Screening of Tuberculosis (TB) and Coronavirus (COVID-19) using Machine Learning

Nurliyana Suhaimi[1], Nor Azimah Khalid[1*] and Marshima Mohd Rosli[1,2]

[1] College of Computing, Informatics & Mathematics, Universiti Teknologi MARA, 40450, Selangor, Malaysia
[2] Institute for Pathology, Laboratory and Forensic Medicine, University Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia
azimahkhalid@uitm.edu.my

Abstract.
Tuberculosis (TB) affects more than 10 million individuals annually, making it a significant global health issue. Due to the emergence of the COVID-19 pandemic, TB services in numerous countries have experienced temporary interruptions, resulting in a possible delay in the detection of TB patients and a significant number of cases going unnoticed. Because of the similarities in symptoms and their impact on the respiratory system, there is a potential for misdiagnosis of both disorders. Misdiagnosis or delayed diagnosis might lead to severe consequences, such as the spread of the diseases and postponed medical intervention. This study attempts to evaluate the performance of three machine learning techniques (Random Forest, Naïve Bayes and XGBoost) to determine whether individuals may be identified as potentially having either TB or COVID-19 based on their symptoms. The study used three measurements: accuracy, mean squared error and root mean squared error. The results indicate the effectiveness of machine learning in differentiating between TB and COVID-19, with XGBoost achieving the highest accuracy of 74.99% compared with Naïve Bayes (65.13%) and Random Forest (70.93%). The study also conducted experiments using feature selection methods to identify the common important symptoms for predicting both TB and COVID-19. The findings indicated that four symptoms are significant. Overall, the effectiveness of diverse machine learning techniques in predicting TB and COVID-19 using electronic health records suggests that machine learning can be effectively employed to determine appropriate therapy and efficient triage.

Keywords: disease prediction, TB; COVID-19; machine learning

1 Introduction

Tuberculosis (TB) and Coronavirus (COVID-19) are both respiratory illnesses that have posed significant challenges for healthcare systems worldwide. Both diseases present comparable symptoms, such as cough, fever, and dyspnea, which create difficulty for healthcare providers and the general public in distinguishing between them [1]. Tuberculosis (TB), caused by the bacteria Mycobacterium tuberculosis, mostly affects the

N. A. S. Abdullah et al. (eds.), *Proceedings of the International Conference on Innovation & Entrepreneurship in Computing, Engineering & Science Education (InvENT 2024)*, Advances in Computer Science Research 117,
https://doi.org/10.2991/978-94-6463-589-8_39

lungs but can also spread to other organs. COVID-19, which originates from the SARS-CoV-2 virus, primarily impacts the respiratory system and can lead to severe outcomes, particularly in individuals with pre-existing medical conditions.

The similarity in symptoms between tuberculosis (TB) and COVID-19 presents a significant challenge in making a diagnosis. Inaccurate or delayed diagnosis can lead to inappropriate treatment, increased spreading of diseases, and elevated mortality rates. The simultaneous presence of both COVID-19 and TB in an individual makes the treatment and control of these diseases more challenging and raises the likelihood of adverse outcomes, particularly in older individuals with pre-existing health conditions. According to Kim et al. [1], individuals with tuberculosis (TB), especially those who are 65 years old and beyond, seem to have greater rates of mortality from COVID-19. Furthermore, those who are advanced in age and those who have additional medical illnesses such as diabetes, cancer, cardiovascular disease, and chronic bronchopulmonary disorders are more susceptible to experiencing severe cases of COVID-19 [2].

In areas with restricted resources, where there is limited availability of modern diagnostic tools and healthcare personnel, the situation is worsened. Traditional diagnostic methods, such as chest X-rays and laboratory tests, can be time-consuming and may not be readily available in remote places [2], [3], [4].

Machine learning algorithms have demonstrated significant potential in analysing complex datasets and making precise predictions. By utilising these qualities, it is feasible to create a tool that can aid healthcare practitioners and individuals in differentiating between tuberculosis (TB) and COVID-19, thereby enabling prompt and suitable medical intervention. Furthermore, the worldwide consequences of the COVID-19 pandemic and the persistent prevalence of tuberculosis in various regions emphasise the immediate need for these developments. Machine learning comes in a variety of forms, including XGBoost, Random Forest (RF), and Naive Bayes (NB). These algorithms are popular machine learning algorithms used for classification problems, but they differ in their approach to making predictions. Substantial efforts have also been made in utilizing various artificial intelligence–based algorithms to address the limitations of human reading of chest radiographs for diagnosing TB [5] and COVID-19 [6]. The findings show promising and higher accuracy of observations with the Artificial Intelligence (AI) - assisted methodology.

This study discusses the performance of machine learning techniques that can effectively differentiate between tuberculosis (TB) and COVID-19 using symptoms. The study involves the collection and integration of precise data on symptoms of both TB and COVID-19. The study applied three feature selection methods including the filter method (Pearson Correlation Coefficient), wrapper method (Recursive Feature Elimination), and embedded method (Random Forest Importance). By identifying key symptoms that differentiate TB from COVID-19, the study implements the machine learning algorithms (Random Forest, Naive Bayes, and XGBoost) using these selected features. The experimental results suggest that these algorithms have a robust ability to predict the presence of symptoms related to TB or COVID-19 for initial triage.

2 Related work

Current evidence indicates that the primary transmission route for both COVID-19 and TB is through respiratory droplets, with the lungs being their primary target. This can result in more severe outcomes for patients with coinfection of COVID-19 and TB (referred to as COVID-TB). Therefore, due to the high prevalence of both of these infectious diseases and the potential for worse prognosis which coinfection, an intensive investigation of COVID-TB cases may be of great clinical significance [7]. This is supported by reviews conducted by [8] using meta-analysis that recommends routine screening for TB among confirmed cases of COVID-19 in countries with high TB. The TB bacteria may also harm other parts of the body, including the kidney, spine, and brain. Even though TB can affect other body parts, some people with TB germs did not become ill [9]. The majority of cases are now treated with antibiotics. To deactivate the germs, the drug must be taken for a lengthy period of time—at least 6 months [10].

According to Koegelenberg et al, COVID-19's indirect effects contributed to an additional 400,000 TB deaths in 2020 [9]. The cases in Asia are 55% and Africa is 30% make them the largest percentages of the TB cases [5]. Typical symptoms of active TB include a cough lasting more than 3 weeks, chest pain, coughing up the blood, feeling tired all the time, night sweats, fever, loss of appetite, weight loss and insomnia [11]. However, many of the signs and symptoms of COVID-19 can also be caused by other disorders, and some persons with COVID-19 have no symptoms at all, making a diagnosis by physical examination challenging [12], [13].

The similarity of the symptoms makes it difficult to identify these two diseases. However, it can be seen that these two diseases still have some distinct symptoms from one another. In addition, the conventional experts' readings have substantial within and between observer variability, indicating poor reliability of human readers [5].

Althenayan et al. conducted a review on deep learning methods to detect COVID-19 [6] while substantial efforts have also been made in utilizing various artificial intelligence–based algorithms to address the limitations of human reading of chest radiographs for diagnosing TB [5]. The findings show promising and higher accuracy of observations with the Artificial Intelligence (AI) - assisted methodology.

Feature selection is crucial for interpretation and prediction, simplifying classification oricess by removing unnesessary [14]. The methods include filter, embedded, and wrapper. Filter methods rank features by relevance using an ordering approach. Important features give useful information about the different classes [16], [15], identified through a correlation matrix using the Pearson Correlation Coefficient (PCC). This method selects features with the highest correlation to the target for the prediction [17].

Wrapper methods analyse training and testing performance iteratively, removing unsignificant features. Embedded methods perform feature selection and algorithm training concurrently [17]. Recursive Feature Elimination (RFE) recursively removes the weakest features until the target number of features is reached [15].

Machine learning comes in a variety of algorithms, including XGBoost. RF, NB, and XGBoost are popular machine learning algorithms used for classification problems, but they differ in their approach to making predictions and their strengths and weaknesses. RF is the most well-known classification algorithm that can do both classification and regression. To find the top features, the Random Forest Importance approach can be used. While regression utilises the mean of all the outputs from each decision tree, classification employs a voting method to determine the class [18].

The XGBoost technique for gradient boosting, which was developed primarily for stimulating machine learning model performance and computational speed, is extremely accurate, scalable, and extends the processing capabilities of boosted tree algorithms. It also performs well on medium-sized, small-scale, subgroup-based, structured datasets with a moderate number of features [19]. The Naive Bayes (NB) algorithm computes the probabilities of an item being assigned to a certain class by considering its constituent elements. Although NB is beneficial in practical scenarios, it has limitations, such as its need for a minimal amount of data for parameter estimate. Furthermore, it solely considers the variance of each class, disregarding the covariance matrices [20].

3 Methodology

This study constructed an analysis workflow model (see Figure 1) that encompasses four phases: data preprocessing, feature selection, model training and performance measurement for predicting TB or COVID-19 based on symptoms. In the Data Preprocessing phase, the dataset was processed through three methods which are data cleaning, data balancing and data scaling. In the Feature Selection phase, the datasets were used for selecting relevant symptoms using three methods known as Pearson Correlation Coefficient, Recursive Feature Elimination and Random Forest Importance. During the Model Training phase, the dataset was partitioned into training and testing sets, with a ratio of 70:30. The training data was utilised in three distinct models, namely RF, NB, and XGBoost. The effectiveness of each model was evaluated using essential measurements such as accuracy, sensitivity, and specificity.

3.1 Data pre-processing

For this analysis, two unprocessed datasets from Kaggle comprised symptoms of individuals diagnosed with tuberculosis (TB) and COVID-19. Significant symptoms such as cough, fever, difficulty breathing, and other distinct markers for each disease were extracted. The data from these two sources were consolidated into a single dataset of 2759 records and 7 features. Missing values have been identified and removed from the dataset. The clean dataset represents 1759 instances of the "Coronavirus" class and 1000 instances of the "Tuberculosis" class that indicate an imbalance issue. This study utilised an oversampling method (SMOTE) to rebalance the dataset, whereby the minority class is augmented by increasing its instances until a balanced representation is achieved. The StandardScaler method is used to normalise and standardise each feature or variable in the dataset. This method removes the mean and scales it to have unit variance. The

StandardScaler applies to the input data, ensuring zero mean and unit variance, and producing the standardized output.

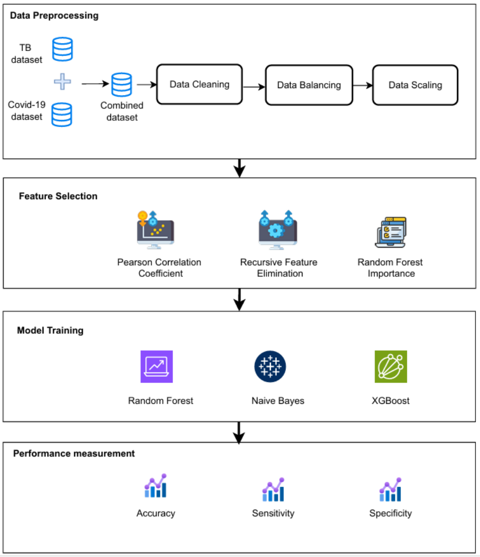


Figure 1. Analysis workflow model

3.2 Feature Selection

Feature selection is a significant step in determining the most relevant symptoms that facilitate differentiating between TB and COVID-19. Three feature selection techniques were compared in this study: (a) Pearson Correlation Coefficient (PCC), (b) Recursive Feature Elimination (RFE) and (c) Random Forest Importance. Figure 2 presents a comparison of the three methods used for feature selection. The maximum level of precision in this comparison is 76.63%. Nevertheless, both the RFE and RF approaches exhibit the same levels of accuracy. Consequently, this study used the Recursive Feature Elimination (RFE) for selecting the most relevant features. By employing RFE, the machine learning models were trained using only the most important features, resulting in enhanced model performance and interpretability.

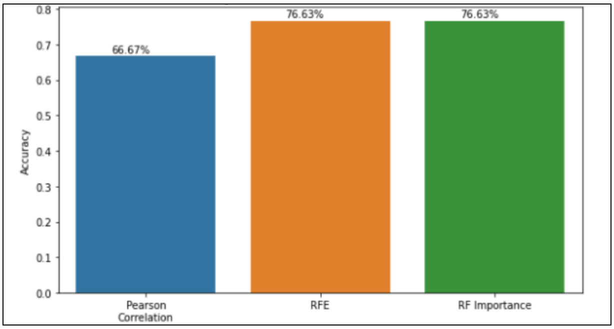


Figure 2. Comparison of feature selection method

3.3 Model Training

This study selected three machine learning algorithms: Random Forest, Naive Bayes, and XGBoost. These algorithms were chosen for their robustness and ability to handle classification problems effectively as described below.

Random Forest: This ensemble learning method constructs multiple decision trees during training and outputs the mode of the classes for classification. Random Forest is known for its high accuracy and ability to handle large datasets with numerous features.

Naive Bayes: This probabilistic classifier applies Bayes' theorem with strong (naive) independence assumptions between the features. Despite its simplicity, Naive Bayes performs well in many real-world situations.

XGBoost: An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. XGBoost uses a gradient boosting framework to solve data science problems accurately and at a scale.

The models were trained using the selected features based on RFE, and their performance was evaluated using metrics such as accuracy, precision, recall, and F1 score. Hyperparameter tuning was conducted using grid search to optimize the model parameters and enhance their performance.

3.4 Performance Measurement

This study evaluated the performance of machine learning techniques using evaluation metrics: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The mean squared error is obtained by averaging the squared differences between the original and anticipated values in the data set. The variance of the residuals is measured. The square root of Mean Squared Error is referred to as Root Mean Squared Error. The standard deviation of the residuals is established. We calculated the MSE and RMSE values using Eq. (1) and Eq. (2).

The formula for Mean Squared Error (MSE) is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

The formula for Roor Mean Squared Error is as follows:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

4 Results and Discussion

The study successfully developed and evaluated machine learning models to differentiate TB and COVID-19 based on user-reported symptoms. The RF and XGBoost models were optimised using a parameter grid. The parameter grid included *n_estimators*, which represents the number of trees, and *max_depth*, which determines the maximum depth of each decision tree. In contrast, the NB model was optimized using *var_smoothing*, which controls the variance of all features. Once the optimization was complete, we employed grid search cross-validation for the training process. This involved using *cv=5*, indicating the number of folds to be used in cross-validation, *n_jobs=-1* to enable parallel processing, and *verbose=2* to control the level of verbosity during the grid search process. Table 1 shows the results of the three machine learning models using hyperparameters.

Table 1. Accuracy using hyperparameter.

Model	**Accuracy of hyperparameter**		**Best**	
	Before (%)	After (%)	Parameter	Score

Random Forest	76.63	76.63	Max_depth =3
			Min_samples_split = 2
			N_estimators = 100
				0.7444

Naïve Bayes	63.95	69.38	Var_smoothing = 0.8	0.6593

XGBoost	76.63	76.63	Max_depth=3
			Learning_rate=0.1
			N_estimators = 100
				0.7444

As shown in Table 1, both Random Forest and XGBoost have an accuracy of 76.63% for hyperparameter tuning. However, this study applied cross-validation during the model training to prevent overfitting. The model was trained using cross validation with five folds to compare the three machine learning models. Figure 3 shows a comparison of the machine learning models.

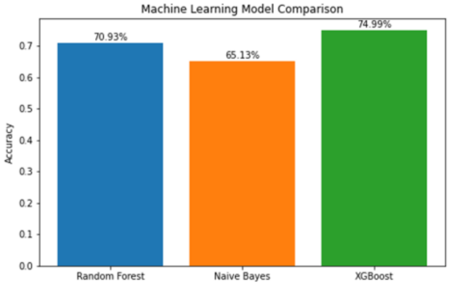
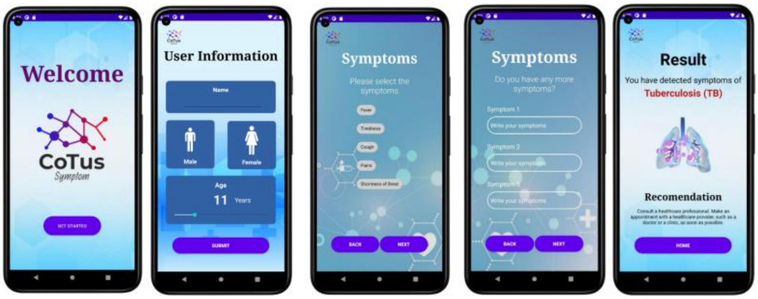


Figure 3: Comparison of machine learning models.

Based on the results shown in Figure 3, XGBoost demonstrated the greatest accuracy rate of 74.99%. The Random Forest achieved an accuracy of 70.93%, while the Naïve Bayes model had the lowest accuracy of 65.13%. The study selected the XGBoost as the optimal model for predicting TB and COVID-19 due to its optimised gradient boosting framework, which improves prediction accuracy and processing efficiency [21]. The findings demonstrate that XGBoost is exceptionally proficient in managing structured datasets of medium size with a moderate number of features.

The models were evaluated using metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). XGBoost exhibited superior performance, achieving an MSE of 0.23 and an RMSE of 0.48 indicating that the predicted values closely matched the actual outcomes. In contrast, predictions from other machine learning models, such Random Forest and Naïve Bayes, were less accurate. These results are consistent with previous research, which shows that XGBoost performs well in a range of healthcare applications, including medical decision support systems [22] and illness risk prediction [23].

This study developed a prototype system to simulate the prediction of Tuberculosis (TB) and Coronavirus (Covid-19) using XGBoost, chosen for its MSE and RMSE driven results. Figure 4 shows the user interface of the system. First, the prototype system requires the user to input information such as name, gender and age. Next, the prototype system requires the user to select symptoms based on the list of symptoms and input additional symptoms if necessary. Finally, the prototype system displays the prediction results and recommendations.



You have detected symptoms of Tuberculosis (TB)

Consult a healthcare professional. Make an appointment with a healthcare provider, such as a doctor or a clinic, as soon as possible.

Fig.4 . A prototype of Tuberculosis (TB) and Coronavirus (Covid-19) screening application.

The study's findings have important consequences for healthcare, particularly in countries with limited resources. Conventional diagnostic screening methods for TB and Covid-19 might be time-consuming and may not always be accessible in remote regions. Machine learning-based screening algorithms provide a practical option, allowing for timely and precise diagnosis using easily accessible data. Healthcare providers can make well-informed decisions about patient management and therapy by accurately distinguishing between TB and COVID-19. This can aid in reducing the transmission of infectious diseases and ensuring prompt medical care, thereby enhancing patient outcomes and public health.

5 Conclusion

This study highlights the significant capacity of machine learning algorithms to distinguish between Tuberculosis (TB) and COVID-19 based on reported symptoms. The research highlighted the effectiveness of algorithms including Random Forest, Naive Bayes, and XGBoost in effectively analysing intricate information and generating dependable diagnostic outcomes. XGBoost proved to be the most precise model, attaining the greatest performance metrics out of all the analysed algorithms. The significance of feature selection strategies was also highlighted, demonstrating their ability to improve model performance by identifying the most pertinent symptoms. This not only streamlines the categorization procedure but also enhances the precision of predictions.

The results of this study are especially significant for healthcare systems in areas with limited resources, where there may be a shortage of conventional diagnostic instruments. Using machine learning algorithms for screening can assist in the timely and accurate detection of tuberculosis (TB) and COVID-19, facilitating appropriate medical interventions and minimising the spread of the diseases. Moreover, this study establishes the foundation for future progress in using artificial intelligence in medical diagnostics, with the possibility of expanding to other contagious illnesses. In general, incorporating sophisticated machine learning models can greatly improve the treatment of diseases and the consequences for public health.

Acknowledgments. The authors would like to thank the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Selangor, Malaysia for their financial support to this research project.

References

[1] B. Kim, Y. A. Kang, and J. Lee, “Heterogeneous impact of Covid-19 response on tuberculosis burden by age group,” Scientific Report 12(1), 13773, (2022)

[2] A. Starshinova, L. Guglielmetti, O. Rzhepishevska, O. Ekaterincheva, Y. Zinchenko, and D. Kudlay, “Diagnostics and management of tuberculosis and COVID-19 in a patient with pneumothorax (clinical case),” Journal Clinical Tuberculosis Other Mycobacterial Disease 24, 100259, (2021)

[3] G. B. Migliori et al., “Tuberculosis and COVID-19 co-infection: description of the global cohort,” European Respiratory Journal 59(3), (2022)

[4] M. Ruhwald, S. Carmona, and M. Pai, “Learning from COVID-19 to reimagine tuberculosis diagnosis,” The Lancet Microbe 2(5), e169-e170, (2021)

[5] S. Hansun, A. Argha, S. T. Liaw, B. G. Celler, and G. B. Marks, “Machine and Deep Learning for Tuberculosis Detection on Chest X-Rays: Systematic Literature Review,” Journal of Medical Internet Research 25, e43154,(2023)

[6] A. S. Althenayan, S. A. AlSalamah, S. Aly, T. Nouh, and A. A. Mirza, “Detection and Classification of COVID-19 by Radiological Imaging Modalities Using Deep Learning Techniques: A Literature Review,” Applied Sciences 12(20), 10535, (2022)

[7] L. A. Callender, M. Curran, S. M. Bates, M. Mairesse, J. Weigandt, and C. J. Betts, “The Impact of Pre-existing Comorbidities and Therapeutic Interventions on COVID-19,” Frontiers in Immunology 11, 1991, (2020)

[8] W. M. Song et al., “COVID-19 and Tuberculosis Coinfection: An Overview of Case Reports/Case Series and Meta-Analysis,” Frontiers in Medicine 8, 657006, (2021)

[9] C. F. N. Koegelenberg, O. D. Schoch, and C. Lange, “Tuberculosis: The Past, the Present and the Future,” Respiration 100 (7), 553-556, (2021)

[10] C. A. Kerantzas and W. R. Jacobs, “Origins of combination therapy for tuberculosis: Lessons for future antimicrobial development and application,” mBio 8(2), 10-1128, (2017)

[11] P. J. Vaidya, M. Munavvar, J. D. Leuppi, A. C. Mehta, and P. N. Chhajed, “Endobronchial ultrasound-guided transbronchial needle aspiration: Safe as it sounds,” Respirology 22(6), 1093-1101, (2017)

[12] S. S. Bhopal, J. Bagaria, B. Olabi, and R. Bhopal, “Children and young people remain at low risk of COVID-19 mortality,” The Lancet Child and Adolescent Health 5(5), e12-e13, (2021)

[13] S. M. H. Israfil et al., “Clinical Characteristics and Diagnostic Challenges of COVID−19: An Update From the Global Perspective,” Frontiers in Public Health 8, 567395, (2021)

[14] K. Touchanti, I. Ezzazi, M. El Bekkali, and S. Maser, “A 2-stages feature selection framework for colon cancer classification using SVM,” in Inter. Conf. on Intelligent Systems & Computer Vision, (2022)

[15] P. S. Nandhini, S. Kuppuswami, M. Harish, S. Gomanishwaran, and S. Bharani, “A Comparison on Feature Selection Methods using Machine Learning Algorithms for improving the Performance Parameters of RPL-BASED IoT Attacks Classification,” in Inter. Conf. on Inventive Research in Computing Applications, (2022)

[16] Y. H. Pullissery and A. Starkey, “Application of Feature Selection Methods for Improving Classifcation Accuracy and Run-Time: A Comparison of Performance on Real-World Datasets,” in Inter. Conf. on Applied Artificial Intelligence and Computing, (2023)

[17] A. Agaal and M. Essgaer, “Influence of Feature Selection Methods on Breast Cancer Early Prediction Phase using Classification and Regression Tree,” in Inter. Conf. on Engineering and MIS, (2022)

[18] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, “Heart disease prediction using machine learning techniques: A survey,” International Journal of Engineering and Technology (UAE) 7(2.8), 684-687, (2018)

[19] Z. Xu and Z. Wang, “A Risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier,” in Inter. Conf. on Advanced Computational Intelligence, (2019)

[20] A. Aldahiri, B. Alrashed, and W. Hussain, “Trends in Using IoT with Machine Learning in Health Prediction System,” Forecasting 3(1), (2021)

[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Inter. Conf. on Knowledge Discovery and Data Mining, pp. 785–794, (2016)

[22] X. Zhang, Y. Wang, and H. Jin, "Predicting the risk of diabetes using structured patient data with XGBoost model," Journal of Medical Systems 43(4), (2019)

[23] Habehh H, Gohel S. Machine Learning in Healthcare. Curr Genomics 22(4),291-300, (2021)