Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Negative binomial mixed model neural network for modeling of pulmonary tuberculosis risk factors in West Java provinces

Restu Arisanti[a*], Resa Septiani Pontoh[a], Sri Winarni[a], Yahma Nurhasanah[a], Silvani Dewi Nura Aini[a], Aissa Putri[a], Nabila Dhia Alifa Rahma[a]

[a]*Department of Statistics, Padjadjaran University, Indonesia*

C H R O N I C L E

Article history:
Received: February 20, 2023
Received in revised format: April 12, 2023
Accepted: June 10, 2023
Available online: June 10, 2023

Keywords:
Pulmonary Tuberculosis
Negative Binomial Mixed Model (NBMM)
Feed-Forward Neural Network (FFNN)
Negative Binomial Mixed Model Neural Network (NBMMNN)

A B S T R A C T

Tuberculosis (TB) is still a major public health concern in many regions of the world, including Indonesia's West Java Provinces. Accurate TB risk factor prediction can enhance overall TB control efforts by directing focused therapies. In this study, utilizing a combination of Negative Binomial Mixed Models (NBMMs) and Feed-Forward Neural Networks (FFNNs), we offer a unique method for the predictive modeling of TB risk variables. A variety of sociodemographic, behavioral, and environmental factors that are known to be linked to TB are included in the dataset utilized in this investigation. To correct for overdispersion and include both fixed and random effects in the model, we first fitted an NBMM major problem in epidemiological investigations is modeling count data with overdispersion, and the NBMM component of the model offers a versatile and effective framework for doing so. Following that, we include an FFNN component in the model, which helps us to detect relevant predictive features and alter the model's weights accordingly. Backpropagation methods are used by the FFNN to adjust model parameters and enhance accuracy. The resulting Negative Binomial Mixed Model Neural Network (NBMMNN) model has a high accuracy value of up to 0.944. Our research suggests that the NBMMNN model outperforms conventional models that are frequently used to predict TB risk factors. By contrast to simpler models, the NBMMNN model can capture complicated and nonlinear interactions between predictors and outcomes. Additionally, the inclusion of random variables in the model enables us to take into account potential sources of variability in the data as well as unmeasured confounding. This work emphasizes the opportunity to enhance TB risk prediction and control efforts by integrating NBMMs with FFNNs. In West Java Provinces and other comparable contexts, the NBMMNN model might be a helpful tool for identifying and resolving TB risk factors, guiding targeted interventions, and enhancing overall TB control efforts.

1. Introduction

Tuberculosis is a contagious disease caused by the bacterium Mycobacterium tuberculosis. This bacterium is also known as Acid Fast Bacteria (BTA). Transmission of tuberculosis can be airborne, for example, when a patient coughs or sneezes, allowing droplets of Mycobacterium bacteria to spread to everyone around the patient (Widjanarko et al., 2019). Cough and fever symptoms deter those affected from further examinations, making early detection of tuberculosis difficult. People infected with tuberculosis bacteria have a 5-10% risk of not recovering from tuberculosis. People with HIV, malnutrition, and people with diabetes are at a higher risk of developing tuberculosis.

* Corresponding author.
E-mail address: r.arisanti@unpad.ac.id (R. Arisanti)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.6.007

In 2016, the Sustainable Development Goals (SDGs), which are more comprehensive and have a global focus, took the place of the Millennium Development Goals (MDGs). Because they encompass a broad range of environmental, economic, and social development activities, the SDG targets are seen as being deeply intertwined, interlaced, and inseparable from one another. Given that tuberculosis has complicated linkages with poverty and its associated social and structural causes, it presents an opportunity to closely explore how the SDGs relate to disease management initiatives. The SDGs and the WHO's End TB Strategy share a conceptual and pragmatic vision. Both the End TB Strategy and the SDG goals are conceptually centered on the fundamental elements that affect human health. The overall objectives of the End TB and SDG agendas should be considered as being connected, and the socioeconomic factors that influence tuberculosis are becoming more widely acknowledged as a top priority for programmatic and research efforts. SDG 3 of the End TB Strategy sets an objective of 80% tuberculosis incidence reduction by 2030 (Carter et al., 2018).

According to the WHO, tuberculosis is the 13th leading cause of death globally. In 2020, around 10 million people worldwide will be diagnosed with tuberculosis. About 5000 people die from tuberculosis every day, 98% of them live in developing countries, especially those of working age (Widjanarko et al., 2019). Indonesia is responsible for 8.4% of tuberculosis cases worldwide after China and India and is the third largest country with the highest number of tuberculosis cases. In Indonesia, around 91,000 people die from tuberculosis every year (Sulistyawati & Ramadhan, 2021). In 2019, the estimated number of cases of tuberculosis in Indonesia reached 845,000 people. Of these estimates, 68% were found and treated in 2018. Those who are not found and treated until recovered have a high potential for spreading the disease to others. In addition, in 2019, the World Health Organization (WHO) declared COVID-19, a coronavirus disease, to be a global pandemic. Large, positive, single-stranded RNA viruses called coronaviruses infect both people and various animals (Velavan et al., 2020). The Ministry of Health announced 857 new cases of COVID-19 in Indonesia and 17 new cases in West Java. To reduce the number of active cases, the Governor of West Java asked the public to record guests from other regions, extended the PSBB, and assigned regents and mayors to develop a plan for implementing the New Normal strategy (Pangestika et al., 2020).

West Java is the province with the most tuberculosis cases in Indonesia in 2021, with 85,681 reported cases. Males dominate tuberculosis cases in West Java, with about 55% of all cases, about 47,053 cases. In West Java, there were almost 60,000 cases of TB in 2019, with a prevalence rate of 280 cases per 100,000 people, according to data from the West Java Provincial Health Office. The three regencies/cities of Bogor Regency, Bandung City, and Bandung Regency in West Java Province, which account for between 7% and 13% of the total number of new cases in West Java, have the highest number of tuberculosis cases. Tuberculosis cases are increasing in West Java almost every year despite government and community efforts in prevention and control.

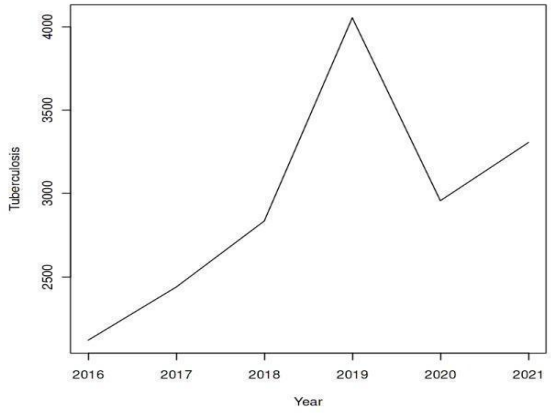

Fig. 1. Annual Average of Tuberculosis Cases in West Java Province, Indonesia

Based on Fig. 1, the graph demonstrates that West Java Province experienced the largest increase in tuberculosis cases, or 4055 cases, in 2019. (Annual average). Despite a 27% decline in cases in 2020, there is still a chance that this trend will continue in 2021 and beyond. The Indonesian government implements tuberculosis-related policies as part of its efforts to lessen the burden of tuberculosis occurrence. The Regulation of the Minister of Health of the Republic of Indonesia No. 67 of 2016 on TB Control sets forth this policy. Health promotion, tuberculosis surveillance, risk factor reduction, identification, and management are all part of the fight against the disease. TB cases offer medications to prevent TB as well as immunity in the form of the BCG vaccine. The Directly Observed Treatment Strategy (DOTS) team Public Health Center collaborates with the TB Program Manager to carry out each of these tasks. As a result, only the implementation of policy determines whether tuberculosis control initiatives are successful or unsuccessful (Notodiputro & Arisanti et al., 2017).

Controlling risk factors is one of the governments of Indonesia's measures to fight tuberculosis cases. In earlier investigations, factors that lead to tuberculosis were noted. The development of tuberculosis is influenced by both individual and

environmental variables. Gender, and nutritional status are individual characteristics that can affect risk for TB. Environmental factors that can affect the occurrence of tuberculosis include population density, sanitary home cleanliness, and proper care services.

Several individuals (cross-sectional units) were repeatedly measured throughout time for this study's longitudinal data collection (Widjanarko *et al*., 2019). Due to the connection between observations within the same unit, the GLMM (Generalized Linear Mixed Model), method is frequently used for longitudinal analysis of data (Suparyanto & Rosad, 2020). A linear predictor is used in the modification of the GLM approach known as GLMM which incorporates fixed effects and random effects (Arisanti *et al*., 2017). In much research, the use of machine learning has also been applied as a GLMM substitute. A statistical modeling method known as generalized linear mixed models (GLMM) may evaluate data whose response variables have a non-normal distribution, such as binary or count data (Arisanti *et al*., 2020). Count data are generally distributed using the Poisson distribution or, in the case of overdispersion, the negative binomial distribution. The negative binomial distribution was used in this study to measure overdispersion. On the other hand, machine learning is a group of computational methods used to find links and patterns in data. Both organized and unstructured data can be used for prediction and classification tasks using machine learning. Machine learning can be used to predict the risk of TB infection, the likelihood that therapy would be effective, and other outcomes in the case of pulmonary tuberculosis.

This study uses the Negative Binomial Mixed Model and Feed Forward Neural Network (FFNN) methods, which are a combination of the generalized linear mixed model (GLMM) and neural network (NN) approaches. Scores such as RMSE are used to measure the accuracy of the prediction results of the FFNN method. In this study, the modeling of tuberculosis case data in West Java is performed by the negative binomial mixed model. Meanwhile, machine learning will be applied to predict the possibility of tuberculosis case infection based on the risk factors identified in the study. Thus, the information gained from this study can be used as a basis for government and societal efforts to improve the quality of tuberculosis prevention and control. This is especially true in West Java province.

2. Literature Review

2.1 Longitudinal Data Modeling

The longitudinal effects of several risk variables on the development of TB were modeled in this study using GLMM. Both fixed and random effects, as well as correlations between repeated observations across time, can be accounted for via GLMM. Furthermore, capable of handling abnormal response variable distributions and missing data, GLMM. Then, investigate the distribution of data using Cullen and Frey.

2.2 Generalized Linear Mixed Models (GLMM)

A method for dealing with response variable distributions that are not normal, generalized linear mixed models (GLMM) combine generalized linear models with linear mixed models to produce precise variance estimates for complex data. Fixed effects and random effects in linear predictors are included in GLMM, an extension of Generalized Linear Models (GLM). A regression model that offers a variety of distributions and connection functions is the GLMM. The link function's objective is to adjust the dependent variable's value to correspond to the linear predictor's scale. The link with the predictor variable will then be linearized as a result. The general form of GLMM is as follows (McCulloch, 2000):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon} \tag{1}$$

with,

$\mathbf{y}$	: vector of the response variable $(N \times 1)$
$\mathbf{X}$	: known matrices according to the fixed effects $(N \times p)$
$\boldsymbol{\beta}$	: column vector for fixed effect parameters $(p \times 1)$
$\mathbf{Z}$	: random effects design matrix $(N \times 1)$
$\mathbf{b}$	: random effect vector
$\boldsymbol{\varepsilon}$	: column vector as a residual $(N \times 1)$

2.3 Negative Binomial Mixed Model

The selection of a suitable distribution and link function for the supplied data is the first stage in the modeling procedure. For count data, the natural distributions are Poisson or, in the case of overdispersion, the negative binomial distribution (McCulloch, 1997; Zhang et al., 2017). The negative binomial distribution was used in this research to account for overdispersion. We presumptively observe the negative binomial distribution for the counting response y_i :

$$\boldsymbol{y_i} \sim \boldsymbol{NB}(\boldsymbol{y_i}|\boldsymbol{\mu_i}, \boldsymbol{\theta}) = \frac{\boldsymbol{\Gamma(y_i+\theta)}}{\boldsymbol{\Gamma(\theta)y_i!}} \cdot \left(\frac{\boldsymbol{\theta}}{\boldsymbol{\mu_i+\theta}}\right)^{\boldsymbol{\theta}} \cdot \left(\frac{\boldsymbol{\mu_i}}{\boldsymbol{\mu_i+\theta}}\right)^{\boldsymbol{y_i}} \tag{2}$$

with,
μ_i : mean
θ : parameter shape
Γ : function gamma

The predictor variable X is as many as units, the random variable Z is as many as random factors, and the total sequence T, as read by the logarithm of the link function, are all related to the mean parameter (μ) in the negative binomial mixed model:

$$\log(\boldsymbol{\mu}) = \log(\boldsymbol{T}) + \boldsymbol{X\beta} + \boldsymbol{Zb} \quad (3)$$

where:

$\log(\boldsymbol{\mu})$: the offset, which accounts for differences in the total number of sequences reads across samples

β : the host factor factors' fixed effects vector X
b : the vector of random effects for Z

The correlation between the samples and the various sources of variation are modeled using random effects, helping to prevent biased inference on the impact of the predictor variable X. Typically, it is believed that the random effects vector will be assumed to be the multivariate normal distribution (Zhang et al., 2017).

2.4 Parameter Estimation for GLMM

In longitudinal research, the same subjects are measured repeatedly over time. As a result, there is a correlation between the measurements, which the statistical analysis must take into consideration. GLMMs are useful for studying longitudinal data because they can model both within-subject correlation and between-subject variation. The likelihood function is maximized given the observed data to estimate the parameters of a GLMM using MLE. The likelihood function for longitudinal data is the combined likelihood of witnessing all the data for each person. Assuming a distribution for the response variable, a link function connecting the response variable's mean to the predictors, and a random effects structure accounting for within-subject correlation and between-subject variation, we can calculate this probability. Because random effects are present, the likelihood function for longitudinal data is often complicated. The likelihood of the fixed effects and the likelihood of the random effects, however, can be separated into two halves. The likelihood of the random effects is determined by integrating across the distribution of the random effects, but the likelihood of the fixed effects is determined using the same formula as for a generalized linear model (GLM). Finding parameters that optimize the total likelihood of a data set is the basic objective of MLE. This is accomplished in the setting of the exponential family by maximizing the log-likelihood function ℓ(θ; y, ϕ), over the canonical parameter θ depending on the observation y and the scale parameter ϕ (McCulloch, 2000; Stroup, 2012). The parameter vector determines the link function based on the distribution of dependent variables; the link function then determines the mean, i.e. μ , where the inverse link function. The canonical parameter is a function of the mean (θ(μ)). So, the following is a general formulation of the exponential family's log-likelihood.

$$l(\boldsymbol{\beta};\ \boldsymbol{y},\boldsymbol{\phi}) = \frac{\boldsymbol{y}\{\boldsymbol{\theta}[\boldsymbol{g^{-1}}(\boldsymbol{X\beta})]\} - \boldsymbol{u}(\{\boldsymbol{\theta}[\boldsymbol{g^{-1}}(\boldsymbol{X\beta})]\})}{\boldsymbol{a}(\boldsymbol{\phi})} + c(\boldsymbol{y},\boldsymbol{\phi}) \quad (4)$$

Simulation has been recommended by Geyer and Thompson (1992) and Gelfand and Carlin (1993) as a direct method of estimating the magnitude of the likelihood. beginning with:

$$\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{D}|\boldsymbol{y}) = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \frac{\boldsymbol{f_{y|u}}(\boldsymbol{y}|\boldsymbol{u}^{(k)},\boldsymbol{\beta},\boldsymbol{\phi})\boldsymbol{f_u}(\boldsymbol{u}^{(k)}|\boldsymbol{D})}{\boldsymbol{h_b}(\boldsymbol{u}^{(k)})} \quad (5)$$

Where N is the number of simulated values and u is chosen from the importance sampling distribution, $h_u(u)$. This provides a fair approximation of the likelihood regardless of the $h_u(u)$ option. Following a single simulation or a series of simulations in an iterative process, the simulated likelihood is then numerically maximized while allowing the importance sampling distribution to depend on the current parameter values.

With a single random effect, the likelihood is relatively easy to evaluate numerically (and hence maximize) (McCulloch, 1997) and for this example it is given by

$$L(\beta,\sigma^2|y) = \prod_{j=1}^{q}\int_{-\infty}^{\infty}\prod_{i=1}^{n}\frac{exp\{y_{ij}(\beta x_{ij}+u_j)\}e^{-u^2{}_j/2\sigma^2}}{1+exp\{y_{ij}\beta x_{ij}+u_j\}(2\pi\sigma^2)^{1/2}}du_j \quad (6)$$

For the metropolis algorithm we chose the candidate distribution, $h_u(u)$, and the acceptance function thus

$$A_k(u, u*) = min\{1, e^{y+k(u_k*-u_k)} \prod \ i \frac{1+e^{\beta x_{ij}+u_k}}{1+e^{\beta x_{ij}+uk*}} \quad (7)$$

while the Newton Raphson iteration is

$$\beta^{(m+1)} = \beta^{(m)} + E[X'W(\beta^{(m)}, U)X|y]^{-1}X'(y - E[\mu(\beta^{(m)}, U)|y] \quad (8)$$

For the σ^2 is,

$$\sigma^{2(m+1)} = \frac{1}{N}\sum_{k=1}^{N}(\sum \ _j u_j{}^{(k)2})/q \quad (9)$$

2.5 Forecasting

Forecasting plays a significant role in decision-making for all companies that are concerned with the future. A conventional time-series forecasting model's standard operating method is to find the pattern that most closely matches the previous data. In other words, a functional form is chosen that best captures the relationship between the input (past observations) and output (prediction) of the system (Charbuty & Abdulazeez, 2021). Large and complicated data sets, like demographic, clinical, and laboratory data sets, can be analyzed using machine learning techniques. Machine learning algorithms can be trained to forecast a variety of outcomes, including the likelihood that a patient will contract TB, the efficacy of their treatment, and other outcomes. The Feedforward Neural Network approach is used in this study to make this prediction. These algorithms can also spot data patterns and linkages that conventional statistical methods might miss out on.

2.6 Feed-Forward Neural Network

Global technological advancements are a result of human innovation, which has led to the invention of many user-friendly machines. These devices are employed to provide for a variety of human needs. Among these endeavors is machine learning. Machine learning's training procedure is its distinguishing characteristic. Machine learning uses two techniques: classification and prediction. SVM (Support Vector Machine) and neural networks are the two machine learning techniques that are most frequently utilized (Batta, 2018).

The neural network is a technique that imitates the structure of the human neural network, which consists of nodes connected to one another. A neural network's layers are collections of neurons that are organized in a specific way. Several network designs may be produced by this arrangement (Sastri & Setiadi, 2018). Time series data are frequently employed with neural networks to forecast future situations (Bolker et al., 2015). A neural network, also known as an artificial neural network, is a forecasting technique with a network structure like to the human brain that is based on a straightforward mathematical model. An artificial neural network is a network made up of several processing units collectively referred to as "nodes" and structured in a certain hierarchy of layers (Yasin, 2018; Endharta, 2009; Suhermi, 2019; Warsito, 2019; Caraka, 2019; Patan, 2019).

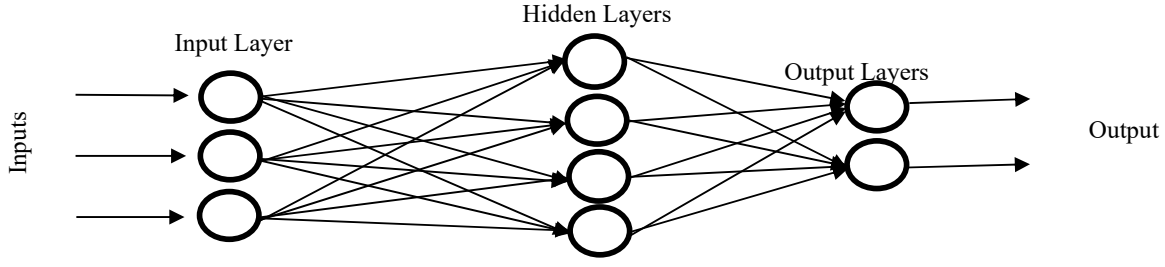

Fig. 2. Feedforward Neural Network Model Structure

McCulloch and Pits published the first Artificial Neural Network in 1943 (McCulloch and Pits, 1943). Natural neural systems, which are acknowledged to be incredibly accurate, are used in a similar manner by ANN, a partner scientific framework (Fausset et al., 1994). Afterwards, concentrated on perceptron machineability for single-layer feed-forward systems. The most esteemed neural network model for use in time arrangement forecasting is FFNN. It envisions a single hidden layer with a few hubs. Each of a set of repetitive frameworks with starting loads that appeared at the middle of handling figures is fitted. While the input nodes are the previously lag-timed observations, the yields guide the forecast for its future attributes. Information transmitted by the input nodes is processed by hidden nodes with material non-direct exchange capabilities (Rosenblatt et al., 1962). The most common prediction model is called a Feedforward Neural Network (FFNN). There is no cycle or rotation in the FFNN because of the connections between the nodes. The feed-forward model is the most basic type of neural network because data is only processed in one direction. Data can go through numerous hidden nodes, but it always moves forward and never backward (Pontoh et al., 2020). The term "multilayer perceptron" is occasionally used imprecisely to refer to feedforward neural networks (Charbuty & Abdulazeez, 2021).

This model has the advantage of producing a minimal error and being able to identify and evaluate complicated issues because its predicted value is very near to the actual value (Utami, 2013). While Feed Forward Neural Networks are pretty straight-forward, some machine learning applications can benefit from their streamlined architecture. For instance, feedforward neural circuits can be constructed with the goal of running them autonomously, yet with few intermediaries to control them. Similar to how the human brain functions, this process uses numerous little neurons to manage and analyze the larger one. Each network completes its tasks independently, allowing the final outputs to be integrated to create a cohesive, synthesized output.

2.7 Model Evaluation

We assess the effectiveness of each forecasting model using R-squared (R^2) and Root Mean Square Error (RMSE). Each train/test pair produced by rolling forward splitting is subjected to evaluation. RMSE measures the difference between actual values and predicted values in a regression model. RMSE provides information about how close the model's predictions are to the actual values, where a lower RMSE value indicates better performance in predicting the actual values, while R^2 is utilized to confirm the accuracy of prediction curve fitting (Yirga et al., 2020). The result is that each metric is run m-p-q+1 times.

$$R^2 = \mathbf{1} - \frac{\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{m}(Y_i - \underline{Y})^2} \tag{10}$$

$$RMSE = \sqrt{\sum_{i=1}^{m}(\frac{Y_i - \hat{Y}_i}{m})} \tag{11}$$

where,
m : number of observations
Y_i : the actual values
$\hat{Y}_i$: the predicted values
$\bar{Y}$: the average of actual value

A potent strategy for analyzing the impact of pulmonary tuberculosis (TB) on various outcomes can be achieved by combining Generalized Linear Mixed Models (GLMM) and machine learning techniques. The advantages of both approaches can be used in this hybrid strategy, which can also offer a more thorough study of the data. Researchers can get a more thorough understanding of how pulmonary TB affects different outcomes by integrating GLMM and machine learning techniques. In this paper, GLMM is used to detect TB risk factors, and machine learning algorithms are then trained to predict the chance of TB infection based on these risk factors. With the use of this hybrid method, new risk factors for pulmonary TB will be discovered, and more potent therapies will be created.

3. Data and Variables

3.1 Data

The data used in this study are secondary data obtained from the Indonesian Public Health Agency website. The observation unit for this data is 27 cities/counties in West Java from 2016 to 2021 with 6 variable factors affecting tuberculosis cases in West Java. Next, the variable factors influencing tuberculosis in West Java are presented in Table 1.

Table 1
Variables and Data Sources

Variable		Source
Dependent Variable		
Number of Tuberculosis Case (y)		West Java Provincial Health Office
Independent Variable		
Population by Age Group (X_1)	0-14 (X_{11})	Statistics Indonesia and Open Data Jabar
	15-44 (X_{12})	
	45-64 (X_{13})	
	65+ (X_{14})	
Infant BCG Immunization Coverage (X_2)		West Java Provincial Health Office and Open Data Jabar
Population Density (X_3)		Statistics Indonesia and West Java Provincial Health Office
Healthcare Facility (X_4)	The Number of Public Hospitals (X_{41})	West Java Provincial Health Office
	The Number of Health Center (X_{42})	
The Implementation of Community-based Sanitation (X_5)		West Java Provincial Health Office

3.2 Tuberculosis case

The dependent variable in this study is the quantity of tuberculosis cases in West Java from 2016 to 2021. There are a number of factors that can affect the occurrence of tuberculosis cases in addition to direct dissemination by airborne droplets.

3.3 Age Group

The prevalence of tuberculosis might vary by age group because an individual's susceptibility to the disease increases with age. Between the ages of 15 and 50, which is the age bracket with the highest economic productivity, around 75% of tuberculosis patients contract the disease (Ronald et al., 2010).

3.4 Vaccination history for BCG

Bacillus Calmette Guerin, also known as BCG, is a bacterium that is used as a vaccination to stop tuberculosis. According to John and Anita's research, administering the BCG vaccine can boost the body's defenses against pulmonary tuberculosis by as much as 80% (Zulaikha, 2018).

3.5 Population Density

Dense environmental variables that are brought on by the expanding population make it easier to become infected and contribute to the rise in pulmonary tuberculosis cases. When a person with tuberculosis coughs, speaks, or sneezes, bacteria-filled saliva splatter and airborne microorganisms can be ingested by others. Population density can raise the risk of tuberculosis since infected droplets can spread up to 1 meter (Prihanti et al., 2015).

3.6 Healthcare Facilities

Facilities for providing healthcare are tools or locations set aside for carrying out efforts to prevent illness, promote wellness, treat patients, and restore care. The prevalence of pulmonary tuberculosis in tuberculosis preventive and treatment initiatives may be impacted by the lack of adequate health services (Liu & Sun, 2023).

3.7 The Implementation of Community-based Sanitation

Pneumopulmonary tuberculosis can develop as a result of an unfavorable environment. The occurrence of tuberculosis can be influenced by environmental variables such as inadequate air circulation, illumination, ventilation, and humidity (Budi et al., 2018). These elements must be in good working order to constitute a healthy environment.

4. Results

4.1 Descriptive Analysis

Table 2 provides an overview of all the research variables used in this study. All variables are measured in person units, with the exception of population density, which is measured in persons per square kilometer, and the number of villages using community-based sanitation. In general, the number of Tuberculosis cases in 27 West Java cities/regencies from 2016 to 2021 averaged 2947 cases, with a standard deviation of 2557.53 cases. Because the number of Tuberculosis cases varies by city/regency, these cities/regencies can be included as a random effect in the GLMM model. Bogor Regency had the highest number of tuberculosis cases in 2019, with 15566 cases, while Cirebon City had the lowest number of Tuberculosis cases in 2017, with only 231 cases. Since the variables in this study have different measurement ranges and scales, standardization can aid in the subsequent analysis stages. Throughout this study's analysis, R software is used.

Table 2
Descriptive Statistics for Variables Used in the Study

Variables		Mean	Median	SD	Min	Max
Number of Tuberculosis Case (y)		2953	2068	2564.231	231	15566
Population by Age Group (X_1)	0-14 (X_{11})	470318	420551	350062.1	42887	1735080
	15-44 (X_{12})	862459	739664	633126.8	75492	3034712
	45-64 (X_{13})	365713	383384	213494.5	44202	1056090
	65+ (X_{14})	101348	106995	54191.36	14307	262351
Infant BCG Immunization Coverage (X_2)		29646.79	20323	24498.80	267	114147
Population Density (X_3)		3935.3	1408	4790.111	389	15798
Healthcare Facility (X_4)	The Number of Public Hospitals (X_{41})	11.07	7	10.79657	0	47
	The Number of Health Center (X_{42})	50.99	39.5	56.0005	8	597
The Implementation of Community-based Sanitation (X_5)		175.34	165	121.0572	10	442

Some values are missing from the available study data, namely the Number of Tuberculosis Cases in Bandung Regency (2017) and Infant BCG Immunization Coverage in West Bandung Regency (2019) and Garut Regency (2019). The mean of these variables in the same city/regency over the two adjacent years is used as an imputation to fill in the gaps left by missing data. Several outliers in the dependent variable were discovered in this study (Number of Tuberculosis Cases). The extreme values discovered are then considered missing values, which are then imputed with the machine learning predicted values (using the Feed Forward Neural Network (FFNN) method) before being used in the analysis process.

Data Exploration

Exploring the Distribution of the Dependent Variable with Cullen and Frey's Method

Figure 3 shows the histogram and Cumulative Distribution Function (CDF) of the dependent variable in this study (number of Tuberculosis cases). The fitdistrplus[x] package and the R program were used to explore the distribution of the variable.

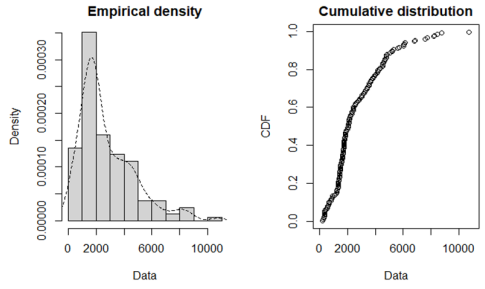

Fig. 3. Histogram and CDF of the Number of Tuberculosis Cases

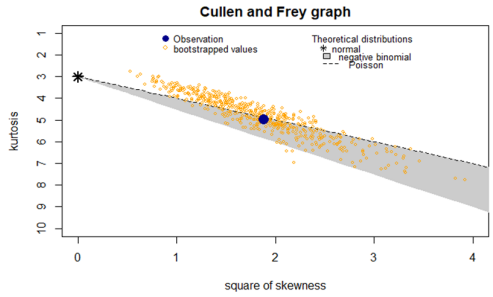

Fig. 4. Cullen and Frey Graph of the Number of Tuberculosis Cases used in the Study

The review of the Cullen-Frey plot and comparison of the Akaike Information Criteria (AIC) values between the distributions can be used to identify the distribution that can best approximate the data. Figure 4 displays the Cullen and Frey plot, which reveals that the variable number of Tuberculosis cases in this study with a bootstrap of 500 is most suited with the negative binomial distribution. Based on the AIC values in Table 3, it is also reasonable to infer that the negative binomial distribution with the lowest AIC value is the most appropriate distribution to fit the data on the number of Tuberculosis cases in West Java Province between the years of 2016 and 2021. Both analyses suggest that the assumption of a negative binomial distribution may be the best to utilize for this variable in the subsequent analytical procedure.

Table 3
Comparison of AIC Values of Various Distributions to Fit the Distribution of Number of Tuberculosis Cases

Distribution	AIC
Normal	2919.524
Poisson	206725.8
Negative Binomial	2852.267

Correlation Between Study Variables

The correlation between each of the predictor variables with the response variables (the number of Tuberculosis cases) can give an indication of the significance of including these factors in the model, whereas correlation values among the predictor variables can be used to assess multicollinearity or a close association between the predictor variables in this study. It can be seen that all of the predictor factors are correlated with the response variable, as shown in Fig. 5. Also, it was discovered that. The various categories in the variable number of populations based on age groups are related to one another.

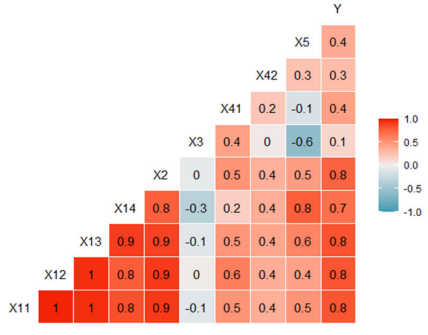

Fig. 5. Correlation Between This Study's Variables

Negative Binomial Mixed Model

In this study, GLMM modeling will be conducted to explore how various Tuberculosis factors affect the number of Tuberculosis cases in various districts/cities in West Java in 2016-2021. Package glmmADMB will be used to help perform the analysis.

According to the findings of a previous investigation into the distribution of the dependent variable, it is assumed that the implementation of the negative binomial distribution assumption in the construction of the GLMM model in this study is more acceptable. With all variables included, Table 4 compares the AIC values of the Poisson and Negative Binomial GLMM models as well as the results of overdispersion testing for the two models. By comparing the two tables, it can be determined that the negative binomial assumption is more suitable to be used on the data of the number of tuberculosis cases in this study because it tends to have a smaller AIC and is also better able to overcome the overdispersion issue found in the Poisson GLMM model when applied to the case of this study. This is consistent with the findings of the distribution exploration stage that came before it, as well as with (Utami, 2013; Yirga et al., 2020) which claims that although discrete data, like disease case numbers, can typically be assumed to have a Poisson distribution, overdispersion (having a variance greater than the mean) can occur in some cases. One of these cases can be overcome by using the negative binomial distribution assumption. As a result, from now on, GLMM with negative binomial assumption will be employed.

Table 4
Comparison of AIC Values for Poisson GLMM and Negative Binomial GLMM Models

GLMM Distribution Assumptions	GLMM Models	AIC	Overdispersion Test		
			Chi Square	df	p-value
Poisson	$g(\mu) = \beta_0 + \beta_1 D_1 X_1 + \beta_2 D_2 X_1 + \beta_3 D_3 X_1 + \beta_4 X_2 + \beta_5 X_3 + \beta_6 D_1 X_4 + \beta_7 X_5 + bZ$	7482.4	1834943999	151	0
Negative Binomial	$g(\mu) = \beta_0 + \beta_1 D_1 X_1 + \beta_2 D_2 X_1 + \beta_3 D_3 X_1 + \beta_4 X_2 + \beta_5 X_3 + \beta_6 D_1 X_4 + \beta_7 X_5 + bZ$	2642	127.2684986	151	0.920

Additionally, insignificant predictor variables were gradually eliminated from the model until a model with all significant variables and the smallest AIC was obtained. This was done to determine the best negative binomial GLMM to model different predictor variables on the number of tuberculosis cases in different districts and cities in West Java. The estimated models of the negative binomial GLMM model with all significant components are shown in Table 5.

Table 5
Best Negative Binomial GLMM Model Selection

Model	Covariates Included	Significant Covariates ($\alpha = 0.05$)	AIC
1	X_{13}	X_{13}	2646.5
2	X_{12},X_{13}	X_{12},X_{13}	2662.6
3	X_{11},X_3,X_5	X_{11},X_3,X_5	2658.9
4	X_2,X_3,X_5	X_2,X_3,X_5	2690.5
5	X_{13}, X_{14}, X_3	X_{13}, X_{14}, X_3	2636.2
6	X_{12}, X_{13}, X_3	X_{12}, X_{13}, X_3	2634

* for each model, random effects of districts/cities are included

According to Table 5, model 6 was chosen as the best model for explaining the relationship between the predictor factors and the number of cases of tuberculosis since it had the minimum AIC value among the GLMM models including all relevant predictor variables. The selected negative binomial GLMM model doesn't show any overdispersion symptoms, as seen further in Table 6. Furthermore, the predictor variables included in the best model did not show symptoms of multicollinearity, as shown by VIF values that did not exceed 10 in Table 7.

Table 6
Overdispersion Testing on a Selected Negative Binomial GLMM Model

Chi Square	Ratio (Chi Square/df)	df	p-value	Description
126.3609508	0.8048468	157	0.9654756	Insignificant, no symptoms of overdispersion

Table 7
VIF of various Predictor Variables in Model 1

Variables		VIF
Population by Age Group	15-44 (X_{12})	1.008359
	45-64 (X_{13})	1.051717
Population Density		1.046202

Table 8
Regression Coefficient and Significance for the Negative Binomial GLMM Model

Variables		Coefficient Estimate	Standard Error	z-score	p-value	Description
Intercept		6.35e+00	1.26e-01	50.63	< 2e-16	Significant ($\alpha < 0.001$)
Population by Age Group	15-44	-5.51e-07	2.22e-07	-2.48	0.013	Significant ($\alpha < 0.05$)
	45-64	4.49e-06	6.61e-07	6.79	1.1e-11	Significant ($\alpha < 0.001$)
Variance				0.06043		
Standard Deviation				0.2458		

According to Table 5, infant BCG vaccination coverage, population density, the number of public hospitals, and the implementation of community-based sanitation are the variables in the chosen model that are significant (with a significance level of =0.05) in predicting the number of tuberculosis cases in West Java for each district/city in it. The other variables have an insignificant effect, so they are not included in the model. The following is the best general linear mixed-model (GLMM) for modeling numerous factors on the number of tuberculosis cases in West Java:

$$\hat{Y} = g(\mu) = \beta_0 + \beta_1 X_{Age15-44} + \beta_2 X_{Age45-64} + \beta_3 X_{Pop.Density} + bZ_{Districts/Cities} \quad (12)$$

with g(.) link function for the mean of the response variable is the natural logarithm function which is appropriate for the negative binomial assumption as described in equation (3) in the preliminaries above and b is the random effect intercept coefficient for each district/city in West Java. The regression coefficient estimation results for these variables can be seen in Table 8.

In this study, it was found that there are two categories of population variables based on age group (that is, for the age range of 15-44 years and also the age range of 45-64 years) that have a significant influence on the number of Tuberculosis cases in each district/city in West Java Province. The number of productive age population (15-44 years old) has a coefficient of -0.000000551. This indicates that the variable has a negative influence on the number of Tuberculosis cases, meaning that the addition of 1 person aged 15-44 years in a district/city will decrease the log of the expected number of Tuberculosis cases in the district/city by 0.000000551 units (if other variables are constant).

Meanwhile, the number of older population (45-64 years old) has a positive and significant influence on the number of Tuberculosis cases with a coefficient of 0.00000449. This indicates that the addition of 1 person aged 45-64 years in a district/city will increase the log of the expected number of Tuberculosis cases in the district/city by 0.00000449 units (if other variables are constant). When compared to the literature review that was conducted (Ronald et al., 2010; L. Pangaribuan et al., 2020), it was mentioned that a person's susceptibility to be infected with Tuberculosis can increase with age, which is quite consistent with the results of this study.

In this study, it was also discovered that, with a coefficient of 0.0000462, population density in each district or city had a positive and significant impact on the number of tuberculosis cases. According to this, assuming other variables remain constant, an increase in population density in a district or city of 1 person/km2 can result in a 0.0000462 rise in the log of the predicted number of tuberculosis cases. This result is in line with the literature study, which found that residential density can contribute to an elevated risk of being infected with tuberculosis (Prihanti et al., 2015).

Then, using the Feed-Forward Neural Network method, the four significant variables will be included in the subsequent analysis step.

Negative Binomial Mixed Model Neural Network

After obtaining the predictor variables that can significantly predict the number of Tuberculosis cases, the lag of these variables along with the lag of the number of Tuberculosis cases are included in the next Feed Forward Neural Network (FFNN) modeling. The number of cases of tuberculosis in the future was predicted using FFNN with the help of bbmle, neuralnet, and R software tools.

The architecture of the FFNN model used in this study is shown in Table 9. The inputs to be entered into the model consist of 5 neurons, namely the district/city, the number of Tuberculosis cases in the prior year (symbolized as Y_{t-1}) as the response variable of interest, and also the prior year data of 4 factors that significantly affect it based on the negative binomial GLMM model selected in the previous stage. These factors are population by age group 15-44, population by age group 45-64, and population density in the prior year (hereafter symbolized as $X_{12,t-1}$, $X_{13,t-1}$, and $X_{3,t-1}$), as well as districts. The output to be predicted in this FFNN model consists of 4 neurons, namely the number of Tuberculosis cases in that year (Y_t), and is

complemented by factors namely population by age group 15-44, population by age group 45-64, and population density in that year (hereafter symbolized by $X_{12,t}$, $X_{13,t}$, dan $X_{3,t}$) which can be reused as input to facilitate forecasting for subsequent years. This enables predictions to be performed again in the following year using the output from this FFNN model. By comparing the RMSE value with the R^2 value, trial-and-error findings are used to determine the number of hidden layers and neurons in each layer. A linear activation function has been used in the FFNN model.

Table 9
The Architecture of FFNN

Input	5 neurons, consist of districts/cities and lags of The Number of TB Cases (Y_{t-1}), Population by Age Group 15-44 ($X_{12,t-1}$), Population by Age Group 45-64 ($X_{13,t-1}$), and Population Density ($X_{3,t-1}$)
Hidden Layers	1, 2, 3 (trial and error)
Hidden Neurons	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, (3-2), (4-3-2) (trial and error)
Output	4 neurons, consist of The Number of TB Cases (Y_t), Population by Age Group 15-44 ($X_{12,t}$), Population by Age Group 45-64 ($X_{13,t}$), and Population Density ($X_{3,t}$)

The input data performs a normalization step initially before being entered into the model. The dataset is then split into two parts, with 80% of the data (years 2016–2020) being used to train the model and 20% of the data (year 2021) being used to test the model by making predictions using a pre–trained model, restoring the standardization process, and comparing the prediction results to actual data. For every trial-and-error hidden layer architecture, this procedure is repeated. Table 10 displays the outcomes of the model assessment.

Table 10
FFNN Evaluation Metrics

Hidden Layer	Neuron	RMSE	R2	Hidden Layer	Neuron	RMSE	R2
1	1	0.606	0.655	1	6	2.851	0.217
	2	0.298	0.918		7	0.589	0.778
	3	0.252	0.944		8	0.563	0.844
	4	9.641	0.136		9	0.796	0.717
	5	1.286	0.541		10	0.928	0.537
2	3-2	0.311	0.913	3	4-3-2	0.567	0.826

According to Table 10, the FFNN model with a 5-3-4 architecture is the most effective model. This is evident from the fact that it has the lowest RMSE and highest R^2 values when compared to other FFNN models. It can be said that the FFNN (5-3-4) model has good predictive capabilities. The architecture visualization of the final model, FFNN (5-3-4), is displayed in Fig. 6.

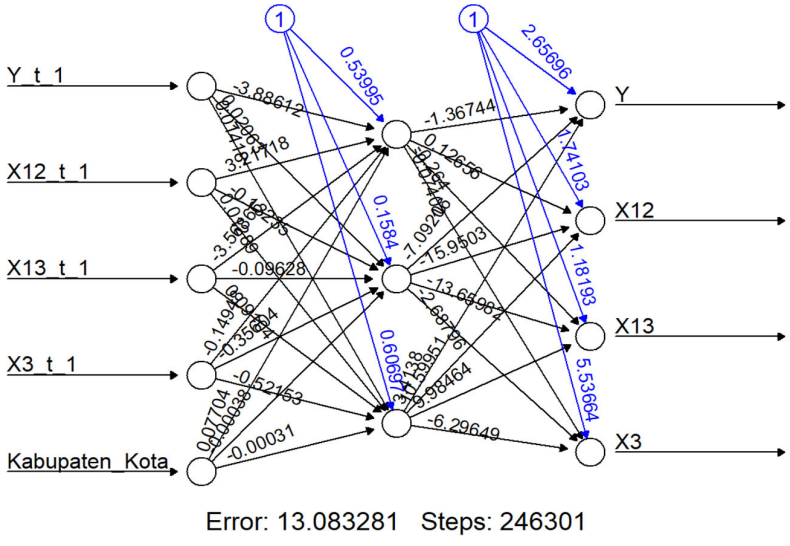

Fig. 6. Architecture of the FFNN (5-3-4) Model

With the best model obtained, forecasts were then made for the next 3 years, which are 2022-2024. The forecast results are shown in Fig. 7.

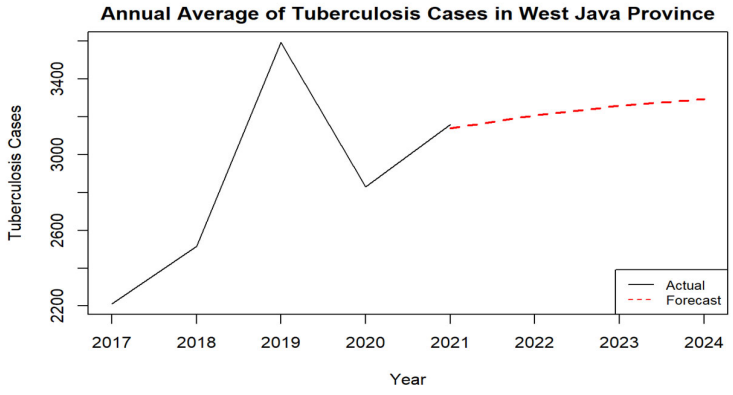

Fig. 7. Annual Average of Tuberculosis (TB) Cases Forecasting Results

Based on the forecast results for the next 3 years, from 2022-2024 in Figure 7, the trend of Tuberculosis cases in various regions in West Java tends to increase significantly in 2022-2024. The persistence of the upward trend in these years indicates the need for vigilance from the government and related parties in various districts and cities in West Java Province against the increase in Tuberculosis cases with various prevention and control programs for Tuberculosis.

5. Conclusions and The Future Research

In this study, longitudinal data on tuberculosis cases and other risk variables were examined over the 2016–2021 period in 27 districts and cities in West Java Province. This study demonstrates that overdispersion in Poisson GLMM can be overcome by GLMM under the assumption of a negative binomial distribution. According to this study, Population by Age Group 15-44, Population by Age Group 45-64, and Population Density all significantly affect the number of tuberculosis cases in each district and city.

Based on the analysis performed in this study, it can be concluded that the GLMM model may be used to explain the number of tuberculosis cases in the West Java Province in 2016–2021 under the assumption of a negative binomial distribution, as shown in the equation:

$$\hat{Y} = g(\mu) = \beta_0 + \beta_1 X_{Age\ 15-44} + \beta_2 X_{Age\ 45-64} + \beta_3 X_{Pop.Density} + bZ_{Districts/Cities} \quad (13)$$

It can then be used to forecast the number of Tuberculosis cases and risk factors again during the following year by using the three risk factors, the number of Tuberculosis cases, and each district's or city's data as input. This study found that the Feed Forward Neural Network (FFNN) model with FFNN architecture (5-3-4) is the most optimal model to predict the number of Tuberculosis cases (Y) as well as the four significant risk factors (X_{12}, X_{13}, and X_3) in the following year. Model evaluation results with RMSE = 0.252 and R^2 = 0.944 were obtained at the testing stage. The model was then used to forecast for the next 3 years. The forecast results showed a significant upward trend in 2022-2024 in the average number of Tuberculosis cases in various cities/districts in West Java. The results of this study indicate the importance of various related parties to continue to be vigilant, pay attention to various related risk factors, and continue to make various efforts to learn, prevent, and control Tuberculosis effectively.

Furthermore, this study used data that did not include extreme values. Future research can continue to study the pattern of Tuberculosis, especially in West Java Province, one of which can be done by conducting an analysis that includes extreme values to compare the best model that can describe Tuberculosis cases in West Java Province.

Acknowledgement

The authors would like to express their sincere gratitude to the Directorate of Research and Community Engagement of Padjadjaran University for their valuable support during the writing of this paper and for providing financial support for publication in this journal.

References

Arisanti, R., Notodiputro, K. A., Sadik, K., & Lim, A. (2017, March). Bias Reduction in Estimating Variance Components of Phytoplankton Existence at Na Thap River Based on Logistics Linear Mixed Models. In *IOP Conference Series: Earth and Environmental Science* (Vol. 58, No. 1, p. 012014). IOP Publishing.

Arisanti, R., Sumertajaya, I.M., Notodiputro, K.A., and Indahwati. (2020). Firth Bias Correction for Estimating Variance Components of Logistics Linear Mixed Model using Penalized Quasi Likelihood Method, *Communication in Mathematical Biology and Neuroscience*, 1–15.

Batta, M. (2018). Machine Learning Algorithms - A Review. *International Journal of Science and Research*, *18*(8), 381–386.

Bolker, B.M. (2015). *Ecological Statistics: Contemporary theory and application*. Oxford University Press.

Budi, I.S., Ardillah, Y., Sari, I.P., & Septiawati, D. (2018). Analisis Faktor Risiko Kejadian penyakit Tuberculosis Bagi Masyarakat Daerah Kumuh Kota Palembang, *Jurnal Kesehatan Lingkungan Indonesia*, *17*(2).

Caraka, R. E., Chen, R. C., Toharudin, T., Pardamean, B., Yasin, H., & Wu, S. H. (2019). Prediction of status particulate matter 2.5 using state Markov chain stochastic process and HYBRID VAR-NN-PSO. *IEEE Access*, *7*, 161654-161665.

Carter, D.J., Glaziou, P., Lonnroth, K., Siroka, A., Floyd, K., Weil, D., Raviglione, M., Houben, R.M.G.J., Boccia, D. (2018). The impact of social protection and poverty elimination on global tuberculosis incidence: a statistical modelling analysis of Sustainable Development Goal 1, *Lancet Global Health*, *6*(5), 514–522.

Charbuty, B., & Abdulazeez, A.M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning, *Journal of Applied Science and Technology Trends*, *2*(01), 20–28.

Endharta, A. J. (2009). Short term electricity load demand forecasting in Indonesia by using double seasonal recurrent neural networks. *International Journal of Mathematical Models and Methods in Applied Sciences*, *3*(3), 171-178.

Fausset, L. (1994). Fundamental of Neural Networks: Architectures, Algorithms, and Applications, *New Jersey: Prentice-Hall.*

Liu, S., & Sun, W. (2023). Attention mechanism-aided data- and knowledge-driven soft sensors for predicting blast furnace gas generation, *ScienceDirect: Energy*, *262*(A).

McCulloch, W. S., & Pits, W. H. (1943). Bulletin of Mathematical Biophysics, 5, 115-133Widjanarko, B., Gompelman, M., Dijkers, M., and van der Werf, M.J. (2009). Factors that influence treatment adherence of tuberculosis patients living in Java, Indonesia, *Patient Prefer. Adherence*, *3*, 231–238.

McCulloch, C.E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models, *Journal of American Statistical Association*, *92*(437), 162–170.

McCulloch, C.E. (2000). An Introduction to Generalized Linear Mixed Models, *Journal of the american Statistical Association*, *95*(452).

Pangaribuan, L., Kristina, K., Perwitasari, D., Tejayanti, T., & Lolong, D. B. (2020). Faktor-Faktor yang Mempengaruhi Kejadian Tuberkulosis pada Umur 15 Tahun ke Atas di Indonesia, *Buletin Penelitian Sistem Kesehatan, 23*(1), 10–17.

Pangestika, D. (2020). West Java extends PSBB in Jakarta's satellite cities until July 2, *The Jakarta Post.*

Patan, K. (2019). Neural Networks. *In Studies in Systems, Decision and Control*, *Switzerland.*

Pontoh, R. S., Toharudin, T., Zahroh, S., & Supartini E. (2020). Effectiveness of the public health measures to prevent the spread of covid-19. *Commun. Math. Biol. Neurosci.*, Article-ID.

Prihanti, G.S., Sulistiyawati, & Rahmawati, I. (2015). Analisis faktor kejadian tuberkulosis paru, *Jurnal Ilmu Kesehatan dan Kedokteran Keluarga*, *11*(2).

Ronald, R.D., Marais, B.J., & Clifton, E.B. (2010). Age and the epidemiology and pathogenesis of tuberculosis, *The Lancet*, *375*(9729), 1852-4.

Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, *Washington D.C: Spartan.*

Sastri, R., & Setiadi, Y. (2018). *Generalized Linear Mixed Model Untuk Data Kematian Bayi di Indonesia*. Jakarta: STIS.

Suhermi, N., Permata, R.P., & Rahayu, S.P. (2019). Forecasting the Search Trend of Muslim Clothing in Indonesia on Google Trends Data Using ARIMAX and Neural Network. *In Proceedings of the Communications in Computer and Information Science, Iizuka, Japan.*

Sulistyawati, S., & Ramadhan, A. W. (2021). Risk Factors for Tuberculosis in an Urban Setting in Indonesia: A Casecontrol Study in Umbulharjo I, Yogyakarta, *Journal of University of Occupational and Environmental Health*, *43*(2), 165–171.

Stroup, W.W. (2012). Generalized Linear Mixed Models: Modern Concepts, CRC Press: A Chapman & Hall Book.

Utami, T.W. (2013). Analisis regresi binomial negatif untuk mengatasi overdispersion regresi poisson pada kasus demam berdarah dengue, *Jurnal Statistika.*, *1*(2), 59–65.

Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic, *Trop. Med. Int. Health*, *25*, 278–280.

Warsito, B., Yasin, H., & Prahutama, A. (2019). *Particle Swarm Optimization to Obtain Weights in Neural Network*, 35.

Widjanarko, B., Gompelman, M., Dijkers, M., & van der Werf, M.J. (2009). Factors that influence treatment adherence of tuberculosis patients living in Java, Indonesia, *Patient Prefer. Adherence, 3*, 231–238.

Yasin, H., Warsito, B., Santoso, R., & Suparti, S. (2018). Soft Computation Vector Autoregressive Neural Network (VAR-NN) GUI-Based. *In Proceedings of the E3S Web of Conferences, Semarang, Indonesia*, *73*, 13008.

Yirga A.A, Melesse, S.F., Mwambi, H.G., & Ayele, D.G. (2020). Negative binomial mixed models for analyzing longitudinal CD4 count data, *Nature Journal.*, *10*(1), 1–15.

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A.K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data, *BMC Bioinformatics*, *18*(1), 1–10.

Zulaikha, E. (2018). Pemetaan dan Analisis Faktor-Faktor Yang Mempengaruhi Tuberkulosis Menggunakan Geographically Weighted.