ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

OPTIMIZING DIABETIC NEUROPATHY SEVERITY CLASSIFICATION USING ELECTROMYOGRAPHY SIGNALS THROUGH SYNTHETIC OVERSAMPLING TECHNIQUES

I Ketut Adi Purnawan[1,5], Adhi Dharma Wibawa[2], Arik Kurniawati[3], Mauridhi Hery Purnomo[1,4]*

[1]Departement of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Indonesia
[2]Department of Medical Technology and Engineering, Institut Teknologi Sepuluh Nopember, Indonesia
[3]Department of Informatics, Universitas Trunojoyo Madura, Indonesia
[4]University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS)
[5]Information Technology of Electrical Engineering, Universitas Udayana, Indonesia

email: 07111960010001@student.its.ac.id[1], adipurnawan@unud.ac.id[1] ,adhiosa@ee.its.ac.id[2], arik.kurniawati@trunojoyo.ac.id[3], hery@ee.its.ac.id[4]*

Abstract

This study investigates the use of Synthetic Minority Oversampling Technique (SMOTE) and Random Over-Sampling (ROS) to improve diabetic neuropathy severity classification based on electromyography (EMG) signals. EMG signals capture electrical activity in muscles, providing critical information about nerve function and muscle health, which are affected in neuropathy. By analyzing EMG signals, we aim to develop a robust method for early detection and accurate classification of neuropathy severity. Our approach utilizes XGBoost in combination with SMOTE to address data imbalance issues, achieving an accuracy of 92%, an F1-score of 0.91, and a recall of 0.93. This study demonstrates that oversampling techniques tailored for EMG data can enhance classification performance, offering a valuable tool for clinical assessments of diabetic neuropathy.

Keywords : electromyography, diabetes neuropathy, ROS, SMOTE, XG Boost

Received: 12-10-2024 | **Revised:** 01-11-2024 | **Accepted:** 02-11-2024
DOI: https://doi.org/10.23887/janapati.v13i3.85675

INTRODUCTION

Diabetes also commonly causes diabetic neuropathy, which can severely affect the nerves and significantly reduce quality of life [1]. While there have been studies discussing the management of DN, this study fills a specific need by classifying the severity of neuropathy using electromyography signals, which are less discussed in previous studies. Accurate estimation and timely grading of neuropathy is crucial because it allows early intervention and care that can delay or limit the deterioration of a nerve 2], [3]. In real clinical practice, reliable grading of severity is difficult due to variability factors and scarcity of EMG data in different patients with unknown severity of neuropathy.

EMG signals are electrical signals generated by muscle activity and are very useful for analyzing the health of muscles and nerves. However, data imbalance is a common problem when processing EMG signal data, especially for datasets involving patients with heterogeneous conditions and limited data. Lack of data along with data imbalance may cause the machine learning model to be overfitted to the majority class and fail to recognize the minority class minority [4], [5], [6], and [7]. If the minority classes are not properly considered, the classification models may not generalize well, leading to misclassification in critical healthcare scenarios. This paucity of data doesn't allow distinguishing relevant patterns, so it became necessary to develop methods to generate synthetic data by augmenting an existing dataset.

Class imbalance in datasets is a challenge for most machine learning algorithms because it sometimes leads to biased predictions towards the majority class at the expense of not considering the representatives of the minority class [5], [6]. Moreover, the limitations of the data sets compound these problems, sometimes causing models to underperform in predicting their targets [7]. In the context of diabetic neuropathy, accurate classification across all severity levels is essential for effective patient management and improved long-term health outcomes [8]. This is very important.

Because of this challenge, various researchers have proposed different resampling techniques, some of which are oversampling and undersampling methods [4]. A few works showed that most of the oversampling methods, such as Synthetic Minority Oversampling Technique, generally performed better than various undersampling techniques to balance the data sets [5]. Geometric SMOTE is an advanced version of SMOTE that showed much better prediction accuracy compared to the original small data set and other sampling methods. Thus, the integration of SMOTE with SVM has solved the problem of a small data set in binary classification. This technique has been effectively used by researchers in integrating SMOTE with the support vector machine for binary classification. Because of this integration, this work focuses on the evaluation of model performance with metrics besides accuracy to include precision, recall and F1 measure essential to ensure proper categorization across all classes [4], [5].

The purpose of this investigation is to [refine the identification of diabetic neuropathy severity] as measured by electromyography (EMG) signals, to provide a more direct, reliable assessment of neuropathy severity based on signal patterns. Most studies to date have focused on classifying diabetic patients using secondary data such as laboratory results and other clinical details [9], [10]. ROS and SMOTE are used in the data augmentation technique to solve class imbalance problems and thus optimize the performance of the classification model in classes with few data points, as suggested by several literatures [10], [11], [12].

The common strategy to handle the class imbalance problem involves the application of various synthetic oversampling techniques [13]. The two most commonly applied techniques for this include the Synthetic Minority Over Sampling Technique and Random Over Sampling; the former generates synthetic samples from existing minority samples, while the latter simply generates multiple copies of the already existing minority samples. This has been a very successful approach to improve the performance of clinical data classification models [14]. Such techniques increase the variation in the data, allowing the model to learn more complex patterns that improve the robustness and accuracy of machine learning models.

In this study, the XGBoost (Extreme Gradient Boosting) model was used for the subsequent phase because of its demonstrated effectiveness in handling datasets with many features. The use of XGBoost in conjunction with oversampling techniques has recently shown promising results in diabetes classification studies, and its adaptability to high-dimensional datasets makes it particularly well-suited for this task. XGBoost is quite efficient for a dataset with a large number of features. It achieves its maximum accuracy of 99% when used in combination with SMOTE at an F1 score of 1.00 [15]. The use of XGBoost in combination with SMOTE effectively works to mitigate the problems of data imbalance, thus providing better classification results for all classes in a balanced manner [16]. Furthermore, the ability of XGBoost to correct errors in previously generated models through weight adjustments is an additional reason for its suitability for the goals of this study [17]. The potential for enhancing diabetes-related classification models using this approach has been clearly demonstrated. We assume that the result of this work will contribute greatly to the construction of a more accurate model for grading the severity of neuropathy in diabetic patients. The novelty of the present work is divided into three parts: first, the focus on direct EMG-based classification; second, the handling of data imbalance by advanced resampling; and third, the high-precision classification by XGBoost with SMOTE. The key contributions of this study are detailed below:

1. Using electromyography (EMG) signals.
2. The Over-Sampling Technique approach used to generate data extracted from time domain and frquency domain.
3. The Over-Sampling Technique approach can enhance accuracy in classifying the severity of diabetic neuropathy.

METHOD

The investigation commences with the acquisition of EMG signal data utilizing the MyoMES instrument. The acquired data undergoes processing through feature extraction in both time and frequency domains, yielding a total of 126 features for analysis. We employ oversampling techniques, specifically random oversampling (ROS) and synthetic minority oversampling technique (SMOTE), to rectify data imbalance prior to inputting into the classification model. We stratify the data and afterwards providing it into the classification model which is XG Boost. In the end, we assessed the performance of the suggested model after performing the above oversampling technique using accuracy, recall, precision, and F1 score etc. Figure 1 provides a summary of the processes followed in this study.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

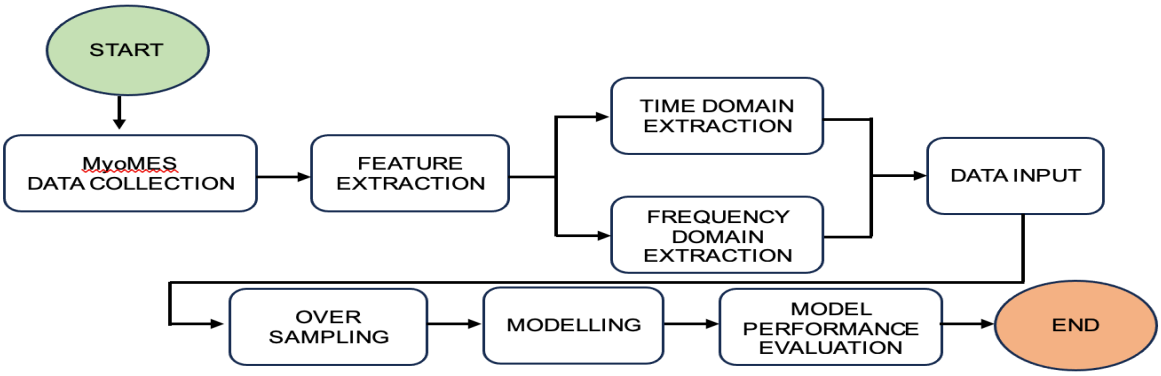

Figure 1. Proposed Method using Oversampling

1. Data Collection

In this study, the MyoMES tool was used. This tool was developed by the Center for Bio Mechanics, Bio Materials, Bio Mechatronics, and Bio Signal Processing (CBIOM3S) at the University of Diponegoro, Semarang [18], to analyze participants' muscular contractions. The participants comprised 32 individuals, encompassing both healthy patients and diabetic patients exhibiting varied levels of muscular stiffness. We employed four classification groups: healthy individuals (normal), diabetic individuals with axonal impairments (axonal), demyelinating conditions (demyelinating), and a combination of both (mixed). We acquired data on MyoMES using gain settings of 3, with a sampling frequency of 1000 Hz and a data collection interval of 50 ms. A non-inverting amplifier amplifies the resultant s'EMG signal to an appropriate level, especially for weak signals (1–10 mV).We utilize an amplification of 500–1000 times, particularly for muscles that demonstrate strong reaction signals, such as the biceps brachii.

2. Data Preprocessing

The analysis that is carried out in the next step can be divided into two parts: the temporal feature extraction of data and its frequency analysis. In the time domain, each data row contains 15 features, which gives a total of 90 features as each respondent provides answers to 6 questions. Moving on, we proceed to feature extraction considering data in the frequency domain, and this is achieved with each data row containing 6 features, thus giving rise to an overall of 36 features in multiplication. We then bring together the two types of feature outputs to offer a more holistic view towards the classification of the diabetic neuropathy severity resulting in additional 126 features 1 target feature 4 classes.

The model's ability to learn from the data is limited by the insufficient number of respondents. Consequently, this research employs two oversampling approaches to evaluate their impact on enhancing the model's learning performance. This research employs two oversampling strategies. The initial method is random oversampling (ROS). We deem this technique appropriate due to its simplicity and efficacy in handling imbalanced or sparse data.

ROS markedly enhances the efficacy of classification models by equilibrating class representations without introducing extra complexity to the model or data, particularly in scenarios where the preservation of the original data's integrity is paramount. Numerous studies substantiate these advantages, illustrating the efficacy of the ROS technique across diverse application fields. Figure 2 illustrates the capability of the ROS algorithm to produce synthetic data, thereby increasing the dataset's volume.

SMOTE (Synthetic Minority Over Sampling Technique) is a commonly used oversampling technique to address the problem of data imbalance. Adding noise to the generated data is another method often used with SMOTE. This is more often the case when employing SMOTE since the technique secondary advantage is producing more data by creating the in-between samples of the minority classes. Figure 3 illustrates how the SMOTE algorithm synthetically generates data to increase the overall size of the available data.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

```
1. INPUT:
   - Dataset D with features X and target labels y
   - Class distribution C (number of samples per class)
   - Target total size for the dataset (target_size)
   - Calculate the number of unique classes U in the dataset
(unique_classes)
   - Calculate the desired number of samples per class (target_class_size)
as target_size / U
2. FOR each class in unique_classes:
   - Identify the current number of samples for the class
(current_class_size)
   - IF the current_class_size is less than the target_class_size:
     - Calculate the number of additional samples needed as:
       additional_samples_needed = target_class_size - current_class_size
     - WHILE additional_samples_needed > 0:
         - Randomly select samples from the current class
         - Duplicate those samples and add them to the class
         - Decrease additional_samples_needed by the number of samples
duplicated
3. COMBINE the newly created samples with the original dataset D
4. RETURN the augmented dataset with the target number of samples
(target_size)
```

Figure 2. Random Over-Sampling Pseudocode

```
Input: minority_data (32 samples), N (target: 200), k (default: 5)
# For each sample xi in the minority_data
For each xi in minority_data:
    # Step 1: Find k-nearest neighbors of xi
    neighbors = find_k_nearest_neighbors(xi, k)
    # Step 2: Randomly pick a neighbor, xj
    xj = randomly_select(neighbors)
    # Step 3: Generate a new sample
    new_sample = xi + random(0, 1) * (xj - xi)
    # Add new_sample to synthetic_data
Repeat this process until we have generated (200 - 32) samples.
Output: augmented_data = original_data + synthetic_data
```

Figure 3. Synthetic Minority Over-Sampling Technique Pseudocode

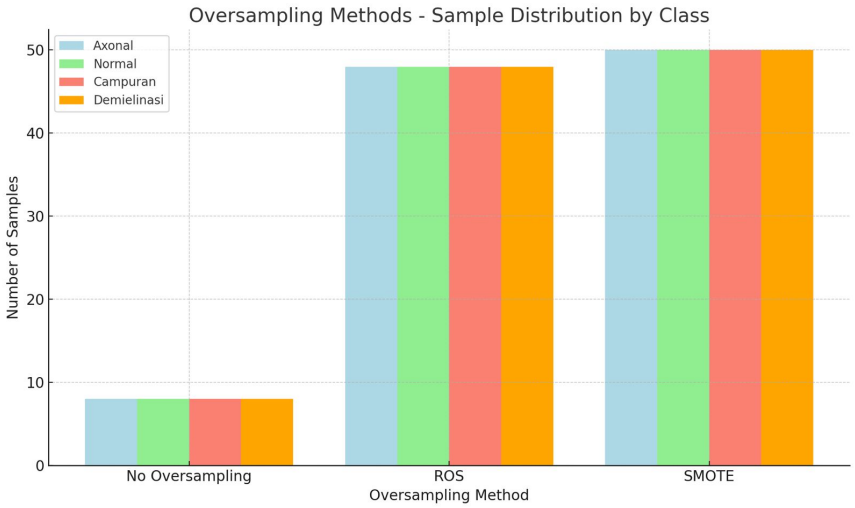

Figure 4. Data Distribution

Tabel 1 Over sampling Result

Oversampling Method	Normal	Axonal	Demie	Mix	Total
Without Oversampling	8	8	8	8	32
ROS	48	48	48	48	192
SMOTE	50	50	50	50	200

Tabel 2 Six best features of the Over Sampling ROS technique

	RTA_tot	LE_tot	RE_tot	RC_tot	LTA_tot	LC_tot
count	192	192	192	192	192	192
max	4462241	3879519	3471291	3171900	3114599	2377253
min	4.221.191	17249.5	22877.11	6.039.356	3.255.423	3.991.617
mean	1068121	1732186	1636148	1266395	1383796	1206180
std	967386.6	890079.8	773849.3	730474.8	1075871	736697

Tabel 3 Six best features of the Over Sampling SMOTE technique

	RTA_tot	LE_tot	RE_tot	RC_tot	LTA_tot	LC_tot
count	200	200	200	200	200	200
max	4462241	3879519	3471291	3171900	3114599	2377252
min	4.221.229	17249.68	22877.14	6.039.425	3.255.381	3.991.713
mean	1052906	1679313	1671133	1243934	1319351	1224318
std	791130.6	783889.5	773211.1	695916	906910.6	636791.3

3. Data Distribution

The use of oversampling in this research incorporated an additional 5% of random variance in the data. The choice of a 5% noise level was based on preliminary experiments and literature indicating that this level introduces sufficient variability to simulate real-world conditions without significantly distorting signal characteristics [19], [20]. The objective of incorporating this degree of variation is to emulate uncertainty or diversity values that reflect reality. Table 1 presents the outcomes of each approach across four variants of neuropathic diabetes.

In the absence of oversampling as illustrated in Figure 4, there exist merely 8 samples within each of the four categories (Normal, Axonal, Demie, and Mix), culminating in a total of 32 samples. Implementing the random oversampling (ROS) technique increases the sample count in each category to 48, yielding a total of 192 samples. Similarly, SMOTE (Synthetic Minority Oversampling Technique) generates 50 samples for each category, resulting in a total of 200 samples. Both oversampling methods substantially augment the sample size, mitigating the problem of class imbalance. SMOTE yields a somewhat greater sample count (200) than ROS (192).

To evaluate whether the ROS and SMOTE oversampling techniques preserve the distribution and variability of the original data while achieving dataset balance, Tables 2 and 3 present statistical data (count, minimum, maximum, mean, and standard deviation) for each feature.

ROS effectively equilibrates the data across all categories while preserving variability in the characteristics. Conversely, SMOTE augments the sample size while yielding somewhat reduced variability, as seen by lower standard deviation values in comparison to ROS. Regarding sample size, SMOTE produces somewhat more samples (200) than ROS (192).

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

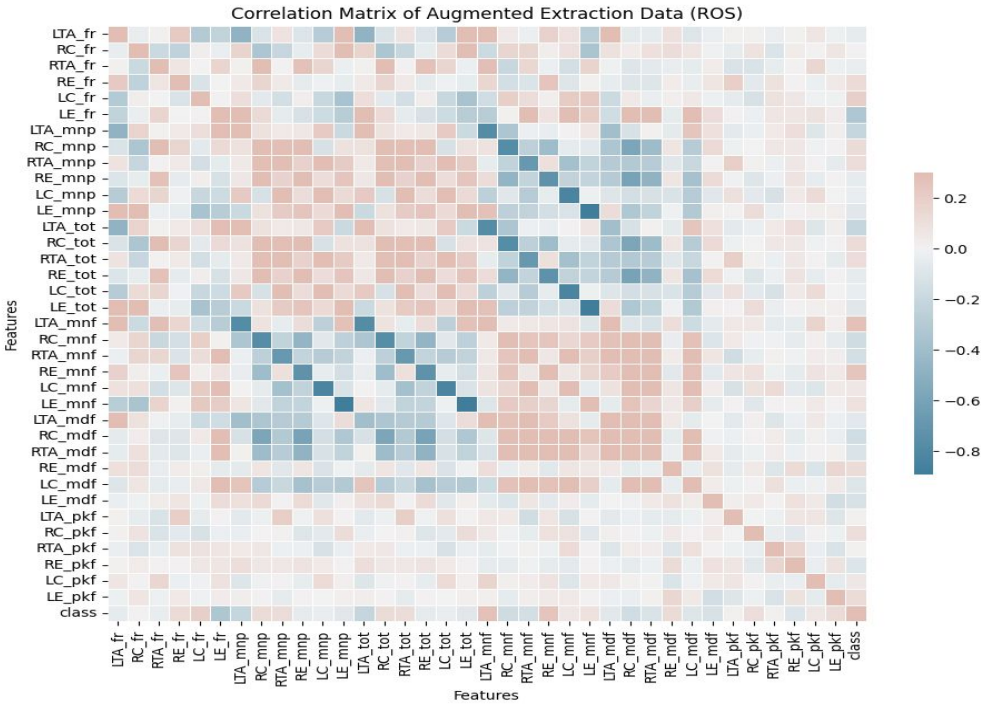

Figure 5. Correlation matrix between features of the ROS technique

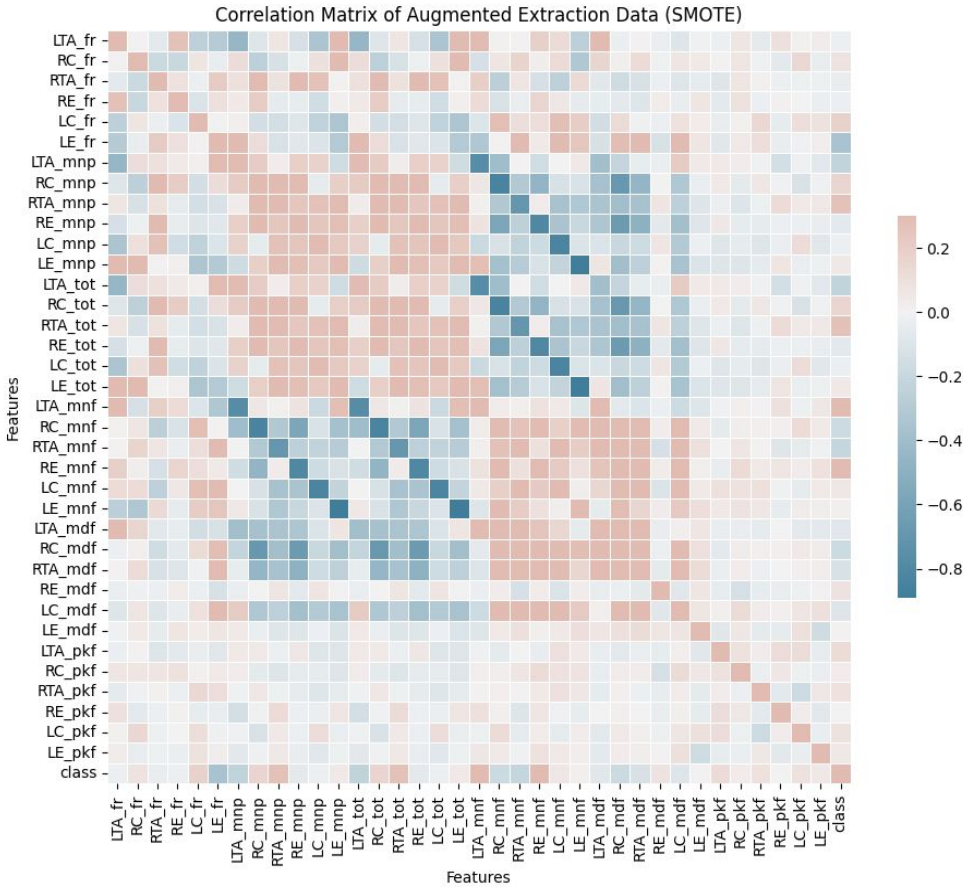

Figure 6. Correlation matrix between features of the SMOTE technique

4. Data Validation

This research analyzes generator data findings through a correlation matrix to ascertain the dependencies among features of the augmented data generated by the ROS and SMOTE procedures. We use this correlation matrix to understand how the two strategies affect the interrelationships among the features in the dataset.

Figures 5 and 6 demonstrate that the correlation patterns between characteristics exhibit notable similarities between ROS and SMOTE. Both strategies preserve a uniform data structure, marked by blocks of significant correlation along the matrix's diagonal. The features LTA_fr, RC_fr, and RTA_fr demonstrate strong correlations in both matrices, indicating a relatively consistent relational pattern between ROS and SMOTE.

Nonetheless, disparities exist in the correlation strength of features in the two matrices. SMOTE typically generates more distinct and pronounced correlations (shown by darker hues) in specific feature blocks, such as the MNF, TOT, and MDF features. On the other hand, ROS tends to establish a more diffuse connection with lower intensity within the same blocks. The operational mechanism of SMOTE, which generates synthetic instances based on nearest neighbors and potentially amplifies certain correlations, explains the disparity, while ROS only replicates existing instances.

Both matrices show a significant degree of similarity in the correlation between characteristics and class labels, indicating that both oversampling strategies influence the class distribution similarly.

5. Classification with ensemble learning models (XGBoost)

This research employs an ensemble learning model that incorporates XGBoost (eXtreme Gradient Boosting) for model construction. Engineered to enhance efficiency and memory use, XGBoost is a scalable tree-boosting framework that excels in managing extensive and intricate datasets with numerous characteristics. This technique operates by building a new model that predicts the residuals of the previous model, gradually integrating them until it minimizes error and stabilizes accuracy [21], [22].

The process of building an XGBoost tree involves the following steps:

1. Initialize the initial prediction probability (Pri_1) for each instance, where i = 1, 2, ..., n.
2. Compute the residuals using the equation:

$$\text{Residual}_i^t = Y_i - P_{ri}^t$$

3. Calculate the cover value of each attribute by using the following equation:

$$\text{Cover}(A) = \sum_{i=1}^{n} \left(P_{ri}^t(1 - P_{ri}^t)\right)$$

4. Determine the similarity score (SS) by using the following equation:

$$SS_{\text{node}} = \frac{(\sum_{i=1}^{n} \text{Residual}_i)^2}{\sum_{i=1}^{n} \left(P_{ri}^t(1 - P_{ri}^t)\right) + \lambda}$$

5. Compute the attribute gain value:

$$\text{Gain}(\text{A}) = SS_{\text{left}} + SS_{\text{right}} - SS_{\text{root}}$$

6. Compute the leaf output:

$$\text{Output}(\text{A})_\text{i} = \frac{\sum_{\text{i=1}}^{\text{n}} \text{Residual}_\text{i}}{\sum_{\text{i=1}}^{\text{n}} \left(\text{P}_\text{ri}(1 - \text{P}_\text{ri})\right) + \lambda}$$

7. Determine the log odds:

$$\log \text{odds}_\text{i}^\text{t} = \log\left(\frac{\text{P}_\text{ri}^\text{t}}{1 - \text{P}_\text{ri}^\text{t}}\right)$$

8. Update the probability value:

$$\text{P}_\text{ri}^{\text{t+1}} = \log \text{odds}_\text{i}^\text{t} + (\eta \times \text{Output}(\text{A})_\text{i})$$

9. Use the sigmoid function to normalize the probability value:

$$\text{Sigmoid}(\text{P}_\text{ri}^{\text{t+1}}) = \frac{\exp(\text{P}_\text{ri}^{\text{t+1}})}{1 + \exp(\text{P}_\text{ri}^{\text{t+1}})}$$

In developing this model, cross-validation with k = 10 was included in the process to assess the performance of this model. The cross-validation technique divides the data into multiple folds to make the model generalize knowledge from patterns in specific subsets of the data in addition to the entire data set. This is important in those datasets that have very large features with few samples, making the application of this boosting method very suitable to improve the performance of the model.

XGBoost has many advanced features; it can prevent overfitting through regularization, find the most important features, and handle multiple hyperparameters. Among the advanced features available in XGBoost, GridSearchCV provides a means to perform hyperparameter optimization. Its hyperparameters include learning_rate, n_estimators, and max_depth, which are tuned through cross-validation by performing grid searches.

GridSearchCV carefully checks each combination by breaking the dataset into many smaller groups and performing validation to make sure the tuning results are accurate and lower the risk of overfitting. GridSearchCV autonomously identifies the optimal combination based on

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

model performance measures, such as accuracy or F1 score, derived from cross-validation outcomes, after assessing all hyperparameter combinations.

RESULT AND DISCUSSION

We use a confusion matrix to evaluate the influence of the model, both with and without the use of data augmentation techniques like ROS and SMOTE. The confusion matrix categorizes predictions into four classifications: true positive (TP), true negative (TN), false positive (FP), and false negative (FN), facilitating a thorough assessment of model efficacy.

The main evaluation methods used to measure model performance include:

1. Accuracy: The percentage of correct predictions out of the total number of predictions.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\backslash\%$$

2. Precision: Precision: The ratio of true positives to all positives predicted.

$$\text{Presisi} = \frac{TP}{TP + FP} \times 100\backslash\%$$

3. Recall: The actual ratio of positive cases that was correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\backslash\%$$

4. F1 Score: A harmonious mean of precision and recall, which provides a balanced measure between both indicators.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivitas}}{\text{Precision} + \text{Sensitivitas}}$$

Figure 7 demonstrates that, without oversampling strategies, the classification model demonstrates a significantly low accuracy level of only 40%. This results from a data imbalance between the majority and minority classes, leading the algorithm to predominantly forecast the majority class while neglecting the minority class. This research encompasses multiple approaches, including the utilization of oversampling techniques such as random oversampling (ROS) and synthetic minority oversampling (SMOTE).

ROS functions by enhancing the dataset through the replication of minority class samples. This approach, while potentially prone to overfitting, provides benefits for smaller datasets by allowing the model to identify new patterns. SMOTE synthesizes new data by creating interpolations between minority class samples. This new variation increases the diversity of patterns in a dataset without simply replicating existing data. In this work, we show that this approach significantly increases accuracy and yields a high F1 score, which is a balance between recall and precision.

In this paper, XGBoost is optimized by an oversampling strategy to improve model performance. By using appropriate oversampling strategies, models can learn from both the majority and often ignored minority patterns of the data, resulting in more accurate and fairer classification. Figure 7 shows the complete model performance results and shows that SMOTE gives much better results compared to other methods.

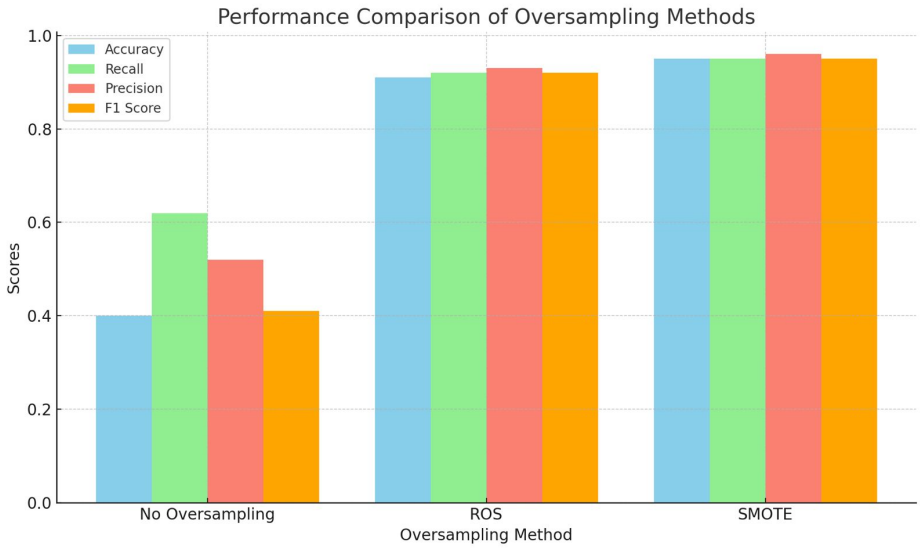

Figure 7. Performance accuracy metrics

CONCLUSION

This study demonstrates the effectiveness of using XGBoost with oversampling techniques like SMOTE and ROS to improve the classification of diabetic neuropathy severity based on EMG signals. By directly addressing data imbalance and enhancing the model's ability to identify minority classes, our approach provides a reliable tool for diagnosing and monitoring neuropathy severity. In real-world applications, this model could be integrated into clinical diagnostic tools to assist healthcare providers in making early and accurate assessments of diabetic neuropathy. This has the potential to improve patient outcomes by enabling timely interventions and personalized treatment strategies based on the severity of the condition. Future studies could further validate this model in diverse clinical settings, ensuring its adaptability and effectiveness across various patient populations

Therefore, with oversampling, the performance of the proposed model is expected to improve the performance of diabetic neuropathy severity classification by a large margin. While performing class balancing using SMOTE and ROS, we note the inherent limitations of both methods: SMOTE includes artificial samples, which can lead to a decrease in variability compared to real samples, and ROS runs the risk of overfitting, as the duplicated samples can be remembered by the model. This could also be compared in the future with more sophisticated techniques, such as ADASYN or cost-sensitive learning methods, which may better represent minority classes without overfitting. In addition, studying the risks of overfitting by ROS may improve the reliability of the model in clinical practice. We believe that this may be one of the important lines that can be taken to robustify model applicability and classification strength in this direction.

ACKNOWLEDGMENT

This research was financially supported by the Indonesian Endowment Fund for Education (LPDP) Republic of Indonesia under Grand No. 0004666/TRP/D/BUDI-2019.

REFERENCES

[1] R. Pop-Busui *et al.*, "Diabetic Neuropathy: A Position Statement by the American Diabetes Association," *Diabetes Care*, vol. 40, no. 1, pp. 136–154, Oct. 2017, doi: 10.2337/dc16-2042.

[2] F. Shakeel, A. S. Sabhitha, and S. Sharma, "Exploratory review on class imbalance problem: An overview," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, Oct. 2017, pp. 1–8. doi: 10.1109/ICCCNT.2017.8204150.

[3] N. Abdelhamid, A. Padmavathy, D. Peebles, F. Thabtah, and D. Goulder-Horobin, "Data Imbalance in Autism Pre-Diagnosis Classification Systems: An Experimental Study," *Journal of Information & Knowledge Management*, vol. 19, no. 01, p. 2040014, Oct. 2020, doi: 10.1142/S0219649220400146.

[4] T. S. Amelia, M. N. S. Hasibuan, and R. Pane, "Comparative analysis of resampling techniques on Machine Learning algorithm," *Sinkron*, vol. 7, no. 2, pp. 628–634, Oct. 2022, doi: 10.33395/sinkron.v7i2.11427.

[5] A. S. Ashraf and T. Ahmed, "MACHINE LEARNING SHREWD APPROACH FOR AN IMBALANCED DATASET CONVERSION SAMPLES," *Journal of Engineeringand Technology*, vol. 11, no. 1, 2020.

[6] N. U. Niaz, K. M. N. Shahariar, and M. J. A. Patwary, "Class Imbalance Problems in Machine Learning: A Review of Methods And Future Challenges," in *Proceedings of the 2nd International Conference on Computing Advancements*, ACM, Oct. 2022, pp. 485–490. doi: 10.1145/3542954.3543024.

[7] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 6, no. 3, p. 379, Dec. 2020, doi: 10.26418/jp.v6i3.42896.

[8] S. V. Narwane and S. D. Sawarkar, "Is handling unbalanced datasets for machine learning uplifts system performance?: A case of diabetic prediction," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 16, no. 9, p. 102609, Oct. 2022, doi: 10.1016/j.dsx.2022.102609.

[9] M. A. Wiratama and W. M. Pradnya, "Optimasi Algoritma Data Mining Menggunakan Backward Elimination untuk Klasifikasi Penyakit Diabetes," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 1, p. 1, Apr. 2022, doi: 10.23887/janapati.v11i1.45282.

[10] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, "The Impact of Oversampling with SMOTE on the

Performance of 3 Classifiers in Prediction of Type 2 Diabetes," *Medical Decision Making*, vol. 36, no. 1, pp. 137–144, Jan. 2016, doi: 10.1177/0272989X14560647.

[11] A. T. Akbar, R. Husaini, and H. Prapcoyo, "Preprocessing Using SMOTE and K-Means for Classification by Logistic Regression on Pima Indian Diabetes Dataset," *Telematika*, vol. 20, no. 2, p. 238, Jun. 2023, doi: 10.31315/telematika.v20i2.9676.

[12] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[13] T. Riston *et al.*, "Oversampling Methods for Handling Imbalance Data in Binary Classification," 2023, pp. 3–23. doi: 10.1007/978-3-031-37108-0_1.

[14] F. Mohd, M. A. Jalil, N. M. M. Noora, S. Ismail, W. F. F. Yahya, and M. Mohamad, "Improving Accuracy of Imbalanced Clinical Data Classification Using Synthetic Minority Over-Sampling Technique," 2019, pp. 99–110. doi: 10.1007/978-3-030-36365-9_8.

[15] N. M. Nayan, A. Islam, M. U. Islam, E. Ahmed, M. M. Hossain, and M. Z. Alam, "SMOTE Oversampling and Near Miss Undersampling Based Diabetes Diagnosis from Imbalanced Dataset with XAI Visualization," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ISCC58397.2023.10218281.

[16] K. S. Gill, V. Anand, D. Upadhyay, and S. Dangi, "Diabetes Classification Using XG Boost Classification Techniques Through Machine Learning based SMOTE Analysis," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, IEEE, Oct. 2024, pp. 1–4. doi: 10.1109/INOCON60754.2024.10512046.

[17] K. D. K. Wardhani and M. Akbar, "Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost)," *Jurnal Online Informatika*, vol. 7, no. 2, pp. 244–250, Oct. 2022, doi: 10.15575/join.v7i2.970.

[18] R. Ismail, "Muscle Power Signal Acquisition Monitoring Using Surface EMG," *J Biomed Res Environ Sci*, vol. 3, no. 5, pp. 663–667, May 2022, doi: 10.37871/jbres1493.

[19] M. Momeny *et al.*, "Learning-to-augment strategy using noisy and denoised data: Improving generalizability of deep CNN for the detection of COVID-19 in X-ray images," *Computers in Biology and Medicine*, vol. 136, p. 104704, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104704.

[20] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 2, pp. 539–550, Apr. 2009, doi: 10.1109/TSMCB.2008.2007853.

[21] M. Arslan, M. Guzel, M. Demirci, and S. Ozdemir, "SMOTE and Gaussian Noise Based Sensor Data Augmentation," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/UBMK.2019.8907003.

[22] M. W. Dwinanda, N. Satyahadewi, and W. Andani, "CLASSIFICATION OF STUDENT GRADUATION STATUS USING XGBOOST ALGORITHM," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1785–1794, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.