InPrime: Indonesian Journal of Pure and Applied Mathematics
Vol. 6, No. 2 (2024), pp. 98 – 111, doi: 10.15408/inprime.v6i2.41463
p-ISSN 2686-5335, e-ISSN: 2716-2478

Enhancing Tuberculosis Diagnosis: Effective Naive Bayes Classification using SMOTE and Tomek Links for Imbalanced Data

Naflah Faulina, Khoirin Nisa*, and Warsono
Department of Mathematics, University of Lampung, Bandar Lampung, Indonesia
Email: *khoirin.nisa@fmipa.unila.ac.id

Abstract

Naive Bayes classification, grounded in Bayes' theorem, is a well-established probabilistic and statistical method. However, it often faces challenges when dealing with datasets that have skewed class distributions. A common issue with unbalanced data is that the classifier tends to predict the majority class more accurately, leading to high accuracy for the majority class but low accuracy for the minority class. Resampling techniques such as oversampling, undersampling, or a combination of both can be employed to address this. This research introduces a novel approach to balancing training data using a hybrid method that combines SMOTE (Synthetic Minority Oversampling Technique) and Tomek Links by applying this method to tuberculosis (TB) diagnosis data from Mayjend HM Ryacudu Kotabumi Hospital. We evaluate the Naive Bayes classifier's performance on the original and newly balanced data. We used 826 patient data for training and 207 for testing out of 1,033. Of the 826 records in the training dataset, 306 patients had a TB diagnosis, whereas 520 patients did not. To achieve a better balance between the majority and minority classes, we oversampled 214 data in the minority class to match the number in the majority class. If necessary, we also reduce 214 data from the majority class. The results demonstrate that this hybrid approach significantly enhances the performance of the Naive Bayes model in terms of data balancing and overall accuracy. Specifically, the hybrid method achieves an average specificity of 96%, sensitivity of 88%, false positive fraction (FPF) of 4%, and false negative fraction (FNF) of 12%. These findings highlight the effectiveness of combining SMOTE and Tomek Links, providing a robust solution for improving classification performance in unbalanced datasets.

Keywords: Naive Bayes classification; SMOTE; Tomek Links; SMOTE+Tomek Links; Tuberculosis.

Abstrak

Klasifikasi Naive Bayes*, yang didasarkan pada Teorema* Bayes*, adalah metode probabilistik dan statistik yang sudah mapan. Namun, metode ini sering menghadapi tantangan ketika berhadapan dengan kumpulan data yang memiliki distribusi kelas yang miring (tidak seimbang). Masalah umum pada data yang tidak seimbang adalah bahwa pengklasifikasi cenderung memprediksi kelas mayoritas dengan lebih akurat, yang mengarah pada akurasi tinggi untuk kelas mayoritas namun menghasilkan akurasi rendah untuk kelas minoritas. Untuk mengatasi masalah ini, teknik resampling seperti oversampling, undersampling, atau kombinasi keduanya dapat digunakan. Penelitian ini memperkenalkan pendekatan baru untuk menyeimbangkan data pelatihan menggunakan metode hibrida yang menggabungkan* SMOTE *(*Synthetic Minority Oversampling Technique*) dan* Tomek Links. *Dengan menerapkan metode ini pada data diagnosis* tuberculosis (TB) *dari Rumah Sakit Mayjend HM Ryacudu Kotabumi. Kami mengevaluasi kinerja pengklasifikasi Naive Bayes pada data yang tidak seimbang asli dan data yang sudah seimbang. Kami menggunakan 826 data pasien untuk pelatihan dan 207 untuk pengujian dari total 1.033. Dari 826 catatan dalam dataset pelatihan, 306 pasien didiagnosis dengan TB, sedangkan 520 pasien tidak. Untuk mencapai keseimbangan yang lebih baik antara kelas mayoritas dan minoritas, kami melakukan oversampling sebanyak 214 data pada kelas minoritas agar jumlahnya seimbang dengan kelas mayoritas. Selain itu, kami juga mengurangi 214 data dari kelas mayoritas. Hasilnya menunjukkan bahwa pendekatan hibrida ini secara signifikan meningkatkan kinerja model* Naive Bayes *dalam hal keseimbangan data dan akurasi keseluruhan. Secara spesifik, metode hibrida ini mencapai spesifisitas rata-rata sebesar 96%, sensitivitas sebesar 88%, fraksi positif palsu (FPF) sebesar 4%, dan fraksi negatif palsu (FNF) sebesar 12%. Temuan ini menyoroti efektivitas penggabungan* SMOTE *dan* Tomek

* Corresponding author

Submitted September 24th, 2024, Revised October 20th, 2024,
Accepted for publication October 22nd, 2024, Published Online October 28th, 2024

Links*, serta memberikan solusi yang tangguh untuk meningkatkan kinerja klasifikasi pada kumpulan data yang tidak seimbang.*

Kata Kunci: *Klasifikasi* Naive Bayes; SMOTE; Tomek Links; SMOTE+Tomek Links*; Tuberkulosis.*

2020MSC: 68T05, 62R07.

1. INTRODUCTION

Naive Bayes classification method is a machine learning technique that uses probability and statistics to infer future probabilities from prior experiences known as Bayes' Theorem which was developed by the English scientist Reverend Thomas Bayes. Following the theorem classification has been further developed by researchers in machine learning [1]. Naive Bayes has many advantages, including speed, efficiency, and performance in various classification tasks. For this reason, Naive Bayes is still a popular method in many areas of machine learning, such as text categorization, healthcare diagnosis, and managing system performance [2].

There is often an uneven distribution among classes in datasets that are used by researches. Unbalanced data occurs when there is a huge disparity in the number of training samples between two classes, with a large number of samples representing the majority class and a small number of samples representing the minority class [3]. A common problem with imbalanced data is that classification tends to predict the class with a larger data composition. As a result, prediction accuracy is high for the majority class data, while it is poor for the minority class data [4]. One method to address imbalanced data is through resampling techniques. Resampling is a preprocessing technique that algorithmically equalizes class distributions to improve the imbalance ratio and reduce the effects of imbalanced class distribution in machine learning processes. Resampling techniques can be performed using oversampling, undersampling, and hybrid methods [5][6].

The minority class is the target of oversampling, which aims to bring their numbers closer to the majority class by repeatedly sampling from the minority class [7]. One method that helps to even out data is the Synthetic Minority Oversampling Technique (SMOTE). It does this by making up instances of the minority class to obtain statistical parity [8]. To ensure the dataset is balanced, undersampling lowers the number of observations from the majority class [9]. Undersampling using Tomek Links involves excluding data from the majority class that has comparable traits [10]. Hybrid techniques address imbalanced data by combining oversampling and undersampling techniques [11]. By combining these two techniques, a dataset is expected to avoid excessive information loss, i.e. a negative effect of undersampling, and overfitting, i.e. a negative effect of oversampling. One such hybrid technique is SMOTE + Tomek Links [12].

The healthcare industry often deals with imbalanced data. Many tasks in healthcare rely on Naive Bayes classification, such as illness diagnostics, risk assessment, and outcome prediction. Its simplicity and efficiency make it particularly suitable for medical applications where rapid decision-making is based on probabilistic models [13]. Regarding the tuberculosis (TB) situation in Indonesia, as of January 2nd 2024, there are estimated to be approximately 1,060,000 TB cases. Data from 2023 show the detection of 792,404 TB cases, indicating an annual increase. In response to the high number of TB cases, a movement of TB prevention was established in Indonesia, an approach aimed at finding, diagnosing, treating, and curing TB patients with the primary goal of stopping the transmission of the disease in the community [14].

In comparison to previous research, our study offers several distinctive features that enhance the understanding of imbalanced data in TB diagnosis. While Tyagi et al. [5], explored various methods, including K-Nearest Neighbors and Support Vector Machines, they identified the ADASYN method as the most effective for data balancing. However, our research diverges by specifically implementing SMOTE, Tomek Links, and their hybrid combination (SMOTE+Tomek Links) within the context of Naive Bayes classification. This approach aims to tackle the dual challenges of overfitting associated with oversampling and information loss due to undersampling.

Additionally, previous studies, such as those by Sastrawan et al. [3], assessed combined sampling methods but did not focus on the unique challenges posed by healthcare datasets, particularly in TB diagnostics. Our research not only emphasizes the necessity for accurate classification in medical applications but also incorporates a detailed analysis of multiple factors, including gender, age, smoking status, body mass index (BMI), and family history of TB, which are critical for improving diagnostic accuracy.

Moreover, the healthcare landscape in Indonesia, characterized by approximately 1,060,000 TB cases as of January 2024, underscores the urgency of our work. By focusing on these specific factors and employing advanced hybrid resampling techniques, our study seeks to fill the gap in the existing literature regarding effective TB diagnosis in imbalanced datasets. This differentiation highlights the potential for our methods to provide a more robust solution in the face of real-world challenges associated with TB detection and management.

2. METHODS

The data used in this study is notably medical records of TB diagnoses collected from Mayjend HM Ryacudu Kotabumi Hospital. The data is available from January to December 2023. There were 1,033 people tested for TB in the dataset. Gender (X_1), age (X_2), smoking status (X_3), BMI (X_4), TB family history (X_5), and Molecular Rapid Test (TCM) findings (X_6) are the six aspects of the dataset for this research, with the diagnosis (Y) serving as the goal variable. Detailed variables are presented in Table 1.

Table 1. Description of Feature and Target Variables for TB Diagnosis Model

No.	Variable	Description	Type	Details
		Feature Variables		
1.	X_1	Gender	Categorical	0 = Female 1 = Male
2.	X_2	Age	Numeric	
3.	X_3	Smoking Status	Categorical	0 = Not at risk (if the patient is a non-smoker) 1 = At risk (if the patient is a smoker)
4.	X_4	Body Mass Index (BMI)	Numeric	$IMT = \frac{\text{weight (kg)}}{\text{height } (m)^2}$
5.	X_5	Family History of TB	Categorical	0 = No family members with TB 1 = Family members with TB
6.	X_6	Molecular Rapid Test (TCM)	Categorical	0 = Test results show MTB not detected 1 = Test results show MTB detected

Table 1. (Cont.)

		Target Variable		
7.	Y	Diagnosis	Categorical	0 = Not diagnosed with TB 1 = Diagnosed with TB

The analysis procedures of the research are as follows:

1. Preprocessing stage. Prepare the data by encoding any categorical variables, creating two sets of data namely one for training and one for testing with an 80:20 ratio and using five folds of cross-validation on the training set.
2. Confusion matrix evaluation on testing data for naive Bayes classification (unbalanced). Construct the naive Bayes model using training data, classify with testing data, and assess the model's performance.

The general notation for posterior probability can be written as follows:

$$P(C_k|x_i) = \frac{P(C_k)\,P(\,x_i|C_k)}{P(x_i)}, \tag{1}$$

where $P(C_k)$ is the prior probability of class C_k. This represents the initial belief about the likelihood of class C_k occurring before observing any data, $P(x_i|C_k)$ is the likelihood of observing feature x_i given that the instance belongs to class C_k (this indicates how likely the feature x_i is for the specific class) and $P(x_i)$ is the marginal probability of the feature x_i . This is the overall probability of observing the feature x_i across all classes. Thus, the Naive Bayes formula can be expressed as follows:

$$\begin{aligned} P(C_k|x_i) &= P(C_k)\,P(\,x_i|C_k) \\ &= P(C_k)\,.P(x_1|C_k).P(x_2|C_k)\,...\,P(x_n|C_k) \\ &= P(C_k)\,\textstyle\prod_{i=1}^{n} P(x_i|C_k). \end{aligned} \tag{2}$$

To get the prior probability, which is the chance of class C_k happening before the sample is seen, one may use the following formula:

$$P(C_k) = \frac{N_{C_k}}{N}, \tag{3}$$

where N_{C_k} represents the number of samples in class C_k and N is the total number of samples. To calculate the likelihood $P(x_i|C_k)$, there are two rules:

a. If the data from the attribute x_i is categorical, then $P(x_i|C_k)$ is the number of occurrences where feature x_i occurs in class C_k fractioned by the sum of all instances in the class C_k:

$$P(x_i|C_k) = \frac{P(x_i \cap C_k)}{P(C_k)}. \tag{4}$$

b. If the data from the attribute $x_i$is continuous, then $P(x_i|C_k)$ it is presumed to adhere to a normal distribution with a mean of μ and a standard deviation of σ, as calculated by:

$$g(x, \mu\,, \sigma) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{5}$$

Thus, $P(x_i|C_k) = g\left(x_i, \mu_{C_k}, \sigma_{C_k}\right),\ \mu = \frac{\sum_{i=1}^{n} x_i}{n}, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}$ [15].

The naive Bayes classification result is determined by selecting the class x_i that maximizes $P(C_k|x_i)$ among all possible classes for each variable x_i [1]. The evaluation metrics such as accuracy, sensitivity, specificity, False Positive Fraction (FPF), and False Negative Fraction (FPF) [16]. Table 2 presents the confusion matrix along with the evaluation metrics, which are calculated using Eq. (6) – (10).

Table 2. Confusion matrix

Prediction	**Actual**		**Total**
	Positive ($\boldsymbol{A}$)	**Negative ($\boldsymbol{\bar{A}}$)**	
Positive ($\boldsymbol{P^+}$)	a	b	$a+b$
Negative ($\boldsymbol{P^-}$)	c	d	$c+d$
Total	$a+c$	$b+d$	n

where

$$\text{Accuracy } = \frac{\frac{a}{n}+\frac{d}{n}}{\frac{a}{n}+\frac{d}{n}+\frac{b}{n}+\frac{c}{n}} = \frac{\frac{a+d}{n}}{\frac{a+d+b+c}{n}} = \frac{a+d}{n} \times 100\%, \tag{6}$$

$$\text{Sensitivity} = \text{True Positive Fraction} = \mathrm{P(P^+|A)} = \frac{P(P^+ \cap A)}{P(A)} = \frac{\frac{a}{n}}{\frac{a+c}{n}} = \frac{a}{a+c} \times 100\%, \tag{7}$$

$$\text{Specificity} = \text{True Negative Fraction} = \mathrm{P(P^-|\bar{A})} = \frac{P(P^- \cap \bar{A})}{P(\bar{A})} = \frac{\frac{d}{n}}{\frac{b+d}{n}} = \frac{d}{b+d} \times 100\%, \tag{8}$$

$$\text{False Positive Fraction (FPF)} = \mathrm{P(P^+|\bar{A})} = \frac{P(P^+ \cap \bar{A})}{P(\bar{A})} = \frac{\frac{b}{n}}{\frac{b+d}{n}} = \frac{b}{b+d} \times 100\%, \tag{9}$$

$$\text{False Negative Fraction (FNF)} = \mathrm{P(P^-|A)} = \frac{P(P^+ \cap A)}{P(A)} = \frac{\frac{c}{n}}{\frac{a+c}{n}} = \frac{c}{a+c} \times 100. \tag{10}$$

3. Data balancing using SMOTE. In order to determine the SMOTE percentage ($N\%$), take the following steps: first, count the number of instances in the majority and minority classes. Then, divide that number by the number of instances in the minority classes. For categorical data, use the Value Difference Metric (VDM) to find the k-nearest neighbors. Here is the definition of the distance V between two feature values:

$$\Delta(A,B) = \textstyle\sum_{i=1}^{N} \delta(V_{1i}, V_{2i}), \tag{11}$$

with,

$$\delta(V_1, V_2) = \textstyle\sum_{i=1}^{n} \left|\frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2}\right|, \tag{12}$$

where n is the number of categories in the *i*-th variable, C_{1i} is the number of category-1 occurrences in the *i*-th variable, C_{2i} is the number of category-2 occurrences in the *i*-th variable, C_1 is the number of category-1 occurrences, and C_2 is the number of category-2 occurrences.

For numerical data, use the nearest euclidean distance from each minority data point. Suppose there are two data points given with p dimensions, namely:

$$x^T = [x_1, x_2, \dots, x_p] \text{ dan } y^T = [y_1, y_2, \dots, y_p]. \tag{13}$$

Then, the Euclidean distance $d(x,y)$ between the two data vectors is as follows:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}. \tag{14}$$

Finally, create synthetic data based on the majority vote of the features being considered and their k-nearest neighbors. Whereas synthetic data is generated using the following equation:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \beta, \qquad i = 1,2,\dots, \tag{15}$$

where x_{syn} is replication data, x_i is data to be replicated, x_{knn} is data that is closest to the data to be replicated, and β is random numbers between 0 and 1.

4. Data balancing using Tomek Links. When working with Tomek Links it is important to count how many occurrences fall into each class. In order to do this, the Euclidean distance and difference value metrics are used for numerical and categorical data respectively. Data points from the dominant class are eliminated from the training data set in the event that two data points from different classes are determined to be Tomek Links. Undersampling using Tomek Links involves excluding data from the majority class that has comparable traits [17]. In order to find Tomek Links, given there are two observations, a and b, where the distance between them is denoted by $\delta(a,b)$. If there is no further observation and assuming c, and a and b are from distinct classes, and

$$\delta(a,c) < \delta(a,b) \text{ or } \delta(b,c) < \delta(b,a), \tag{16}$$

then a and b are called Tomek link observations [11]

5. Data balancing using Hybrid SMOTE + Tomek Links. Determining the number of instances in the majority and minority classes, increasing the number of samples in the minority class using SMOTE, and identifying Tomek Links in the data created by SMOTE are all tasks that may be accomplished utilizing Hybrid SMOTE + Tomek Links.
6. A balanced dataset may be used for naive Bayes classification in three steps: training data construction, testing data classification, and testing data evaluation using a confusion matrix.
7. Compare the evaluation results from unbalanced and balanced data by using the average values of the classification accuracy (see Equation 6).

The research flow, depicted in Figure 1, was implemented using RStudio software (version 4.2.2) to process and analyze the data efficiently.

3. RESULTS

The 1,033 observations in the data used in this study were classified into two classes namely TB diagnosed and not diagnosed. We used 826 patient data for training and 207 for testing. Out of the 826 records in the training data set, 306 patients had a TB diagnosis, whereas 520 patients did not. R Studio software was used to aid in the categorization process. We conducted data balancing using

SMOTE, Tomek Links, and a mix of SMOTE Tomek Links. After the data was balanced, a naive Bayes model was trained to identify TB based on its attributes. The accuracy of the model was then compared on both the original and balanced sets of data. Separate sets of information are stored in the database: training data and testing data. Also, in order to get more accurate findings, we used 5-fold cross-validation so the train:test rasio to be 80:20. Both the original and balanced TB data classifications using the Naive Bayes technique are presented in the following parts.

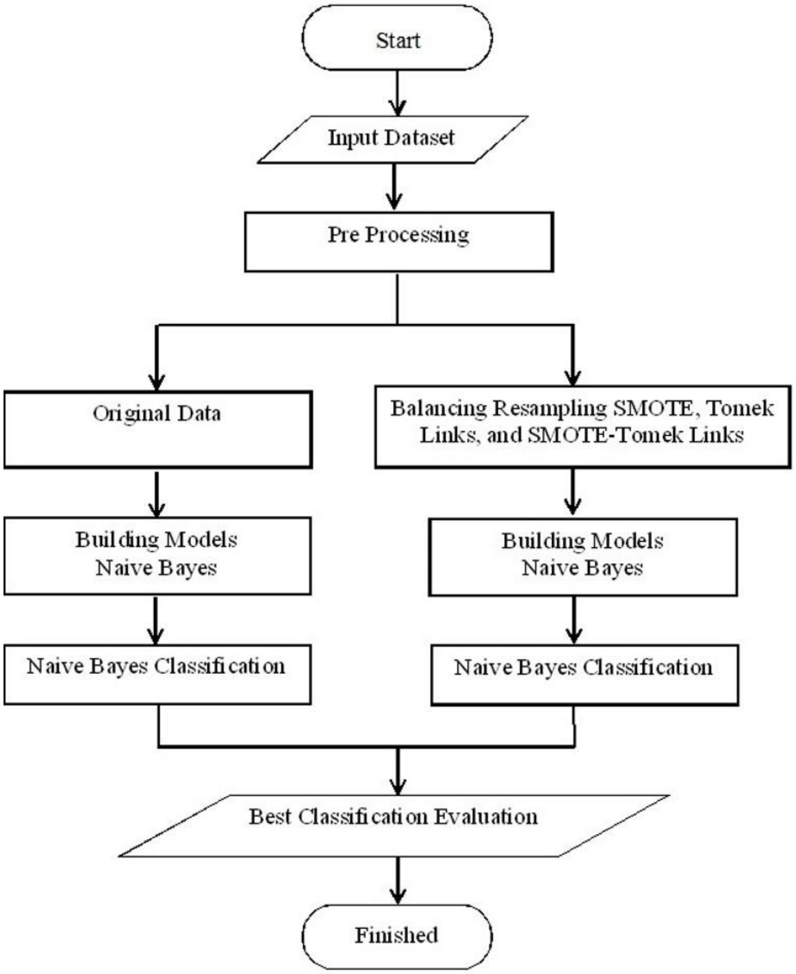
Start
Input Dataset
Pre Processing
Original Data
Balancing Resampling SMOTE, Tomek Links, and SMOTE-Tomek Links
Building Models Naive Bayes
Building Models Naive Bayes
Naive Bayes Classification
Naive Bayes Classification
Best Classification Evaluation
Finished

Figure 1. Flowchart of Naive Bayes Classification with Data Resampling Techniques

3.1. Naive Bayes Classification on Original Data

In order to determine the likelihood of a class for fresh data, Naive Bayes employs Bayes' Theorem, supposing that the characteristics in the data are unrelated to one another in light of the class [18]. Table 3 displays the Average Evaluation of 5-Fold Cross- Validation on the Original Data, which was obtained by applying the Naive Bayes algorithm on a sample of 826 training data and 207 testing data that used 6 features.

Table 3. Cross-Validation Results: Model Performance Metrics Across 5 Folds

Fold	Accuracy	Sensitivity	Specificity	FPF	FNF
1	90 %	80%	99%	1%	20%
2	87 %	76%	97%	3%	24%
3	91%	83%	97%	3%	17%
4	91%	85%	96%	4%	15%
5	89%	82%	94%	6%	18%
Average	89%	81%	96.6%	3.4%	19%

This model was able to correctly classify 89% of the total data, including both diagnosed TB cases and non-diagnosed cases. An average sensitivity of 81% indicates that this model is able to detect 81% of all actual TB cases. An average specificity of 96.6% indicates that this model is highly effective in identifying non-diagnosed TB cases, with 96.6% of all actual non-diagnosed cases being correctly classified. A False Positive Fraction (FPF) of 3.4% indicates that approximately 3.4% of all actual non-diagnosed TB cases were incorrectly classified as TB. A False Negative Fraction (FNF) of 19% indicates that about 19% of all actual TB cases were incorrectly classified as non-TB diagnosed. This relatively high rate of false negatives suggests that nearly 1 in 5 TB cases were missed by the model. Reducing the FNF is crucial to ensure that all TB cases are accurately detected, which is essential for effective diagnosis and treatment of the disease. Consequently, in order to enhance the model's performance, resampling methods were used to balance the data.

3.2. Balancing Method

In order to achieve data parity, resampling methods are used. By repeatedly sampling from the minority group, oversampling brings the minority group's observational count closer to that of the majority group. One method that helps to even out data is the Synthetic Minority Oversampling Technique (SMOTE). It does this by making up instances of the minority class to ensure statistical parity [8]. Finding nearby data points is how the SMOTE approach finds patterns. Undersampling using Tomek Links involves excluding data from the majority class that has comparable traits [17]. Hybrid techniques are methods for addressing imbalanced data by combining both oversampling and undersampling methods. By combining these two techniques, a dataset is expected to avoid excessive information loss (a negative effect of undersampling) and overfitting (a negative effect of oversampling). One such hybrid technique is SMOTE + Tomek Links [12][19]. The table 4 displays the outcomes of the data balance.

Table 4. Class Distribution of Tuberculosis Diagnosis Dataset After Applying Resampling Techniques

Datasets	Class	Original	SMOTE	Tomek Links	SMOTE+ Tomek Links
Tuberculosis diagnosis	not tuberculosis	520	509	306	464
	tuberculosis	306	508	306	504

From the 826 training data, there are 306 patients diagnosed with TB and 520 patients not diagnosed with TB. After applying SMOTE, the number of samples for both classes became balanced, with 509 patients for not diagnosed with TB and 508 patients for diagnosed with TB. After the Tomek

Links process, the number of samples for each class changed to 306 patients for not diagnosed with TB and 306 patients for diagnosed with TB. After applying SMOTE to increase samples in the minority class (TB diagnosed) and Tomek Links to clean the dataset, the proportions changed to 464 patients for not diagnosed with TB and 504 patients for diagnosed with TB. The number of training data points before and after balancing illustrated in Figure 2.

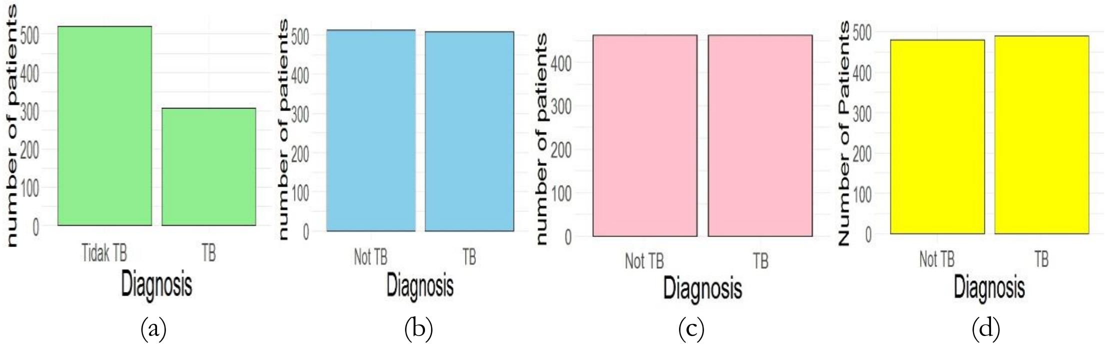

Figure 2. Number of Training Data Points Before and After Balancing (a) Original Data, (b) SMOTE Data, (c) Tomek Links Data, (d) SMOTE+Tomek Links Data.

Conducting Naive Bayes classification analysis follows the completion of the data balancing procedure. We will use 5-fold cross-validation to train our Naive Bayes model on the training data, and then we will use the testing data to measure the model's performance in terms of accuracy, sensitivity, specificity, FPF, and FNF. There will be an 80:20 split in the data. Table 5 – 7 show the results of the model assessment.

Table 5. Performance Metrics of the TB Diagnosis using SMOTE

Fold	Accuracy	Sensitivity	Specificity	FPF	FNF
1	90 %	85 %	95%	5%	15%
2	92 %	87 %	96%	4%	13%
3	88 %	85 %	92%	8%	15%
4	89 %	85 %	94%	4%	15%
5	90 %	85 %	96%	4%	15%
Average	90%	85.4 %	94%	5%	14.6%

Table 6. Performance Metrics of the TB Diagnosis using Tomek Links

Fold	Accuracy	Sensitivity	Specificity	FPF	FNF
1	90 %	85 %	99%	1%	15%
2	88 %	82 %	97%	3%	18%
3	90 %	88 %	96%	4%	12%
4	91 %	89 %	96%	4%	11%
5	90 %	88 %	94%	6%	12%
Average	89%	86 %	96%	4%	14%

Table 7. Performance Metrics of the TB Diagnosis using SMOTE + Tomek Links

Fold	Accuracy	Sensitivity	Specificity	FPF	FNF
1	93 %	88 %	97%	3%	12%
2	91 %	87 %	93%	7%	13%
3	94 %	89 %	97%	3%	11%
4	94 %	90 %	97%	3%	10%
5	94 %	87 %	97%	3%	13%
Average	93%	88 %	96%	4%	12%

Based on Table 5 – 7, the SMOTE model achieved an average accuracy of 90%, the Tomek Links model 89%, and the SMOTE+Tomek Links model 93%. In other words, whether a case of TB is identified or not, the model can accurately categorize 90%, 89%, and 93% of the data, respectively. The average sensitivity of the models is 85.4%, 86%, and 88%. This indicates the model's ability to detect 85.4%, 86%, and 88% of all actual TB diagnosed cases. The average specificity of the models is 94%, 96%, and 96%. This indicates that the model has a good ability to accurately identify most of the non-TB diagnosed cases, with an average False Positive Fraction (FPF) of 6%, 4%, and 4%. The low FPF indicates that the model rarely misclassifies healthy individuals as having TB, with an average False Negative Fraction (FNF) of 14.6%, 14%, and 12%.

3.3. Comparison of Naive Bayes Classification Before and After Balancing

Table 8 presents the results of naive Bayes classification using the original and balanced data. In Table 8, the performance of the classifications resulted from each scenario is compared using the four metrics: accuracy, sensitivity, specificity, false positive fraction (FPF), and false negative fraction (FNF).

Table 8. Performance Metrics of the TB Diagnosis using Naive Bayes Classification.

	Original	SMOTE	Tomek Links	SMOTE+Tome k Links
Accuracy	89 %	90 %	89 %	93 %
Sensitivity	81 %	85.40 %	86 %	88 %
Specificity	96.60 %	94 %	96 %	96 %
FPF	3, 40 %	5 %	4 %	4 %
FNF	19 %	14.60 %	14 %	12 %

In terms of performance, Table 8 shows that the SMOTE+Tomek Links technique came out on top with a 93% accuracy rate. This suggests that undersampling with Tomek Links to decrease overlap between minority and majority classes and oversampling with SMOTE to solve class imbalance may enhance the Naive Bayes model's classification accuracy. The SMOTE method also showed a significant improvement with an accuracy of 92%, while both the original data and data after applying Tomek Links had the same accuracy of 89%. Nevertheless, in this evaluation context, SMOTE+Tomek Links proves to be the better choice for enhancing classification accuracy on imbalanced data. For more clarity Figure 3 shows the comparison of the accuracy values of the four data.

4. DISCUSSION

The results of this study demonstrate that using a hybrid approach combining SMOTE and Tomek Links significantly improves the performance of Naive Bayes in classifying TB cases, with an accuracy increase from 89% to 93%. This improvement is consistent with findings from other studies, such as those by Tyagi et al. [5], who highlighted the importance of oversampling techniques like ADASYN in managing data imbalance for TB diagnosis. However, while ADASYN showed improved performance in some cases, our study highlights the added benefit of combining SMOTE with Tomek Links for removing noisy samples. Previous research, such as the study by Sastrawan et al. [3], examined the effects of hybrid sampling methods on predictive accuracy, yet did not specifically address their application in TB diagnosis. Our focus on TB adds value to the field by addressing the unique diagnostic challenges posed by this disease, particularly in high-burden areas like Indonesia, where early and accurate diagnosis is critical. Studies like those by Sejie et al. [20] have also emphasized the role of accurate classification in reducing the TB burden in resource-limited settings.

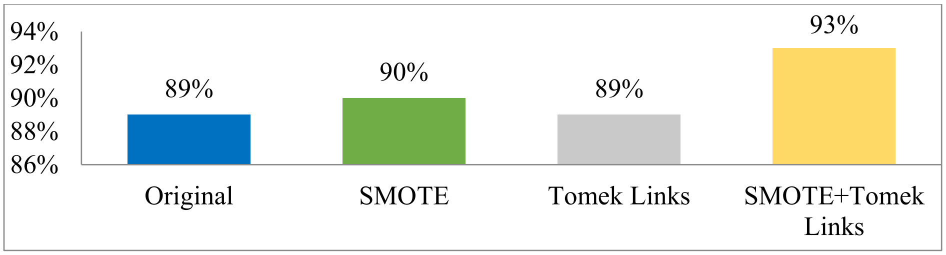

Figure 3. Comparison of Model Accuracy with Different Data Balancing Techniques

In terms of performance metrics, our study achieved a sensitivity of 88% and specificity of 96%, aligning closely with research by Singh et al. [21], which showed that balanced data improved sensitivity in TB detection using machine learning methods. The reduction of false positives and false negatives in our approach addresses the challenges identified in studies like those by Yadav et al. [22], where false negatives posed significant risks to patient health in TB management. Furthermore, the combination of SMOTE and Tomek Links addresses the dual challenges of overfitting and information loss, a balance that has been difficult to achieve in previous studies. For instance, Zhang et al. [23] found that while SMOTE effectively balanced data, it often led to overfitting when used alone. Our study overcomes this by integrating Tomek Links, which cleanses the dataset of noise, as also suggested by Singh et al. [21] in their exploration of hybrid resampling for medical diagnoses. This study's focus on Indonesia's TB burden is particularly relevant given the estimated 1,060,000 TB cases as of January 2024. Studies by Noviyani et al. [24] have highlighted the increasing TB incidence in Indonesia and the pressing need for reliable diagnostic methods. Our results suggest that integrating hybrid resampling techniques into diagnostic protocols could improve the identification and management of TB cases, ultimately aiding initiatives like the TOSS TB movement in Indonesia. Moreover, unlike studies that focus solely on molecular testing, our study incorporates additional factors like age, gender, smoking status, and BMI. This multifactorial approach provides a more comprehensive understanding of TB diagnosis, as emphasized in studies. The improvement in model accuracy through hybrid methods ensures better use of such diverse data, aligning with the findings of Wang et al. [25], who advocated for robust preprocessing techniques in enhancing medical classification outcomes. In summary, the hybrid SMOTE + Tomek Links approach proves to be a valuable method for addressing data imbalance, offering significant advantages in terms of accuracy, sensitivity, and specificity. This study contributes to the existing body of literature by providing a robust solution to the challenges of diagnosing TB in imbalanced datasets, a crucial step for improving patient outcomes in high-incidence regions.

4. CONCLUSIONS

This study found that combining the SMOTE oversampling technique with the Tomek Links undersampling technique enhances the accuracy of the Naive Bayes model. This conclusion is based on comparing the Naive Bayes classifier's performance on original and balanced data for TB diagnosis at RSD Mayjend HM Ryacudu Kotabumi. Training data was balanced using SMOTE, Tomek Links, and a combination of both. Statistically, the hybrid SMOTE and Tomek Links method significantly improved classification accuracy, achieving an average of 93% accuracy, 88% sensitivity, 96% specificity, 4% FPF, and 12% FNF. This hybrid approach effectively addresses data imbalance, leading to a more reliable model for distinguishing between diagnosed and non-diagnosed TB cases.

Future studies should consider exploring additional data balancing techniques, such as Adaptive Synthetic Sampling (ADASYN), to compare their effectiveness with the SMOTE + Tomek Links method. Moreover, integrating advanced machine learning algorithms, like ensemble methods or deep learning models, could enhance predictive accuracy. Investigating the influence of additional variables, such as socio-economic factors or comorbidities, on TB diagnosis may provide a more comprehensive understanding of patient outcomes. Finally, longitudinal studies assessing the long-term impacts of accurate TB diagnosis on public health outcomes would be beneficial.

AKNOWLEDGMENT

Author(s) would like to thank the Indonesian Ministry of Education, Culture, Research and Technology, which has funded this research with grant number 057/E5/PG.02.00.PL/2024. This research is part of a thesis submitted in fulfilment of the requirements for a Master of Science degree at Universitas Lampung, Indonesia.

REFERENCES

[1] J. Han, M. Kambe, and J. Pe, *Data Mining Concepts and Techniques*. 2012. doi: 10.1016/C2009-0-61819-5.

[2] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 103–130, 1997, doi: 10.1023/a:1007413511361.

[3] A. S. Sastrawan *et al.*, “Analisis Pengaruh Metode Combine Sampling Dalam Churn Prediction Untuk Perusahaan Telekomunikasi,” *Semin. Nas. Inform. 2010 (semnasIF 2010) UPN*, vol. 1, no. 1, pp. 14–22, 2010.

[4] H. Sain and S. W. Purnami, “Combine Sampling Support Vector Machine for Imbalanced Data Classification,” *Procedia Comput. Sci.*, vol. 72, pp. 59–66, 2015, doi: 10.1016/j.procs.2015.12.105.

[5] S. Tyagi and S. Mittal, “Sampling Approaches For Imbalanced Data Classification Problem In Machine Learning,” *Lect. Notes Electr. Eng.*, vol. 597, no. 7, pp. 209–221, 2020, doi: 10.1007/978-3-030-29407-6_17.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4, no. 1. 2006. doi: 10.53759/7669/jmc202404020.

[7] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 2, pp. 321–357, 2002, doi: 10.1613/jair.953.

[9] C. Drummond and R. C. Holte, "Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling," *Phys. Rev. Lett.*, vol. 91, no. 3, 2003.

[10] I. Tomek, "An Experiment with the Edited Nearest-Neighbor Rule," *IEEE Trans. Syst. Man, Cybern. SMC*, vol. 6, no. 6, pp. 448–453, 1973.

[11] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[12] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study Of The Behavior Of Several Methods For Balancing Machine Learning Training Data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004, doi: 10.1145/1007730.1007735.

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

[14] Kemenkes RI, *Petunjuk Teknis Pemeriksaan TB Menggunakan Tes Cepat Molekuler*. 2017.

[15] K. Fithriasari, I. Hariastuti, and K. S. Wening, "Handling Imbalance Data in Classification Model with Nominal Predictors," vol. 6, no. 1, pp. 33–37, 2020.

[16] L. M. Sullivan, *Essentials of Biostatistics in Public Health*. United States of America: Jones & Bartlett Learning, 2018.

[17] R. M. Pereira, Y. M. G. Costa, and C. N. Silla, "MLTL: A Multi-Label Approach For The Tomek Link Undersampling Algorithm," *Neurocomputing*, vol. 383, pp. 95–105, 2020, doi: 10.1016/j.neucom.2019.11.076.

[18] K. Murphy, *Machine Learning A Probabilistic Perspective*. London, England: The MIT Press, 2012.

[19] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.

[20] G. A. Sejie and O. H. Mahomed, "Mapping The Effectiveness of The Community Tuberculosis Care Programs: A Systematic Review," *Syst. Rev.*, vol. 12, no. 1, pp. 1–15, 2023, doi: 10.1186/s13643-023-02296-0.

[21] M. Singh *et al.*, "Evolution of Machine Learning in Tuberculosis Diagnosis: A Review of Deep Learning-Based Medical Applications," *Electron.*, vol. 11, no. 17, 2022, doi: 10.3390/electronics11172634.

[22] S. Yadav, G. Rawal, M. Jeyaraman, and N. Jeyaraman, "Advancements in Tuberculosis Diagnostics: A Comprehensive Review of the Critical Role and Future Prospects of Xpert MTB/RIF Ultra Technology," *Cureus*, vol. 16, no. 3, 2024, doi: 10.7759/cureus.57311.

[23] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *Mathematics*, vol. 12, no. 11, pp. 1–17, 2024, doi: 10.3390/math12111709.

[24] A. Noviyani, T. Nopsopon, and K. Pongpirul, "Variation of Tuberculosis Prevalence Across Diagnostic Approaches and Geographical Areas of Indonesia," *PLoS One*, vol. 16, no. 10 October, pp. 1–12, 2021, doi: 10.1371/journal.pone.0258809.

[25] Y. Wang, L. Liu, and C. Wang, "Trends in Using Deep Learning Algorithms in Biomedical Prediction Systems," *Front. Neurosci.*, vol. 17, no. November, pp. 1–32, 2023, doi: 10.3389/fnins.2023.1256351.