# Instal : Jurnal Komputer

**E-ISSN: 2808-683X**
**Edisi : Volume 15 Nomor 01 | June 2024**

Available online at https://journalinstal.cattleyadf.org/index.php/Instal/index

# Handling Data Imbalance Problem in Hybrid Resampling Approach to Improve Accuracy of K-Nearest Neighbors Algorithm

**Novriadi Antonius Siagian[1], Sardo Pardingotan Sipayung[2]**
Universitas Katolik Santo Thomas[1,2]

**ARTICLE INFO**

***Article history:***

Received : 26 May 2024
Revised : 27 May 2024
Accepted : 27 June 2024

***Keywords:***

Data Imbalance, Undersampling, Oversampling, K-Nearest Neighbors, Smote, Nearmiss

**ABSTRACT**

Data imbalance is a problem where the classes or labels in a dataset are not evenly distributed. This often occurs in cases where one class has a much larger number of samples than another class. This can be a problem in various contexts, especially in the context of medical data such as stroke detection. This research proposes a hybrid resampling approach using SMOTE (Synthetic Minority Over-sampling Technique) and NearMiss to improve the accuracy of the K-Nearest Neighbors (KNN) algorithm on imbalanced stroke datasets. This approach aims to overcome the weaknesses of each technique, with SMOTE generating minority class samples and NearMiss reducing samples from the majority class. The use of SMOTE with K-1 neighbors provided 100% accuracy improvement, followed by 97% improvement with K-2, and 93% accuracy with K-3. Meanwhile, the undersampling approach using NearMiss showed 100% accuracy improvement with K-1, followed by 74% improvement with K-2, and 76% accuracy with K-3. In conclusion, the SMOTE approach provides a better accuracy improvement than the nearmiss and conventional approaches. So that this research contributes in handling data imbalance with the approach used. So it can be seen from the use of the SMOTE approach, with k-1 and k-2 neighborhoods showing good results from the previous approach.



***Corresponding Author:***
**Novriadi Antonius Siagian**
Universitas Katolik Santo Thomas Medan
Email: novriadi.antonius95@gmail.com

## INTRODUCTION

In the era of big data, the problem of data imbalance is often a significant challenge in data processing [1]. Data imbalance occurs when the number of samples in one class is much higher than that of another class [2]. This problem can result in prediction models that are biased towards the majority class, reducing the accuracy and overall performance of the model [3].

In stroke detection, the issue of data imbalance is crucial as the prevalence of patients who did not have a stroke is much higher than that of patients who did [4]. This imbalance can cause machine learning models to over-predict the majority class, resulting in a lack of

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

sensitivity in identifying stroke cases [5]. Models trained on unbalanced datasets may result in biased predictions towards the majority class, making accurate classification difficult [6]. Therefore, it is important to develop an effective approach to balance the class distribution in stroke detection datasets to improve the accuracy and sensitivity of the models in precise and rapid diagnosis [7].

In areas such as fraud detection, medical diagnosis, and speech recognition, imbalanced data is a common problem [8]. Early diagnosis and accurate prediction are essential to reduce the adverse effects of stroke [9]. However, in many medical cases, patient data is often unbalanced. For example, the number of patients who did not have a stroke is much larger compared to patients who had a stroke [10]. This imbalance may result in biased prediction models and less accurate detection of stroke cases [11].

Handling data imbalance plays an important role in improving the accuracy of machine learning models, especially in healthcare applications such as disease detection [12]. In the context of disease detection, including stroke detection, an unbalanced distribution between positive and negative patient classes can result in bias in model predictions [13]. By balancing the dataset, the model can better learn from both classes so that sensitivity to rare disease cases can be improved [14]. This allows the system to more accurately identify disease cases that are important for early diagnosis and appropriate treatment. Thus, addressing data imbalance not only improves the accuracy of the model in disease detection, but can also potentially save lives and improve the overall quality of healthcare [15].

Improve the accuracy of the K-Nearest Neighbors model on unbalanced datasets. By integrating SMOTE which is an oversampling method to increase the number of samples from the minority class and NearMiss which is an undersampling method to reduce samples from the majority class, it is expected that this hybrid resampling approach can create an optimal balance between the classes in the dataset. Thus, this study aims to improve the sensitivity and accuracy of the K-Nearest Neighbors model in detecting cases of stroke or other medical conditions on unbalanced datasets [16].

The highly competitive telecommunications industry, customer retention is the key to success for companies. With the advancement of information and communication technology, telecommunications companies have access to large amounts of customer data. Predicting customer churn is not easy due to various factors that can influence a customer's decision to stop using a service. In addition, class imbalance in customer churn datasets often makes it difficult for models to achieve high prediction performance. The resampling approaches used are SMOTE (Synthetic Minority Over-sampling Technique), SMOTE Tomek-Links, and SMOTE-ENN (SMOTE Edited Nearest Neighbors) [17].

Chronic kidney disease is a serious health problem and early detection of the disease is crucial to reduce the mortality rate caused by the disease. Explored the effect of resampling in chronic kidney disease detection using an imbalanced dataset. The problem of class imbalance and optimal feature selection in medical datasets are the main focus of this research to improve the performance of the prediction model. The resampling approaches used are Synthetic Minority Oversampling Technique (SMOTE) and Near Miss Undersampling Technique [18].

The approach used in this study is hybrid resampling, which uses oversampling and undersampling techniques to overcome data imbalance. Oversampling adds samples to the minority class, while undersampling reduces samples to the majority class. The aim of this approach is to achieve a more balanced class distribution without losing important information from the original data. So that it can be seen how the comparison of the approaches taken to improve the accuracy of the K-Nearest Neighbors (KNN) algorithm. By seeing good accuracy in the classification of data imbalance in the stroke dataset used [19].

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

This research makes a significant contribution to the literature related to data imbalance handling and the development of hybrid resampling techniques. In the context of data imbalance handling, a hybrid resampling approach combining SMOTE and NearMiss successfully improves the accuracy of K-Nearest Neighbors models on imbalanced datasets, particularly in disease detection such as stroke. This contribution provides new insights in addressing the challenges of data imbalance in medical applications. Moreover, the development of hybrid resampling techniques with the integration of two different resampling methods shows the potential to improve dataset balance and machine learning model accuracy in general. Thus, this research not only provides practical benefits in improving disease detection accuracy, but also makes theoretical contributions in the development of innovative and effective hybrid resampling approaches for data imbalance handling in various application domains [20].

## METHODS

A hybrid resampling approach combining SMOTE and NearMiss is used to address data imbalance. Oversampling and undersampling steps are performed to add samples to the minority class and reduce samples to the majority class to achieve a balanced class distribution [21].

The workflow used in this research is as follows:

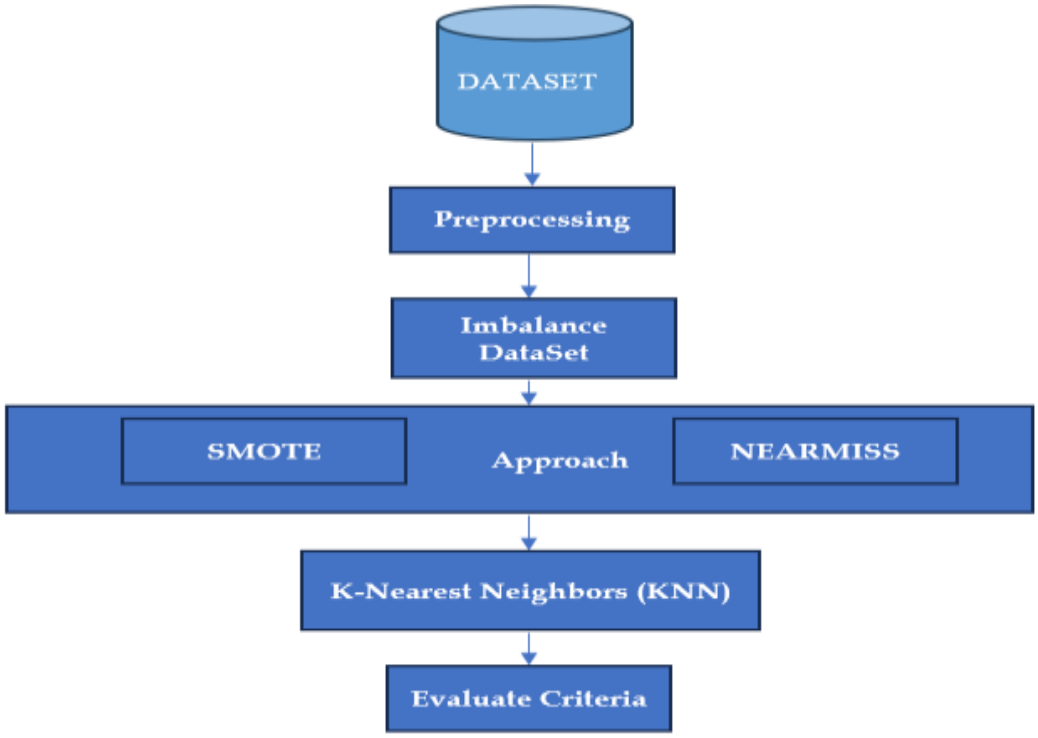


**Figure 1.** Workflow Method

### Dataset

This research requires a dataset that has class imbalance characteristics, where one or more classes have a much smaller number of samples than other classes. This research uses the

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

Stroke Dataset which is obtained from the Kaggle repository. The dataset details have 12 Attributes and 5110 total data [22].

**Preprocessing**

In this study, data pre-processing was performed as a first step before applying the model approach. Some of the data pre-processing operations performed include:

1. **Data entry**
   Identification and handling of missing values and Detection and handling of outliers that may affect the analysis.
2. **Data Normalization**
   Normalization of numerical features to ensure uniform scaling and prevent dominance of certain features in the model.
3. **Data Separation**
   Division of the dataset into a training subset and a testing subset to train and test the model.

**Dataset Analisis**

In this study, there is an imbalance in the Stroke dataset which has an impact on accuracy.

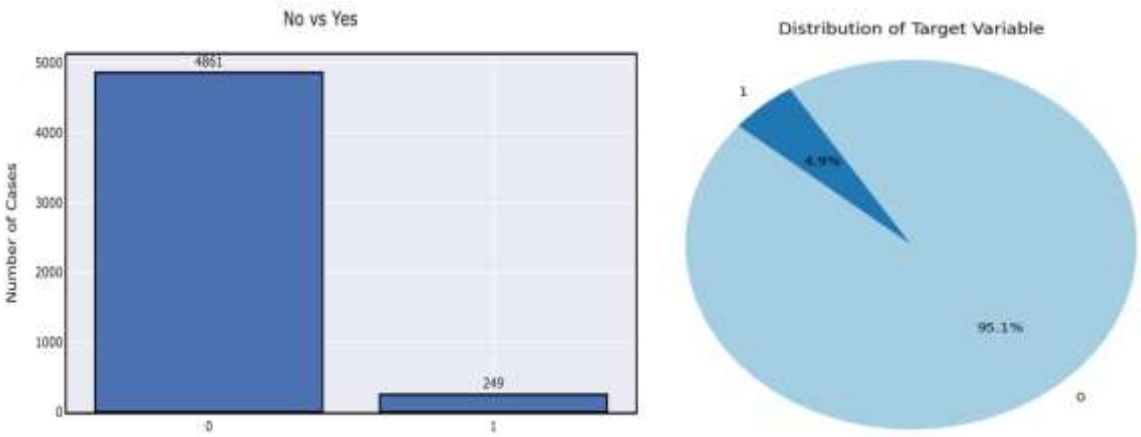


**Figure 2.** Dataset Imbalance

**Imbalance Data Problem**

The unbalanced data problem occurs in many real-world datasets where the distribution of data classes is highly unbalanced. The imbalanced data problem causes the majority class to dominate the minority class; therefore, the classifier is more inclined to the majority class, and its performance is unreliable. The Imbalance dataset shows an imbalance of 95.1% in the majority class and 4.9% in the minority class. So it is necessary to take an approach to evaluate the imbalance of the data to improve the KNN algorithm used. The approach used is Hybrid Resampling as follows.

**SMOTE (Synthetic Minority Over-sampling Technique)**

As one of the oversampling techniques to handle data imbalance in datasets. SMOTE works by creating new synthetic samples of minority classes by combining features from existing minority samples. In the context of this research, SMOTE helps in increasing the number of samples from minority classes without eliminating existing data, thus enabling a more representative data model. The synthetic minority oversampling approach is mathematically described as follows.

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

$$B_r = \left|\frac{\alpha}{y}\right|$$

Br = the balancing ratio given by Equation
B = represents an unbalanced dataset,
Alpha = denotes the subset of the sample that belongs to the minority class
y = indicates the portion of the sample that belongs to the majority class.

**Nearmiss**

One of the techniques for handling class imbalance in the dataset used in this study. The nearmiss technique is an undersampling method that is performed by selecting samples from the majority class based on heuristic rules that are based on the nearest neighbor algorithm. By using nearmiss, samples from the majority class are selected that have the closest distance to samples from the minority class. By applying the nearmiss technique, the dataset will be balanced by reducing the number of samples from the majority class so that the proportion between the majority class and the minority class becomes more balanced. This can help improve the performance of the model in better classifying the minority class without sacrificing important information from the majority class.

**K-Nearest Neighbors (KNN)**

K-NN is one of the simplest and most widely used classification algorithms. It is used to address both classification and regression problems. K-NN is a non-parametric algorithm, also called a learning algorithm. When used to make predictions, it searches for the nearest neighbor in the nearest across the data set[4]. The distance of the new data that will participate in the sample dataset is calculated according to the available data and looking at the nearest neighborhood. Three types of distance functions are commonly used for distance calculation; Euclidean distance, Manhattan Manhattan distance, and Minkowski distance[5]. In this paper, we use Euclidean Distance, which is a commonly used measure to determine the distance between input features.

Euclidean distance formula between two points $P(x1,y1)$ dan $Q(x2,y2)$ in two dimensional space is [8]:

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

In the KNN algorithm, this Euclidean distance is calculated between the unknown point (which we want to predict) and all points in the training dataset[10]. The closest points (usually determined by the value of K) are then used to determine the class or value of the unknown point based on the majority of its neighbors' class or average value.

**Evaluation Criteria**

To validate and test the model, test datasets are processed to ensure that the model produces correct results based on evaluation metrics[8]. Algorithm evaluation is generally done using various metrics, such as accuracy, precision, recall, and F-1 score. The confusion matrix is used to assess the performance of the classification model. This matrix displays the number of true positives, false positives, and false negatives. True positives are cases where the model correctly predicts a positive outcome, while true negatives are cases where the model correctly predicts a negative outcome. The number of false positives is the number of cases where the model predicts a positive outcome but the actual outcome is negative.

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

**Table 1.** Confusion Matrix

| | | Actual Class | |
|---|---|---|---|
| | | **Positive (Fraud)** | **Negative (Non-Fraud)** |
| **Predicted Class** | **Positive** | True positive (TP) | False positive (FP) |
| | **Negative** | False negative (FN) | True negative (TN) |

The formula for the confusion matrix accuracy measure is given:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}.$$

The formulas for recall and precision accuracy measures are given:

$$Recall = \frac{TP}{TP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

The formula for the accuracy measure F-1 Score is given:

$$F1 = 2 \times (Recall \times Precision)/(Recall + Precision)$$

## RESULTS AND DISCUSSION

The result of data imbalance in the Stroke dataset gives different results in the approaches used. The accuracy of K-Nearest Neighbors (KNN) only involves K-1 to K-3 neighbors. Of course, the results given in the imbalance approach are very different. Here are the results:

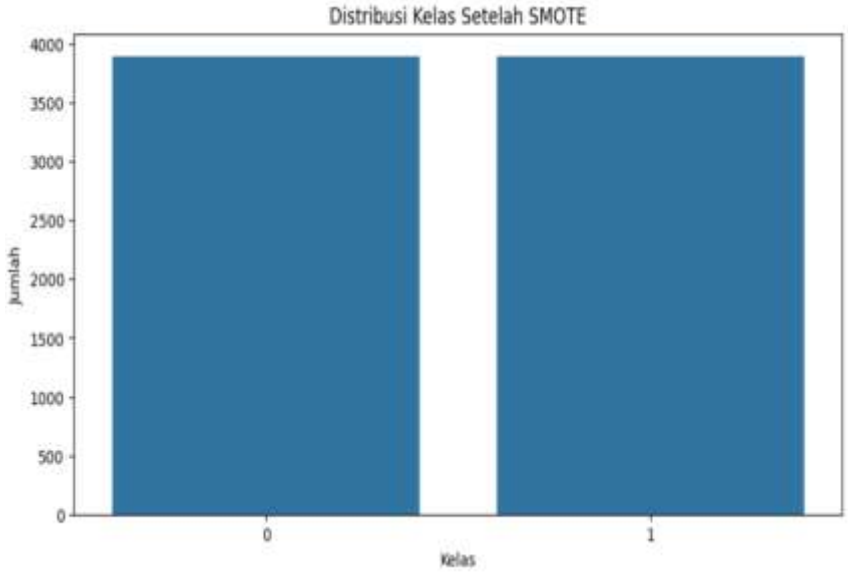


**Figure 3.** Imbalance SMOTE

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

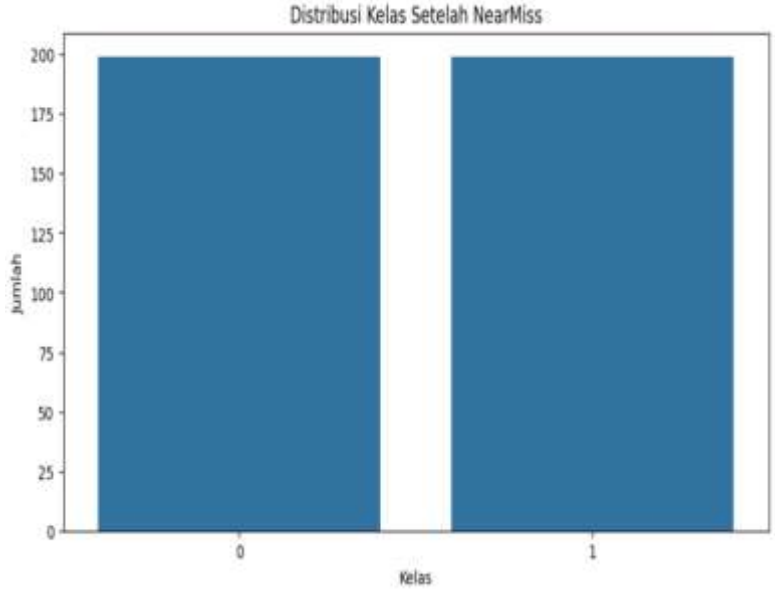


**Figure 4.** Imbalance NearMiss

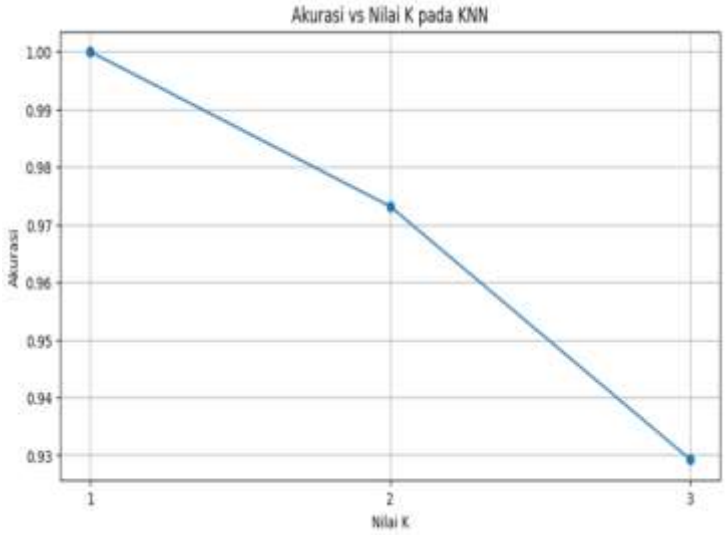


**Figure 5.** KNN + SMOTE Value Accuracy

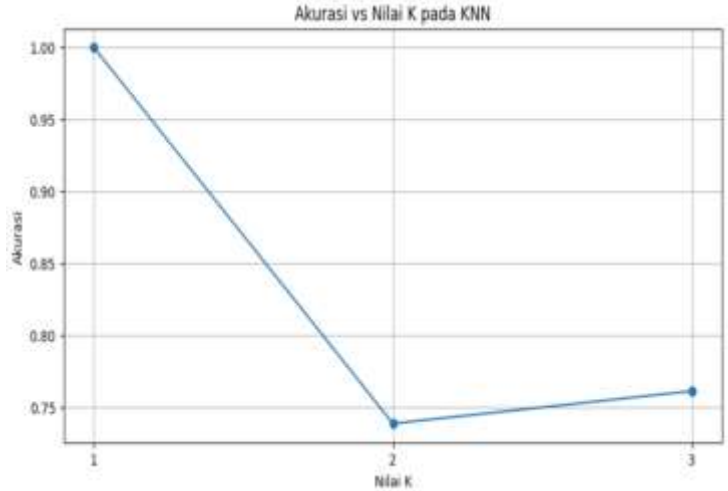


**Figure 6.** KNN + NearMiss Value Accuracy

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

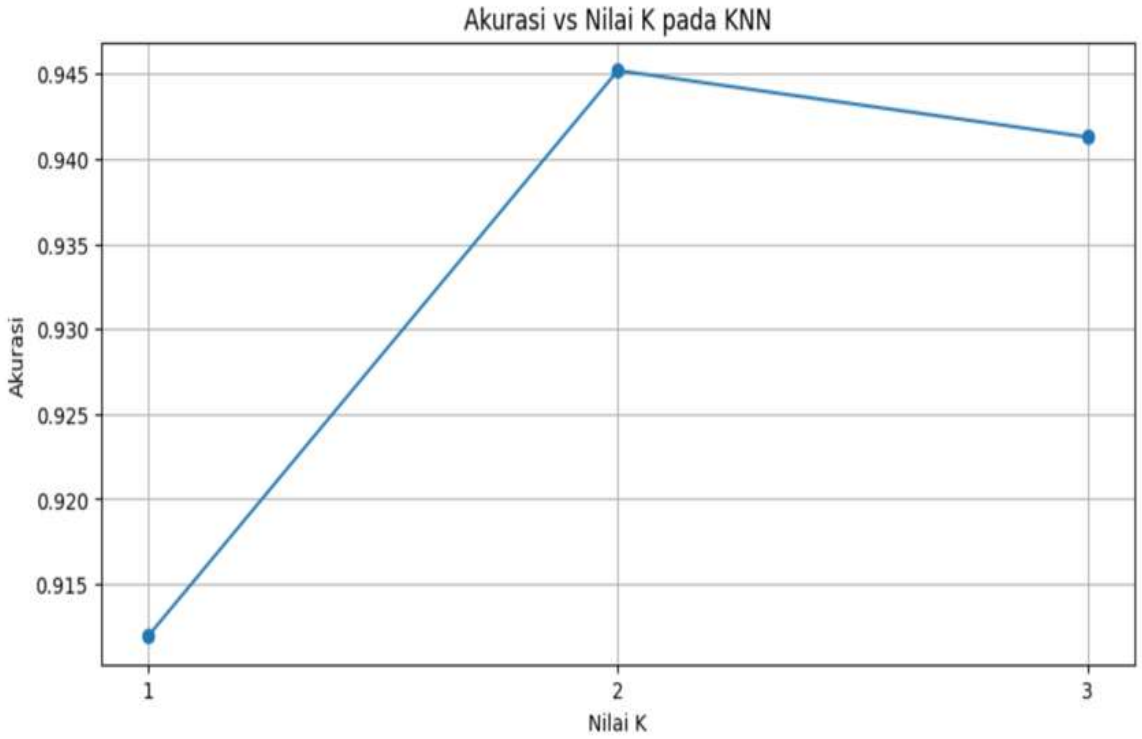


**Figure 7.** Conventional KNN Value Accuracy

In Figures 5, 6, 7, we can see the different accuracy values, we can see in the following table:

**Table 2.** K-Value Accuracy

| Nilai K | Smote | Nearmiss | Convensional |
|---|---|---|---|
| K-1 | 100% | 100% | 91% |
| K-2 | 97% | 74% | 95% |
| K-3 | 93% | 76% | 94% |

## CONCLUSION

Research on handling data imbalance in the Stroke dataset provides an increase in KNN accuracy, namely where using the SMOTE approach to handle data imbalance is very significant in the K-1 - K2 neighborhood data, which is 100% and 97% respectively but in K-3 has decreased by 93%. In this study, we also see the approach of Nearmiss which has increased in the K-1 neighborhood which is 100% but in the K-2 and K-3 neighborhoods has decreased, inversely proportional to the conventional KNN whose K-1 and K-2 accuracy movements are 91% and 95% respectively. So in this study, the effective approach obtained to improve the accuracy of KNN is the SMOTE approach, which is found at the K-1 and K3 values. In future research, it can be seen from different neighboring values, of course, it can provide different accuracy values.

## REFERENCES

[1] O. Volk and G. Singer, "An adaptive cost-sensitive learning approach in neural networks to minimize local training–test class distributions mismatch," *Intell. Syst. with*

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

*Appl.*, vol. 21, 2024, doi: 10.1016/j.iswa.2023.200316.

[2] S. N. Kalid, K. C. Khor, K. H. Ng, and G. K. Tong, "Detecting Frauds and Payment Defaults on Credit Card Data Inherited with Imbalanced Class Distribution and Overlapping Class Problems: A Systematic Review," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3362831.

[3] A. Ahmad, A. A. Rather, A. M. Gemeay, M. Nagy, L. P. Sapkota, and A. H. Mansi, "Novel sin-G class of distributions with an illustration of Lomax distribution: Properties and data analysis," *AIP Adv.*, vol. 14, no. 3, 2024, doi: 10.1063/5.0180263.

[4] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121549.

[5] C. C. Lin, D. J. Deng, C. H. Kuo, and L. Chen, "Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2912631.

[6] O. Gonzalez, "Summary Intervals for Model-Based Classification Accuracy and Consistency Indices," *Educ. Psychol. Meas.*, vol. 83, no. 2, 2023, doi: 10.1177/00131644221092347.

[7] N. G. Siddappa and T. Kampalappa, "Imbalance Data Classification Using Local Mahalanobis Distance Learning Based on Nearest Neighbor," *SN Comput. Sci.*, vol. 1, no. 2, 2020, doi: 10.1007/s42979-020-0085-x.

[8] M. Seera, C. P. Lim, A. Kumar, L. Dhamotharan, and K. H. Tan, "An intelligent payment card fraud detection system," *Ann. Oper. Res.*, vol. 334, no. 1–3, 2024, doi: 10.1007/s10479-021-04149-2.

[9] X. Zhu *et al.*, "Intelligent financial fraud detection practices in post-pandemic era," *Innovation*, vol. 2, no. 4. 2021, doi: 10.1016/j.xinn.2021.100176.

[10] A. Ali *et al.*, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Applied Sciences (Switzerland)*, vol. 12, no. 19. 2022, doi: 10.3390/app12199637.

[11] M. Uppal *et al.*, "Enhancing accuracy in brain stroke detection: Multi-layer perceptron with Adadelta, RMSProp and AdaMax optimizers," *Front. Bioeng. Biotechnol.*, vol. 11, 2023, doi: 10.3389/fbioe.2023.1257591.

[12] I. Kosmidis, E. C. Kenne Pagui, and N. Sartori, "Mean and median bias reduction in generalized linear models," *Stat. Comput.*, vol. 30, no. 1, 2020, doi: 10.1007/s11222-019-09860-6.

[13] C. Sweeney, E. Ennis, M. Mulvenna, R. Bond, and S. O'neill, "How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups," *Computers*, vol. 11, no. 5, 2022, doi: 10.3390/computers11050083.

[14] Q. Nguyen, I. Diaz-Rainey, A. Kitto, B. I. McNeil, N. A. Pittman, and R. Zhang, "Scope 3 emissions: Data quality and machine learning prediction accuracy," *PLOS Clim.*, vol. 2, no. 11, 2023, doi: 10.1371/journal.pclm.0000208.

[15] Z. Nasreddine, V. Garibotto, S. Kyaga, and A. Padovani, "The Early Diagnosis of Alzheimer's Disease: A Patient-Centred Conversation with the Care Team," *Neurology and Therapy*, vol. 12, no. 1. 2023, doi: 10.1007/s40120-022-00428-7.

[16] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Inf. Softw. Technol.*, vol. 139, 2021, doi: 10.1016/j.infsof.2021.106662.

[17] A. Li, M. Liu, and S. Sheather, "Predicting stock splits using ensemble machine learning and SMOTE oversampling," *Pacific Basin Financ. J.*, vol. 78, 2023, doi: 10.1016/j.pacfin.2023.101948.

[18] M. Alauthman *et al.*, "Enhancing Small Medical Dataset Classification Performance

Page : 78-87

Doi. https://doi.org/10.54209/jurnalinstall.v16i02.207

Using GAN," *Informatics*, vol. 10, no. 1, 2023, doi: 10.3390/informatics10010028.

[19] X. Wang, H. Zhang, S. Bai, and Y. Yue, "Design of agile satellite constellation based on hybrid-resampling particle swarm optimization method," *Acta Astronaut.*, vol. 178, 2021, doi: 10.1016/j.actaastro.2020.09.040.

[20] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[21] G. AlMahadin, A. Lotfi, M. M. Carthy, and P. Breedon, "Enhanced Parkinson's Disease Tremor Severity Classification by Combining Signal Processing with Resampling Techniques," *SN Comput. Sci.*, vol. 3, no. 1, 2022, doi: 10.1007/s42979-021-00953-6.

[22] A. Azab, M. Khasawneh, S. Alrabaee, K. K. R. Choo, and M. Sarsour, "Network traffic classification: Techniques, datasets, and challenges," *Digital Communications and Networks*. 2023, doi: 10.1016/j.dcan.2022.09.009.