http://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Ser. A2-A3
Volume 65, Number 1, Pages 52–68 (2023)
DOI:10.33769/aupse.1215962
ISSN 1303–6009 E-ISSN 2618–6462

Research Article; Received: December 09, 2022; Accepted: April 11, 2023

DISEASE PROGNOSIS USING MACHINE LEARNING ALGORITHMS BASED ON NEW CLINICAL DATASET

Melike COLAK[1], Talya TUMER-SIVRI[2], Nergis PERVAN-AKMAN[1], Ali BERKOL[1] and Yahya EKICI[3]

[1]Defense and Information Systems, BITES, Ankara, TÜRKİYE
[2]Middle East Technical University, Informatics Institute, Ankara, TÜRKİYE
[3]General Surgery Department, Medicana Health Point, Istanbul Beylikdüzü International Hospital, Istanbul, TÜRKİYE

Abstract. Today, artificial intelligence-based solutions are produced to facilitate human life in almost every field. The healthcare sector is one of the sectors which took advantage of these solutions. Due to reasons such as the world's ever-expanding population, ongoing epidemics, and the emergence of new disease types, it is becoming increasingly difficult for a patient to benefit from health services quickly and to make an accurate diagnosis. At this juncture, artificial intelligence reduces the patient density in hospitals, enables patients to access accurate information, and allows medical students to practice by seeing new cases. In this study, a new and reliable dataset was created with disease information obtained from various sources under the supervision of a specialist medical doctor. Then, new patient histories were added to the dataset used in the previous study, the experiments were repeated with the same algorithms, and the accuracy score comparison was presented. The created dataset includes 2006 unique patient histories, 358 symptoms, and 141 diseases and we think it will be a valuable dataset for researchers who make developments using machine learning in the field of healthcare. Various machine learning algorithms have been used in the training process to predict diseases belonging to different branches of medicine, such as diabetes, bronchial asthma, and covid. Besides, Support Vector Machine, Naive Bayes, K-Nearest Neighbors, Multilayer Perceptron, Decision Tree, and Random Forest algorithms, we also studied popular boosting algorithms such as XGBoost and LightGBM. All algorithms were validated with cross-validation and performance comparisons were made with different performance metrics such as accuracy, precision, recall, and f1-score. It is also the first study to achieve an accuracy score of 99.33% with a dataset that involves a greater number of diseases than the datasets used in the studies examined.

Keywords. Healthcare symptom checker, clinical decision support systems, machine learning.

nergis.pervan@bites.com.tr-Corresponding author; 0000-0003-3241-6812
melike.colak@bites.com.tr; 0000-0002-7779-4756
talyatumer@gmail.com; 0000-0003-1813-5539
ali.berkol@bites.com.tr; 0000-0002-3056-1226
yahya.ekici@medicana.com.tr; 0000-0002-8518-8967.

1. Introduction

There are thousands of diseases that people faced all over the world, and approximately 69 million people all over the world died from various diseases or accidents in 2021 [1]. Due to the hectic pace of life, people began to pay little attention to the symptoms they felt and to not take the time to go to the hospital for a possible diagnosis of a disease. Such problems brought by the accelerating life, unfortunately, reduce the quality of life of people and shorten their lives. With the development of Artificial Intelligence (AI) and data analytics in the healthcare sector, these datasets, which store the history of thousands of patients, make sense with these technologies that produce automated approaches.

Medical organizations around the world have data on a variety of health-related topics. This data is too large for the human mind to grasp and it must be freed from noise by exploring a variety of data analytic methods to be used in various Machine Learning (ML) algorithms. Recent advances in data analytics tools and methods now enable the comprehensive use of data such as demographics, clinical diagnosis, health habits, test results, prescriptions, and service usage. As a result of processes such as data cleaning, categorization, and analysis by data analysts, this data becomes meaningful and contributes to the development of various AI studies to predict disease or identify patients at risk for other health problems. Today, cancer detection by processing X-ray images with deep learning techniques [2–5] and heart disease and stroke risk estimated with data such as heart rate and oxygen ratio in the blood [6–9]. Also, there is a study that classifies X-ray images with ML algorithms and helps orthopedists in the determination of shoulder implant types before performing revision surgery [10]. With the help of these studies, the cost required to reach health services will be reduced and the density in the hospital will be prevented. In the literature, the number of studies with ML [11–13], developed with datasets containing both categorical and image data containing the symptoms of a specific disease is quite high.

Nowadays, people want to quickly access any type of information they require from websites. When a person feels symptoms, they may be directed to false diseases unrelated to the person by searching websites containing incorrect or incomplete information. In this study, deep learning-based, ensemble-based, and tree-based approaches are presented to aim to protect people against the information pollution they will be exposed to from the internet and gives reliable results when people enter the symptoms they feel without the need to go to the hospital to get information about their disease. In this way, it is aimed to save people's time and health costs and to direct them to a medical doctor with an early diagnosis. We have seen that Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), and Decision Tree (DT) algorithms are frequently used in disease prediction developments. In addition to these algorithms, we tested the Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Multilayer Perceptron (MLP) algorithms that

we did not see used in other studies. We succeeded in surpassing the accuracy scores obtained from the SVM, random forest, decision tree, KNN, naive Bayes, and LightGBM algorithms in our first study [11]. This study contributed to the literature at two important points, under the supervision of a specialist medical doctor, a new dataset was created with disease information obtained from various sources. Secondly, it is also the first known study to reach an accuracy score of 99.33% with a dataset with a greater number of diseases than the datasets used in the studies examined.

2. Literature Review

Developments based on ML, which predict disease based on patient symptoms, have received a lot of attention recently due to the challenges associated with accessing healthcare services. With the DBMI dataset [14], which had 133 symptoms and 42 disease types, Gandhi et al. [15] experimented with supervised and unsupervised algorithms. In experiments with the Linear Discriminant Analysis (LDA), random forest, naive Bayes, SVM, KNN, Classification and Regression Trees (CART), and logistic regression algorithms, the accuracy score for logistic regression was the lowest of the group, coming at 80.85%. Agrawal et al. [16] suggested a new ML model in this research that combines a support vector machine and a genetic algorithm. Additionally, they attempted to reduce the number of features in the dataset, and by using their ML models, they were able to achieve adequate accuracy for all three datasets. They attained the best accuracy of 78.6% for the liver dataset consisting of categorical data. They claim that unstructured medical text data from sources including diagnoses, doctor-patient interactions, medical records, etc. would also be used in future research despite using only structured data in this study.

To get over the limitations of ML, Vinitha et al. [17], proposed leveraging big data to predict diseases based on ML. The idea is to gather information from a hospital that used the Map Reduce (MR) approach and Machine Learning Decision Tree (MLDT) algorithm to analyze data from a forum referred to as structured and unstructured data. The MR algorithm detects the possibility of disease occurrences faster than CNN-UDRP, reaching 94.8% with the standard speed. Kumar, Sharma, and Prakash [18], created a Django-based online application that uses ML algorithms to predict and provide clinical guidance for general disease, heart disease, diabetes, and liver disease. While the results of predictions for common diseases are the names of the diseases, results for predictions for specific diseases, such as heart disease, diabetes, and liver disease, are true or false. In general disease prediction, it is seen that the highest accuracy score of 90.2% among the KNN, logistic regression, random forest, and naive Bayes algorithms was achieved in random forest. The highest accuracy in heart disease was seen in logistic regression, with 92.3%. While the KNN algorithm gave the highest accuracy with 74% in the liver, it was seen that logistic regression gave the highest accuracy with 78% in diabetes. Mallela, Bhavani, and Ankayarkanni [19], developed a GUI to get the

symptoms from the user and they used ML models such as naive Bayes and decision trees. The outputs are the disease, the accuracy of the model, its definition, and the treatment of the particular disease based on the symptoms given by the individual. This paper shows a detailed explanation of how to find the diseases from symptoms; so that the individual can contact the respective doctor of medicine and stay healthy at an early stage. A sample of 4920 patient records with diagnoses for 41 disorders was chosen by Grampurohit and Sagarnal [20] for analysis. 41 diseases made up a dependent variable. There were 132 independent variables, 95 of which were symptoms closely associated with diseases. The disease prediction system created utilizing ML techniques including decision tree, random forest, and naive Bayes is demonstrated in this research project.

Dhabarde et al. [21] use not only structured data but also textbook data, and the dataset used has 230 conditions consisting of the individual's symptoms, age, and gender. In the paper, they conducted experiments with logistic regression, naive Bayes, SVM, random forest, and decision tree algorithms, and the decision tree gave the highest accuracy score, 93.24%. Alanazi [22], proposes a method for chronic disease prediction using ML algorithms such as Convolutional Neural Network (CNN) and KNN. The proposed system used both structured and unstructured data from real life which were used for dataset preparation. The performance of the proposed model in the study shows that it is higher than the naive Bayes, decision tree, and logistic regression algorithms and provides 95% accuracy. Uddin et al. [23], conducted a study on different KNN variants (classical one, adaptive, locally adaptive, K-means clustering, fuzzy, reciprocal, ensemble, Hassanat, and generalized mean distance) and their performance comparison for disease prediction. For accuracy measurement, Hassanat KNN shows the highest average accuracy with 83.62%, followed by ensemble approach KNN with 82.34%.

For disease prediction with big data in healthcare, Joel and Priya [24], employed extended CNN. The hospital is built using this approach, which offers great accuracy, performance, and convergence speed in the medical industry. The unstructured data is employed with the CNN algorithm, which automatically selects the features, to choose a specific location and then assesses the chronic diseases that contain the structured data which extracted valuable features. The medical data and illness risk model were proposed by the innovative CNN. The suggested approach seeks to forecast the likelihood of liver-focused illness. Therefore, the hospital dataset is concerned with diseases that affect the liver, and it exclusively collects structured data from information on liver diseases. The proposed approach obtains accuracy by using disease risk modeling. Ibrahim et al. [25] proposed a method for predicting the defervescence day of fever in dengue patients using an artificial neural network. The suggested method primarily depends on clinical symptoms and indicators for detection. Data from 252 patients were collected, of which 4 patients had Dengue Fever (DF) and 248 had Dengue Haemorrhagic Fever (DHF). The neural network toolkit in MATLAB is utilized and the Multi-layer Feed-Forward Neural

Network (MFNN) technique is used in this experiment. 90% of the time, MFNN in DF and DHF correctly predicts the day of defervescence of fever. Venkatesh et al. [26], worked on five algorithms, such as random forest, KNN, naive Bayes, SVM, and decision tree; the highest accuracy score was decision tree, with 95.13%. They also have developed a user interface for patients to input their symptoms and see the disease prediction. Chauhan et al. [27], performed preprocessing on the dataset and then performed experiments on naive Bayes, decision tree, and random forest. When the experiments performed on the non-preprocessed dataset were compared with the results of the preprocessed data, it was seen that the accuracy score of the random forest was the highest in both, increasing to 95.28% in raw data and 97.64% after processing.

Maram, Kumar, and Gampala [28], stored data including 400 symptoms and 147 diseases collected from various repositories in the Hadoop Distributed File System (HDFS). Among decision trees, random forest, naive Bayes, and a new algorithm proposed in the article, the accuracy of the proposed algorithm showed the best result, with 97.60%. Through the analysis of performance measures, Ferjani [29], identifies patterns among several supervised ML model types for disease diagnosis. The supervised ML algorithms, naive Bayes, decision trees, and KNN, received the greatest attention. According to research, a support vector machine is most effective at spotting Parkinson's illness and kidney ailments. They found that the logistic regression performed well for heart disease prediction. Additionally, CNN and random forest made accurate predictions for common diseases and breast disorders, respectively. For accurate prediction, naive Bayes and KNN algorithms were used in [30] to process the person's life behaviors and check-up data. The accuracy of heart disease prediction using naive Bayes was shown to be 94.5% greater than KNN. Furthermore, compared to naive Bayes, KNN requires more memory and time. In this work, heart disease was first predicted, and then a risk prediction system using the CNN algorithm was developed to assess the risk of heart disease.

The CNN-based Multimodal Disease Prediction (CNN-MDRP) method was developed by Shirsath and Patil [31] to address the limitations of their CNN-based Unimodal Disease Prediction (CNN-UDRP) algorithm, which only analyzes structured data. In CNN-MDRP, which focuses on both structured and unstructured data, the accuracy of disease prediction is higher and faster compared to CNN-UDRP, with an accuracy score of 94.80%. Nearly 230 diseases were listed by Keniya et al. [32] with over 1000 distinctive symptoms. Various ML algorithms receive as input a person's symptoms, age, and gender. About 230 diseases were predicted using 11 different ML algorithms. The weighted KNN model had a 93.5% accuracy score, which was the highest. For disease prediction, Dahiwade, Patle, and Meshram [33] used KNN and CNN algorithms. The model accepts information from

the person's checkups and daily routine as input. With 84.5% accuracy, CNN outperforms the KNN algorithm in general disease prediction. The time and memory requirements for KNN are also higher than for CNN.

3. Methodology

3.1. **Utilized Machine Learning Algorithms.**

3.1.1. *K-Nearest neighbors.* KNN, which is used in both classification and regression problems, is an algorithm used in supervised learning. The basic logic is to search for K data, which are called neighbors and have the closest properties throughout the dataset, for data whose class is unknown, and assign it to the most appropriate class. There are several methods for calculating inter-data distance, the most well-known being Euclidean, Manhattan (for continuous), and Hamming distance (for categorical). The mathematical representation of the methods is shown in Eq. 1, Eq. 2 and Eq. 3:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

$$\text{Manhattan Distance} = \sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

$$\text{Minkowski Distance} = (\sum_{i=1}^{n}|x_i - y_i|^p)^{1/p} \tag{3}$$

In Eqs. 1, 2 and 3, n refers to the number of dimensions, x_i refers to the data point, and y_i refers to the new data point that is wanted to predict for all i, where $i, \in \{1, 2, ..., n\}$, n is the size of the data points. KNN applies one of these formulas to calculate the distance between each data point and the test data. It then finds the probability that these points are similar to the test data and classifies the data according to which points share the highest probability.

3.1.2. *Support vector machine.* SVM is one of the linear supervised learning models used in classification and regression problems. When determining the class of new data, it tries to determine a dividing line (hyperplane) that best separates the available data from each other. Hyperplane can be formulated as follows,

$$f(x) = ax + c \tag{4}$$

In Eq. 4, a equals to dimensional coefficient and c equals the offset. Then, the a point closest to this line is selected from both classes, and these points are called

support vectors. The algorithm aims to provide the maximum distance difference between the support vectors and the hyperplane. There are two different types of SVM algorithms, Linear and Nonlinear. Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and the classifier is used as Linear SVM classifier. Non-linear SVM is used in cases where the dataset cannot be classified with a straight line and contains non-linear data.

3.1.3. *Naive bayes.* Naive Bayes is an algorithm based on Bayes' theorem used in classification problems, and it is known as a probabilistic classifier. It assumes that the value of probability ω_j is independent of the probability of any other event x, which means that the dependencies between the data are neglected. The simple form of calculation for Bayes' theorem is shown below.

$$P(\omega_j|x) = \frac{p(x|\omega_j) \cdot p(\omega_j)}{p(x)} \tag{5}$$

In Eq. 5, $P(\omega_j|x)$ is the posterior probability of class (target) given predictor (attribute). $p(\omega_j)$ is the prior probability of class. $p(x|\omega_j)$ is the likelihood which is the probability of the predictor given class and $p(x)$ is the prior probability of the predictor.

3.1.4. *Decision tree.* A decision tree classifier creates a model that will predict the target variable by learning the simple decision rule extracted from the feature data. In this algorithm, which has two types of nodes, decision, and leaf, the decision nodes play a decisive role in reaching the leaf, while the leaf node is the result node. Due to its nature, it is more suitable for multi-class problems. The mathematical process starts with the $D = X, y$ dataset, where each node must have a tree structure and decision rules. Each node divides the dataset into two or more discrete subsets and $D_(a, b)$, in which D represents subscript(a, b), where a is the layer number and b denotes each subset. If all tags in this subset belong to the same class, the subset is said to be pure and this node is declared as a leaf node, and this part of the tree has come to an end. Otherwise, the partitioning process continues. Data is considered pure or homogeneous if it contains only one class, and impure or heterogeneous if it contains several classes. There are various indices such as entropy and Gini to quantify the degree of impurity. Entropy is the amount of information required to accurately describe some samples. That is, if the sample is homogeneous, all elements are similar with entropy 0, otherwise if the sample entropy is divided evenly by a maximum of 1. The other index, the Gini index, is defined as a measure of inequality in the data and has a value between 0 and 1. If the Gini index is 0, the data is considered completely homogeneous and all elements are similar. A Gini index of 1 means the maximum inequality between the elements. It is the sum of the squared probabilities of each class. Mathematical expressions for the Gini index and entropy are shown in Eq. 6 and Eq. 7.

$$\text{Gini index} = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{6}$$

where p_i is the probability of an object being classified to a particular class.

$$\text{Entropy} = \sum_{i=1}^{c} -p_i log_2 p_i \tag{7}$$

where p denotes the probability.

3.1.5. *Random forest.* Random forest, one of the supervised learning algorithms, combines multiple classifiers to solve a classification or regression problem and improve the model's performance. Instead of relying on a single decision tree to reach the result, it combines the output of multiple decision trees to obtain safer results. High variance and low bias are characteristics of decision trees, and by averaging decision trees, the variance component of the model is reduced. It is possible to create the unknown samples by averaging the prediction with Eq. 8 and Eq. 9,

$$I = \frac{1}{N} \sum_{N}^{n=1} f(x) \tag{8}$$

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N} \left(f(x) - \hat{f} \right)^2}{N-1}} \tag{9}$$

Where σ denotes the uncertainty and N denotes the sample number.

Hence, random forest is a bagging algorithm, which is a method of generating different training subsets from training data with replacement. A final result is reached by calculating the feature importance in each decision tree that makes up the random forest. Feature importance is calculated as the reduction in node impurity weighted by the probability of reaching that node. The node probability can be calculated by dividing the number of samples reaching the node by the total number of samples. The higher the value, the more important the feature. For each decision tree, a node's importance is calculated using Gini importance. Then the importance of each feature on a decision tree is then calculated and these can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values. At the random forest level final feature importance is its average over all the trees. The sum of the feature's importance value on each tree is calculated and divided by the total number of trees. The feature importance describes which features are relevant and it sometimes leads to model improvements by employing the feature selection.

3.1.6. *Extreme gradient boosting.* XGBoost is a specially optimized version of the gradient boosting algorithm for high performance and speed and uses decision trees as "weak" predictors. Since it contains many parameters that can be optimized, it provides the opportunity to improve the model. In addition to system optimizations such as parallel operation, tree pruning, and hardware optimization, algorithmic improvements such as regularization, cross-validation, weighted quantile sketch, and sparsity-aware split make XGBoost a more efficient algorithm compared to other models. Mathematically in the XGBoost algorithm, the objective function is shown in Eq. 10.

$$O(t) = \sum_{i=0}^{n} Q(y_i, y^{'t-1} + f_t(x_i)) + K \tag{10}$$

Normalization function can be defined as

$$Nor(f_t) = \kappa T + 0.5\lambda \sum_{i=0}^{T} {W_j}^2 \tag{11}$$

κ =Controlling factor for the leaf node number, T =Leaf node number, W_j =Weightage of the j leaf nodes, λ =Overfitting controlling factor, K =Constant, in Eq. 10 and Eq. 11.

XGBoost performs exceptionally well on structured tabular data rather than data like images and audio.

3.1.7. *Light gradient boosting machine.* LightGBM is a gradient-assisted decision tree algorithm that is very similar to XGBoost, increasing the model's accuracy and reducing memory usage. LightGBM develops models that have lower error rates and learn faster by using a leaf-oriented strategy instead of a level-oriented strategy in decision trees. According to the [34] where the model is introduced, it has been concluded that LightGBM is 20 times faster than other models. The decision tree growth process used by XGBoost and LightGBM differs significantly. Decision trees in XGBoost grow horizontally with a method called level-wise, but decision trees in LightGBM grow vertically with a leaf-wise approach. The leaf-wise approach is an effective method as it makes LightGBM an efficient technique for high-dimensional datasets. XGBoost and LightGBM decision tree growth processes are shown in Figure 1. LightGBM's multithreaded optimization and leaf growth technique with depth restriction help to reduce excessive XGBoost memory usage so that big data processing can be done more quickly, with fewer false alarms, and with lower missed detections.

3.1.8. *Multilayer perceptron.* Multilayer perceptron is mainly used in recognition, prediction, regression, and pattern classification and it is the most basic type of feed-forward network architecture, compared to other major types. Here, the units are arranged into a set of layers, and each layer contains some number of identical units. The network is considered fully connected when each unit in a layer connects to each unit in the layer above it. The input layer is the top layer, and its units use

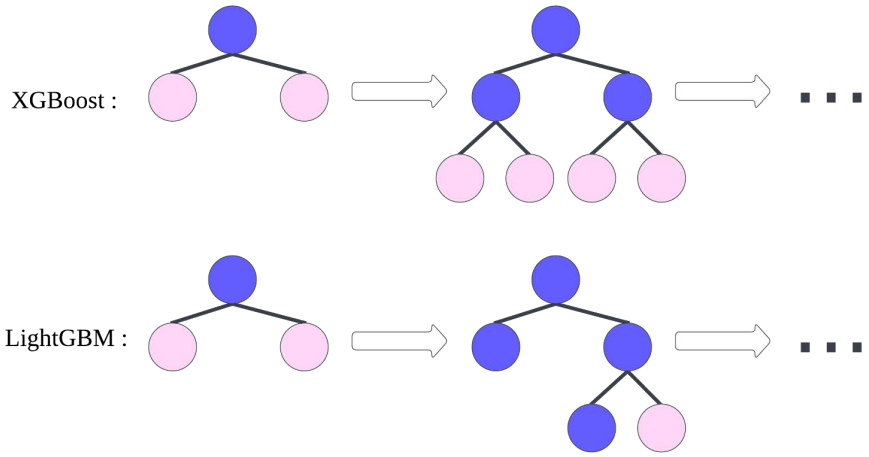

FIGURE 1. XGboost and LightGBM decision tree growth processes.

the values of the input properties as input. For each value produced by the network, there is one unit in the output layer in case of regression or binary classification, and K units in classification cases involving K classes. All the levels in between are called hidden layers because it is not known what these units have to calculate in advance and must learn this while training. Input units, output units, and hidden units are the names of units located at these levels, respectively. Backpropagation MLPs can be used to solve problems that cannot be classified linearly, such as the XOR problem in sensors.

3.2. **Dataset.** In our previous study [11], the dataset prepared by DBMI [14] was used. The previous dataset consisted of 133 symptoms and 42 diseases and included 306 patient histories. In this study, we add disease and symptom data from various sources under the supervision of a specialist medical doctor; and use a new dataset in which we increased the total number of cases approximately 14 times with data augmentation.

3.2.1. *Data collection.* The diseases in the dataset were searched on the Internet and matched with the relevant symptoms. After this process, a dataset containing 150 diseases and 383 symptoms was created. The created dataset was checked by a medical doctor, and if there was incorrect symptom information, it was removed from the dataset and the reliability of the dataset was ensured. After making sure that the data were correct, the dataset was ready for preprocessing with 141 diseases and 358 symptoms. A part of the dataset created at the end of this stage is shown in Table 1.

Table 1. Raw dataset.

disease	symptom 1	symptom 2	symptom 3	...
acne	thirst	blackheads	skin rash	...
acute	pancreatitis	vomiting	diarrhoea	...
addison's disease	fatigue	lethargy	low mood	...
adenovirus	diarrhoea	cough	runny nose	...
aids	extra marital contacts	patches in throat	high fever	...
...	...	...	...	...

3.2.2. *Preprocessing.* Any manipulation applied to raw data is called data preprocessing. Through this process, scattered data becomes organized, and problem-appropriate, and transforms into a format that can be processed effectively in ML developments. To provide reliable, accurate, and robust findings for enterprise applications, practically every sort of data analysis, data science, or ML development requires some kind of data preprocessing. At this stage, the columns in the dataset were changed to consist of 358 symptom names and target columns, and the rows to consist of 141 patient cases. Data preprocessing was carried out so that the value at the intersection of the row and column is "1" if the disease contains the relevant symptom, and "0" if it does not. Except for the target column values, disease, which is a textual type that is converted to numerical form using a label encoder, all of the column values in the dataset are numbers. A part of the dataset formed after preprocessing is shown as an example in Table 2.

Table 2. Preprocessed dataset.

...	yellow urine	yellowish eyes	yellowish skin	disease
...	0	0	0	acne
...	0	0	0	acute
...	0	0	0	addison's disease
...	0	0	0	adenovirus
...	0	0	0	aids
...	...	...	...	...

3.2.3. *Data augmentation.* By creating additional data points from existing data, a group of techniques known as data augmentation can be used to artificially enhance the amount of data. This includes making minor adjustments to the data or creating new data points using ML models. By creating additional and distinct instances for training datasets, data augmentation helps ML models perform better and more accurately with large datasets. For this reason, it is aimed to obtain more realistic results by adding new data to the dataset used in our previous study [11] and applying data augmentation. Some symptoms play a key role in the prediction of

diseases. An example of this is the symptom of loss of smell for Covid disease. For this reason, while creating a new patient history with data augmentation, attention was paid to including these symptoms in each patient's history. In addition, the data augmentation process is aimed to make the data more suitable for real-world cases by completely randomly determining how many times each disease will increase with this process and what symptoms it will contain. As a result of this process, the dataset was increased approximately 15 times, and a new dataset was created with 2006 patient histories.

3.2.4. *Data splitting.* When starting an ML project, one of the first considerations to be discussed is how to use existing data. Typically referred to as training and test sets, dividing data into two groups is a standard strategy. When making predictions on data that was not used to train the model, ML algorithms perform as predicted using the train-test separation process. The training set is used for estimating parameters, comparing models, and all other activities required to arrive at a final model. The test set is used to predict a final, unbiased assessment of the model's performance only at the end of these activities Since the test set is used to measure the performance of the model after the training is over, it reveals a high but erroneous performance result that the model had seen before. The most common technique for splitting the dataset into training and test sets is random splitting. The used dataset, which was randomly divided by 85-15%, 1713 samples were determined as training data and 303 samples as test data. The cross-validation method was utilized in the study in addition to the traditional method of creating test data at random to evaluate the model's performance.

3.3. **Training.** In the previous study, k-fold cross-validation was used with the k value chosen as 5 to avoid overfitting, and the average accuracy of the classifier was taken. The choice of k is usually 5 or 10, but there is no formal rule. As k gets larger, the difference in size between the training set and the re-sampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller [35]. K-fold cross-validation is not suitable for evaluating unbalanced classifiers because the data is divided into k-folds with a uniform probability distribution. This method may work for data with a well-balanced class distribution, but it is dangerous when the distribution is severely dispersed, with one or more of the folds having few or no samples from the minority class. It means that most model evaluations produce misleading results, as the model only needs to predict the majority class correctly. Accuracy scores with high deviations at the end of the training k times help to understand that the data in the dataset is not evenly distributed [36]. In this study, the k value remained constant to make a reliable comparison of the change in performance metrics with the expansion of the dataset.

In the experiments, the hyperparameter optimization library hyperopt was used to find the parameters showing the highest accuracy score, and the learning rate in the LightGBM model was determined as 0.1175, max bin 316, max depth 3, num

leaves 200, objective parameters binary. In the XGBoost model, hyperparameters are determined as one drop true, learning rate 0.3, colsample by tree 0.5698, gamma 0.5296, max depth, and objective multi:softprob. In random forest, criterion parameter is entropy, max-depth is 14,056 with and the n-estimators parameter is 340. In SVM, the C parameter was determined as 1.144, gamma, 0.278 and kernel rbf. In KNN, we determined the n-neighbors parameter value as 3. The decision tree's criterion value is determined by Gini and in naive Bayes, the alpha parameter is 1.0. As a result of the experiments, the highest accuracy score was achieved in the MLP model by using the activation function tanh, hidden layer sizes 32 and max iter 3000 hyperparameters. For other classifier parameters not mentioned above, the default parameter values of the scikit-learn library are based.

4. Results

In the experiments conducted in the study, the lowest accuracy score was obtained as 92% among all algorithms. The fact that this value was 79.3% in the previous study, clearly showed the positive effect of the data augmentation techniques we applied to the disease detection experiments. Among all algorithms, it was seen that the highest scores in all metrics were obtained in the ensemble learning method, the random forest algorithm, which consists of many decision trees. The most important feature of the random forest algorithm is that it can work with high performance in regression and classification problems with datasets containing continuous or categorical variables. The decision trees that made up the random forests are prone to errors in classification problems with a large number of classes and relatively few training examples. The decision tree, which had the lowest accuracy score of 79.3% in the first dataset containing a small number of samples from each class, increased to 95.37% as a result of increasing the number of data belonging to each class. However, SVM appears to perform as well as random forest in its measurements across all metrics, achieving the second-highest precision, recall, and F1-score results. It also appears that the SVM achieves the highest average accuracy in the k-fold algorithm. In addition to the remarkable performance of random forest and SVM, we observed that the MLP algorithm was the 3rd highest-performing algorithm in the experiments. Although the lightGBM algorithm has a high accuracy score compared to the studies in the literature, it gave the lowest result among the algorithms in all performance metrics in the experiments conducted in our study. The most valuable output of the experimental results is that we obtained higher accuracy scores in SVM, lightGBM, random forest, naive Bayes, decision tree, and KNN algorithms compared to our previous study.

All algorithms were fitted using the hyperparameters mentioned in the training section; where estimators were evaluated using k-fold cross-validation where k is 5 and average accuracy scores are shown in Table 3. Precision, recall, and f1-score performance measures were used to examine the performance of the models. The blue values in Table 3 show the best results for that metric among the models,

while the red color indicates the worst result. The model results are shown in Table 3 with the specified performance measures and methods, and the comparison of accuracy scores of the methods is shown in Figure 2.

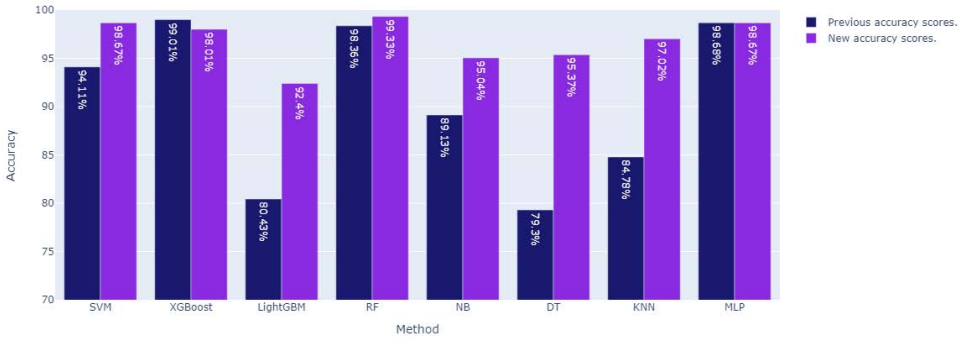

FIGURE 2. Comparison of accuracy scores of the utilized methods.

TABLE 3. Model results with the specified performance measures and methods.

Method	5-Fold Average Accuracy	Accuracy	Precision	Recall	F1-Score
SVM	99.25	98.67	99.02	98.81	98.92
XGBoost	97.83	98.01	97.47	96.37	96.92
LightGBM	88.39	92.40	90.80	90.56	89.72
RF	99.20	99.33	99.51	99.30	99.40
NB	93.50	95.04	92.79	93.70	93.24
DT	93.55	95.37	96.08	95.55	95.81
KNN	98.31	97.02	97.52	96.76	97.14
MLP	99.05	98.67	98.48	97.86	98.17

5. CONCLUSION

Technological developments in healthcare are highly important as they are directly related to human life. In this paper, a disease prediction study is developed where people can obtain free, reliable, and fast information about their health. The development we offer serves important purposes such as reducing the patient density in hospitals and early diagnosis of viral diseases. In addition, with our study, which we developed under the supervision of a specialist medical doctor, it is possible to

prevent people from obtaining information about their health from websites that contain information pollution. As an alternative to the dataset used in our previous study, a new dataset consisting of clinical data was created under the supervision of medical doctors, and the data size was increased. The new study includes combinations of different types of symptoms and diseases. The highest accuracy score is reached with the random forest classifier at 99.33%, while the lightGBM has the lowest accuracy score among the algorithms with 92.40%. Although our study consisted of a high number of disease classes compared to the studies in the literature, all algorithms obtained reliable results with an accuracy score of over 92%. For future studies, the determination of the severity of the symptoms in the dataset by the doctors and the learning of the model during the training process, taking into consideration the symptom weights of each disease, are among our further studies.

Author Contribution Statements The authors jointly worked on the study. All authors read and approved the final copy of the manuscript.

Declaration of Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement In this study, we made progress from our previous work [11] in terms of both increasing the number of data and improving the algorithms used. The study that forms the basis of this paper was presented at the 5. International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE).

References

[1] Our World in Data, (2022). Available: https://ourworldindata.org/births-and-deaths/. [Accessed: December 2022].

[2] Cantalay, P. J., Uçan, O. N., Zontul, M., Diagnosis of breast cancer from X-ray images using deep learning methods, *J. Ponte*, 77 (6), (2021), https://doi.org/10.21506/j.ponte.2021.6.1.

[3] Wang, Y., Yang, F., Zhang, J., Yue, X., Liu, S., Application of artificial intelligence based on deep learning in breast cancer screening and imaging diagnosis, *Neural Comput. & Applic.*, (2021), 9637–9647, https://doi.org/10.1007/s00521-021-05728-x.

[4] Mobark, N., Hamad, S., Rida, S. Z., CoroNet: Deep neural network-based end-to-end training for breast cancer diagnosis, *Appl. Sci.*, 12 (14), (2022), 7080, https://doi.org/10.3390/app12147080.

[5] Manishkumar, S. H. and Saranya, P., Detection and classification of breast cancer from mammogram images using adaptive deep learning technique, *2022 6th Int'l Conf. on Dev., Circ. & Syst.*, (2022), 327-331, https://doi.org/10.1109/ICDCS54290.2022.9780770.

[6] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., Pranavanand, S., Heart disease risk prediction using machine learning classifiers with attribute evaluators, *Appl. Sci.*, 11 (18) (2021), 8352, https://doi.org/10.3390/app11188352.

[7] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., Singh, P., Prediction of heart disease using a combination of machine learning and deep learning, *Comp. Intell. & Neurosci.*, (2021), 8387680, https://doi.org/10.1155/2021/8387680.

[8] Mehmood, A., Iqbal, M., Mehmood, Z. et al., Prediction of heart disease using deep convolutional neural networks, *Arab. J. for Sci. & Eng.*, 46 (2021), 3409–3422, https://doi.org/10.1007/s13369-020-05105-1.

[9] Puri, H., Chaudhary, J., Raghavendra, K. R., Mantri, R., Bingi, K., Prediction of heart stroke using support vector machine algorithm, *21 8th Int'l Conf. on Sm. Comput. & Comm.*, (2021), 21-26, https://doi.org/10.1109/ICSCC51209.2021.9528241.

[10] Sivari, E., Güzel, M. S., Bostanci, E., Mishra, A., A novel hybrid machine learning based system to classify shoulder implant manufacturers, *Healthcare*, (2022), 10, 580, https://doi.org/10.3390/healthcare10030580.

[11] Çolak, M., Sivri, T. T., Akman, N. P., Berkol, A., Ekici,Y., A study of disease prediction on weighted symptom data using deep learning and machine Llearning algorithms, *Int'l Conf. on Theor. & Appl. Comput. Sci. & Eng.*, (2022), 116-119, https://doi.org/10.1109/ICTACSE50438.2022.10009857.

[12] Xie, S., Yu, Z., Lv, Z., Multi-disease prediction based on deep learning: A survey, *Comput. Mod. in Eng. & Sci.*, 128 (2) (2021), 489-522, https://doi.org/10.32604/cmes.2021.016728.

[13] Ahsan, M., Siddique, Z., Machine learning-based heart disease diagnosis: A systematic literature review, *Artif. Intel. in Med.*, 128 (2022), https://doi.org/10.1016/j.artmed.2022.102289.

[14] Disease-Symptom Knowledge Database, (2022). Available: https://people.dbmi.columbia.edu/. [Accessed: September 2022].

[15] Gandhi, K. , Mittal, M., Gupta, N. and Dhall, S., Disease prediction using machine learning, *Int'l J. for Res. in Appl. Sci. & Eng. Tech.*, 8 (2020), 500-507, http://doi.org/10.22214/ijraset.2020.6077.

[16] Agrawal, A., Agrawal, H., Shivam, M., Sharma, M., Disease prediction using machine learning, *Proc. of 3rd Int. Conf. on IoT & Connected Tech.*, (2018), http://dx.doi.org/10.2139/ssrn.3167431.

[17] Vinitha, S., Sweetlin, S., Vinusha, H., Sajini, S., Disease prediction using machine learning over big data, *Comput. Sci. & Eng.: An Int'l J.*, 8 (1) (2018), http://dx.doi.org/10.2139/ssrn.3458775.

[18] Kumar, A., Sharma, G. K. and Prakash, U. M. , Disease prediction and doctor recommendation system using machine learning approaches, *Int'l J. for Res. in Appl. Sci. & Eng. Tech.*, 9 (2021), 34-44, https://doi.org/10.22214/ijraset.2021.36234.

[19] Mallela, R. C., Bhavani., R. L., Ankayarkanni, B., Disease prediction using machine learning techniques, *2021 5th Int. Conf. on Trends in Electronics & Informatics*, (2021), 962-966, https://doi.org/10.1109/ICOEI51242.2021.9453078.

[20] Grampurohit, S., Sagarnal, C., Disease prediction using machine learning algorithms, *2020 Int. Conf. for Emerging Tech.*, (2020), 1-7, https://doi.org/10.1109/INCET49848.2020.9154130.

[21] Dhabarde, S., Mahajan, R., Mishra, S., Chaudhari, S., Manelu, S., Shelke, N. S., Disease prediction using machine learning algorithms, *Int. Res. J. of Mod. in Eng. Tech. & Sci.*,4 (2022), 379-384.

[22] Alanazi, R., Identification and prediction of chronic diseases using machine learning approach, *J. of Healthc. Eng.*, 2022 (2022), 1-9, https://doi.org/10.1155/2022/2826127.

[23] Uddin, S., Haque, I., Lu, H., Moni, M. A., Gide, E., Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction, *Sci. Rep.*, 12 (2022), 1-11, https://doi.org/10.1038/s41598-022-10358-x.

[24] Joel, G. N., Priya, S. M., Improved ant colony on feature selection and weighted ensemble to neural network based multimodal disease risk prediction (WENN-MDRP) classifier for disease prediction over big data, *Int. J. of Eng. & Tech.*, 7 (2018), 56-61, https://doi.org/10.14419/ijet.v7i3.27.17654.

[25] Ibrahim, F., Taib, M. N., Abas, W. A., Guan, C. C., Sulaiman, S., A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), *Comput. Methods Programs Biomed.*, 79 (3) (2005), 273-281, https://doi.org/10.1016/j.cmpb.2005.04.002.

[26] Venkatesh, K., Dhyanesh, K., Prathyusha, M. and Teja, C. H. N. , Identification of disease prediction based on symptoms using machine learning, *JAC: J. Comp. Theory*, 14 (2021), 86-93, https://doi.org/10.1155/2022/2826127.

[27] Chauhan, R. H., Naik, D. N., Halpati, R. A., Patel, S. J. and Prajapati, A. D., Disease prediction using machine learning, *Int. Res. J. of Eng. & Tech.*, 7 (2020), 2000-2002.

[28] Maram, B., Kumar, K. S. and Gampala, V., Symptoms based disease prediction using bigdata analytics, *Turk. J. of Phys. Ther. & Rehab.*, 32 (3) (2021), 3228-3234.

[29] Ferjani, M., Disease prediction using machine learning, *Bournemouth Univ.*, (2020), http://dx.doi.org/10.13140/RG.2.2.18279.47521.

[30] Awari, S. V., Diseases prediction model using machine learning technique, *Int. J. of Sci. Res. in Sci. & Tech.*, 8 (2) (2021), 461-467, https://doi.org/10.32628/IJSRST.

[31] Shirsath, S. S., Patil, S., Disease prediction using machine learning over big data, *Int. J. of Innov. Res. in Sci. Eng. & Tech.*, 7 (6) (2018), 6752-6757, https://doi.org/10.15680/IJIRSET.2018.0706059.

[32] Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M. and Mehendale, N., Disease prediction from various symptoms using machine learning, *SSRN Electron. J.*, (2020), https://dx.doi.org/10.2139/ssrn.3661426.

[33] Dahiwade, D., Patle, G., Meshram, E., Designing disease prediction model using machine learning approach, *3rd Int. Conf. on Comput. Methodol. & Comm.*, (2019), 1211-1215, https://doi.org/10.1109/ICCMC.2019.8819782.

[34] Ke, G., Meng, Q., Finley, T., Wang, T. , Chen, W., Ma, W., Ye, Q., Liu, T., LightGBM: A highly efficient gradient boosting decision tree, *Adv. in Neural Inf. Process. Syst.*, (30) (2017), 3146–3154.

[35] Kuhn, M., Kjell, J., Applied Predictive Modeling, Springer, New York, 2013.

[36] He, H., Ma, Y., Imbalanced Learning: Foundations, Algorithms, and Applications, Willey, 2013, 188, http://dx.doi.org/10.1002/9781118646106.