Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/130308

Prediction and classification of Tuberculosis using machine learning

Azhari A. Elhag

Department of Mathematics and Statistics, College of Science, P.O. Box 11099, Taif University, Taif 21944, Saudi Arabia

Received: 17 Aug. 2023, Revised: 20 Dec. 2023, Accepted: 7 Jan. 2024.
Published online: 1 May 2024

Abstract: Tuberculosis can be fatal and is an infectious disease if it is not treated, that primarily affects the lungs but can also affect other organs in the body. According to the World Health Organization (WHO), tuberculosis is second only to the Covid-19 in terms of the number of deaths it causes and is the thirteenth largest cause of death overall. As a result, it is required to construct predictive models for the incidence and classification of tuberculosis, which aid in identifying the groups and places in which tuberculosis spreads and monitoring the various trends and patterns of tuberculosis. Developing these models is necessary because they help in identifying the groups and locations in which tuberculosis is spread. Artificial network models and decision tree models were used to predict and classify tuberculosis cases in the United States of America using tuberculosis cases data. The results showed that the decision tree model (DT) is more accurate than the artificial neural network (ANN).

Keywords: Decision tree model, Artificial neural network, predictive models, Trends, and patterns.

1 Introduction

Tuberculosis is a highly consequential communicable illness mostly impacting the respiratory system [1]. The etiology of this condition is attributed to a bacterial pathogen known as Mycobacterium tuberculosis. Tuberculosis (TB) can be transmitted via airborne transmission when an individual with active TB expels respiratory droplets through coughing, sneezing, or speaking [2]. Tuberculosis, a prevalent worldwide health concern, can be effectively prevented and treated with the administration of antibiotics. In accordance with the World Health Organization (WHO), tuberculosis claimed the lives of 1.4 million individuals in the year 2020, while an estimated 10 million individuals were afflicted with the disease [3]. The number of recorded tuberculosis (TB) cases in the United States increased from 7,874 in 2021 to 8,300 in 2022[4]. The incidence of tuberculosis (TB) also shown a minor increase in the year 2022, with a rate of 2.5 cases per 100,000 individuals. The incidence of tuberculosis (TB) cases and reported TB cases in the United States are reverting to levels observed prior to the COVID-19 pandemic. This trend follows a significant decrease in 2020, which can be attributed to many causes connected with the COVID-19 pandemic, such as the occurrence of missed or delayed diagnosis [5]. The relationship between HIV/AIDS and tuberculosis is complex as the presence of one condition can significantly increase the risk and severity of the other. The incidence of tuberculosis has seen an upward trend in conjunction with the heightened prevalence of human immunodeficiency virus (HIV). The prevalence of tuberculosis (TB) infection is estimated to affect around one-third of the global population, while a significant proportion of individuals remain asymptomatic. The diagnosis of tuberculosis (TB) patients may be compromised due to the complex interaction between TB and co-infections, which can result in atypical reactions and resemble the symptoms of TB. Individuals infected with the human immunodeficiency virus (HIV) or acquired immunodeficiency syndrome (AIDS) exhibit a significantly heightened susceptibility, around tenfold, to tuberculosis (TB) in comparison to those who are not afflicted with the aforementioned viral infection. The primary etiology of this condition is the extreme immunodeficiency resulting from HIV infection. Tuberculosis (TB) has the potential to manifest at any stage of HIV infection, however the probability of acquiring TB tends to rise as the HIV illness advances [6]. An artificial neural network is a type of information processing system that exhibits unique skills similar to those shown in biological neural networks [7]. Artificial neural networks can be used as a diagnostic tool for tuberculosis prediction and can help to enhance the role of computer technology in diagnostics for speedy tuberculosis management [8]. A novel blood testing device including artificial intelligence (AI) and nanotechnology has been developed by researchers affiliated with Tulane University, with the aim of improving the diagnostic procedure for pediatric tuberculosis [9]. Artificial intelligence (AI), particularly deep learning, has the potential to be utilized in the diagnosis and management of tuberculosis (TB) in pediatric patients. Additionally, it can be employed in chest imaging to offer computer-assisted diagnosis, hence enhancing workflow efficiency and screening endeavors. AI models have been constructed utilizing medical imaging and genetic data for the purpose of detecting pulmonary tuberculosis, differentiating the infection from other pulmonary diseases, and determining medication resistance in tuberculosis. It is expected to be employed more in biological systems in the next years. The use of artificial neural networks is becoming more common as neural network technology advances, and their application sectors are expanding. The key feature of ANN is self-learning, which occurs without prior understanding of the complicated non-linear correlations that exist between input and

[*] Corresponding author e-mail: a.alhag@tu.edu.sa; azharielhagn@gmail.com

output variables [10]. The treatment duration for drug-resistant tuberculosis necessitates an extended period of therapy, ranging from nine to twenty-four months, transitioning from a recently formulated short-term regimen (STR) to a long-term regimen (LTR). This is necessary due to the challenging nature of treating the disease. Second-line anti tuberculosis medicines (SLDs) are commonly employed as a reserve therapy [11] for drug-resistant tuberculosis (DR-TB). When comparing first-line anti tuberculosis pharmaceuticals (FLDs) to second-line anti tuberculosis drugs (SLDs), it is evident that SLDs pose greater challenges in terms of accessibility, affordability, and safety. Nevertheless, when considering the negative impacts of drug usage, it has been established that the clinical results of SLDs are not satisfactory. Several studies have demonstrated that the percentages of treatment completion ranged from 60 to 70 percent [12]. We successfully identified people with smear-negative pulmonary tuberculosis (SNTB) using the Classification and Regression Tree (CART) algorithm [13]. The use of CART in patients with SNTB resulted in positive results. According to the present literature, the only study that used the Classification and Regression Trees (CART) methodology to predict tuberculosis (TB) in hospitalized patients was conducted in the United States. The researchers presented a succinct technique that significantly reduced unnecessary resource use by 40%. [14] The contact tuberculin skin testing (TST) model is a statistical methodology employed for the analysis of data obtained from TST conducted on individuals who have been in contact with individuals diagnosed with infectious tuberculosis (TB) disease. The objective is to ascertain the risk factors associated with tuberculosis (TB) infection and disease among individuals who have had contact with TB cases. These risk factors include variables such as age, gender, ethnicity, duration of exposure, closeness to the TB case, and the kind of TB case. [15], CART analysis has been found to be a valuable tool for CART analysis has the capability to address missing values in the predictor variables through the utilization of surrogate splits. These surrogate splits serve as alternative splits that aim to closely replicate the principal split. The user's text is incomplete and does not provide any information. Additionally, CART analysis has the capability to generate rules that are easily comprehensible, as they may be articulated in the form of if-then statements [16].

The present study is organized into the following sections: The introductory piece, as well as succeeding sections, particularly sections two and three, provide a comprehensive explanation of the research methodology utilized in the study, with a strong emphasis on networks and decision trees. The mathematical formulae utilized for calculating the statistical measures employed in model comparison are outlined in the fourth section of the paper. Subsequently, section five presents the acquired results, whilst section six offers an extensive analysis and interpretation of these findings. In summary, the aforementioned points collectively provide evidence in support of the research objective.

2 Multi-layer Perceptron

The Multi-layer Perceptron (MLP) is a neural network architecture that consists of an input layer, an output layer, and one or more hidden layers of several stacked neurons. In the Perceptron model [17], the neuron is required to incorporate an activation function that enforces a threshold, such as the sigmoid function [18]. However, in the case of a Multi-layer Perceptron, the neurons have the flexibility to employ any arbitrary activation function [19]. The diagram presented illustrates the structure and components of the Multilayer Perceptron Network Fig.1.

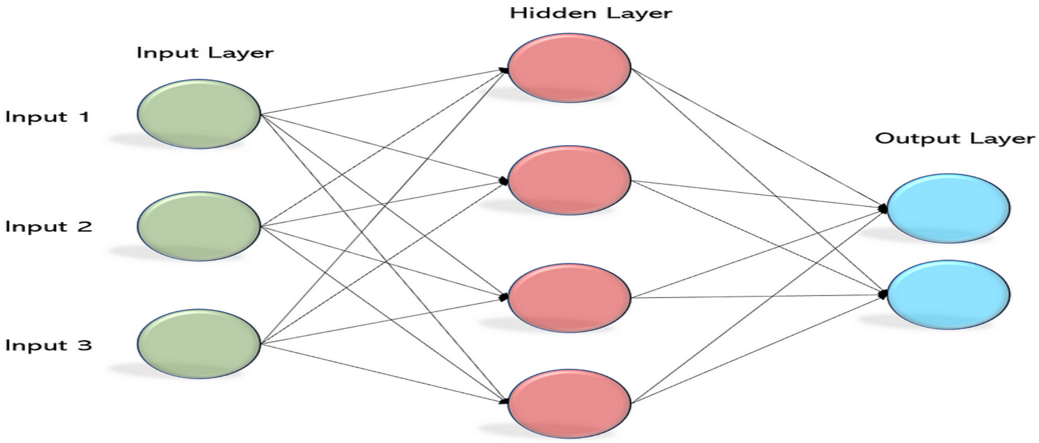


Fig. 1 Multilayer Perceptron Network.

3 Decision tree

A classification and regression modeling tool known as a decision tree is an example of an algorithm that uses supervised learning [20]. Because regression is a technique for carrying out predictive modeling, these trees can either classify the data or make forecasts for what will occur in the future.[21] Decision trees are diagrams that look like flowcharts, and they begin at the root node with a particular data inquiry. In the field of machine learning, decision tree algorithms are utilized for both prediction and classification purposes. With the use of the decision tree and a predetermined collection of inputs, one is able to map the numerous outcomes that are the result of the decisions or consequences [22]. The structure of DT follows a hierarchical tree pattern, wherein the central root node is surrounded on all sides by branches, internal nodes, and leaf nodes. Root nodes are the initial nodes situated at the inception of a decision tree. At this juncture within the tree, the population will commence to undergo segmentation into multiple groups based on the shared features they possess. The nodes that persist in the tree structure subsequent to the severance of the root nodes are commonly denoted as internal nodes. The term "Decision Node" is employed to designate these nodes. the leaf that Possesses Nodal Structures The plant's leaf nodes or terminal nodes are defined as the nodes located on the leaves that lack the ability to undergo further division into additional divisions [23].

4 The Accuracy Measurement

The accuracy of the forecast has been assessed using four statistics metrics: The mean square error (MSE)the root mean square error (RMSE), the symmetric mean absolute percentage error (SMAPE), the mean absolute percentage error (MAPE) [24]. The formulas for MSE, RMSE, SMAPE, and MAPE are as follows [25].

$$mean\ square\ error\ (MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

$$Root\ mean\ square\ error(RMSE) = \sqrt{MSE} \qquad (2)$$

$$Symmetric\ Mean\ Absolute\ percentage\ \ Error$$

$$(SMAPE) = \frac{1}{N}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\frac{(|y_i|+|\hat{y}_i|)}{2}} \qquad (3)$$

Mean Absolute percent error

$$(MAPE) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \qquad (4)$$

5 Results and discussion

To classify and predict the incidence of Tuberculosis based on historical from Centers for Disease Control and Prevention website reports from 1953 to 2021 data in the USA [26]

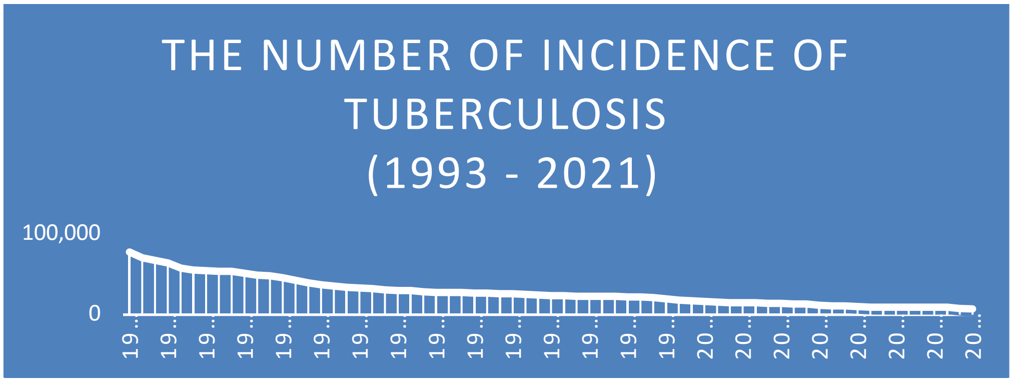


Fig. 2 The graph depicts the time series of the incidence of Tuberculosis (1993 - 2021).

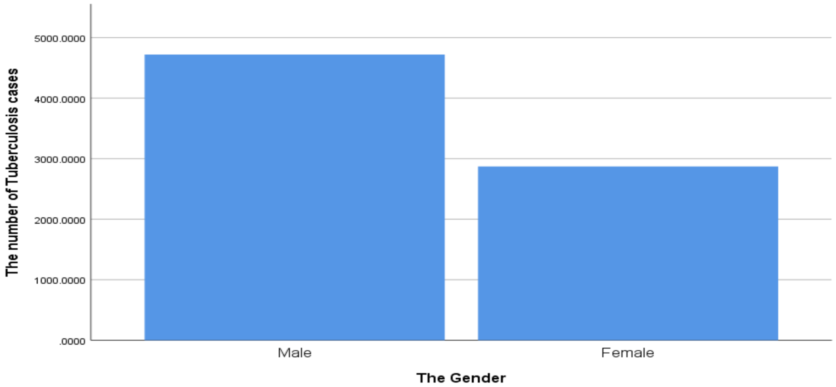


Fig. 3 The number of Tuberculosis cases distributed by

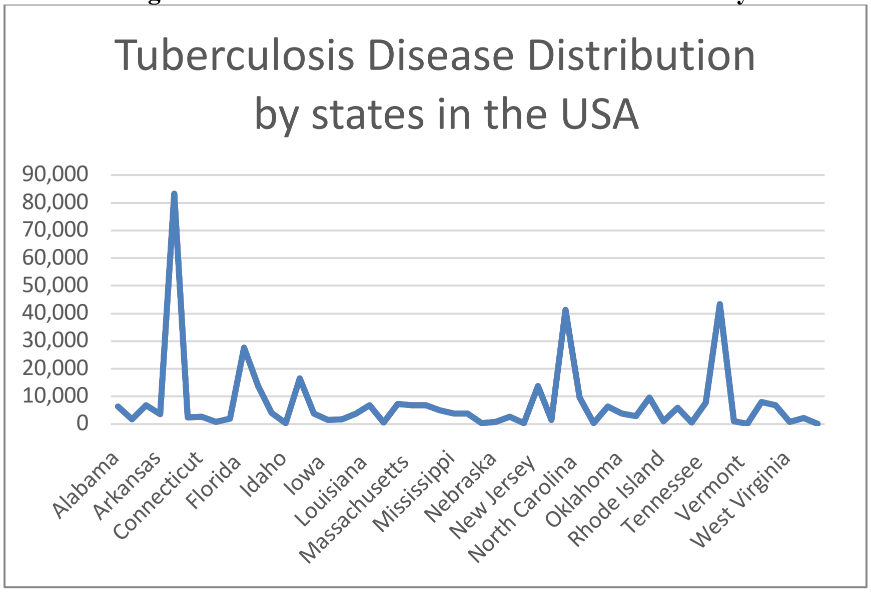


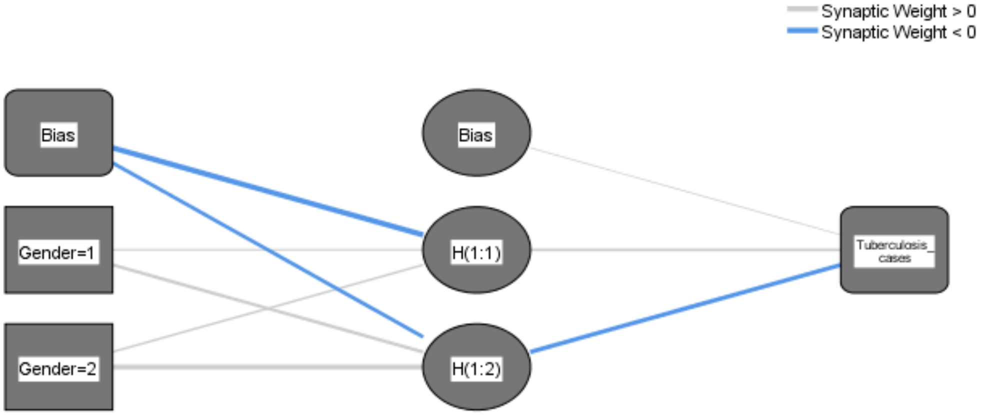
Fig. 4 Tuberculosis cases distribution by states

5.1 Artificial Neural Network (ANN)

The data employed in this investigation, according to the ANN, included a total of 102 records (Table 1), of which 71 (69.6%) were used as training sets and 31 (30.4%) as test sets.

Table 1 Case Processing Summary

		N	Percent
Sample	Training	71	69.6%
	Testing	31	30.4%
Valid		102	100.0%
Excluded		0	
Total		102	



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

Fig. 5 Displays the prediction results using the proposed Neural Network

Units in hidden Layers for the Hidden Layer the activation function is a Hyperbolic tangent, for the output Layer the dependent variable (concentration) used the Sigmoid function and the Sum of Squares are used as the error Function

Table 2 The Model Summary

Model Summary		
Training	Sum of Squares Error	34.765
	Relative Error	0.993
	Stopping Rule Used	Training error ratio criterion (.001) achieved
	Training Time	0:00:00.01
Testing	Sum of Squares Error	13.483
	Relative Error	1.000

Dependent variable: the number of Tuberculosis cases.

5.2 A Classification and Regression Tree (CART)

A Classification and Regression Tree (CRT) is employed as the methodology for growth, where the dependent variable is the count of Tuberculosis cases (102). The training set consists of 54 cases, while the testing set has 48 cases.

Table 3 Gain Summary for Nodes

Sample	Node	N	Percent	Mean
Training	1	29	53.70%	3679.758621
	2	25	46.30%	3046.28
Test	1	22	45.80%	6090.818182
	2	26	54.20%	2701.423077

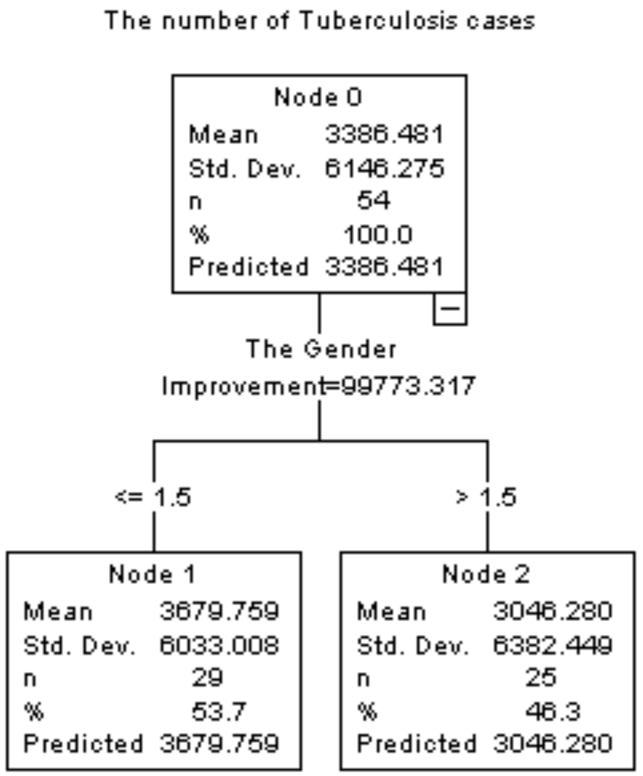
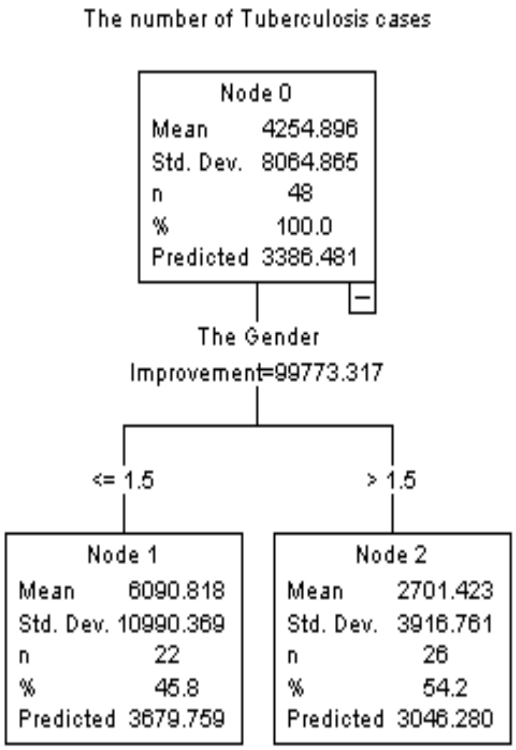
Figure 6 (a) Structure of a Training sample a decision tree with two nodes	Figure 6 (b) Structure of a Testing sample a decision tree with two nodes
 The number of Tuberculosis cases Node 0 Mean 3386.481 Std. Dev. 6146.275 n 54 % 100.0 Predicted 3386.481 The Gender Improvement=99773.317 <= 1.5 > 1.5 Node 1 Mean 3679.759 Std. Dev. 6033.008 n 29 % 53.7 Predicted 3679.759 Node 2 Mean 3046.280 Std. Dev. 6382.449 n 25 % 46.3 Predicted 3046.280	 The number of Tuberculosis cases Node 0 Mean 4254.896 Std. Dev. 8064.865 n 48 % 100.0 Predicted 3386.481 The Gender Improvement=99773.317 <= 1.5 > 1.5 Node 1 Mean 6090.818 Std. Dev. 10990.369 n 22 % 45.8 Predicted 3679.759 Node 2 Mean 2701.423 Std. Dev. 3916.761 n 26 % 54.2 Predicted 3046.280

Fig. 6 shows the that the mean of node 1 is 3679.759 and for node 2 is 3046.280 for the Training. Sample and the mean of node 1 is 6090.818 and for node 2 is 2701.423 for the Testing Sample and the Predicted value 3679.759 for node 1 and 3046.280 for node 2.

Table 4 The statistical metrics

	DT	ANN
$mean\ square\ error\ (MSE)$	0.9487822	0.95798956
$Root\ mean\ square\ error(RMSE)$	0.9740545	0.9787694
$Symmetric\ Mean\ Absolute\ Percetage\ \ Error$	0.209825	0.214907
$Mean\ Absolute\ Percent\ Error$	1.382815129	1.440561368

6 Conclusions

The number of reported tuberculosis cases has gradually decreased over time as a result of high-level political commitment at the global and national levels to achieving these goals, as seen in Figure 2. This information was provided in this paper in the form of a graphic description of tuberculosis cases in the United States. As can be seen in Figure 3, the proportion of males who suffer from this illness is significantly larger than that of females. In addition to this, it was found that California is the state in which the disease is spread to the greatest number of people (refer to Figure 3). Machine learning-based models were utilized in order to build a model that could predict and classify the time series to report new tuberculosis cases in the United States. In order to determine which model had superior performance, a comparison was carried out between the ANN model and the DT model. For this purpose, a number of statistical metrics were employed, including MSE, RMSE, sMAPE, and MAPE. Since a low value for a metric indicates that the model in question is highly accurate, the comparison revealed that the DT model had the best performance of all the models that were examined (see table 4). The findings of this article may provide policymakers with helpful information that can inform vaccine use not only in the United States but also elsewhere in the world. Research tools can also serve as instructional materials for public engagement and be incorporated into a variety of different types of research endeavors.

References

[1] Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2022). A Neural Network-Based Method for Respiratory Sound Analysis and Lung Disease Detection. Applied Sciences, 12(8), 3877.

[2] Alemie, G. A., & Gebreselassie, F. (2014). Common types of tuberculosis and co-infection with HIV at private health institutions in Ethiopia: a cross sectional study. BMC public health, 14(1), 1-5.

[3] Abate, G. (2000). Anti-tuberculosis activity of-lactam antibiotics: prospects for the treatment of multi-drug-resistant tuberculosis. Ethiopian Journal of Health Development, 14(3).

[4] Lestari, V., Mawengkang, H., & Situmorang, Z. (2023). Artificial Neural Network Back-propagation Method to Predict Tuberculosis Cases. Sinkron: jurnal dan penelitian teknik informatika, 8(1), 35-47.

[5] Khan, M. T., Kaushik, A. C., Ji, L., Malik, S. I., Ali, S., & Wei, D. Q. (2019). Artificial neural networks for prediction of tuberculosis disease. Frontiers in microbiology, 10, 395.

[6] Lickona, T. (2009). Educating for character: How our schools can teach respect and responsibility. Bantam.

[7] El-Solh, A. A., Hsiao, C. B., Goodnough, S., Serghani, J., & Grant, B. J. (1999). Predicting active pulmonary tuberculosis using an artificial neural network. Chest, 116(4), 968-973.

[8] Kramer, F., Modilevsky, T., Waliany, A. R., Leedom, J. M., & Barnes, P. F. (1990). Delayed diagnosis of tuberculosis in patients with human immunodeficiency virus infection. The American journal of medicine, 89(4), 451-456.

[9] Alhumaid, S., Al Mutair, A., Al Alawi, Z., Alsuliman, M., Ahmed, G. Y., Rabaan, A. A., ... & Al-Omari, A. (2021). Knowledge of infection prevention and control among healthcare workers and factors influencing compliance: a systematic review. Antimicrobial Resistance & Infection Control, 10(1), 1-32.

[10] Ali, M. H., Khan, D. M., Jamal, K., Ahmad, Z., Manzoor, S., & Khan, Z. (2021). Prediction of multi drug-resistant tuberculosis using machine learning algorithms in swat, Pakistan. Journal of healthcare engineering, 2021.

[11] Aguiar, F. S., Almeida, L. L., Ruffino-Netto, A., Kritski, A. L., Mello, F. C., & Werneck, G. L. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. BMC pulmonary medicine, 12, 1-8.

[12] Preisser, J. S., & Qaqish, B. F. (1996). Deletion diagnostics for generalized estimating equations. Biometrika, 83(3), 551-562.

[13] Goldman, L., Weinberg, M., Weisberg, M., Olshen, R., Cook, E. F., Sargent, R. K., ... & Medical House Staffs at Yale–New Haven Hospital and Brigham and Women's Hospital*. (1982). A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. New England journal of medicine, 307(10), 588-596.

[14] Sauerbrei, W., Madjar, H., & Prömpeler, H. J. (1998). Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. Methods of information in medicine, 37(03), 226-234.

[15] El-Solh, A., Mylotte, J., Sherif, S., Serghani, J., & Grant, B. J. (1997). Validity of a decision tree for predicting active pulmonary tuberculosis. American journal of respiratory and critical care medicine, 155(5), 1711-1716.

[16] Prasitpuriprecha, C., Jantama, S. S., Preeprem, T., Pitakaso, R., Srichok, T., Khonjun, S., ... & Nanthasamroeng, N. (2022). Drug-resistant tuberculosis treatment recommendation, and multi-class tuberculosis detection and classification using ensemble deep learning-based system. Pharmaceuticals, 16(1), 13.

[17] Hrizi, O., Gasmi, K., Ben Ltaifa, I., Alshammari, H., Karamti, H., Krichen, M., ... & Mahmood, M. A. (2022). Tuberculosis disease diagnosis based on an optimized machine learning model. Journal of Healthcare Engineering, 2022.

[18] Abu Al-Haija, Q., Krichen, M., & Abu Elhaija, W. (2022). Machine-learning-based darknet traffic detection system for IoT applications. Electronics, 11(4), 556.

[19] Miller, S., Curran, K., & Lunney, T. (2018, June). Multilayer perceptron neural network for detection of encrypted VPN network traffic. In 2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA) (pp. 1-8). IEEE.

[20] Hamdi, M., Hilali-Jaghdam, I., Elnaim, B. E., & Elhag, A. A. (2023). Forecasting and classification of new cases of COVID 19 before vaccination using decision trees and Gaussian mixture model. Alexandria Engineering Journal, 62, 327-333.

[21] Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. Central European journal of operations research, 26, 135-159.

[22] Povhan, I. F. (2020). Logical recognition tree construction on the basis of a step-to-step elementary attribute selection. Radio Electronics, Computer Science, Control, (2), 95-105.

[23] Povkhan, I. (2020, April). A constrained method of constructing the logic classification trees on the basis of elementary attribute selection. In CMIS (pp. 843-857).

[24] Reda Abonazel, M. (2018). Different estimators for stochastic parameter panel data models with serially correlated errors. Journal of Statistics Applications & Probability, 7(3), 423-434.

[25] Torsen, E., Mwita, P. N., & Mung'atu, J. K. (2018). Nonparametric estimation of the error functional of a location-scale model.

Centers for disease control and prevention web side

[26] Centers for disease control and prevention web side.