Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:85

https://doi.org/10.28919/cmbn/8748

ISSN: 2052-2541

INTEGRATING GENERALIZED LINEAR MIXED MODELS WITH EXTREME NEURAL NETWORK: ENHANCING PULMONARY TUBERCULOSIS RISK MODELING IN WEST JAVA, INDONESIA

RESTU ARISANTI[1,*], RESA SEPTIANI PONTOH[1], SRI WINARNI[1], YAHMA NURHASANAH[2], AISSA PUTRI PERTIWI[2], SILVANI DEWI NURA AINI[2]

[1]Department of Statistics, Padjadjaran University, Indonesia

[2]Bachelor Program of Statistics Department, Padjadjaran University, Indonesia

Abstract: Mycobacterium tuberculosis is the primary cause of tuberculosis (TB), a global public health concern that primarily affects the respiratory system. This disease is also prevalent in the West Java Province of Indonesia. The study aims to analyze tuberculosis (TB) patterns using extreme values to identify the most accurate models for forecasting illness cases. The study uses advanced machine learning methods, the Generalized Linear Mixed Model (GLMM), and the Extreme Neural Network (ENN), to investigate how environmental and personal factors affect the number of tuberculosis (TB) cases. Results indicate that the population density and age groups of 45–64 and over 65 years strongly influence the occurrence of tuberculosis in West Java. We utilize the FFNN model to predict the future number of TB cases and risk variables. To effectively prevent and manage the spread of tuberculosis in the community, it is crucial for all parties to be watchful and aware of the different risk factors linked to the disease, as revealed by the findings.

Keywords: pulmonary tuberculosis; negative binomial mixed model; extreme learning feed-forward neural network.

2020 AMS Subject Classification: 62P10.

*Corresponding author

E-mail address: r.arisanti@unpad.ac.id

Received July 08, 2024

1. INTRODUCTION

Tuberculosis (TB) is an infectious disease posing a public health problem worldwide. This disease, caused by Mycobacterium tuberculosis primarily affects the lungs (pulmonary TB) but can also manifest in other sites [1]. Transmission occurs when infected individuals release bacteria into the air through activities such as coughing, sneezing, or talking [1]. A persistent cough, blood in the cough, chest pain, fever, night sweats, appetite loss, weight loss, and acute exhaustion are common signs of TB [2]. Given that indicators, such as exhaustion, weight loss, and abdominal discomfort, are similar to those of TB or HIV, it is often chal-lenging to detect adrenal insufficiency in dangerously sick patients [2]. Urgent medical treatment after the onset of symptoms is important, as the disease can spread and potentially become fatal when left untreated. Approximately 85% of TB patients recover with the 4-to-6-month anti-TB treatment regimen recom-mended by World Health Organization (WHO) [3].

The numerous social and structural elements that impact the interaction between TB and poverty are closely related to the objective of Sustainable De-velopment Goals (SDGs) [4]. This interconnection underscores the need for a comprehensive strategy to address socioeconomic variables influencing TB oc-currence. All WHO and UN Member States successfully overcame TB epidemic in 2014 and 2015 by adopting End TB Strategy and SDGs [1]. This comprehen-sive method comprises specific targets, including an 80% reduction in TB oc-currence by 2030, as outlined in SDG 3, along with substantial decreases in TB-related deaths and financial burdens [5]. Indonesia is one of the countries with a high burden of TB [6]. According to WHO (Global Tuberculosis Report 2023), 30 high TB burden countries ac-counted for 87% of the world's cases in 2022 with Indonesia ranking among the top 8 responsible for two-thirds of the global cases. These countries include India (27%), Indonesia (10%), China (7.1%), the Philippines (7.0%), Pakistan (5.7%), Nigeria (4.5%), Bangladesh (3.6%) and the Democratic Republic of the Congo (3.0%) [1,7]. In 2022, TB occurrence increased by 5% in 39 countries, including Indonesia [1]. These countries were mostly found in WHO Area of the Americas, but four high-TB burden countries in Asia were also included, namely Indonesia, Mongolia, Myanmar, and the Philippines [1]. In Indonesia, an estimated 969,000 new cases were reported, or 354 per 100,000 population

with a mortality of 144,000, or 52 per 100,000 population [8]. Therefore, WHO aims to reduce TB deaths and cases by 90% and 80%, respectively, in 2030 (2030 Sustainable De-velopment Goals) [5].

Aside from East and Central Java, which collectively account for approxi-mately 47% of all TB cases in Indonesia, West Java Province is one of the most populous provinces with the largest number of cases [9]. According to the West Java Health Profile in 2022, approximately 85.681 TB cases were reported, out of the 301.682 total cases in 2021. There were 47,053 more cases in males than in women, or 54.9% higher occurrence in men. The highest number of cases was recorded in three regencies/cities, namely Bogor Regency, Bandung City, and Bandung Regency, constituting between 7% to 13% of the new cases in the province [8].

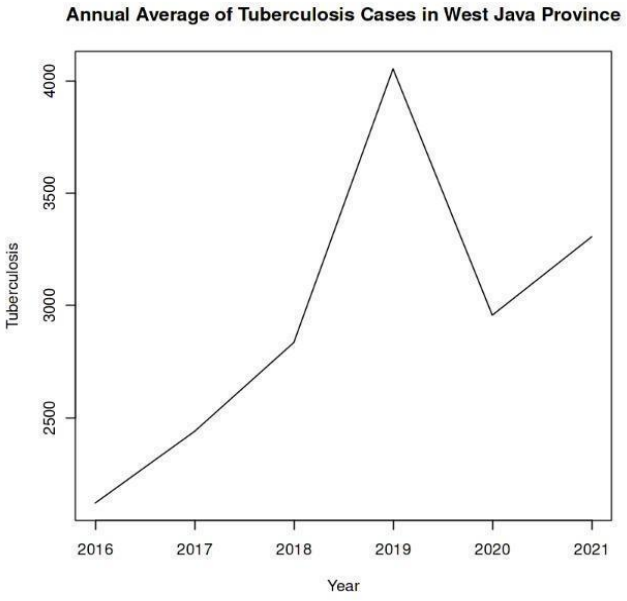

Figure 1. Annual Average of TB Cases in West Java Province, Indonesia

Figure 1 shows that the annual average TB cases in West Java Province (2016-2021) witnessed the highest increase in 2019, reaching 4,055. Despite a subsequent 27% decrease in 2020, the possibility of a resurgence in cases beyond 2021 remained plausible.

The high prevalence of TB in West Java is attributed to various factors, in-cluding a high population density, poor living conditions, limited access to healthcare services, and the presence of other risk factors namely malnutrition and HIV/AIDS. In addition, the use of traditional remedies and delayed treatment behavior are also common in some communities, contributing to

the spread of the disease. In other words, there are individual and environmental factors that play a very important role in influencing TB occurrence.

This study expands on a previous report by Arisanti et al. (2023) [10], of-fering a comprehensive understanding of TB disease patterns through an analysis conducted with extreme values to evaluate several models aimed at describing TB cases in West Java. Although the procedure was similar to that of the previous study, extreme learning was used in the analysis. GLMM (Generalized Linear Mixed Model) and extreme machine learning method, specifically Feed-Forward Neural Network (FFNN), were used to investigate and model the impact of indi-vidual and environmental factors on TB occurrence.

GLMM is a statistical modeling method particularly suitable for analyzing data with a non-normal distribution of the response variable, such as count or binary data. It is advantageous for analyzing longitudinal data with repeated measurements over time. GLMM allows for the inclusion of both fixed and ran-dom effects in the model, making it a flexible tool for analyzing complex data [11]. On the other hand, extreme machine learning methods are a set of compu-tational methods used to automatically detect patterns and relationships in data of extreme values, suitable for prediction and classification tasks, as well as han-dling structured and unstructured data. Extreme machine learning can be used to predict the risk of TB infection, the likelihood of treatment success, and other outcomes. Moreover, it is used to analyze large and complex data sets containing demographic, clinical, and laboratory data.

The combination of GLMM and FFNN will provide a powerful method for analyzing the effects of TB on various outcomes. This hybrid method has the the strengths of both methods and can provide a more comprehensive analysis of the data, identifying risk factors, and predicting the likelihood of TB infection. This study was conducted using TB case data, with the analysis unit covering all cit-ies/governments in West Java from 2016–2021, alongside other variables identi-fied as individual and environmental factors with GLMM and FFNN. The results are expected to offer insights for government and public health officials in the efforts to control TB, as well as enhance prevention and control measures in West Java.

2. PRELIMINARIES

TB remains a pressing public health problem in Indonesia, specifically in densely populated areas such as West Java. To better understand the diverse dynamics, this study adopts a methodological method centered on the integration of GLMM and Extreme Machine Learning methods. GLMM offers a powerful framework for analyzing longitudinal TB data, accommodating both fixed and random effects to spot different patterns. Complementing this, the computational capabilities of extreme machine learning enable the detection of complex relationships and predictive modeling in TB datasets. Using the combined method, this study aimed to explain the interaction between individual and environmental factors in shaping TB occurrence in West Java.

2.1. Generalized Linear Mixed Model (GLMM)

GLMM is a statistical modeling method applied to data with a non-normal distribution for the response variable, such as count or binary data. It is particularly useful in the analysis of longitudinal data, which consists of repeated measurements performed over time. The model calculates the effect of predictor factors on response variables (linear parameters), with some predictor variables fixed and others random [12]. This method can be categorized as a combination of GLM (Generalized Linear Model) and LMM (Linear Mixed Model) models. GLMM is used for non-normally distributed responses, allowing for nonlinear relationships between response averages and predictors. It also incorporates random effects to represent overdispersion and correlation [13]. According to UCLA (Statistical Consulting Group) and Arisanti et al (2017) [14,15], the general formula of GLMM model is as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G}), \qquad \boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\mu}), \qquad \boldsymbol{y}|\boldsymbol{u} \sim \boldsymbol{G}(\boldsymbol{\mu}, \boldsymbol{R}) \tag{1}$$

Where:

$\boldsymbol{y}$: column vector $N \times 1$ as response variable

$\boldsymbol{g}(.)$: link function transforming the linear predictor $\boldsymbol{\mu}$ into the scale of the response variable $\boldsymbol{y}$

$\boldsymbol{X}$: matrix $N \times p$ as fixed effect

$\boldsymbol{\beta}$: column vector $p \times 1$ as fixed effect coefficients

$\boldsymbol{Z}$: design matrix $N \times p$ as random effect

$\boldsymbol{u}$: random effect vector, with random effects **u** follow a normal distribution with mean 0 and covariance matrix $\boldsymbol{G}$

$\boldsymbol{G}(\boldsymbol{\mu}, \boldsymbol{R})$: distribution of the response variable given random effects $\boldsymbol{u}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{R}$

$\boldsymbol{\varepsilon}$: column vector $N \times 1$ as residual

2.1.1. Negative Binomial Mixed Model (NBMM)

The initial stage of modeling is to select an appropriate distribution and link function for the input data. In this context, the Poisson regression model examines the relationship between response (discrete data) and predictor variables (discrete, continuous, categorical, or mixed). However, the equidispersion assumption is rarely met when using Poisson regression models as discrete data frequently shows situations of overdispersion (high deviation) with variance values that are not equal to or larger than the mean [16,17]. This condition produces a high model deviation value, and the resulting model tends to be inaccurate.

In cases of overdispersion, a viable solution is to substitute the Poisson distribution assumption with a more flexible type, such as the negative binomial regression model. The negative binomial distribution for the count response y is represented by the formula:

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

Where:

μ_i: mean

$\boldsymbol{\theta}$: parameter shape

$\boldsymbol{\Gamma}$: function gamma

In the negative binomial mixture model, the predictor variable X represents multiple units, while the random variable Z denotes various random factors. These factors, along with the logarithm of the link function, contribute to the determination of the mean parameter (μ):

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

Where:

$log\ (\boldsymbol{\mu})$: adjustment factor, accommodating means in the overall count of sequence reads among different samples.

$\boldsymbol{\beta}$: the coeficients of fixed effects

$\boldsymbol{b}$: the vector of random effects for $\mathbf{Z}$

The correlation between the samples and the various sources of variation is represented by random effects, which helps to avoid erroneous inferences about the impact of the predictor variable X. It is often assumed that the random effects vector will follow the multivariate normal distribution.

2.1.2. Maximum Likelihood Estimation (MLE)

GLMM fixed parameter estimation uses Maximum Likelihood Estimation (MLE) method, referred to as an asymptotic method that yields an unbiased parameter estimate in cases where the sample size is high or approaching infinity [18]. According to Hosmer and Lemeshow [19], MLE method can be used to estimate parameters in nonlinear models such as logistic regression [20]. It represents the estimated value of β obtained by maximizing the probability function (L(β)). This is accomplished in the context of exponential families by maximizing the log-likelihood function, l(θ; y,ϕ) for the canonical parameter θ based on the observation, y, and the scale parameter, ϕ [21]. The parameter vector β determines the link function (η) based on the model definition; the combination function then determines the mean, namely μ = g^(-1) (η), where g^(-1) (.) is the inverse combination function, and the mean is a function of the canonical parameter (θ(μ)). Therefore, the general form of the log-likelihood for the exponential family can be written as follows:

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

Geyer and Thompson (1992) as well as Gelfand and Carlin (1993) recommend simulation for directly estimating the magnitude of the likelihood [22,23]. The equation is as follows:

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

Where N is the number of simulated values and u is a random number drawn from the importance sampling distribution $h_{\boldsymbol{u}}(\boldsymbol{u})$. Regardless of the $h_{\boldsymbol{u}}(\boldsymbol{u})$ option, this provides a reasonable approximation of the likelihood. Following a single or a series of simulations in an iterative process, the simulated likelihood is numerically maximized, allowing the importance sampling distribution

to depend on the current parameter values. Using a single random effect, the likelihood is relatively easy to calculate numerically and maximize, as shown by [13] using the formula:

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

The candidate distribution, $h_{\boldsymbol{u}}(\boldsymbol{u})$ and the acceptance functions were selected for the metropolis algorithm contrasting with the Newton-Raphson iteration:

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

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

For the σ^2, the equation is as follows:

$$\boldsymbol{\sigma^{2(m+1)} = \frac{1}{N}\sum_{k=1}^{N}\left(\sum_j u_j^{(k)2}\right)/q} \tag{9}$$

2.2. *Extreme Learning Feed Forward Neural Network*

In this study, extreme machine learning was used to handle the dataset with extreme values. In contrast to typical machine learning methods, this method integrated extreme values into the analysis, allowing for a more comprehensive understanding of complexities and uncovering valuable patterns or relationships that might otherwise remain obscured when using conventional analytical methods. Machine learning methods can be used as a tool to tackle complicated problems, including function approximation, prediction, classification, and regression. Among these methods, Artificial Neural Network (ANN) is the most widely used in the field of machine learning [24].

ANN or neural network (NN) is a sophisticated computer model inspired by the intricate operations of the human brain [25,26]. This network is made up of interconnected neurons that function similarly to the neuronal structure of the brain, allowing for complicated information processing. ANN has the unique capacity to learn from data through a process known as training, in which the strengths of connections between neurons are modified to enhance predicted performance [27]. This ability facilitates the detection of patterns and generates predictions, that mirror human

cognitive processes.

For modeling nonlinear relationships, the ideal choice is NN, which allows for the examination and modification of data patterns learned repeatedly, transforming into forecasting tools without requiring attention to the original pattern [28]. A specific type is FFNN, distinguished by the lack of loops in the graph architecture. In contrast to Recurrent Neural Network (RNN), which incorporates loops due to feedback, FFNN produces only a set of output values from an input rather than a sequence of values [29]. This characteristic renders feedforward networks as static systems, where responding to an input is independent of previous network conditions.

FFNNs typically consist of three layers as shown in Figure 2 including an input, one or more hidden, and an output layer. Each neuron in one layer is connected to every neuron in the next layer, and these connections have weights that define the intensity of the association. A unit is considered an input type when it is present in the input layer. Meanwhile, an input type is constant when it appears in successive layers due to the range of units in the preceding layer [30]. Weight plays a crucial role in modifying each input-to-unit and unit-to-unit link. Every unit includes an extra input, for which a constant value of one is taken. The bias, represented by the weight, adjusts this extra input to influence the neuron's activation in subsequent layers [30]. During the training process, the weights are systematically adjusted to minimize the disparity between the actual and desired output, ensuring the network learns to make accurate predictions.

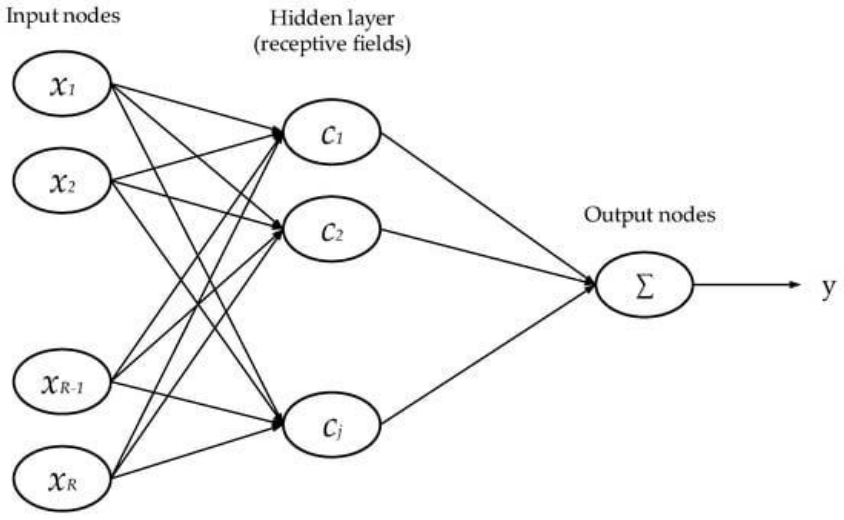

Figure 2. FFNN Architecture [24]

2.3. Model Evaluation

In the selection of GLMM models, deviance plays a crucial role in evaluating the fit of the observed data. It measures goodness-of-fit, with lower values indicating better consistency between the model and the data. Typically, this comparison is made between the fitted (restricted) and the saturated (reference) model. Deviance closer to the saturated model suggests a superior fit. By using deviance, it is possible to identify the most suitable GLMM model for explaining the data, balancing model simplicity with data fit.

$$\text{Deviance} = -2\left(\log L\left(\hat{\theta}_{saturated}\right) - \log L\left(\hat{\theta}_{full}\right)\right) \tag{10}$$

In forecasting using FFNN, the model performance is evaluated by two statistical measurements, namely the coefficient of determination (R^2) and root-mean-square error (RMSE). Wright (1921) defined the coefficient of determination as the proportion of the variance in the dependent variable predictable from the independent variables [31]. The formula for coefficient of determination is as follows:

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n}(u_i - \overline{u_i})^2}{\sum_{i=1}^{m}(u_i - \overline{u})^2}} \tag{11}$$

(worst value = -∞, best value = +1)

Besides R-squared, root mean square error (RMSE) can also be used to evaluate the performance of the model. Lower values of RMSE designate a higher performance of the given machine learning algorithm. RMSE can be calculated using the formula:

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u_i})^2}} \tag{12}$$

(best value = 0; worst value = +∞)

Where $\overline{u_i}$ is the predicted output value, u_i is the measured target value, and *n* is the number of samples.

2.4. *Data and Variables*

Secondary data used in this study comprised the number of TB cases in West Java Province,

specifically 27 cities/regencies from 2018 to 2023 [32]. The variables collected through literature reviews from multiple sources are shown in Table 1.

Table 1. Variable 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})	

a. Tuberculosis (TB) Case

The number of TB cases by district/city in West Java province from 2016 to 2021, expressed in person units, was used as the dependent variable.

b. Factor of Age

A crucial factor that significantly influences the number of TB cases is age. A previous study reported that participants in the 55+ age group have a 1.73 times higher risk of developing TB than those in the 15-34 age group [33].

c. History of Bacillus Calmette Guerin (BCG) Immunization

BCG is a bacterium used as a vaccine to reduce the risk of TB spreading. A one-month-old baby is immunized once with BCG and according to a study conducted by John in 2001 and Anita in 2006, immunization can strengthen the body's defense against TB by up to 80% [34]. In this study, BCG vaccination history factor comprised data on infant immunization coverage in West Java, which witnessed a decrease of 94.79% to 84.2% between 2017 and 2021 [21].

d. Factor of Population Density

TB occurrence is also affected by population density. Based on a study by Rusnoto et al. (2005), a settlement area that is not proportional to the number of inhabitants will lead to overcrowding. This creates an unhealthy environment and facilitates disease transmission from an infected person to the surrounding area due to a lack of oxygen consumption [34].

e. Factor of Healthcare Facility

A healthcare facility is a location or tool dedicated to healthcare efforts, such as prevention, outreach, promotion, healing, and recovery. To prevent and cure TB, a lack of health facilities can impair the prevalence of the disease. The health facility factor in this study used data on the number of public hospitals and health centers available in each city/district of West Java.

3. MAIN RESULTS

3.1. Descriptive Analysis

The R software was used for all analyses, and Table 2 provides an overview of the variables. Except for population density measured in persons per square kilometer, and the number of villages with community-based sanitation, all variables were measured in person units. From 2016 to 2023, the average number of TB cases across 27 cities/governments in West Java was 2,947, with a standard deviation of 2557.53. Bogor Regency had the highest number in 2019, with 15566 cases, while Cirebon City had the lowest number in 2017, with 231 cases. Given that the number of TB cases varied, these cities/regencies can be included as a random effect in GLMM model. This study used the same data as Arisanti [10] except for incorporating extreme values.

Table 2. Descriptive Statistics for Variables Used in the Study

Variabel		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

3.2. Handling of Missing and Extreme Values

Data for the variable Number of TB Cases in Bandung Regency (2017) were missing, hence, the mean over two adjacent years (1 year before and 1 year after) was imputed. For example, to compute the number of TB cases in 2019, the mean values for 2018 and 2020 were used.

Despite using the same data as previous studies, the extreme values found in the data were still included in the analysis process. Figure 3 shows outliers found in the dependent variable (Number of TB Cases), which included data for Bogor Regency (2017-2023) and Bandung City (2019, 2023).

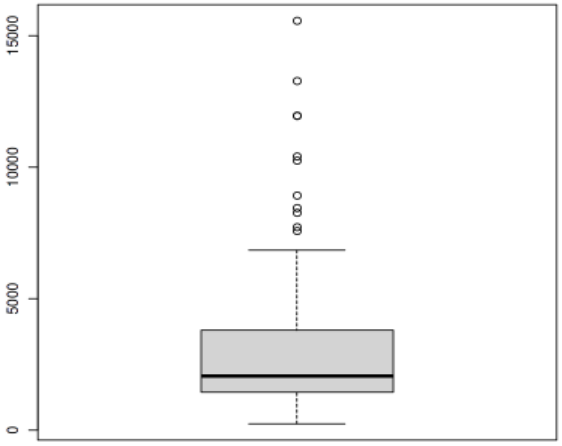

Figure 3. Boxplot for Number of Tuberculosis (TB) Cases

3.3. Data Exploration

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

The fitdistrplus package and the R software were used to verify the distribution of the dependent variable (Number of TB Cases). Figure 4 shows a histogram and the Cumulative Distribution Function (CDF).

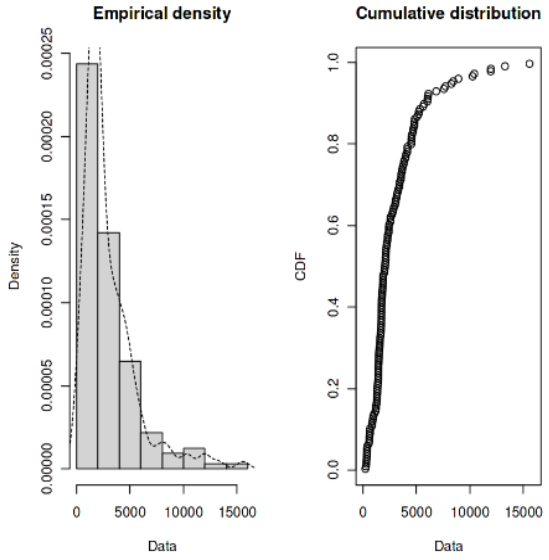

Figure 4. Number of TB Case Histogram and CDF

Based on the histogram and CDF, the response variable consisted of discrete data. Common distributions suitable for modeling discrete data are normal, negative binomial, and Poisson distributions. The distributions were further verified using the Cullen and Frey graph, which used residual values obtained from a linear model.

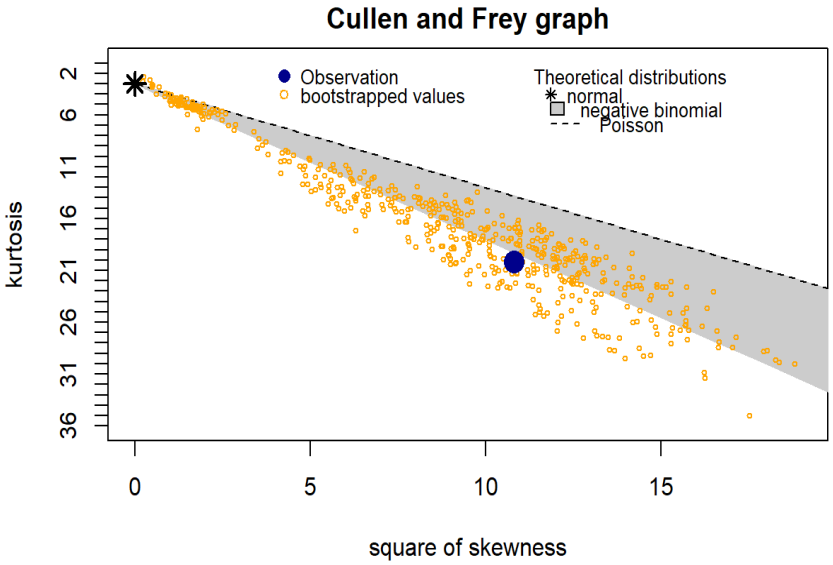

Figure 5. Cullen and Frey Graph of Number of TB Case

By comparing the Akaike Information Criteria (AIC) values between the distributions, the best fit for the data was determined. The Cullen and Frey plot, as shown in Figure 5, demonstrates that the variable Number of TB Cases with a bootstrap of 500 has a distribution close to Poisson and negative binomial distribution. This is consistent with the initial assumption that the number of disease cases follows a Poisson distribution. However, in some cases, discrete data often has more variance than the average (overdispersion) [35] and the applicable solution is to assume a negative binomial distribution [35,36]. This was supported by a review of AIC values in Table 3, showing that the negative binomial distribution with the smallest AIC value was the best fit for data on the number of TB Cases in West Java Province from 2016 to 2023.

Table 3. Comparison of AIC Values in Different Distributions for the Number of TB Cases

Distribution	AIC
Normal	2654.469
Poisson	105679.7
Negative Binomial	2521.159

3.3.2. Correlation Between Study Variables

An investigation into the association between variables was first performed before modeling GLMM. This was carried out to aid in selecting the appropriate independent variables for the model, as well as to better comprehend the relationship between the independent and dependent variables. Figure 6 shows the relationship between the independent and dependent variables. A high correlation value was observed between the categories in the variable X1 (population by age group).

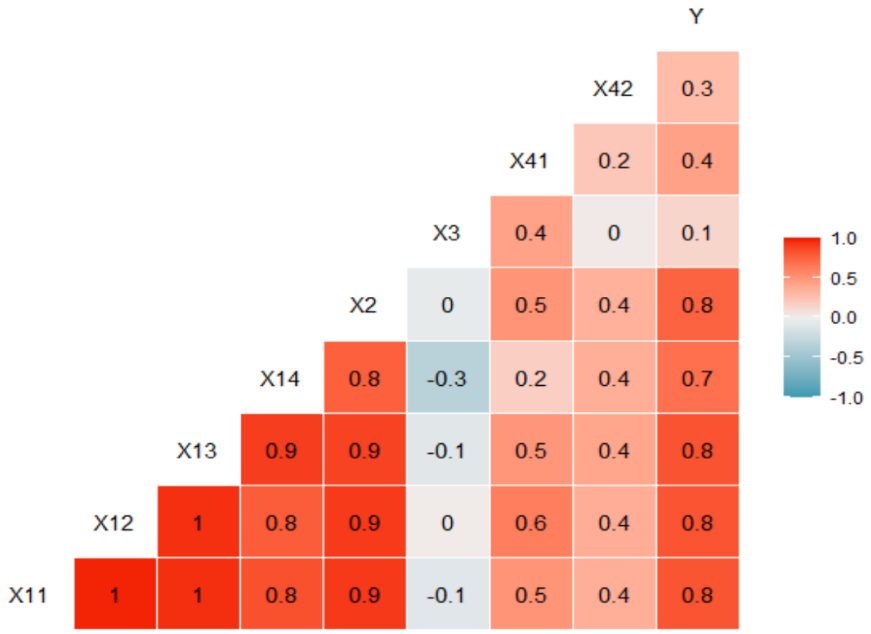

Figure 6. Correlation among the Study Variables

3.4. Negative Binomial Mixed Model (NBMM)

An estimate of GLMM model was obtained using the R programming language and the glmmADMB package. GLMM was used to model how various factors affect the number of TB cases in various regencies/cities in West Java. Based on the results for the distribution of dependent variables, GLMM, with the assumption of a negative binomial distribution, was considered the most suitable choice to model the effect of various predictor factors on the Number of TB Cases in the West Java Province from 2016-2023. The best NBMM was selected by removing the non-significant predictor variables iteratively until only the significant predictor variables remained in the model, as shown in Table 4. When more than one model with all significant predictor variables was found, then analysis of deviance was used for comparison. This was identical to the method used in [10], but analysis of deviance was used and extreme data was taken into account.

Table 4. Analysis of Deviance for NBMM with All Significant Predictor Variables

Model	Covariates Included	Significant Covariates	AIC	Deviance	P-value	Description
1	X_{13}, X_3	X_{13}, X_3	2644.3	0		
2	X_{12},X_3	X_{12},X_3	2681.3	-37.00	1	Not significant
3	X_{13},X_{14}, X_3	X_{13},X_{14}, X_3	2641.8	41.44	1.215e-10	Significant ($\alpha < 0.05$)
4	X_{12}, X_{13},X_3	X_{13}, X_3,X_5	2642.2	-0.42	1	Not significant

*for each model, random effects from regencies/cities were included

According to Table 4, there are three NBMM with all predictor variables significant. Model 3 was found to be significant, hence, it was selected to explain the relationship between predictor variables and the number of TB cases. Table 5 shows that VIF values for all predictor variables in the selected model did not exceed 10, suggesting there was no evidence of multicollinearity in the predictor variables.

Table 5. VIF of various predictor variables in the Selected Model (Model 1)

Variable	VIF
Population by Age Group 45-64 (X_{13})	1.000069
Population by Age Group 65+ (X_{14})	1.025475
Population Density (X_3)	1.025540

Table 6 shows that the selected NBMM has no symptoms of overdispersion. This is consistent with the results of the previous stage and further demonstrates the suitability of negative binomial assumption used for the response variable.

Table 6. Test of Overdispersion on the Selected NBMM

Variable	VIF
Population by Age Group 45-64 (X_{13})	1.000069
Population by Age Group 65+ (X_{14})	1.025475
Population Density (X_3)	1.025540

Based on the results, the best model selected for this study has the equation:

$$\hat{\boldsymbol{y}} = g(\boldsymbol{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 X_{Age45-64} + \hat{\beta}_2 X_{Age65+} + \hat{\beta}_3 X_{Pop.Density} + \boldsymbol{Z}_{Regency/City}\boldsymbol{b} \quad (13)$$

where $g(.)$ is the natural logarithmic link function for the mean of the response variable (as written in Equation (13)) and $\boldsymbol{b}$ is the intercept coefficient of the random effect for each regency/city in West Java.

According to Table 4, the predictor variables in the selected model most significant for predicting the number of TB cases include age, particularly population aged 45-64 and 65+ as well as population density. Meanwhile, other variables had insignificant effects and were excluded from the model. This was in contrast to a previous study that did not include extreme values [10], where the significant predictors comprised population age (15-44 and 45-64) and population density. Table 7 shows the results of the parameter estimate for the model.

Table 7. Regression Coefficients and Significance for the Selected NBMM

Variable	Coefficient Estimate	Standard Error	Z-score	p-value	Description
Intercept	6.23e+00	1.27e-01	49.12	< 2e-16	Significant ($\alpha < 0.001$)
Population Age Group 45-64 (X_{13})	2.08e-06	5.74e-07	3.62	0.00029	Significant ($\alpha < 0.001$)
Population Age Group 65+ (X_{14})	5.06e-06	2.34e-06	2.16	0.03077	Significant ($\alpha < 0.05$)
Population Density (X_3)	5.76e-05	1.19e-05	4.85	1.3e-06	Significant ($\alpha < 0.001$)
Variance			0.0422		
Standard Deviation			0.2054		

Based on the results, the number of populations in the age range of 45 to 64 years and above 65 years was found to have a positively significant impact on the number of TB cases in each regency/city in West Java. The coefficient was 0.00000208 and 0.00000506 for the age range of 45 to 64 years and above 65 years, respectively. This demonstrates that a 1-year increase in the population aged 45 to 64 affects the increase in the log of expected TB cases by 0.00000208 units. Meanwhile, a 1-year increase in the population aged above 65 years affects the increase in the log of expected TB cases by 0.00000506 units, when other variables remain constant. This is in line with [33], stating that participants over 55 have a higher risk of developing TB.

The results showed that the population density in each regency/city had a positive and significant effect on the number of TB cases, with a coefficient of 0.0000583. This implies that an increase in the population density by 1 person per km^2 can lead to a 0.0000576 rise in the log of expected TB cases. Similarly, a previous study [34] found that more densely populated living areas increase the risk of TB. Three significant variables ($\boldsymbol{X_{13}}$, $\boldsymbol{X_{14}}$, and $\boldsymbol{X_3}$) were then included in the subsequent analysis stage which used the extreme learning machine method.

3.5 Negative Binomial Mixed Extreme Neural Network

Using the data from the previous year (1-year lag), extreme learning FFNN was used to forecast the number of TB cases for the following year. The R programming language together with bbmle and neuralnet packages were used to perform this analysis.

Table 8 shows the architecture of the extreme learning FFNN model used in this study. The input consists of five neurons including regencies/cities, the number of TB cases in the preceding year (represented by $\boldsymbol{Y_{t-1}}$), and data on the three risk factors from the preceding year. The three risk factors included in the model were the predictor variables with a considerable impact on the number of TB cases according to the negative binomial GLMM model selected in the previous stage. These include the past values for population aged 45-64 ($\boldsymbol{X_{13,t-1}}$), population aged 65+ ($\boldsymbol{X_{14,t-1}}$), and population density ($\boldsymbol{X_{3,t-1}}$). The predicted output of FFNN model includes four neurons namely the number of TB cases for the year ($\boldsymbol{Y_{t-1}}$), as well as the values for the significant factors in the population aged 45-64 ($\boldsymbol{X_{13,t}}$), population aged 65+ (represented by $\boldsymbol{X_{14,t}}$), and population density (represented by $\boldsymbol{X_{3,t}}$). These features make the output of FFNN model suitable for forecasting occurrences in subsequent years. In addition, the number of hidden layers and neurons was determined through trial and error by comparing RMSE and $\mathbf{R^2}$ values.

Table 8. Architecture of FFNN Model Used

Input	5 neurons, consisting of the name of regency/city and 1 year lag (past) values of the Number of Tuberculosis Cases (Y_{t-1}), Population Aged 45-64 years old ($X_{13,t-1}$), Population Aged 65+ years old ($X_{14,t-1}$), 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, consisting of the Number of Tuberculosis Cases (Y_t), Population Aged 45-64 years old ($X_{13,t}$), Population Aged 65 years old ($X_{14,t}$), and Population Density ($X_{3,t}$)

The data used as input was subjected to standardization process before being included in the model. The dataset was split into two parts with the 80:20 ratio of train and test. Data from 2016 to 2020 was used to train the model, while those from 2021 were used for testing purposes by providing predictions using the previously trained model, reverting standardization process, and comparing the predicted results to the actual data. This process was repeated for each hidden layer architecture, using trial and error to determine the best model. Table 9 shows the evaluation results of FFNN model.

Table 9. FFNN Evaluation Metrics

Hidden Layer	Neuron	RMSE	R^2	Hidden Layer	Neuron	RMSE	R^2
1	1	0.615	0.650	1	6	0.455	0.855
	2	0.304	0.931		7	1.075	0.655
	3	0.278	0.936		8	0.866	0.624
	4	0.276	0.946		9	0.889	0.488
	5	1.208	0.617		10	24.179	0.162

According to Table 9, model 4 was found to be the most optimal FFNN. This model had the lowest RMSE and the highest R2 values. The architectural visualization of the best model is shown in Figure 7.

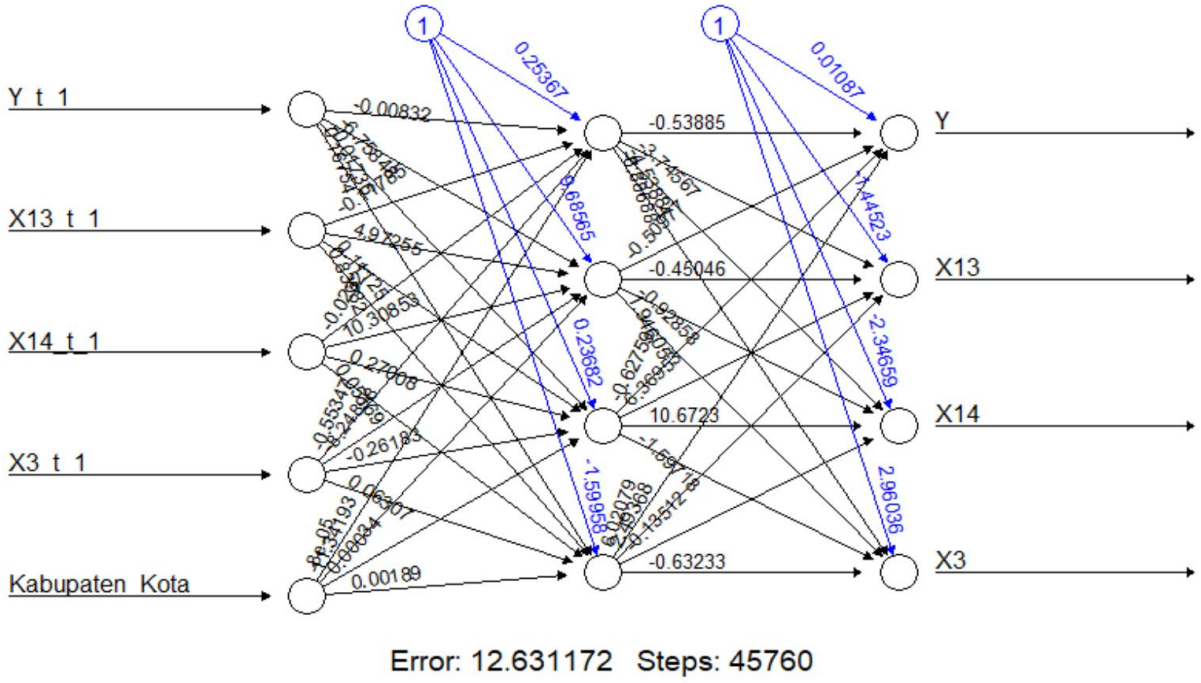

Figure 7. Architecture of FFNN Model

4. CONCLUSION

In conclusion, GLMM model, with the assumption of a negative binomial distribution for the response variable, could be used to explain the trend of TB cases in West Java between 2016 and 2021 based on the equation:

$$\hat{y} = g(\mu) = 6.26 + 0.00000208X_{Pop.Aged\ 45-64} + 0.00000506X_{Pop.Aged\ 65+} + 0.0000576X_{Pop.Density} + \boldsymbol{Z}_{Regency/City}\boldsymbol{b} \quad (14)$$

The results further confirm the significance of risk factors including the number of residents aged 45-64 years and above 65 years, as well as population density, all of which have a positive and significant impact on the number of TB cases. The integration of past values for the three significant factors as well as the number of TB cases can be used to forecast future occurrence. Extreme learning FFNN model was used to forecast the number of Tb cases (Y) as well as the three risk factors ($\boldsymbol{X_{13}}$, $\boldsymbol{X_{14}}$, $\boldsymbol{X_3}$) for the forthcoming year. In the testing phase, the best model discovered was FFNN (5-4-4) with RMSE = 0.276 and R^2 = 0.946.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to 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.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] WHO, Global tuberculosis report 2023, World Health Organization, (2023).

[2] D. Kibirige, N. Owarwo, A.P. Kyazze, et al. Prevalence, clinical features, and predictors of adrenal insufficiency in adults with tuberculosis or HIV: A systematic review and meta-analysis, Open Forum Infect. Dis. 11 (2024), ofae098. https://doi.org/10.1093/ofid/ofae098.

[3] WHO, Tuberculosis, (2023). https://www.who.int/news-room/fact-sheets/detail/tuberculosis.

[4] A. Odone, A.C. Crampin, V. Mwinuka, et al. Association between socioeconomic position and tuberculosis in a large population-based study in rural Malawi, PLoS ONE. 8 (2013), e77740. https://doi.org/10.1371/journal.pone.0077740.

[5] WHO, Global tuberculosis report 2022, World Health Organization, (2022).

[6] Dinas Kesehatan Provinsi Jawa Barat, Profil kesehatan Provinsi Jawa Barat tahun 2022, (2022).

[7] Kementerian Kesehatan Republik Indonesia, Profil kesehatan Indonesia 2022, (2022). https://kemkes.go.id/id/profil-kesehatan-indonesia-2022.

[8] D.J. Carter, P. Glaziou, K. Lönnroth, et al. The impact of social protection and poverty elimination on global tuberculosis incidence: a statistical modelling analysis of Sustainable Development Goal 1, Lancet Glob. Health. 6 (2018), e514–e522. https://doi.org/10.1016/s2214-109x(18)30195-5.

[9] A. Devi, J. Jalius, U. Kalsum, Pengaruh faktor sosial, ekonomi dan lingkungan terhadap kejadian tuberkulosis paru pada anak di kota jambi, J. Pembang. Berkelanjutan. 3 (2020), 1-6. https://doi.org/10.22437/jpb.v3i2.9655.

[10] R. Arisanti, R.S. Pontoh, S. Winarni, et al. Negative binomial mixed model neural network for modeling of pulmonary tuberculosis risk factors in West Java provinces, Int. J. Data Network Sci. 7 (2023), 981–994. https://doi.org/10.5267/j.ijdns.2023.6.007.

[11] R. Arisanti, I.M. Sumertajaya, K.A. Notodiputro, et al. The application of firth bias correction in variance components estimation of clustered random intercept binary model, J. Phys.: Conf. Ser. 1863 (2021), 012028. https://doi.org/10.1088/1742-6596/1863/1/012028.

[12] F. Nirmala, Kuntoro, H.B. Notobroto, Aplikasi, *General Linear Mixed Model* (GLMM) pada data longitudinal kadar trombosit demam berdarah dengue, J. Unair: J. Biometrika Kependudukan, 2 (2013), 131-139.

[13] C.E. McCulloch, Maximum likelihood algorithms for generalized linear mixed models, J. Amer. Stat. Assoc. 92 (1997), 162–170. https://doi.org/10.1080/01621459.1997.10473613.

[14] C.E. McCulloch, Introduction to generalized linear mixed models, UCLA: Statistical Consulting Group, (1997). https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models.

[15] R. Arisanti, K.A. Notodiputro, K. Sadik, et al. Bias reduction in estimating variance components of phytoplankton existence at Na Thap River based on logistics linear mixed models, IOP Conf. Ser.: Earth Environ. Sci. 58 (2017), 012014. https://doi.org/10.1088/1755-1315/58/1/012014.

[16] R.T. Simarmata, D. Ispriyanti, Penanganan overdispersi pada model regresi poisson menggunakan model regresi binomial negatif, Media Stat. 4 (2011), 95-104. https://doi.org/10.14710/medstat.4.2.95-104.

[17] A.S.N. Zaina, R.S. Pontoh, B. Tantular, Pemodelan dan pemetaan penyakit TB paru di kota bandung menggunakan geographically weighted negative binomial regression, in: Prosiding Seminar Nasional Statistika, pp. 2087-2590, (2021).

[18] R. Arisanti, I.M. Sumertajaya, K.A. Notodiputro, et al. Firth bias correction for estimating variance components of logistics linear mixed model using penalized quasi likelihood method, Commun. Math. Biol. Neurosci. 2020 (2020), 65. https://doi.org/10.28919/cmbn/4955.

[19] D.W. Hosmer, S. Lemeshow, Applied logistic regression, Wiley, 2000. https://doi.org/10.1002/0471722146.

[20] R. Sastri, Y. Setiadi, Generalized linear mixed model untuk data kematian bayi di Indonesia, Pusat Penelitian dan Pengabdian Masyarakat, Politeknik Statistika STIS, (2018).

[21] W.W. Stroup, Generalized linear mixed models: Modern concepts, methods and applications, CRC Press, (2016).

[22] C.J. Geyer, E.A. Thompson, Constrained Monte Carlo maximum likelihood for dependent data, J. R. Stat. Soc. Ser. B: Stat. Methodol. 54 (1992), 657-683. https://doi.org/10.1111/j.2517-6161.1992.tb01443.x.

[23] A.E. Gelfand, B.P. Carlin, Maximum likelihood estimation for constrained or missing data models, Defense Technical Information Center, Fort Belvoir, VA, 1993. https://doi.org/10.21236/ADA266563.

[24] L. Zhou, Q. Sun, S. Ding, et al. A machine-learning-based method for ship propulsion power prediction in ice, J. Mar. Sci. Eng. 11 (2023), 1381. https://doi.org/10.3390/jmse11071381.

[25] S.S. Haykin, Neural networks and learning machines, Prentice Hall, (2009).

[26] R.S. Pontoh, T. Toharudin, B.N. Ruchjana, et al. Bandung rainfall forecast and its relationship with Niño 3.4 using nonlinear autoregressive exogenous neural network, Atmosphere. 13 (2022), 302. https://doi.org/10.3390/atmos13020302.

[27] C.M. Bishop, Pattern recognition and machine learning, Springer, (2016).

[28] R. Arisanti, M.D. Puspita, Non-linear autoregressive neural network with exogenous variable in forecasting USD/IDR exchange rate, Commun. Math. Biol. Neurosci. 2022 (2022), 5. https://doi.org/10.28919/cmbn/6931.

[29] A.K. Jain, J. Mao, K.M. Mohiuddin, Artificial neural networks: a tutorial, Computer 29 (1996), 31–44. https://doi.org/10.1109/2.485891.

[30] R.S. Pontoh, S. Zahroh, H.R. Nurahman, et al. Applied of feed-forward neural network and facebook prophet model for train passengers forecasting, J. Phys.: Conf. Ser. 1776 (2021), 012057. https://doi.org/10.1088/1742-6596/1776/1/012057.

[31] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Computer Sci. 7 (2021), e623. https://doi.org/10.7717/peerj-cs.623.

[32] Dinas Kesehatan Provinsi Jawa Barat, Profil kesehatan Jawa Barat tahun 2021. (2022).

[33] L. Pangaribuan, K. Kristina, D. Perwitasari, et al. Faktor-faktor yang mempengaruhi kejadian tuberkulosis pada umur 15 tahun ke atas di Indonesia, Bul. Penelit. Sist. Kesehat. 23 (2020), 10-17. https://doi.org/10.22435/hsr.v23i1.2594.

[34] G.S. Prihanti, Sulistiyawati, I. Rahmawati, Analisis faktor risiko kejadian tuberkulosis paru, Saint. Med. 11 (2015), 127. https://doi.org/10.22219/sm.v11i2.4207.

[35] T.W. Utami, Analisis regresi binomial negatif untuk mengatasi overdispersion regresi poisson pada kasus demam berdarah dengue, J. Stat. Univ. Muhammadiyah Semarang, 1 (2013), 59-65.

[36] A.A. Yirga, S.F. Melesse, H.G. Mwambi, et al. Negative binomial mixed models for analyzing longitudinal CD4 count data, Sci. Rep. 10 (2020), 16742. https://doi.org/10.1038/s41598-020-73883-7.