Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 1 (2024) 85 - 93 e-ISSN: 2580-0760

Impact of Adaptive Synthetic on Naïve Bayes Accuracy in Imbalanced Anemia Detection Datasets

Muhammad Khahfi Zuhanda[1*], Lisya Permata[2], Hartono[3], Erianto Ongko[4], Desniarti[5]
[1,3]Department of Informatics, Faculty of Engineering, Universitas Medan Area, Medan, Indonesia
[2]Department of Medical Education, Faculty of Medicine, Universitas Islam Sumatera Utara, Medan, Indonesia
[4]Department of Computer Science, Faculty of Engineering, Institut Modern Arsitektur dan Teknologi, Medan, Indonesia
[5]Department of Mathematics Education, Faculty of Teaching and Education, Universitas Muslim Nusantara Al Washliyah, Medan, Indonesia
[1]khahfi@staff.uma.ac.id

Abstract

This research aims to analyze the impact of the Adaptive Synthetic (ADASYN) oversampling technique on the performance of the Naïve Bayes classification algorithm on datasets with class imbalance. Class imbalance is a common problem in machine learning that can cause bias in prediction results, especially in minority classes. ADASYN is one of the oversampling methods that focuses on adaptively synthesizing new data for minority classes. In this study, the performance of the Naïve Bayes algorithm was tested on Anemia Diagnosis datasets before and after the application of ADASYN. This dataset contains 104 instances, 5 attributes, and 2 classes, and has an imbalance ratio of 3. The evaluation was carried out by comparing accuracy, confusion matrix, precision, recall, and F1-score to obtain a more comprehensive picture of the effectiveness of ADASYN in improving Naïve Bayes. The results of the study show that the performance of the oversampling method depends on the imbalance ratio so it is important to ensure that the oversampling method does not cause overfitting and this can be overcome by using ADASYN which only selects Selected Neighbors. The results showed that ADASYN significantly increased accuracy from 0.57 to 0.78, precision from 0.17 to 0.74, recall from 0.20 to 0.88, and F1-Score from 0.18 to 0.80. In this study, we also compared the application of ADASYN and SMOTE on the Naïve Bayes algorithm. The results show that ADASYN outperforms SMOTE across all key metrics—accuracy, precision, recall, and F1-Score—while the accuracy improvements were statistically significant (p-value = 0.00903).

Keywords: ADASYN; Class Imbalance; Oversampling; Machine Learning; Naïve Bayes;

How to Cite: M. K. Zuhanda, Lisya Permata, Hartono, Erianto Ongko, and Desniarti, "Impact of Adaptive Synthetic on Naïve Bayes Accuracy in Imbalanced Anemia Detection Datasets", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 1, pp. 85 - 93, Jan. 2025.
DOI: https://doi.org/10.29207/resti.v9i1.6031

1. Introduction

Classification problems in machine learning have been widely used in disease detection[1]. The identification of anemia is one of the applications of machine learning. Anemia is the most prevalent blood disorder worldwide. Anemia is a condition in which the body's physiological requirements are not met due to an insufficient number of red blood cells and, as a result, an insufficient oxygen-carrying capacity, as defined by the World Health Organization (WHO)[2]. Datasets are essential components of machine learning techniques. Machine learning methods can achieve optimal performance When processing high-quality datasets free of noise, outliers, missing values, and unbalanced data. However, class imbalance issues are actually present in anemia detection, as negative classes are significantly more numerous than positive classes[3].

A feature of the class imbalance problem is the existence of a class that has a significantly greater instance count than other classes [4]. The problem of class imbalance can cause a 10% decrease in the minority class's accuracy [5]. The problem of class imbalance is lessened by the data balancing procedure [6]. Class imbalance data classification is still a relatively new challenge[7], [8], particularly when dealing with binary classification problems when one class outnumbers the other [9]. When models are skewed toward dominating classes, it might result in

Received: 05-09-2024 | Accepted: 04-12-2024 | Published Online: 27-01-2025

imbalanced classification[10], which makes it difficult to forecast the minority class[11], [12].

The most often used processing approaches for unbalanced data nowadays are data improvement methods based on oversampling [10], [13], [14]. To address the issue of class imbalance, numerous classification techniques have been put forth, with minority oversampling techniques playing a key part [15]. Classifiers produced on uneven training sets show a prediction bias linked with poor performance in the minority class, which is the primary cause of class imbalance [16].

The goal of oversampling is to create artificial minority examples that minimize their distance from the border between minority and majority instances while also optimizing their real-valued classification potential as described [8], [17]. By boosting the amount of minority class samples through interpolation, the oversampling method balances datasets [18]. A disparity in class can also have a detrimental effect on learning and reduce accuracy [19].

In order to obtain an approximately balanced number of samples in each category, oversampling is a technology that produces minority samples [20]. Conversely, over-sampling techniques enhance the training dataset by incorporating a collection of minority data. These samples may consist of pre-existing samples from the initial minority class, synthetic samples produced by linear interpolation, or samples identified through active learning [21]. Clustering of the cases in the majority class is the fundamental process of the over-sampling approach used here. The closest neighbors who belong to the minority class are chosen based on the centers of the clusters created in this manner, which serve as reference examples[22]. Ultimately, all of the closest neighbors chosen from the minority class are created as synthetic instances [23].

Adaptive synthetic sampling (ADASYN) is a different kind of SMOTE than borderline or neighbor sampling[24]. Rather, it builds synthetic data based on data density considerations [25]. Adaptive Synthetic (ADASYN) automatically determines how many synthetic samples are required for each minority class sample and can adaptively synthesis a portion of the samples based on the distribution[26], [27]. The main idea behind the Adaptive Synthetic (ADASYN) algorithm is to automatically calculate the number of synthetic samples that should be generated for each minority data example by using the density distribution as a criterion[28], [29].

Machine learning algorithms like KNN, Random Forest, and Naive Bayes work incredibly well in the current era of machine learning approaches because of their low complexity and acceptable computation times [30]. Because Naive Bayes is a quick and effective technique for classification modeling[31], it was selected [32]. To maximize the effectiveness of the Naïve Bayes classifier, the data must first undergo extensive preparation[33], [34]. The "naive" belief that all features are independent of one another given the class label is where the classifier gets its name [35]. Naive Bayes is a popular machine-learning algorithm due to its effectiveness and simplicity [36]. One well-known machine learning technique, Naive Bayes, is based on a Bayesian network and is typically used for classification tasks. It performs exceptionally well [37]. Naive Bayes, a straightforward linear classifier that assumes all characteristics are independent given the class, has already demonstrated astonishing classification performance and is regarded as one of the top 10 data mining techniques [38].

A comparison of five oversampling techniques—SMOTE, K-means-SMOTE, BS-SMOTE, ADASYN, and DPC-SMOTE—was generated in the study by [18], and all of the evaluations showed a considerable rise in the index. This suggested strategy reduces sampling risk and uncertainty by incorporating past knowledge about cases of the minority class, according to research by [21] Furthermore, this suggested method is developed into adaptive unbalanced learning through the use of an error-bound model in conjunction with multi-objective optimization. Numerous tests have been conducted on unbalanced problems, and the outcomes show that this approach can enhance the functionality of several classification algorithms.

In the study by [25] four different oversampling techniques—the Synthetic Minority Oversampling Technique (SMOTE), SVM-SMOTE, Adaptive Synthetic Sampling (ADASYN), and borderline-SMOTE—were examined in order to produce an optimized dataset and address the imbalanced issue of the dataset. A PSO technique is also used to optimize the weights of the features.

In the study by [27] a two-step feature-selection strategy was used to optimize the feature set for training the prediction model's accuracy. Edited nearest-neighbor undersampling method and adaptive synthetic oversampling approach were used to solve dataset imbalance. Comparative empirical research between Inspector's performance and that of current technologies shows that Inspector could distinguish lysine succinylation sites with competitive prediction performance. Researchers [30] classified them without the use of oversampling techniques by employing various iterations of the Naive Bayes classifier.

Based on the explanation, it is very interesting to conduct research using the Adaptive Synthetic (ADASYN) oversampling technique on datasets and compare the accuracy results before and after applying ADASYN using the Naïve Bayes algorithm. To date, there has been no research specifically focusing on this particular comparison using ADASYN in conjunction with Naïve Bayes for handling imbalanced datasets. This gap in the literature presents an opportunity to explore how ADASYN can enhance the performance of Naïve Bayes in different scenarios. The findings could

contribute valuable insights into the effectiveness of oversampling techniques in improving classification model accuracy, particularly for datasets with significant class imbalances.

2. Research Methods

2.1 Process

The flow of this research can be seen in Figure 1.

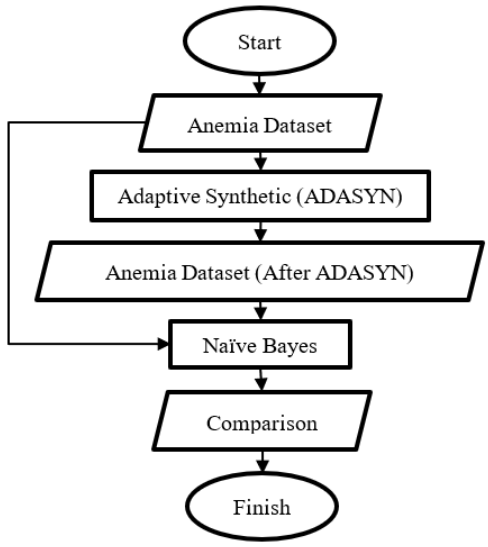

Figure 1. Flowchart of Research Process

Figure 1 shows the overall research flow, comparing the results of the Naïve Bayes algorithm applied to an anemia dataset before and after the Adaptive Synthetic (ADASYN) oversampling technique. The process starts with data pre-processing, followed by the application of Naïve Bayes to the original, imbalanced dataset. The results are recorded and analyzed. Next, the Adaptive Synthetic (ADASYN) method is applied to address the class imbalance, after which Naïve Bayes is run again on the balanced dataset. Adaptive Synthetic Sampling (ADASYN) is a data resampling technique designed to address class imbalance in machine learning. It generates synthetic data points for the minority class by focusing more on the difficult-to-learn samples that are near the decision boundary. The final step is comparing the performance metrics, such as accuracy, precision, recall, and F1-score, to evaluate the impact of ADASYN on the model's effectiveness.

2.2 Datasets

In this study, the performance of the Naïve Bayes algorithm was tested on Anemia Diagnosis datasets before and after the application of Adaptive Synthetic (ADASYN). This dataset contains 104 instances, and 5 attributes, and has an imbalance ratio of 3. The datasets used are sourced from [39] which can be seen in Table 1.

Table 1 shows that the dataset consists of 104 instances, with 6 columns representing various features such as Sex, Red, Green, and Blue pixel values, hemoglobin (Hb) levels, and whether the individual has anemia. Out of the 104 instances, there are 78 entries labeled as "No" (indicating no anemia) and 26 entries labeled as "Yes" (indicating the presence of anemia). Pixel values (RGB) and hemoglobin were chosen for anemia detection because RGB values in medical images can reflect changes in blood oxygenation and red blood cell concentration, which are linked to anemia, while hemoglobin levels provide a direct measure of the blood's oxygen-carrying capacity, a key indicator of anemia. Combining these features improves detection accuracy by using both visual and physiological data[40]. This imbalance between the two classes sets the stage for applying techniques such as Adaptive Synthetic (ADASYN) to improve classification accuracy.

Table 1. Anemia Datasets

No	Sex	%Red Pixel	%Green pixel	%Blue pixel	Hb	Anemia
1	M	432555	308421	259025	63	Yes
2	F	456033	2819	262067	135	No
3	F	450107	289677	260215	117	No
4	F	445398	289899	264703	135	No
5	M	43287	306972	260158	124	No
6	M	450994	279645	269361	162	No
7	F	431457	301628	266915	86	Yes
8	F	436103	291099	272798	103	No
9	F	450423	29166	257918	13	No
10	F	465143	274282	260575	97	Yes
…	…	…	…	…	…	…
104	F	435706	298094	266199	122	No

2.3. Adaptive Synthetic (ADASYN)

Creating synthetic data is a useful way to increase the number of examples available for testing and training models. Equations 1 and 2 illustrate the ADASYN method's functionality. An imbalance ratio was first created by establishing a collection of k-nearest neighbors for the majority and minority classes. This imbalance ratio serves as the basis for calculating the quantity of synthetic samples that must be produced [41].

$$IR_{xi} = \frac{|N_{maj(xi)}|}{|N_{min(xi)}|} \quad (1)$$

Equation 1 computes the ratio for a class of xi samples by taking into account the nearest neighbors of the majority and minority, as determined by the k-nearest neighbor technique. Equation 2 is then used to calculate the number of synthetic samples that need to be generated using this imbalance ratio IR_{xi} [41].

$$NS_{xi} = int(IR_{xi} * NS_{xi} - NS_{xi}) \quad (2)$$

Equation 2 uses NS_{xi} to stand for the synthetic samples that will be produced by multiplying the original input samples NO_{xi} in the minority class by the imbalance ratio determined in Equation 1. Equation 3 can be used to represent the process of creating fresh samples [41].

$$x_{new} = x_{orig} + \delta * (x_{neig} - x_{orig}) \quad (3)$$

Equation 3 denotes the recently produced samples. The difference between the minority class neighbors as determined by the KNN method and the original x feature is multiplied by the random number δ, which ranges from 0 to 1. A new sample is created by combining this result with the initial minority class sample [41].

The pseudocode of Adaptive Synthetic (ADASYN) can be seen as follows:

Input: *X_min, X_maj, K, G*
Output: *X_synth*
1. *r ← calculate_imbalance_ratio(X_min, X_maj)*
2. *G ← calculate_synthetic_samples_needed(r, |X_maj|, |X_min|)*
3. ***For each*** *x_i* ***in*** *X_min* ***do***
   a. *Neighbors ← find_K_nearest_neighbors(x_i, X_maj, K)*
   b. *r_i ← calculate_r_i(Neighbors, K)*
4. *Normalize r_i values such that Σ r_i = G*
5. ***For each*** *x_i* ***in*** *X_min* ***do***
   a. *N_synthetic ← calculate_number_of_synthetic_samples(r_i)*
   b. ***For*** *j = 1* ***to*** *N_synthetic* ***do***
      x_j ← randomly_select_neighbor(Neighbors)
      λ ← generate_random_number(0, 1)
      x_synth ← generate_synthetic_sample(x_i, x_j, λ)
      Add x_synth ***to*** *X_synth*
6. ***Return*** *X_synth*

End

The ADASYN pseudocode calculates the class imbalance ratio to determine the number of synthetic samples needed. It uses KNN to find neighbors for minority class samples, computes the ratio of majority neighbors, and normalizes these values. Synthetic samples are generated by interpolating between minority samples and their neighbors with a random factor. The algorithm returns these synthetic samples to balance the dataset

2.4 Naïve Bayes

The naive Bayes approach is a common statistical methodology in machine learning that addresses classification issues based on the Bayes Theorem [42]. Naïve Bayes has the advantage of only requiring a small amount of training data to determine the range of parameters used in the classification process because an independent variable only takes the variant of a variable in a class that is required to determine the classification, not the entire covariance matrix [43]. The most widely used and well-liked Naïve Bayes classifier is the Gaussian. The Gauss distribution is the source of the decision function which can be seen in Equation 4 [44].

$$P(v|y) = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(v-\mu)^2}{2\sigma^2}} \quad (4)$$

Equation 4 shows μ and σ are the estimated mean and standard deviation using the maximum likelihood principle [44]. Pseudocode naïve bayes can be seen as follows:

Input: *X_train, y_train, X_test*
Output: *y_pred (predicted labels* ***for*** *X_test)*
1. *Classes ← unique_classes(y_train)*
2. ***For each class in*** *Classes* ***do***
   a. *X_class ← subset* ***of*** *X_train* ***where*** *y_train = **class***
   b. *Mean_class ← calculate_mean(X_class)*
   c. *Variance_class ← calculate_variance(X_class)*
   d. *Prior_class ← calculate_prior_probability(y_train,* ***class****)*
3. ***For each*** *sample x* ***in*** *X_test* ***do***
   a. ***For each class in*** *Classes* ***do***
      Likelihood_class ← calculate_likelihood(x, Mean_class, Variance_class)
      *Posterior_class ← Likelihood_class * Prior_class*
   b. *y_pred ←* ***class with*** *highest Posterior_class*
4. ***Return*** *y_pred*

End

Pseudocode Gaussian Naïve Bayes classifies data by first calculating the mean, variance, and prior probability for each class in the training data, assuming a Gaussian (normal) distribution for each feature. For each test sample, the algorithm computes the likelihood of the sample belonging to each class using the Gaussian probability density function, then multiplies the likelihood by the class's prior probability to get the posterior probability. The class with the highest posterior probability is chosen as the predicted label.

2.5 Confusion Matrix

Adopted metrics can be characterized, according to the confusion matrix which can be seen in Equations 5, 6, 7, 8 and Table 2 [10].

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F-Score = \frac{(1+\beta^2)*Recall*Precision}{\beta^2*Precision+Recall} \quad (8)$$

Table 2. Confusion Matrix

	Predicted	
Actual	Class_X	Class_NX
Class_X	TP	FN
Class_NX	FP	TN

Equations 5, 6, 7, 8 and Table 2 shows defines $TP, FN, TN,$ and FP. The value β is often chosen by 1, indicating that Precision is equally crucial to accuracy as recall. By taking Precision and Recall into account, F-Score offers a thorough evaluation metric that is only large when both Precision and Recall are large. In summary, we anticipate that FP and FN will be near zero, so that the evaluation metrics will be near one [10].

3. Results and Discussions

3.1 Adaptive Synthetic (ADASYN)

Before oversampling using Adaptive Synthetic (ADASYN) datasets are first preprocessed by changing the contents of the dataset into numbers which can be seen in Table 3.

Table 3. Pre-Processing Datasets

No	Sex	%Red Pixel	%Green pixel	%Blue pixel	Hb	Anaemic
1	0	432555	308421	259025	63	1
2	1	456033	2819	262067	135	0
3	1	450107	289677	260215	117	0
4	1	445398	289899	264703	135	0
5	0	43287	306972	260158	124	0
6	0	450994	279645	269361	162	0
7	1	431457	301628	266915	86	1
8	1	436103	291099	272798	103	0
9	1	450423	29166	257918	13	0
10	1	465143	274282	260575	97	1
…	…	…	…	…	…	…
104	1	435706	298094	266199	122	0

Table 3 illustrates that the data in the Sex column has been modified, where "M" is converted to "0" and "F" to "1." Similarly, in the Anaemic column, the values "Yes" and "No" are replaced with "1" and "0," respectively. This transformation standardizes the categorical data for further analysis and machine learning applications. Before applying the Adaptive Synthetic (ADASYN) technique, the Anaemic column contained 78 instances of Class 0 (No Anaemic) and 26 instances of Class 1 (Anaemic). These original proportions highlight the significant imbalance between the two classes, which can negatively impact model performance.

Table 4 presents a comparison of the data before and after the ADASYN process was applied. Initially, the dataset comprised 104 instances, with a majority of 78 instances categorized as "No Anaemic" and only 26 instances as "Anaemic."

Table 4. Before and After Adaptive Synthetic

Category	Before ADASYN	After ADASYN
No Anaemic	78	80
Anaemic	26	78

After ADASYN, the dataset increased to 158 instances, with the number of "Anaemic" instances rising substantially from 26 to 78. This significant boost in minority class instances demonstrates ADASYN's role in creating synthetic data points to balance the dataset. Such balance is crucial for improving the accuracy of predictive models, particularly in cases with high-class imbalances.

The application of Adaptive Synthetic (ADASYN) effectively addresses the issue of class imbalance by generating synthetic samples for the minority class. By increasing the number of Class 1 (Anaemic) from 26 to 78, ADASYN equalizes the distribution of both classes. This adjustment reduces the risk of bias in machine learning models, which may otherwise overfit the majority class. Furthermore, the increase in data size, from 104 to 158 instances, enhances the model's ability to learn more effectively. The balanced dataset ultimately results in better generalization when applied to unseen data.

3.2 Naïve Bayes

Both datasets were trained and tested using the Naïve Bayes algorithm, with 80% of the data used for training and 20% for testing. In the case of the dataset before applying Adaptive Synthetic (ADASYN), as shown in Figure 2, the Naïve Bayes model achieved an accuracy of 0.57, with a Precision of 0.17, Recall of 0.20, and an F1-Score of 0.18. The confusion matrix indicated 11 True Negatives, 1 True Positive, 5 False Positives, and 4 False Negatives. These results highlight the difficulty of predicting the minority class in an imbalanced dataset, leading to low recall and precision values. Overall, the model struggled to correctly identify the minority class, which resulted in poor overall performance.

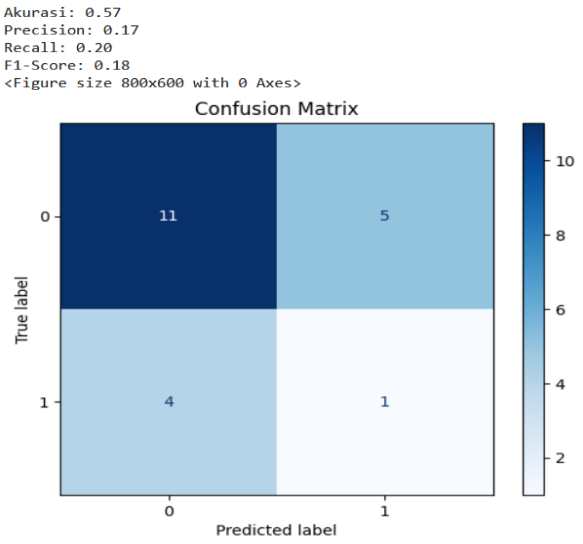

Figure 2. Training and Testing with Dataset Before Adaptive Synthetic (ADASYN)

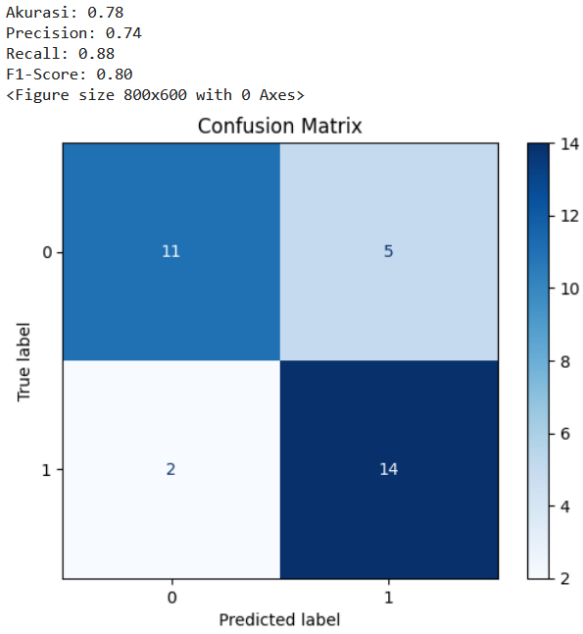

Figure 3. Training and Testing with Dataset After Adaptive Synthetic (ADASYN)

After applying Adaptive Synthetic (ADASYN) to balance the dataset, the Naïve Bayes model demonstrated significant improvements, as shown in Fig. 3. The accuracy increased to 0.78, with a Precision of 0.74, Recall of 0.88, and an F1-Score of 0.80. The confusion matrix for this model showed 11 True Negatives, 14 True Positives, 5 False Positives, and 2 False Negatives. These results demonstrate that ADASYN helped improve the model's ability to predict the minority class more accurately, reflected by the higher recall and precision scores. This balanced dataset allowed the model to perform better overall, reducing the impact of class imbalance on prediction accuracy.

3.3 Discussion

After applying the Adaptive Synthetic (ADASYN) method to the dataset, a comparison was made to observe the increase in data volume, as shown in Table 5. The number of instances increased significantly, from 104 before ADASYN to 158 after ADASYN. More importantly, the number of instances in Class 1, which initially had only 26 data points, increased to 78. This increase in the minority class ensures that the dataset is more balanced, allowing the model to train more effectively on the previously underrepresented class. The change in data distribution highlights the effectiveness of ADASYN in addressing class imbalance issues.

Table 5. Results Comparison for Pre-Processing Dataset

No	Pre-Processing Dataset	Instance	Class 0	Class 1
1	Before Adaptive Synthetic (ADASYN)	104	78	26
2	After Adaptive Synthetic (ADASYN)	158	80	78

Following this increase in data, both the pre-ADASYN and post-ADASYN datasets were used to train and test a Naïve Bayes model, and the results are compared in Table 6. Before applying ADASYN, the model's performance was suboptimal, with an accuracy of only 0.57. The precision and recall values were also low, at 0.17 and 0.20, respectively, indicating that the model struggled to correctly classify instances from the minority class. This poor performance was reflected in the F1-Score of 0.18, showing that the balance between precision and recall was weak in this imbalanced dataset.

Table 6. Comparison Naïve Bayes

No	Dataset	Accuracy	Precision	Recall	F1 - Score
1	Before Adaptive Synthetic (ADASYN)	0.57	0.17	0.20	0.18
2	After Adaptive Synthetic (ADASYN)	0.78	0.74	0.88	0.80

However, after applying ADASYN to balance the dataset, the performance of the Naïve Bayes model improved substantially across all metrics. Accuracy increased to 0.78, showing that the model's overall ability to classify the data became much better. Similar research results were obtained by Ozdemir et al.[45] who used the Xuzhou HYSPEX dataset sourced from the IEEE-Dataport Machine Learning Repository. The classification results on One vs All without balancing, the accuracy was 93.15 and after balancing with ADASYN, the accuracy increased to 95.57. There is a tendency for ADASYN to be applied to a classification method in machine learning in handling class imbalance. Research conducted by Assegie et al.[46] in breast, cancer identification showed that ADASYN applied to Logistic Regression (LR) significantly improved performance on LR to reach 99.46% and was much better when compared to SVM which only obtained 97.87%. The same research results were obtained in this study where accuracy increased from 0.57 (57%) to 0.78 (78%). Research conducted by Malhotra and Kamal [47] comparing a number of oversampling methods showed that ADASYN provided the best performance compared to other methods.

The precision saw a remarkable improvement, jumping from 0.17 to 0.74, indicating a substantial reduction in false positives. Similarly, recall increased dramatically from 0.20 to 0.88, meaning the model became much more capable of identifying the minority class. The F1-Score, which balances precision and recall, also increased significantly from 0.18 to 0.80, indicating that the model now performs well in both aspects. The results obtained show that ADASYN can improve the balance of the number of positive and negative samples in determining anemia using the Anemia Dataset. The classification results obtained using Naïve Bayes show an increase in accuracy, precision, Recall, and F1-Score so that the model can properly determine Anemia.

ADASYN proficiently mitigates class imbalance by producing synthetic data points for the minority class. It is essential to acknowledge the potential risk of overfitting. Noise can be introduced into the dataset through synthetic data generation methods like ADASYN, especially when the minority class is oversampled near decision boundaries. This could result in the model becoming excessively sensitive to the synthetic data, which can lead to overfitting. In this situation, the model exhibits strong performance on the training data but underperforms on novel test data.

Several distinctions are emphasized when ADASYN is compared with other oversampling methodologies, including SMOTE. In scenarios with intricate decision boundaries, SMOTE produces synthetic samples by interpolating between instances of the minority class, which may not consistently correspond with the actual data distribution. Conversely, ADASYN focuses on generating an increased quantity of synthetic samples in areas where the minority class is more difficult to learn, potentially enhancing performance in those regions. Both methods are prone to overfitting if not executed

judiciously, especially in datasets characterized by a significant imbalance ratio or limited size.

The results of applying ADASYN to Naive Bayes will be compared with SMOTE using 10-Folds cross-validation and can be seen in Table 7.

Table 7. Results Comparison with SMOTE

No	Accuracy		Precision		Recall		F1 -Score	
	ADASYN	SMOTE	ADASYN	SMOTE	ADASYN	SMOTE	ADASYN	SMOTE
1	0.78	0.74	0.74	0.73	0.88	0.82	0.8	0.65
2	0.81	0.73	0.72	0.71	0.87	0.86	0.81	0.79
3	0.82	0.81	0.75	0.80	0.85	0.89	0.77	0.81
4	0.73	0.72	0.76	0.75	0.86	0.83	0.71	0.74
5	0.75	0.74	0.81	0.76	0.79	0.77	0.82	0.81
6	0.74	0.74	0.82	0.81	0.81	0.79	0.78	0.77
7	0.81	0.76	0.79	0.77	0.82	0.82	0.77	0.76
8	0.82	0.81	0.81	0.79	0.79	0.81	0.84	0.83
9	0.76	0.73	0.77	0.78	0.83	0.79	0.77	0.74
10	0.75	0.73	0.84	0.81	0.84	0.77	0.73	0.71

Table 7 compares ADASYN and SMOTE in terms of Accuracy, Precision, Recall, and F1-Score across 10 tests. ADASYN shows a clear advantage over SMOTE, particularly in both Accuracy and Recall, where it consistently scores higher, indicating not only better identification of positive instances but also overall stronger model performance. While Precision differences are minimal between the two, ADASYN generally performs slightly better. The F1-Score also tends to favor ADASYN, though SMOTE matches closely in certain tests. Overall, ADASYN outperforms SMOTE, especially in key metrics like Accuracy and Recall.

To test the significance, it is done using Wilcoxon Signed-Rank. The results of the Wilcoxon Signed-Rank Test can be seen in Table 8.

Table 8. Significance Test using Wilcoxon Signed-Rank Test

No	Parameter	P-Value	Hypothesis
1	Accuracy	0.00903	Significant score difference
2	Precision	0.1504	No significant difference
3	Recall	0.1537	No significant difference
4	F-1 Score	0.2826	No significant difference

Table 8 compares different evaluation metrics such as Accuracy, Precision, Recall, and F-1 Score, with their respective p-values and hypothesis results. Based on the p-value column, only Accuracy has a significant score difference (p-value = 0.00903), indicating that the difference in accuracy is statistically significant. For Precision, Recall, and F-1 Score, the p-values are higher than the typical threshold of 0.05, suggesting no significant differences for these metrics.

4. Conclusions

From the results of training and testing the Naïve Bayes model carried out using the original unprocessed dataset and the dataset that has been processed using the Adaptive Synthetic (ADASYN) method, it is proven that the application of this oversampling technique can have a positive impact on increasing the number of data instances overall, especially in increasing the number of instances in the minority class which is often overlooked in standard analysis. This not only has an impact on balancing the data distribution, but also improves model performance in terms of accuracy, precision, recall, and f1-score, which reflect the model's ability to perform better classification in Anemia. In this study, we also compared the application of ADASYN and SMOTE on the Naïve Bayes algorithm. The results show that ADASYN outperforms SMOTE across all key metrics—accuracy, precision, recall, and F1-Score—and while the accuracy improvements were statistically significant (p-value = 0.00903), the differences in precision, recall, and F1-Score were not, indicating that ADASYN's advantage is primarily driven by its impact on accuracy. Thus, for further research, it is highly recommended that this approach be expanded by increasing the size of the dataset used for training so that the model can be trained on more diverse data. In addition, the application of other oversampling techniques is expected to be explored to see how far it can improve model performance and employ a statistical significance test, such as a t-test or a Wilcoxon signed-rank test. However, ADASYN may introduce the risk of overfitting, particularly in smaller datasets or when synthetic data does not align well with real data distribution. Additionally, while Naïve Bayes is efficient, its assumption of feature independence can limit performance in real-world scenarios with correlated features, suggesting the need for future research to explore more advanced algorithms, such as Random Forest or Support Vector Machines (SVM), which can better handle feature interactions and complex decision boundaries.

Acknowledgements

This work was supported by Universitas Medan Area and the Ministry of Higher Education, Science and Technology.

References

[1] M. Ahammed, Md. A. Mamun, and M. S. Uddin, “A machine learning approach for skin disease detection and classification using image segmentation,” *Healthcare Analytics*, vol. 2, p. 100122, Nov. 2022, doi: 10.1016/j.health.2022.100122.

[2] T. Karagül Yıldız, N. Yurtay, and B. Öneç, “Classifying anemia types using artificial learning methods,” *Engineering*

Science and Technology, an International Journal, vol. 24, no. 1, pp. 50–70, Feb. 2021, doi: 10.1016/j.jestch.2020.12.003.

[3] H. Hairani, T. Widiyaningtyas, and D. D. Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1310–1318, Sep. 2024, doi: 10.62527/joiv.8.3.2283.

[4] Hartono and R. B. Y. Syah, "Hybrid Approach with Membership-Density Based Oversampling for handling multi-class imbalance in Internet Traffic Identification with overlapping and noise," *ICT Express*, p. S2405959524000444, Apr. 2024, doi: 10.1016/j.icte.2024.04.007.

[5] Q. D. Nguyen and H.-T. Thai, "Crack segmentation of imbalanced data: The role of loss functions," *Engineering Structures*, vol. 297, p. 116988, Dec. 2023, doi: 10.1016/j.engstruct.2023.116988.

[6] A. Noor, N. Javaid, N. Alrajeh, B. Mansoor, A. Khaqan, and S. H. Bouk, "Heart Disease Prediction Using Stacking Model With Balancing Techniques and Dimensionality Reduction," *IEEE Access*, vol. 11, pp. 116026–116045, 2023, doi: 10.1109/ACCESS.2023.3325681.

[7] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, "RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, Jun. 2022, doi: 10.1016/j.jksuci.2022.06.005.

[8] J. Jedrzejowicz and P. Jedrzejowicz, "Bicriteria Oversampling for Imbalanced Data Classification," *Procedia Computer Science*, vol. 207, pp. 245–254, 2022, doi: 10.1016/j.procs.2022.09.057.

[9] E. B. Fatima, B. Omar, E. M. Abdelmajid, F. Rustam, A. Mehmood, and G. S. Choi, "Minimizing the Overlapping Degree to Improve Class-Imbalanced Learning Under Sparse Feature Selection: Application to Fraud Detection," *IEEE Access*, vol. 9, pp. 28101–28110, 2021, doi: 10.1109/ACCESS.2021.3056285.

[10] F. Dai, Y. Song, W. Si, G. Yang, J. Hu, and X. Wang, "Improved CBSO: A distributed fuzzy-based adaptive synthetic oversampling algorithm for imbalanced judicial data," *Information Sciences*, vol. 569, pp. 70–89, Aug. 2021, doi: 10.1016/j.ins.2021.04.017.

[11] P. Sadhukhan and S. Palit, "Adaptive learning of minority class prior to minority oversampling," *Pattern Recognition Letters*, vol. 136, pp. 16–24, Aug. 2020, doi: 10.1016/j.patrec.2020.05.020.

[12] D. Appasani, C. S. Bokkisam, and S. Surendran, "An Incremental Naive Bayes Learner for Real-time Health Prediction," *Procedia Computer Science*, vol. 235, pp. 2942–2954, 2024, doi: 10.1016/j.procs.2024.04.278.

[13] H. Cang *et al.*, "Jujube quality grading using a generative adversarial network with an imbalanced data set," *Biosystems Engineering*, vol. 236, pp. 224–237, Dec. 2023, doi: 10.1016/j.biosystemseng.2023.11.002.

[14] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, "CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification," *Neural Comput & Applic*, vol. 33, no. 7, pp. 2839–2851, Apr. 2021, doi: 10.1007/s00521-020-05130-z.

[15] J. Liu, "A minority oversampling approach for fault detection with heterogeneous imbalanced data," *Expert Systems with Applications*, vol. 184, p. 115492, Dec. 2021, doi: 10.1016/j.eswa.2021.115492.

[16] A. S. Tarawneh, A. B. A. Hassanat, K. Almohammadi, D. Chetverikov, and C. Bellinger, "SMOTEFUNA: Synthetic Minority Over-Sampling Technique Based on Furthest Neighbour Algorithm," *IEEE Access*, vol. 8, pp. 59069–59082, 2020, doi: 10.1109/ACCESS.2020.2983003.

[17] S. Korkmaz, M. A. Şahman, A. C. Cinar, and E. Kaya, "Boosting the oversampling methods based on differential evolution strategies for imbalanced learning," *Applied Soft Computing*, vol. 112, p. 107787, Nov. 2021, doi: 10.1016/j.asoc.2021.107787.

[18] W. Wang and F. Liu, "ADDPC-SMOTE: An Oversampling Algorithm Based on Density Difference Peak Clustering and Spatial Distribution Entropy," *IEEE Access*, vol. 11, pp. 108152–108166, 2023, doi: 10.1109/ACCESS.2023.3320265.

[19] I. Czarnowski, "Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams," *Journal of Computational Science*, vol. 61, p. 101614, May 2022, doi: 10.1016/j.jocs.2022.101614.

[20] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM," *Computers & Security*, vol. 106, p. 102289, Jul. 2021, doi: 10.1016/j.cose.2021.102289.

[21] T. Zhang, Y. Li, and X. Wang, "Gaussian prior based adaptive synthetic sampling with non-linear sample space for imbalanced learning," *Knowledge-Based Systems*, vol. 191, p. 105231, Mar. 2020, doi: 10.1016/j.knosys.2019.105231.

[22] B. Mirzaei, B. Nikpour, and H. Nezamabadi-pour, "CDBH: A clustering and density-based hybrid approach for imbalanced data classification," *Expert Systems with Applications*, vol. 164, p. 114035, Feb. 2021, doi: 10.1016/j.eswa.2020.114035.

[23] I. Czarnowski, "Agent-based population learning algorithm for over-sampling in the classification of imbalanced data streams," *Procedia Computer Science*, vol. 225, pp. 686–692, 2023, doi: 10.1016/j.procs.2023.10.054.

[24] R. Mitra, A. Bajpai, and K. Biswas, "ADASYN-assisted machine learning for phase prediction of high entropy carbides," *Computational Materials Science*, vol. 223, p. 112142, Apr. 2023, doi: 10.1016/j.commatsci.2023.112142.

[25] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[26] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Applied Soft Computing*, vol. 132, p. 109853, Jan. 2023, doi: 10.1016/j.asoc.2022.109853.

[27] Y. Zhu, C. Jia, F. Li, and J. Song, "Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling," *Analytical Biochemistry*, vol. 593, p. 113592, Mar. 2020, doi: 10.1016/j.ab.2020.113592.

[28] Z. Qing, Q. Zeng, H. Wang, Y. Liu, T. Xiong, and S. Zhang, "ADASYN-LOF Algorithm for Imbalanced Tornado Samples," *Atmosphere*, vol. 13, no. 4, Art. no. 4, Apr. 2022, doi: 10.3390/atmos13040544.

[29] T. Xu, G. Coco, and M. Neale, "A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning," *Water Research*, vol. 177, p. 115788, Jun. 2020, doi: 10.1016/j.watres.2020.115788.

[30] M. Vishwakarma and N. Kesswani, "A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection," *Decision Analytics Journal*, vol. 7, p. 100233, Jun. 2023, doi: 10.1016/j.dajour.2023.100233.

[31] D. Petschke and T. E. M. Staab, "A supervised machine learning approach using naive Gaussian Bayes classification for shape-sensitive detector pulse discrimination in positron annihilation lifetime spectroscopy (PALS)," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 947, p. 162742, Dec. 2019, doi: 10.1016/j.nima.2019.162742.

[32] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments," *Procedia Computer Science*, vol. 234, pp. 349–356, 2024, doi: 10.1016/j.procs.2024.03.014.

[33] D.-H. Vu, "Privacy-preserving Naive Bayes classification in semi-fully distributed data model," *Computers & Security*, vol. 115, p. 102630, Apr. 2022, doi: 10.1016/j.cose.2022.102630.

[34] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Computer Science*, vol. 179, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[35] A. V. D. Sano, A. A. Stefanus, E. D. Madyatmadja, H. Nindito, A. Purnomo, and C. P. M. Sianipar, "Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes classifier," *Procedia Computer Science*, vol. 227, pp. 482–489, 2023, doi: 10.1016/j.procs.2023.10.549.

[36] S. Wang, J. Ren, and R. Bai, "A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes," *Expert Systems with Applications*, vol. 225, p. 120094, Sep. 2023, doi: 10.1016/j.eswa.2023.120094.

[37] W. Guo, G. Wang, C. Wang, and Y. Wang, "Distribution network topology identification based on gradient boosting decision tree and attribute weighted naive Bayes," *Energy Reports*, vol. 9, pp. 727–736, Sep. 2023, doi: 10.1016/j.egyr.2023.04.256.

[38] H. Zhang, L. Jiang, and G. I. Webb, "Rigorous non-disjoint discretization for naive Bayes," *Pattern Recognition*, vol. 140, p. 109554, Aug. 2023, doi: 10.1016/j.patcog.2023.109554.

[39] Shahzad Aslam, "Anemia Diagnosis." [Online]. Available: https://www.kaggle.com/datasets/zeesolver/uhygtttt/data

[40] S. Suner *et al.*, "Prediction of anemia and estimation of hemoglobin concentration using a smartphone camera," *PLoS ONE*, vol. 16, no. 7, p. e0253495, Jul. 2021, doi: 10.1371/journal.pone.0253495.

[41] A. Alabrah, "Scientific Elegance in NIDS: Unveiling Cardinality Reduction, Box-Cox Transformation, and ADASYN for Enhanced Intrusion Detection," *CMC*, vol. 79, no. 3, pp. 3897–3912, 2024, doi: 10.32604/cmc.2024.048528.

[42] Y. Shang, "Prevention and detection of DDOS attack in virtual cloud computing environment using Naive Bayes algorithm of machine learning," *Measurement: Sensors*, vol. 31, p. 100991, Feb. 2024, doi: 10.1016/j.measen.2023.100991.

[43] A. Yudhana, D. Sulistyo, and I. Mufandi, "GIS-based and Naïve Bayes for nitrogen soil mapping in Lendah, Indonesia," *Sensing and Bio-Sensing Research*, vol. 33, p. 100435, Aug. 2021, doi: 10.1016/j.sbsr.2021.100435.

[44] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108972, Oct. 2024, doi: 10.1016/j.engappai.2024.108972.

[45] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Systems with Applications*, vol. 178, p. 114986, Sep. 2021, doi: 10.1016/j.eswa.2021.114986.

[46] T. A. Assegie, A. O. Salau, K. Sampath, R. Govindarajan, S. Murugan, and B. Lakshmi, "Evaluation of Adaptive Synthetic Resampling Technique for Imbalanced Breast Cancer Identification," *Procedia Computer Science*, vol. 235, pp. 1000–1007, Jan. 2024, doi: 10.1016/j.procs.2024.04.095.

[47] R. Malhotra and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," *Neurocomputing*, vol. 343, pp. 120–140, May 2019, doi: 10.1016/j.neucom.2018.04.090.

