Risk Factors for Lymphatic Filariasis in Endemic Areas of Papua Using Binary Logistic Regression Based on Synthetic Minority Over-sampling Technique

Sri Rohmanisa Simangunsong[1] **and Siskarossa Ika Oktora**[1*]
[1]Politeknik Statistika STIS, Jakarta, Indonesia
*Corresponding author: siskarossa@stis.ac.id

Received: 15 April 2024
Revised: 15 June 2024
Accepted: 1 July 2024

ABSTRACT –Neglected tropical diseases (NTDs), such as lymphatic filariasis (LF), are a significant issue in Indonesia. The high percentage of LF in Papua highlights the urgency of addressing LF in the area due to its devastating impact on the health and economy of the poor. Moreover, imbalanced outcome variable categories are a common issue in logistic regression analysis using medical data. One of the solutions to this problem is using Synthetic Minority Over-sampling Technique (SMOTE). Therefore, this study aims to provide an overview of LF cases in endemic areas of Papua and identify the factors that influence its occurrence using binary logistic regression analysis and the SMOTE method. The data utilized was the LF diagnosis status of individuals in endemic areas of Papua Province, Indonesia as contained in the Riset Kesehatan Dasar (Riskesdas) 2018. It was found that the SMOTE approach in binary logistic regression analysis can be used to address data imbalance. The following factors are significant: sex, age, occupation, education level, use of mosquito bite preventive measures, use of latrines for defecation, and participation in Mass Drug Administration (MDA).

Keywords– lymphatic filariasis, endemic areas, synthetic minority over-sampling technique, binary logistic regression

I. INTRODUCTION

One of the goals of Sustainable Development Goals (SDGs) is to ensure healthy lives and promote well-being for all at all ages. To achieve this goal, countries around the world set a target to end the epidemics of *acquired immunodeficiency syndrome* (AIDS), tuberculosis, malaria and neglected tropical diseases (NTDs), and combat hepatitis, waterborne diseases, and other communicable diseases by 2030. Among these diseases, NTDs continue to be overlooked [1], despite their significant impact on global health. NTDs in developing countries result in billions of dollars in healthcare costs, lost productivity, and reduced socioeconomic and educational opportunities [2]. NTDs are a group of diseases that disproportionately affect populations living in impoverished areas, particularly in tropical and subtropical regions. These diseases are also often stigmatized and result in social ostracization within affected communities. NTDs comprise 20 distinct diseases. These diseases are caused by agents, including bacteria, viruses, protozoa, parasitic worms, and external parasites. They can be spread through water, soil, food, droplets, direct contact with patients, and vectors, such as mosquitoes and flies. To achieve inclusive development, it is important to give more attention to tackling NTDs. This will ensure that no individual is left behind by the principle of the SDGs.

Indonesia is ranked third in the world for the highest number of people requiring intervention against NTDs, with 79.9 million individuals [3]. According to the Kementerian Kesehatan Republik Indonesia (Kemenkes RI), one of the NTDs that is still a public health problem in Indonesia is lymphatic filariasis (LF). The mapping of endemic areas revealed that LF is endemic in 236 districts across 28 provinces in Indonesia. LF endemic areas are determined based on the results of a baseline survey of the percentage of microfilariae that shows a microfilariae rate of 1% or higher [4].

LF is an infectious disease caused by filarial worms that are transmitted by mosquitoes. This disease can damage the lymphatic system, resulting in swelling of the hands, feet, breasts, and scrotum. Based on data from Riset Kesehatan Dasar (Riskesdas), the prevalence of LF in Indonesia has increased from 0.05% in 2007 to 0.8% in 2018. In 2018, Papua Province had the second-highest percentage of LF in Indonesia [5]. Although Papua did not have the highest percentage of cases, it has the highest number of endemic areas in Indonesia. In Papua, there are 23 districts, which account for 79.3% of the total districts in the region, that are considered endemic areas and are the primary focus of the national health program.

Kemenkes RI aims to reduce the microfilariae rate to less than 1% in all LF endemic areas and achieve LF elimination status in 190 endemic districts by 2024. By the program of the Kemenkes RI, Dinas Kesehatan (Dinkes) Papua has also set a goal of eliminating LF, as outlined in Rencana Strategis Dinkes Papua 2019-2023. However, the Papua Province has consistently underperformed in this program, achieving only between 25% to 31.25% of the elimination target set by Dinkes Papua each year.

In 2022, there were 8,742 cases of chronic LF spread across 34 provinces in Indonesia [6]. In Papua, there were 3,615 cases of chronic LF. Since 2017, Papua has had the highest number of chronic cases in Indonesia. The existence of these chronic cases indicates that the infection has been present in the Papua population for a long period. If left untreated, this

DOI: 10.12962/j27213862.v7i2.20283

can lead to a higher disease burden and an increased risk of transmission to non-endemic areas [7]. Chronic LF do not significantly impact mortality rates in Papua. Since 2019, only 14 individuals with chronic cases passed away. However, in the chronic stage, the swelling experienced by most patients cannot be cured and persists even after completing treatment [8]. Chronic cases with increased swelling may result in permanent disabilities [9].

According to the World Health Organization (WHO), NTDs, including LF, tend to affect poorer areas. This statement is pertinent to the conditions in Papua, the province with the highest poverty rate in Indonesia every year. Chronic LF in Papua have the potential to exacerbate poverty by negatively impacting the ability of patients to engage fully in economic activities. This is because the condition of the patients can result in job loss or a reduction in work hours. In addition, the stigma associated with the disease prevents individuals from starting businesses, such as trading [10]. Therefore, a high incidence of LF in a community with a high poverty rate can worsen the economic burden of the area.

Regression analysis was used to explain the relationship between the outcome and predictor variables. If the outcome variable is categorical data, the logit regression model can be used to explain the relationship between the outcome and predictor variables [11]. However, logistic regression analysis using medical data often faces the problem of imbalanced outcome variable categories [12]. The distribution of the estimated parameters is influenced by the percentage of 0/1 in the data used to fit a logistic regression model. An extreme proportion of zeros (or ones) in the data tends to increase the variability of the estimated parameters [13]. To address this issue, there are two main approaches: over-sampling and under-sampling. Under-sampling decreases the sample of the majority class while over-sampling increases the sample of the minority class. The goal of both strategies is to achieve a balanced distribution of classes [14]. However, the use of random under-sampling and over-sampling techniques has several drawbacks. Random under-sampling may eliminate important examples, while random over-sampling can lead to overfitting [15]. To circumvent overfitting and expand the decision region of minority class examples, [16] have introduced the Synthetic Minority Over-sampling Technique (SMOTE), a novel technique to generate synthetic examples. The minority class can be over-sampled by generating synthetic examples along the line segments that connect each minority class sample to its nearest neighbors.

Earlier studies has shown that SMOTE can produce synthetic samples that are useful for classifying unbalanced data and increasing the precision of predicting the efficacy of cardiac implantable electronic device infection [17]. Meanwhile, another study found that the model using the SMOTE approach was more accurate in classifying the poverty status of rural and urban households in East Java Province [18]. The SMOTE method has also been shown to be effective in addressing inferential and classification quality issues. SMOTE model produced more statistically significant variables with little differences in estimated regression coefficients.

Research on LF in Papua Province has been conducted in Yahukimo using only descriptive analysis [19]. Another study, conducted in Sarmi, used inferential analysis but did not address the issue of imbalanced outcome variable category [20]. Therefore, this study was conducted to overview LF patients in endemic areas of Papua Province and to identify the variables that affect it using binary logistic regression based on the SMOTE approach.

II. LITERATURE REVIEW

A. Lymphatic Filariasis (LF)

LF is a chronic infectious disease caused by filarial worms that attack the lymphatic system [4]. LF in Indonesia is caused by three species of filarial worms, namely *Wuchereria bancrofti*, *Brugia malayi*, and *Brugia timori*. In Indonesia, 23 species of mosquitoes belonging to 5 genus (*Mansonia, Anopheles, Culex, Aedes,* and *Armigeres*) have been identified as vectors of LF. LF is typically endemic in lowland areas, particularly in rural, coastal, inland, rice fields, swamps, and forests. The diagnostic method for determining the presence or absence of filarial worms in the blood depends on the type of filarial worm. There are two methods: the Immunochromatographic Test (ICT)/Rapid Test for *Wuchereria bancrofti*, which detects the presence of *Wuchereria bancrofti* antigens in the blood, and the Rapid Test for *Brugia*, which detects the presence of *Brugia malayi* or *Brugia timori* antibodies.

B. Epidemiology Triad

Diseases arise from the interaction of agents, hosts, and the environment. The onset of infectious diseases depends on the specific situation and the roles played by each of these factors [21]. The patterns of infectious diseases are shaped by various factors that influence the contact between infectious agents and vulnerable hosts. Disease spread depends on factors, such as the excretion of agents, environmental conditions affecting the survival of agents, entry points of agents into hosts, and the presence of alternative reservoirs. The availability of susceptible hosts is determined by immunity levels from previous infections, population mobility, and interactions. A framework was presented by [22] for identifying classifications of agents, hosts, and environmental factors that are relevant when investigating the factors that influence the occurrence of disease in a population. Based on that, age, sex, prior immunologic experience, human behavior, occupation, urbanization, and economic development are some of the factors that influence the occurrence of a disease.

C. Binary Logistic Regression

Logistic regression is used to describe the relationship between the outcome variable and one or more predictor variables and to analyze data with discrete outcome variables [23]. Several distribution functions have been proposed for analyzing dichotomous outcome variables, which are coded as 0 or 1, representing the absence or the presence of the characteristic, respectively. However, the logistic distribution is the most recommended due to its mathematical flexibility and ease of use, along with its ability to generate clinically meaningful estimates of effect. Consider a collection of p predictor variables denoted by the vector $\boldsymbol{x}' = (x_1, x_2, \dots, x_p)$, the general form of the logistic regression model is:

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}. \qquad (1)$$

A transformation of $\pi(\boldsymbol{x})$ in logistic regression is the logit transformation. This transformation is defined, in terms of $\pi(\boldsymbol{x})$, as:

$$g(\boldsymbol{x}) = \ln\left[\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \qquad (2)$$

To fit the logistic regression model to a set of data, it is necessary to estimate the values of the unknown parameters $\boldsymbol{\beta}$. The method of estimation that results in the least squares function under the linear regression model (when the error terms are normally distributed) is called maximum likelihood. This method requires the construction of the likelihood function first, which expresses the probability of the observed data as a function of the unknown parameters. As the observations are assumed to be independent, the likelihood function is obtained as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i}. \qquad (3)$$

The maximum likelihood estimators of the parameters are the values that maximize this function. However, it is easier mathematically to work with the log of equation. This expression, the log-likelihood, is defined as:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}. \qquad (4)$$

The value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ is found by differentiating $L(\boldsymbol{\beta})$and setting the resulting expressions equal to zero. The expressions are nonlinear in $\boldsymbol{\beta}$, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into logistic regression software.

D. Synthetic Minority Over-sampling Technique (SMOTE)

The SMOTE algorithm employs an over-sampling approach to rebalance the original training set. In contrast to the straightforward replication of minority class instances, the fundamental concept of SMOTE entails the introduction of synthetic examples. The new data is generated through interpolation between multiple minority class instances within a specified neighborhood. Consequently, the procedure is regarded as being focused on the "feature space" rather than on the "data space." In other words, the algorithm is based on the values of the features and their relationship, rather than on the data points as a whole. This has also led to a detailed examination of the theoretical relationship between the original and synthetic instances, including the data dimensionality [24].

The SMOTE resampling method has been demonstrated to markedly enhance the performance of classification algorithms, thereby facilitating the development of more effective solutions for addressing class imbalance issues [25]. Furthermore, the issue of learning from class-imbalanced data remains a prevalent and complex challenge in the field of supervised learning. Standard classification algorithms are designed to operate within a balanced class distribution, making it difficult to apply them to data sets with an imbalanced class ratio. While various strategies have been proposed to address this issue, those that generate artificial data, such as SMOTE, to achieve a balanced class distribution tend to be more versatile than modifications to the classification algorithm [26]. Furthermore, linear interpolation between selected minority examples can prevent the generation of redundant and replicated examples, as observed in random over-sampling algorithms. Consequently, SMOTE is an effective approach for overcoming overfitting problems and enhancing the learning task of most classifiers [27].

On numerical data, these methods are used to create synthetic samples:

1. Calculate the difference between the feature vector being analyzed and its closest neighbor.
2. Multiply the difference by a random number between 0 and 1.
3. Add the result to the feature vector being considered.

SMOTE can also be extended to nominal features by computing the nearest neighbors using a modified Value Difference Metric. The VDM takes into consideration the overall similarity of classification of all instances for each possible value of each feature [28]. The distance between two values for a specific feature is defined as:

$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k . \tag{5}$$

In the equation, V_1 and V_2 represent the two corresponding feature values. C_1 represents the total number of instances in which feature value V_1 was observed, whereas C_{1i} denotes the number of instances in which feature value V_1 was observed for class i. A similar convention can be applied to C_{2i} and C_2. The constant k is typically set to the value of 1. Then, the total distance between two instances is given by:

$$\Delta(X, Y) = w_x w_y \sum_{i=1}^{N} \delta(x_i, y_i)^r . \tag{6}$$

where X and Y represent two instances. The variables x_i and y_i are values of the i^{th} feature for X and Y, where each example has N features. w_x and w_y are negligible weights. r = 1 yields the Manhattan distance, and r = 2 yields the Euclidean distance [28].

III. METHODOLOGY

This study utilizes secondary data, specifically raw data of Riskesdas 2018 conducted by Kemenkes RI. This study focuses on individuals aged 10 years and above, totaling 14,821 participants in 23 districts of Papua Province that are endemic for LF, including Merauke, Jayawijaya, Jayapura, Nabire, Kepulauan Yapen, Biak Numfor, Puncak Jaya, Mimika, Boven Digoel, Mappi, Asmat, Yahukimo, Pegunungan Bintang, Sarmi, Keerom, Waropen, Supiori, Mamberamo Raya, Nduga, Mamberamo Tengah, Puncak, Intan Jaya, and Kota Jayapura. The outcome variable was the diagnosis status of LF, categorized as either diagnosed or undiagnosed. The predictor variables included sex, age, occupation, education level, use of mosquito bite prevention measures, use of latrines for defecation, and participation in Mass Drug Administration (MDA), as shown in Table 1.

Table 1 Categories of Variables Used

Variable Notation	Variable Description	Categories
Y	LF diagnosis status	1 = diagnosed 0 = undiagnosed*
X_1	sex	1 = male 0 = female*
X_2	age	numeric variable
X_3	occupation	1 = farmers 0 = non-farmers*
X_4	education level	1 = middle school and lower 0 = high school and higher*
X_5	use of mosquito bite preventive measures	1 = no 0 = yes*
X_6	use of latrines for defecation	1 = no 0 = yes*
X_7	participation in MDA	1 = no 0 = yes*

*reference categories

This study utilized a descriptive-analytical approach, employing crosstabulation and binary logistic regression with and without SMOTE methods for inferential analysis. The stages of the SMOTE method applied in this research are as follows:

1. Calculate the distance between observations based on the value of the predictor variable. All variables used in this study are categorical variables, therefore the inter-observation distance used is VDM, as stated in equation (6).
2. Determine the number of nearest neighbors used. The number of nearest neighbors used in this study is five.
3. Randomly select a single individual diagnosed with LF.
4. Determine the five nearest neighbors of the selected individual based on the VDM measure. The selected observation is the one with the smallest VDM value against the target observation. Subsequently, a random selection is made to ascertain whether the values are identical.
5. Generate new data by assigning values to each predictor variable. In the case of categorical data, the value of each predictor variable is the mode value of its five nearest neighbors.
6. Repeat steps 3-5 until a total of 14,462 new observations of individuals diagnosed with LF were obtained.

To investigate whether the strength and direction of the relationship between behavioral variables and LF diagnosis status could be affected by education level, interaction terms were added to the regression model. Specifically, the interaction terms were from education level with mosquito bite prevention behavior and latrine use behavior. To evaluate the model's goodness of fit, this study utilized the classification table and the area under the receiver operating characteristic curve (AUC-ROC) [23].

IV. RESULTS AND DISCUSSIONS

A. Descriptive Analysis

According to the Riskesdas 2018 shown in Figure 1, 2,42% of individuals were diagnosed with LF, while 97,58% were not diagnosed with the disease. This number is relatively high because it is still above 1%, which means that the LF elimination in the endemic areas of Papua has not been achieved. Papua Province boasts Indonesia's largest natural forest area, covering 25 million ha [29]. In addition, Papua Province has 670 kilometers of river that flows south of the Foja Mountains and crosses six districts. Forests and watersheds provide suitable habitats for mosquitoes that transmit LF [30], [31].

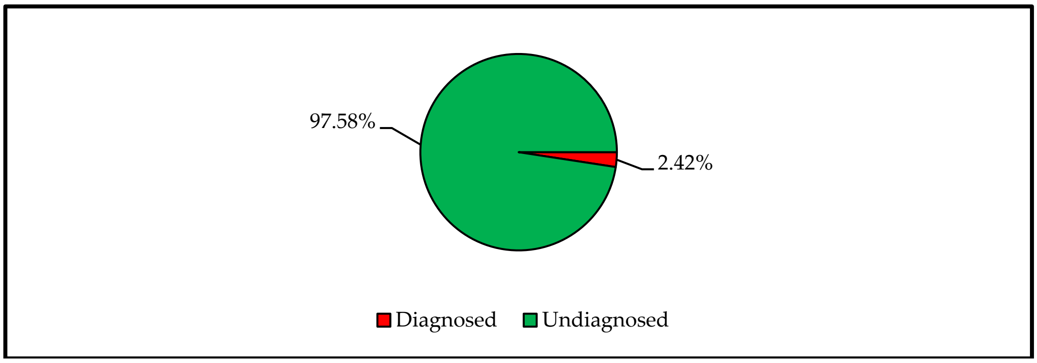

Figure 1 Percentage of Individuals by LF Diagnosis Status

Table 2 shows the characteristics of individuals and the percentage of LF based on each predictor variable. The proportion of females to males was nearly equivalent. The percentage of females with LF was observed to be higher than males. Furthermore, the proportion of individuals engaged in farming was less than that of individuals not engaged in farming. People who did not work in agriculture had a higher percentage of LF than people who did. Further investigation revealed that 96.35% of individuals engaged in agricultural work reside in rural areas.

Table 2 LF Diagnosis Status according to Each Variable

Variable	Categories	Percentage	LF Diagnosis (%)	
			Diagnosed	Undiagnosed
sex	male	49.29	2.19	97.81
	female	50.71	2.65	97.35
occupation	farmers	28.53	1.58	98.42
	non-farmers	71.47	2.76	97.24
education level	middle school or lower	70.89	2.40	97.60
	high school or higher	29.11	2.48	97.52
use of mosquito bite preventive measures	no	21.29	1.55	98.45
	yes	78.71	2.66	97.34
use of latrines for defecation	no	29.88	3.14	96.86
	yes	70.12	2.12	97.88
participation in MDA	no	62.63	0.95	99.05
	yes	37.37	4.89	95.11

DOI: 10.12962/j27213862.v7i2.20283

In the aspect of education level, the majority of individuals had not received education in accordance with the 12-year compulsory education program. Furthermore, only 20.40 percent of individuals were found to be under 19 years old. This indicates that the majority of the individual had reached the age for completion of secondary education, but did not complete their studies at this level. Those with high school education or higher had a higher percentage of LF compared to those with middle school education or lower. It was discovered that 56.72% of individuals with high school education or higher lived in areas where the percentage of cases was above 2%.

The percentage of individuals who used mosquito bite preventive measures was greater than the percentage of individuals who did not. In addition, 62.64 percent used mosquito nets when sleeping, 26.64 percent used mosquito repellent, and 5.84 percent used electric mosquito repellents. The percentage of individuals who used mosquito bite preventive measures with LF was higher than those who did not. Furthermore, there were still individuals who did not use latrines for defecation. However, the majority of individuals used latrines for defecation. A greater proportion of individuals diagnosed with LF are found among those who defecate in the open than among those who defecate in latrines.

Finally, MDA coverage remained low. The percentage of individuals participating in MDA was 37.37 percent. Meanwhile, Kemenkes RI recommends the implementation of MDA for five consecutive years with a minimum coverage of 65 percent to break the chain of LF transmission. In addition, it was found that the people who participated in MDA had a higher percentage of LF than people who did not.

Moreover, a three-way analysis is conducted between the behavioral variables, specifically use of mosquito bite preventive measures and use of latrines for defecation against LF, with consideration of the education level to ascertain whether there are differences in the effects of the two behaviors at varying education levels. Table 3 shows the phenomenon of Simpson's paradox, where the marginal association is in opposition to the conditional association. In the partial/conditional condition, it was found that the percentage of LF in individuals who did not use mosquito bite preventive measures was greater than individuals who did. This is observed in the group of individuals with high school education or higher. The observed result differs from that obtained for all individuals without distinguishing their education group, which is at the marginal condition as shown in Table 2.

Table 3 Three-way Contingency Table between Use of Moquito Bite Preventive Measures, LF Diagnosis Status, and Education Level

Education Level	Use of Mosquito Bite Preventive Measures	LF Diagnosis (%)	
		Diagnosed	Undiagnosed
middle school or lower	no	1.01	98.99
	yes	2.81	97.19
high school or higher	no	3.24	96.76
	yes	2.31	97.69

Table 4 shows that the percentage of LF was higher among individuals who did not use latrine for defecation than among those who did, both among individuals with high school education or higher and among those with middle school education or lower (partial condition). This finding is consistent with the results observed in the overall population or the marginal condition.

Table 4 Three-way Contingency Table between Use of Latrine for Defecation, LF Diagnosis Status, and Education Level

Education Level	Use of Latrines for Defecation	LF Diagnosis (%)	
		Diagnosed	Undiagnosed
middle school or lower	no	3.14	96.86
	yes	1.95	98.05
high school or higher	no	3.16	96.84
	yes	2.40	97.60

B. Inferential Analysis

The identification of factors that influence the occurrence of LF is accomplished by creating a binary logistic regression model. Table 5 displays the parameter estimation results of the binary logistic regression model.

Table 5 Parameter Estimation and Variable Significance for Model without SMOTE

Variable	Category	Estimate	Standard Error	p-value	Odds Ratio
intercept		-2.852	0.185	0.000**	
sex (X_1)	male	-0.136	0.109	0.213	0.873
	female*				
age (X_2)	-	-0.003	0.004	0.486	0.997
occupation (X_3)	farmers	-0.649	0.154	0.000**	0.523
	non-farmers*				
education level (X_4)	middle school or lower	-0.059	0.153	0.699	0.943
	high school or higher*				
use of mosquito bite preventive measures (X_5)	no	0.473	0.236	0.045**	1.605
	yes*				
use of latrines for defecation (X_6)	no	0.326	0.297	0.273	1.385
	yes*				
participation in MDA (X_7)	no	-1.617	0.126	0.000**	0.199
	yes*				
interaction term 1 (use of mosquito bite prevention measures and education level) (X_8)		-1.266	0.322	0.000**	0.282
interaction term 2 (use of latrines and education level) (X_9)		0.480	0.323	0.138	1.616

*reference categories
**significant at $\alpha = 5\%$

According to the classification table in Table 6, the model is unable to accurately classify the success category. This indicates an imbalance problem in the category of the outcome variable used. Previous studies found that unbalanced data can lead to biased conclusions and errors in identifying significant variables [13]. Thus, it is advisable to balance the data during the process of inferential analysis. A binary logistic regression model was formed using the SMOTE approach.

Table 6 Classification Table for Model without SMOTE

Classified	Observed		Total
	Diagnosed = 1	Undiagnosed = 0	
Diagnosed = 1	0	0	0
Undiagnosed = 0	359	14,462	14,861
Total	359	14,462	14,861

sensitivity = 0%; specificity = 100%; AUC = 0.7478

Table 7 displays the results of parameter estimation for the binary logistic regression model using the SMOTE approach. According to the classification table in Table 8, the imbalance problem has been successfully resolved.

Table 5 and Table 7 display the difference in the number of significant predictor variables between the model with and without SMOTE. The findings indicate that the model with SMOTE does not yield significantly different regression parameters compared to the model without SMOTE, yet exhibits a smaller standard error of estimation. This suggests enhanced parameter estimation [18]. The enhancement of the model with SMOTE in the estimation standard error resulted in an increase in the number of variables that significantly influenced the LF diagnosis status of individuals. Furthermore, logistic regression without and with the SMOTE approach has the same classification performance based on the AUC value. However, the logistic regression without SMOTE appears to be less effective in terms of sensitivity than the logistic regression with SMOTE approach, yet more effective in terms of specificity. This result is attributed to the inherent bias of the model towards the majority (negative) class when the data set is unbalanced. This is a challenge that the regression model without SMOTE is unable to overcome.

Therefore, the binary logistic regression model formed to explain the diagnosis status of LF is:

$$\hat{g}(x) = 1.022 - 0.175X_1 - 0.002X_2 - 0.828X_3 - 0.089X_4 + 0.297X_5 + 0.316X_6 - 1.921X_7 - 1.637X_8 + 0.569X_9 \tag{7}$$

Table 7 Parameter Estimation and Variable Significance for Model with SMOTE

Variable	Category	Estimate	Standard Error	p-value	Odds Ratio
intercept		1.022	0.048	0.000**	
sex (X_1)	male	-0.175	0.029	0.000**	0.839
	female*				
age (X_2)	-	-0.002	0.001	0.038**	0.998
occupation (X_3)	farmers	-0.828	0.040	0.000**	0.437
	non-farmers*				
education level (X_4)	middle school or lower	-0.089	0.038	0.019**	0.915
	high school or higher*				
use of mosquito bite preventive measures (X_5)	no	0.297	0.066	0.000**	1.346
	yes*				
use of latrines for defecation (X_6)	no	0.316	0.086	0.000**	1.371
	yes*				
participation in MDA (X_7)	no	-1.921	0.029	0.000**	0.146
	yes*				
interaction term 1 (use of mosquito bite prevention measures and education level) (X_8)		-1.637	0.087	0.000**	0.195
interaction term 2 (use of latrines and education level) (X_9)		0.569	0.093	0.000**	1.767

*reference categories
**significant at $\alpha = 5\%$

Table 8 Classification Table for Model with SMOTE

Classified	Observed		Total
	Diagnosed = 1	Undiagnosed = 0	
Diagnosed = 1	246	4,312	4,558
Undiagnosed = 0	113	10,150	10,263
Total	359	14,462	14,861

sensitivity = 68.80%; specificity = 69.35%; AUC = 0.7463

Based on the test results, the values of $G^2 = 7{,}456.892 > \chi^2_{0.05(9)} = 16.919$ and p-value = 0.000 are obtained, thereby causing H_0 to be rejected. It can be concluded that there is at least one predictor variable that significantly affects the diagnosis of LF in individuals living in endemic areas of Papua Province. According to Table 7, with a significance level of 5%, it can be concluded that sex, age, occupation, education level, use of mosquito bite prevention measures, use of latrines for defecation, and participation in MDA are the predictor variables that have a significant effect on the status of individual LF diagnosis in endemic areas of Papua. Furthermore, it was discovered that the interaction terms between education level with both preventive behavior and latrine use were significant.

The interpretation of the results from the estimation of the binary logistic regression model is based on the odds ratio of each variable. Males were less likely to be infected with LF than females. The study revealed that females were 1.192 (1/0.839) more likely to be diagnosed with LF in comparison to males. This finding is consistent with [32], but contradicts the results of [33]. On the other hand, [19] stated that both sexes are equally exposed to mosquitoes based on their daily habits, such as nighttime activities. Furthermore, the odds ratio of age was 0.998. Consequently, if the age of an individual was reduced by one year, the probability of being diagnosed with LF increased by 1.002 (1/0.998) times. According to [32], adults who spend time outdoors are more likely to be exposed to mosquitoes. Conversely, the elderly tend to have less interaction with mosquitoes as they spend more time indoors [34].

Individuals who worked as farmers had a lower risk of LF infection compared to those who did not work in farming. Individuals who did not engage in agricultural work were 2.288 (1/0.437) times more likely being diagnosed with LF than that of individuals who did. These finding contradicts the research conducted by [35] in the Republic of Congo. However, in rural areas where farmers reside, the population density is typically lower than in urban areas. In line with this, [36] discovered that high LF cases are influenced by population density. Furthermore, the odds ratio of the education level was 0.915. It was determined that individuals with high school education or higher were 1.093 times (1/0.915) more likely to be diagnosed with LF compared to individuals with middle school education or lower. This was observed in both individuals who used mosquito bite preventive measures and individuals who used latrines for defecation. This

DOI: 10.12962/j27213862.v7i2.20283

result is consistent with those of [37], who examined individual and contextual factors associated with adult malaria cases in Eastern Indonesia. The study demonstrated that adults in Papua with a higher educational background were more likely to be diagnosed with malaria. This indicates that residing in an area with high transmission rates will also elevate the risk of contracting the disease, despite individuals having higher educational levels.

The study revealed that individuals who did not use moquito bite preventive measures were 1.346 times more likely to be diagnosed with LF compared to individuals who did. This finding was observed in individuals with high school education or higher. This finding is consistent with the results of research conducted in Yahukimo District [19]. The results demonstrated that individuals with a negative attitude toward LF prevention exhibited a higher likelihood of infection compared to those with a positive attitude. Furthermore, the study revealed that individuals who did not use latrines were 1.371 times more likely to be diagnosed with LF compared to those who use latrines. This finding was observed in individuals with high school education or higher. Finally, the odds ratio of the participation in MDA was obtained as 0.146. This indicates that the tendency of individuals who participated in MDA to be diagnosed with LF was 6.489 times greater than that of individuals who did not participate in MDA. One of the challenges in implementing MDA in Papua is encouraging community participation, given the lack of observable symptoms in local communities [38]. Therefore, individuals who were already infected were more likely to participate in the MDA due to the presence of symptoms or awareness of their infection status.

The interaction between preventive behavior against mosquito bites and education level is significant. This implies that the effect of preventive behavior on LF diagnosis depended on the education level of the individual. In individuals with middle school education or lower, the odds ratio of the use of mosquito bite preventive measures was obtained as 0.262 (0.195 × 1.346). This indicates that the propensity of individuals who used mosquito bite preventive measures was 5.128 (1/0.262) times greater than that of individuals who did not. Individuals with lower level of education are less likely to engage in consistent preventive behaviors. Additionally, they are less knowledgeable about bed net maintenance. The efficacy of insecticide-treated bed nets can be diminished by inadequate maintenance. For instance, washing mosquito nets with laundry soap and using bleach in conjunction with scrubbing can reduce the insecticide content in the nylon fibers of the nets, resulting in diminished effectiveness and potential resistance in mosquitoes [39].

The interaction between use of latrine for defecation and education level was significant, indicating that the effect of defecation behavior on LF diagnosis depended on the education level of the individual. In individuals with middle school education or lower, the odds ratio of the use of latrines for defecation was 2.423 (1.767 × 1.371). This value indicates that the tendency of individuals who engaged in defecation behavior outside of a toilet was 2.423 times greater than individuals who engaged in defecation behavior inside a toilet. This suggests that a lack of education may contribute to an increased influence of non-latrine defecation practices on LF infection. The effect of education on defecation behavior was different from the effect of education on mosquito bite prevention behavior.

V. CONCLUSIONS AND SUGGESTIONS

The evaluation results indicate that the binary regression analysis of the SMOTE approach can be used to address data imbalance. Furthermore, the logistic regression model with SMOTE exhibits a greater number of significant predictor variables due to the smaller standard error of estimation. Binary logistic regression analysis with SMOTE identified variables affecting the diagnosis of LF in endemic areas of Papua are included sex, age, occupation, education, mosquito bite prevention, latrine use, and participation in the MDA. Interactions were found between use of mosquito bite preventive measures and use of latrines for defecation with education. The results indicated that the tendency of LF was higher among females, younger individuals, non-farmers, and those with high school education or higher. Use of mosquito bite preventive measures was associated with a decreased tendency of LF. Individuals who did not use latrines for defecation and who participated in MDA were also susceptible to LF.

The study revealed interactions between education level and both mosquito bite prevention practices and the use of latrine behaviors. Thus, the government may consider incorporating health education and sanitation-related materials into the school curriculum in endemic areas to enhance the delivery of health education within the existing formal education system. Furthermore, the government may also assist in the construction of latrines in areas with substandard housing.

REFERENCES

[1] WHO, “Neglected Tropical Diseases,” 2023. https://www.who.int/news-room/questions-and-answers/item/neglected-tropical-diseases (accessed 19 Oktober 2023).

[2] W. K. Redekop *et al.*, “The Socioeconomic Benefit to Individuals of Achieving the 2020 Targets for Five Preventive Chemotherapy Neglected Tropical Diseases,” *PLoS Negl. Trop. Dis.*, vol. 11, no. 1, hal. 1–27, 2017, doi: 10.1371/journal.pntd.0005289.

[3] WHO, “Reported Number of People Requiring Interventions Against NTDs,” 2021. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/reported-number-of-people-requiring-interventions-against-ntds (accessed 26 November 2023).

[4] Kemenkes RI, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 94 Tahun 2014 tentang Penanggulangan Filariasis*. 2014.

[5] Kemenkes RI, “Laporan Nasional Riskesdas 2018,” Jakarta, 2019.

[6] Kemenkes RI, “Profil Kesehatan Indonesia 2022,” Jakarta, 2023.

[7] B. Widjanarko, L. D. Saraswati, dan P. Ginandjar, “Perceived Threat and Benefit toward Community Compliance of Filariasis’

Mass Drug Administration in Pekalongan District, Indonesia," *Risk Manag. Healthc. Policy*, vol. 11, hal. 189–197, 2018, doi: 10.2147/RMHP.S172860.

[8] Kemenkes RI, "Penyakit Kaki Gajah Masih Ada di Indonesia, Kenali agar Bisa Mencegahnya," *Redaksi Sehat Negeriku*, 2018. https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20180926/5028023/penyakit-kaki-gajah-masih-ada-indonesia-kenali-agar-mencegahnya/ (accessed 26 November 2023).

[9] Kemenkes RI, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 21 Tahun 2020 tentang Rencana Strategis Kementerian Kesehatan Tahun 2020-2024*. 2020.

[10] S. O. Asiedu, A. Kwarteng, E. K. A. Amewu, P. Kini, B. C. Aglomasa, dan J. B. Forkuor, "Financial Burden Impact Quality of Life among Lymphatic Filariasis Patients," *BMC Public Health*, vol. 21, no. 1, hal. 1–10, 2021, doi: 10.1186/s12889-021-10170-8.

[11] D. N. Gujarati dan D. C. Porter, *Basic Econometrics*, 5th ed. New York: McGraw-Hill Irwin, 2009.

[12] P. K. Kondeti *et al.*, "Applications of Machine Learning Techniques to Predict Filariasis Using Socio-Economic Factors," *Epidemiol. Infect.*, vol. 147, hal. e260, 2019, doi: 10.1017/S0950268819001481.

[13] C. Salas-Eljatib, A. Fuentes-Ramirez, T. G. Gregoire, A. Altamirano, dan V. Yaitul, "A study on the Effects of Unbalanced Data when Fitting Logistic Regression Models in Ecology," *Ecol. Indic.*, vol. 85, no. October 2017, hal. 502–508, 2018, doi: 10.1016/j.ecolind.2017.10.030.

[14] M. Nakamura, Y. Kajiwara, A. Otsuka, dan H. Kimura, "LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for Biomedical Data," *BioData Min.*, vol. 6, no. 1, hal. 1–10, 2013, doi: 10.1186/1756-0381-6-16.

[15] N. V Chawla, "Data Mining for Imbalanced Datasets: An Overview," *Data Min. Knowl. Discov. Handb.*, 2010, doi: 10.1007/978-0-387-09823-4.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, dan W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, hal. 321–357, 2002, [Online]. Available at: http://arxiv.org/abs/2003.09788

[17] X. F. Feng, L. C. Yang, L. Z. Tan, dan Y. G. Li, "Risk Factor Analysis of Device-Related Infections: Value of Re-sampling Method on the Real-World Imbalanced Dataset," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, hal. 1–8, 2019, doi: 10.1186/s12911-019-0899-4.

[18] F. R. A. Pratama dan S. I. Oktora, "Synthetic Minority Over-sampling Technique (SMOTE) for Handling Imbalanced Data in Poverty Classification," *Stat. J. IAOS*, vol. 39, no. 1, hal. 233–239, 2023, doi: 10.3233/SJI-220080.

[19] M. K. Puhili, A. L. Rantetampang, B. Sandjaya, dan A. Mallongi, "The Factors Affecting with Filariasis Incidence at Dekai Public Health Regional Yahukimo District," *Int. J. Sci. Healthc. Res.*, vol. 3, no. 4, hal. 234–244, 2018, [Online]. Available at: www.ijshr.com

[20] M. Sipayung, C. U. Wahjuni, dan S. Devy, "Pengaruh Lingkungan Biologi dan Upaya Pelayanan Kesehatan terhadap Kejadian Filariasis Limfatik di Kabupaten Sarmi," *J. Berk. Epidemiol.*, vol. 2, no. 2, hal. 263, 2018, [Online]. Available at: https://e-journal.unair.ac.id/index.php/JBE/article/viewFile/181/51

[21] J. E. Gordon dan H. Le Riche, "The Epidemiologic Method Applied to Nutrition," *Am. J. Med. Sci.*, vol. 219, hal. 321–345, 1950.

[22] D. E. Lilienfeld dan P. D. Stolley, *Foundations of Epidemiology*, 3rd ed. New York: Oxford University Press, 1994.

[23] D. W. Hosmer, S. Lemeshow, dan R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. New Jersey: John Wiley & Sons, 2013.

[24] A. Fernández, S. García, F. Herrera, dan N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, hal. 863–905, 2018, doi: 10.1613/jair.1.11192.

[25] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, hal. 3161–3172, 2020, doi: 10.30534/ijatcse/2020/104932020.

[26] G. Douzas, F. Bacao, dan F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, hal. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[27] N. A. Azhar, M. S. Mohd Pozi, A. M. Din, dan A. Jatowt, "An investigation of SMOTE based methods for imbalanced datasets with data complexity analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 7, hal. 6651–6672, 2023, doi: 10.1109/TKDE.2022.3179381.

[28] S. Cost dan S. Salzberg, "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features," *Mach. Learn.*, vol. 10, no. 1, hal. 57–78, 1993, doi: 10.1007/bf00993481.

[29] Koalisi Indonesia Memantau, "Menatap ke Timur: Deforestasi dan Pelepasan Kawasan Hutan di Tanah Papua," Jakarta, 2021.

[30] T. K. Barik, "Ecologically Sound Mosquito Vector Control in River Basins," *Environ. Manag. River Basin Ecosyst.*, hal. 749–761, 2015, doi: 10.1007/978-3-319-13425-3_33.

[31] R. Pratiwi *et al.*, "Diversity and Abundance Model According to Habitat Characteristics of Filariasis Vector, Mansonia spp. in Banyuasin, South Sumatera, Indonesia," *J. Phys. Conf. Ser.*, vol. 1246, no. 1, 2019, doi: 10.1088/1742-6596/1246/1/012039.

[32] Irfan, N. T. Kambuno, dan Israfil, "Factors Affecting the Incidence of Filariasis in Welamosa Village Ende District East Nusa Tenggara (*Faktor yang Memengaruhi Kejadian Penyakit Filariasis di Desa Welamosa Kabupaten Ende Nusa Tenggara Timur*)," *Glob. Med. Heal. Commun.*, vol. 6, no. 2, hal. 130–137, 2018, [Online]. Available at: https://ejournal.unisba.ac.id/index.php/gmhc/article/view/3208

[33] M. Maifrizal, Teuku Reza Ferasyi, dan Fahmi Ichwansyah, "Risk Factor Analysis of Filariasis in Pidie Regency," *J. Kesehat. Lingkung.*, vol. 15, no. 3, hal. 226–234, 2023, doi: 10.20473/jkl.v15i3.2023.226-234.

[34] J. A. Brown, K. L. Larson, S. B. Lerman, A. Cocroft, dan S. J. Hall, "Resident Perceptions of Mosquito Problems are More Influenced by Landscape Factors than Mosquito Abundance," *Sustainability*, vol. 13, no. 20, hal. 1–17, 2021, doi: 10.3390/su132011533.

[35] C. B. Chesnais *et al.*, "Risk Factors for Lymphatic Filariasis in Two Villages of the Democratic Republic of the Congo," *Parasites and Vectors*, vol. 12, no. 1, hal. 1–13, 2019, doi: 10.1186/s13071-019-3428-5.

[36] C. R. Burgert-Brucker *et al.*, "Risk Factors Associated with Failing Pretransmission Assessment Surveys (Pre-tas) in Lymphatic Filariasis Elimination Programs: Results of a Multi-Country Analysis," *PLoS Negl. Trop. Dis.*, vol. 14, no. 6, hal. 1–17, 2020, doi: 10.1371/journal.pntd.0008301.

[37] P. W. Dhewantara, M. Ipa, dan M. Widawati, "Individual and Contextual Factors Predicting Self-Reported Malaria among Adults in Eastern Indonesia: Findings from Indonesian Community-Based Survey," *Malar. J.*, vol. 18, no. 1, hal. 1–17, 2019, doi: 10.1186/s12936-019-2758-2.

[38] N. Bhullar dan J. Maikere, “Challenges in Mass Drug Administration for Treating Lymphatic Filariasis in Papua, Indonesia,” *Parasites and Vectors*, vol. 3, no. 1, hal. 1–7, 2010, doi: 10.1186/1756-3305-3-70.

[39] S. Sandy dan I. Ayomi, “Gambaran Pengetahuan, Perilaku dan Pencegahan Malaria oleh Masyarakat di Kabupaten Maluku Tenggara Barat dan Maluku Barat Daya,” *J. Heal. Epidemiol. Commun. Dis.*, vol. 4, no. 1, hal. 7–14, 2018, doi: 10.22435/jhecds.v4i1.369.