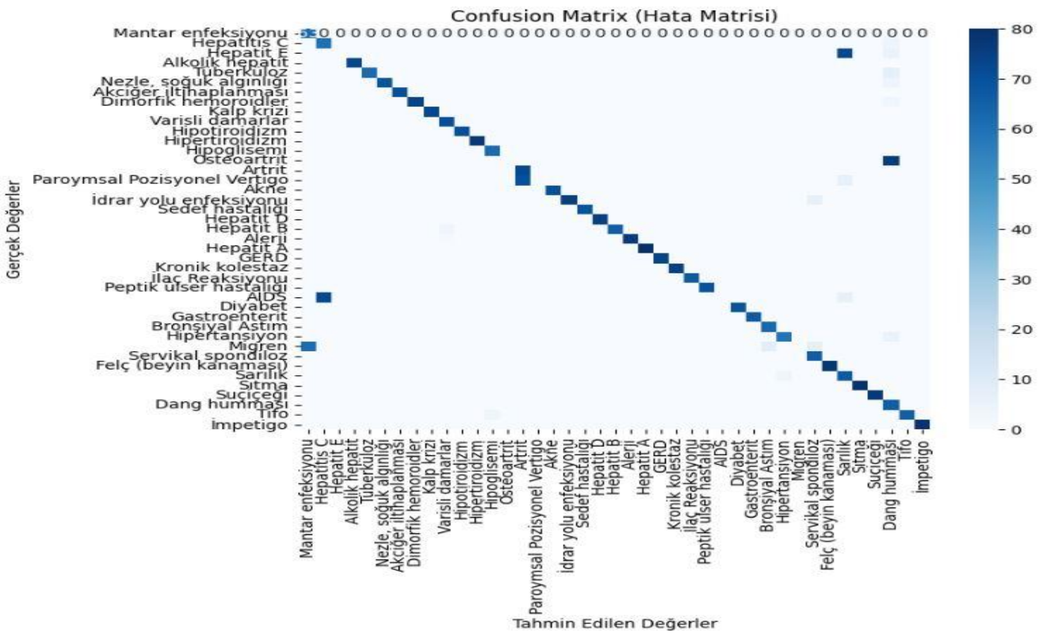

Figure 3. Random forest confusion matrix

3.4. Gradient Boosting

Gradient Boosting is an ensemble machine learning algorithm used to improve the prediction accuracy of previous models through decision trees [14]. In this study, the metrics of the Gradient Boosting algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 0.93. The confusion matrix for the Gradient Boosting algorithm is shown in Figure 4.

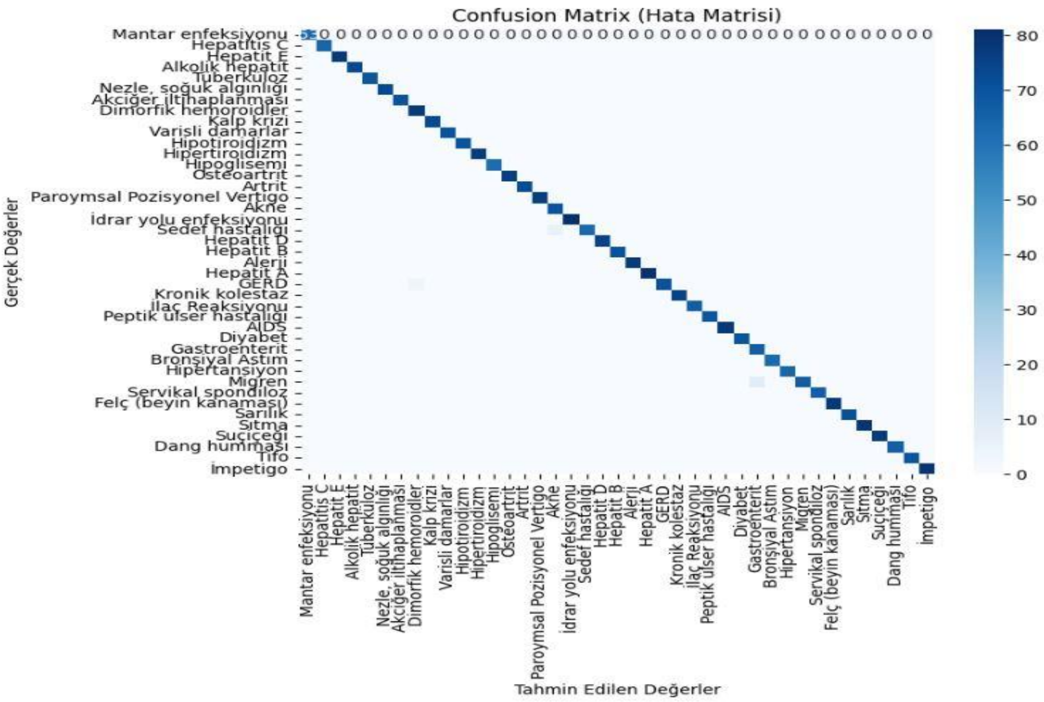

Figure 4. Gradient boosting confusion matrix

3.5. AdaBoost

Ada Boost is an ensemble machine learning algorithm that combines multiple weak classifiers to create a strong classifier [15]. In this study, the metrics of the Ada Boost algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 0.97. The confusion matrix for the AdaBoost algorithm is shown in Figure 5.

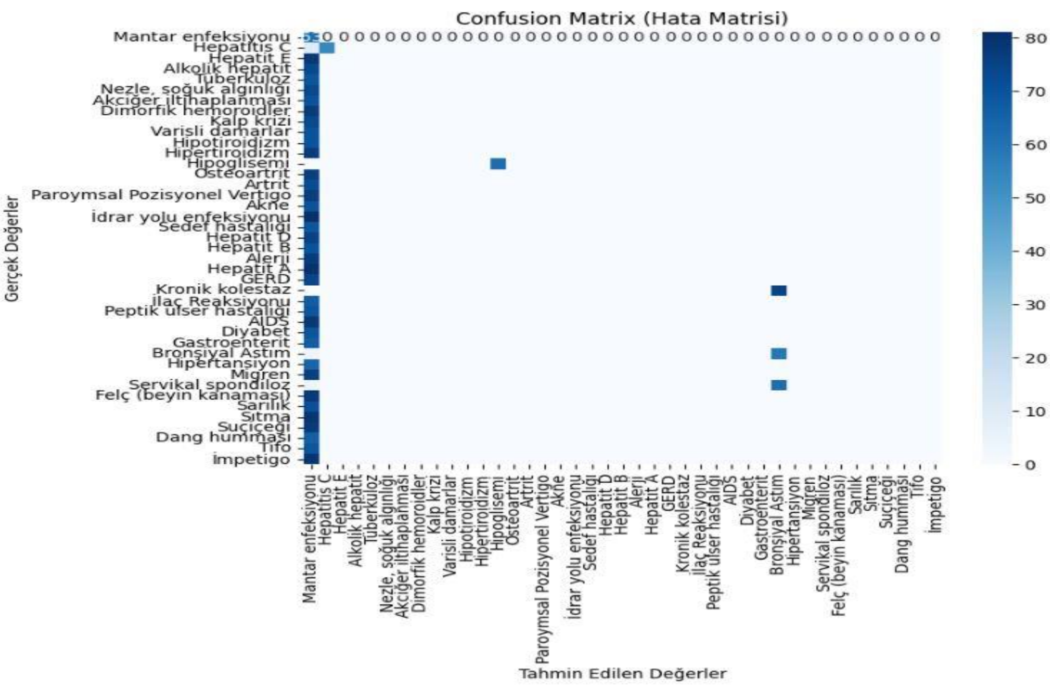

Figure 5. AdaBoost confusion matrix

3.6. *Naive Bayes*

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem, used for classification problems [16]. In this study, the metrics of the Naive Bayes algorithm have been examined to evaluate the performance of the developed model. As a result of this analysis, the model achieved an accuracy score of 1.0. The confusion matrix for the Naive Bayes algorithm is shown in Figure 6.

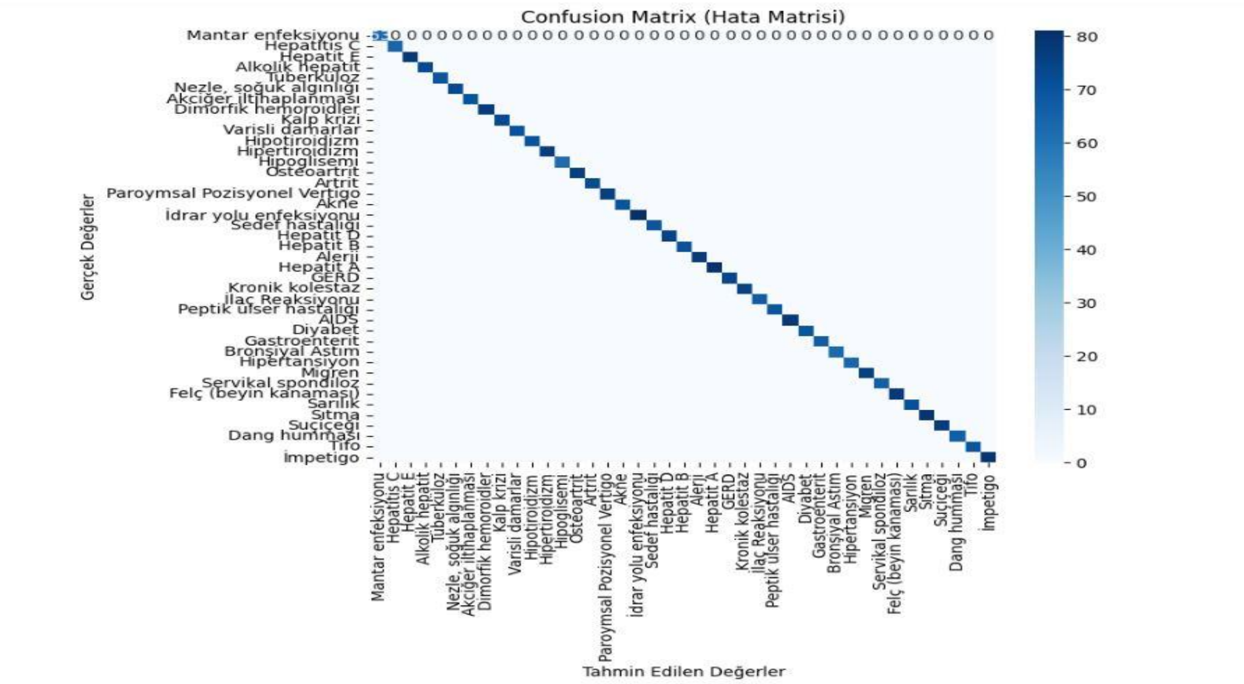

Figure 6. Naive bayes confusion matrix

4. Conclusion

This study focuses on developing a machine learning model that can predict a person's illness based on their symptoms. To calculate the accuracy Support Vector Machines, and AdaBoost algorithms. The detailed performance results of all the algorithms are shown in Table 2.

of the created model, six different machine learning algorithms were used. Among these six algorithms, the best results were obtained from Naive Bayes, Gradient Boosting,

Table 2. The performance scores of the algorithms

Name	Accuracy	F1 Score	Precision	Recall
Decision Tree	0.927845	0.927845	0.927845	0.927855
Support Vector Machine	1.0	1.0	1.0	1.0
Random Forest	0.855691	0.855691	0.855691	0.855691
Gradient Boosting	0.933563	0.933563	0.933563	0.933563
AdaBoost	0.972560	0.972609	0.972560	0.972560
Naïve Bayes	1.0	1.0	1.0	1.0

The results obtained in the study highlight the performance and effectiveness of traditional machine learning algorithms in classification problems. The evaluations show that machine learning methods achieve high accuracy rates in symptom-based disease prediction. In this context, it is planned to enhance the accuracy and generalization capabilities of the developed model and algorithms, and to optimize their performance by applying them to datasets containing more comprehensive symptom and disease data.

Conflict of Interest

No conflict of interest is declared by tehe authors. In addition, no financial support was received.

Ethics Committee

The study does not require ethical approval.

Author Contributions

Conception and design of the study: AGG, VA; Data collection: AGG, VA; Data analysis: AGG, VA;

Data Interpretation: AGG, VA; Drafting the article and/or its critical revision: AGG, VA; All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Anwer, Z., Shaker, A., Ihtiram, A., & Khan, R. (**2024**). Assessing institutional performance using machine learning algorithms. *Wasit Journal of Computer and Mathematics Science*, 3(3), 11–21. [CrossRef]
2. Sayed, M. A., Badruddowza, M., Sarker, U. S. M., Mamun, A. A., Nabi, N., Mahmud, F., Alam, M. K., Hasan, T., Buiya, R., Zaman, M., & Choudhury, E. (**2024**). Comparative analysis of machine learning algorithms for predicting cybersecurity attack success: A performance evaluation. *The American Journal of Engineering and Technology*, 6(9), 81–91. [CrossRef]
3. Poorana Senthikumar, S., Wilfred Blessing, N. R., Rajesh Kanna, R., & Karthik, S. (**2024**). Performance evaluation of predicting IoT malicious nodes using machine learning classification algorithms. *International Journal of Computational and Experimental Science and Engineering*, 10(3), 217–222. [CrossRef]
4. Neethu, S. (**2024**). Comprehensive performance evaluation of machine learning algorithms for detecting DDoS attacks in SDN. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(4), 4488–4499. [CrossRef]
5. Çömez, E., & İnan, O. (**2024**). Performance evaluation of machine learning algorithms in estimating taxi times at Istanbul Airport. *Intelligent Methods in Engineering Sciences*, *3*(3), 82–90. [CrossRef]
6. Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (**2021**). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research*, 9, 652801. [CrossRef]
7. Ryabko, B., & Fionov, A. (**2012**). Estimating the performance of computer systems through computer capacity. *In Proceedings of the International Conference on Research and Education in Development (pp. 74–77). IEEE*. [CrossRef]
8. Aweh, A. (**2022**). Using analytics to transform data into actionable insights. *Climate and Energy*, 39(3).
9. Altuntaş, V. (**2022**). Diffusion alignment coefficient (DAC): A novel similarity metric for protein-protein interaction network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2), 894–903. [CrossRef] [PubMed]
10. Altuntaş, V. (**2024**). NodeVector: A novel network node vectorization with graph analysis and deep learning. *Applied Sciences*, 14(2), 775. [CrossRef]
11. Modarresi, K., & Munir, A. (**2018**). Standardization of featureless variables for machine learning models using natural language processing. *In Natural Language Processing and Information Systems (pp. 234–246)*. [CrossRef]
12. Korpad, D., Satpute, N., Joshi, N., Kulkarni, S., Walgude, K., & Dhadiwal, N. (**2024**). *Numerical data processing by the implementation of trees and graphs*.
13. Suzuki, J. (**2020**). *Support vector machine. In Machine Learning and Data Mining* (pp. 171–192). *Springer, Singapore*.
14. Salah, A., & Yevick, D. (**2024**). A random forest model for predicting and analyzing the performance of CNT TFET with highly doped pockets. *Advanced Theory and Simulations*, *8*(1), 2400607. [CrossRef]
15. Kiran, R. J., Jayamohan, S., & Asharaf, S. (**2024**). A novel approach for model interpretability and domain aware fine-tuning in AdaBoost. *Human-Centric Intelligent Systems, 4*(1), 610–632. [CrossRef]
16. Arfah, D. J., Masrizal, M., & Irmayanti, I. (**2024**). Machine learning to predict student satisfaction level using KNN method and Naive Bayes method. Sinkron: *Jurnal dan Penelitian Teknik Informatika* , *8*(3), 1895–1908. [CrossRef]
17. Kösesoy, İ., (**2023**). Prediction of POMA-G Score from Spatiotemporal Gait Parameters. *Dicle University Journal of Engineering*, 14:4, 555-561. [CrossRef].
18. Altın, S., & Sarı, E. (**2021**). Reliability and validity of an innovative method of ROM measurement using Microsoft Kinect V2. *Pamukkale University Journal of Engineering Sciences*, 24(5), 915–920. [CrossRef].