Indonesian Journal of Infomatics

E-ISSN: 3124-033X

Research Article

The Effect of Data Imbalance on the Interpretation Stability of LIME-Based Explainable AI on Nutritional Status Prediction Models

Sri Nurhayati[1*], Hidayat[2], Siti Ar-Rachmi Ningrum [3], Zainal Arifin Hasibuan[4], Sri Supatmi[5]

1 Doctoral Program in Information System, Universitas Komputer Indonesia, Bandung, Indonesia; e-mail: si.75725002@mahasiswa.unikom.ac.id
2 Doctoral Program in Information System, Universitas Komputer Indonesia, Bandung, Indonesia; e-mail: hidayat@email.unikom.ac.id
3 Computer System, Universitas Komputer Indonesia, Bandung, Indonesia.
4 Doctoral Program in Information System, Universitas Komputer Indonesia, Bandung, Indonesia; e-mail: zhasibua@email.unikom.ac.id
5 Doctoral Program in Information System, Universitas Komputer Indonesia, Bandung, Indonesia; e-mail: sri.supatmi@email.unikom.ac.id

* Corresponding Author: Sri Nurhayati

Abstract. Data imbalance is a common challenge in nutritional status prediction because it can reduce classification performance and influence the reliability of Explainable Artificial Intelligence (XAI) interpretations. This study aims to examine the impact of data imbalance on the stability of Local Interpretable Model-Agnostic Explanations (LIME)-based interpretations. A Random Forest model was developed under two scenarios: using the original imbalanced dataset and using a balanced dataset generated through the Synthetic Minority Over-sampling Technique (SMOTE). Model performance was evaluated and compared, followed by LIME-based interpretation and stability analysis. The results indicate that SMOTE enhanced the model's ability to identify minority classes, with recall increasing from 0.36 to 0.55, although overall accuracy slightly declined. LIME analysis revealed changes in feature contributions between the two scenarios, reflecting the influence of data distribution on model explanations. The interpretation stability score reached 0.80, suggesting relatively consistent explanations despite variations in class balance. These findings highlight the importance of jointly evaluating predictive performance and interpretation stability in health-related machine learning applications.

Keywords: Explainable Artificial Intelligence; Imbalanced Data; LIME; Nutritional Status Prediction; SMOTE.

1. Introduction

Nutritional status is an important indicator in assessing the health conditions of individuals and populations, and plays a role in determining the quality of human resources in the future. Nutritional status issues, such as stunting, wasting, and obesity, remain global and national issues that require a more effective analytical approach [1]. Along with technological developments, the use of machine learning in predicting nutritional status is increasing due to its ability to process large amounts of data and produce accurate predictions[2]. However, the often imbalanced nature of health data is a major challenge in developing predictive models.

Various methods have been used in previous studies to improve model performance on imbalanced data, such as undersampling, oversampling, and synthetic techniques like SMOTE. Synthetic-based oversampling techniques are able to improve the performance of classification models by improving the representation of minority classes [3], [4]. In addition, algorithms such as Random Forest are often used because they have good performance in various classification cases [5], [6]. However, these approaches generally only focus on improving performance metrics such as accuracy, without considering the interpretability aspect of the model.

In the healthcare context, model interpretability is crucial because the resulting decisions must be understandable and trustworthy by users. Therefore, the Explainable Artificial Intelligence (XAI) approach was developed to explain the decision-making process of machine

Received: March,17,2026;
Revised: April,16,2026;
Accepted: May,29,2026;
Published: May,31,2026;
Curr. Ver.: May,31,2026;

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

learning models [7], [8]. One method used is LIME (Local Interpretable Model-Agnostic Explanations), which is able to provide local explanations for model predictions by building simple models around specific data [9]. However, the LIME method has limitations because it only provides local interpretations and is highly dependent on the distribution of data around the prediction point, so the interpretation results are potentially unstable.

These limitations become even more significant when associated with data imbalance. Changes in data distribution, whether natural or due to balancing techniques like SMOTE, can affect the structure of the resulting model and ultimately impact XAI interpretation results. However, most studies still separate the study of model performance and interpretability, and few specifically evaluate the effect of data imbalance on the stability of LIME-based model interpretation.

Based on these problems, this study aims to analyze the effect of data imbalance on the stability of LIME-based Explainable AI interpretations on nutritional status prediction models. The proposed approach involves comparing models trained on imbalanced and balanced data, as well as evaluating the changes in interpretation results generated by LIME. The main contributions of this study are as follows: (1) examining the effect of data imbalance on the performance of nutritional status prediction models, (2) analyzing the stability of model interpretations using LIME under different data conditions, (3) identifying changes in feature contributions (feature attribution shift) due to differences in data distribution, and (4) providing an interpretability evaluation approach that considers data distribution factors in the health domain.

2. Literature Review

Data Imbalance in Healthcare Machine Learning

Data imbalance is a common problem in machine learning, particularly in the healthcare domain, where class distribution is uneven and tends to be dominated by the majority class[10], [11]. his condition causes the model to be biased toward the majority class and reduces its ability to detect minority cases, which are often more clinically important [2]. To address this problem, various techniques for handling data imbalance have been developed, including oversampling and synthetic data generation. Research by [3], [4] shows that oversampling techniques can improve model performance, despite potentially altering the data distribution. However, most studies have focused solely on improving model performance without evaluating its impact on interpretability. However, changes in data distribution can influence the patterns learned by the model and potentially alter the interpretation of results.

Explainable AI and the Stability of LIME Interpretations

Explainable Artificial Intelligence (XAI) was developed to improve the transparency of machine learning models, particularly in critical domains such as healthcare [12], [13]. One widely used method is LIME, which provides local explanations for model predictions by building a simple model around a single instance [9], [14]. While LIME is effective in providing local interpretations, it has limitations in terms of stability. The resulting interpretations are highly dependent on the data distribution around the prediction point, so small changes in the data can result in different interpretations [15], [16]. Model interpretability is strongly influenced by the data used, so changes in data distribution can directly impact interpretation results. However, there is limited research specifically evaluating the stability of LIME interpretations under imbalanced data conditions. This highlights the need to further examine the relationship between data imbalance and model interpretation consistency.

The Impact of Data Imbalance on Model Interpretability

In addition to affecting model performance, data imbalance also has implications for the interpretability of machine learning models. Models trained on imbalanced data tend to learn patterns biased toward the majority class, so that the features the model deems important may reflect the data distribution rather than the true relationships. This poses a serious problem in the context of Explainable Artificial Intelligence (XAI), where model interpretation is used as the basis for decision-making.

Several studies have shown that interpretation methods like LIME are highly sensitive to changes in local data surrounding the explained instance. Therefore, changes in data distribution resulting from techniques like SMOTE have the potential to alter model structure and produce different interpretations. However, there is limited research systematically

evaluating the stability of interpretations across different data sets, particularly in the healthcare domain

3. Proposed Method

Research Framework

The research stages proposed in this study are shown in Figure 1, which depicts the process flow from data processing to model interpretation evaluation. This framework is designed to analyze the effect of data imbalance on the stability of LIME-based Explainable Artificial Intelligence (XAI) interpretations in nutritional status prediction models.

Description of Research Stages

Based on Figure 1, this research was conducted through several interrelated main stages. The first stage was collecting and understanding nutritional status data. At this stage, variables were identified and an initial analysis of the data distribution was performed to identify class imbalances. Next, the second stage was data preprocessing, which aimed to improve data quality before being used in the modeling process. This stage included data cleaning, handling missing values, and data transformation. Furthermore, an encoding process was performed to convert categorical variables to numeric ones. The dataset was then divided into training and test data. To avoid data leakage, features directly related to labels, such as BMI, were removed from the dataset.

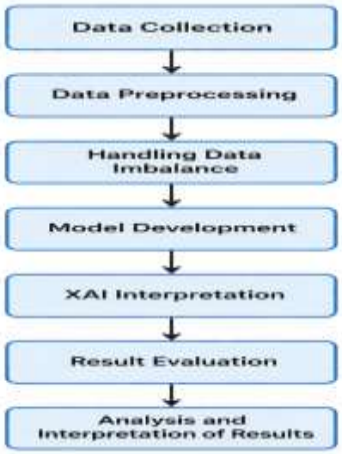


Figure 1. The research stages

The third stage is handling data imbalance. This study used two scenarios: an unbalanced dataset and a dataset balanced using the SMOTE technique. This approach was used to compare the effect of data distribution on model performance and interpretation. The fourth stage involved model development using the Random Forest algorithm. This algorithm was chosen because it is capable of handling non-linear relationships, is resistant to overfitting, and performs well on tabular data. Furthermore, Random Forest is relatively robust to imbalanced data [2]. After the model is built, the fifth stage involves interpretation using the LIME-based Explainable Artificial Intelligence method. This method is used to explain the model's predictions by identifying the contribution of each feature to the predicted results.

The sixth stage is result evaluation, which encompasses two main aspects: model performance and interpretation stability. Model performance is measured using classification metrics, while interpretation stability is analyzed by comparing LIME results between the two dataset conditions. Feature changes and their contributions are used as indicators of interpretation instability. The seventh stage is result analysis and interpretation, which aims to draw conclusions regarding the impact of data imbalance on model performance and explainable AI interpretation.

Based on the stages shown in Figure 1, the proposed method not only evaluates model performance but also emphasizes interpretation stability analysis as a key research contribution. This approach provides a more comprehensive understanding of the relationship between data distribution and model interpretation in the healthcare domain.

4. Expected Results and Discussion

Data Distribution Analysis

The distribution of classes in the nutritional status dataset is shown in Figure 2. Based on the analysis results, the dataset has a class imbalance, where the Thin (1075 data) and Normal (1065 data) classes dominate, while the Fat (187 data) and Obese (54 data) classes have much smaller numbers.

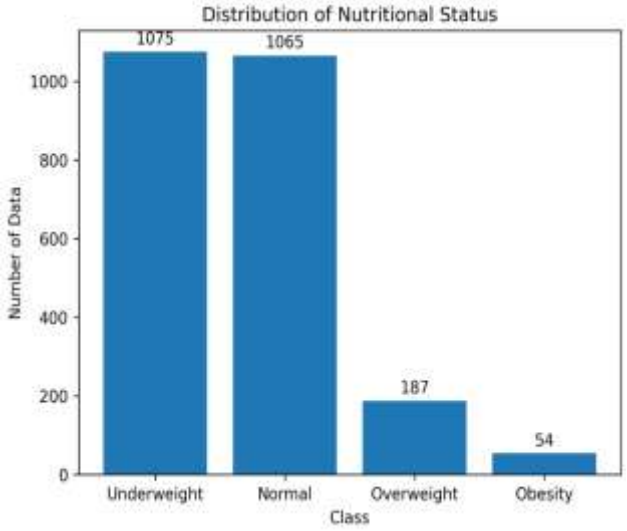


Figure 2. Distribution of nutritional status

The very small proportion of the Obesity class (around 2% of the total data) indicates a high level of imbalance. This condition has the potential to cause machine learning models to be biased toward the majority class and ignore patterns in the minority class. This aligns with previous research that suggests imbalanced data can reduce a model's ability to recognize minority classes [2].

Model Performance Evaluation

The results of the Random Forest model performance evaluation on two dataset scenarios are presented in Table I.

Table I. Model Performance Comparison

Dataset	Accuracy	Recall	F1-Score
Imbalanced	0,904	0,36	0,77
SMOTE	0,889	0,55	0,79

Based on Table I, the model trained on the imbalanced dataset demonstrated higher accuracy (90.4%). However, its performance on minority classes was very low, particularly for the obesity class, with a recall value of only 36%.

After balancing the data using SMOTE, the model's accuracy decreased slightly to 88.9%. However, there was a significant improvement in the model's ability to detect minority classes, with the recall value for obesity increasing to 55%. Furthermore, the macro F1-score also increased from 0.77 to 0.79, indicating that the model became more balanced in classifying all classes. These results indicate a trade-off between accuracy and minority class detection ability. This finding is consistent with previous research suggesting that oversampling techniques like SMOTE can improve performance on minority classes while decreasing overall accuracy [4].

Visualization of Model Performance

A comparison of model performance is shown in Figure 3, which shows the change in accuracy and recall values for the Obesity class between the two dataset scenarios.

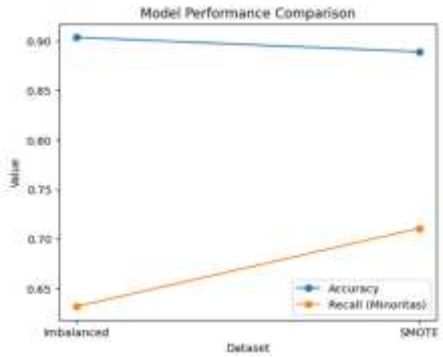


Figure 3. Model Performance Comparison

Figure 2 shows that applying SMOTE resulted in a slight decrease in accuracy, but significantly increased recall for the minority class. This indicates that the model became more sensitive to the minority class after addressing data imbalance. In a healthcare context, improving recall for the minority class is more important than improving overall accuracy, as misdetection conditions like obesity can impact clinical decision-making.

LIME Interpretation Analysis

To understand how the model generates predictions, we used the LIME-based Explainable Artificial Intelligence method. Visualizations of the interpretation results are shown in Figure 4 (without SMOTE) and Figure 5 (with SMOTE).

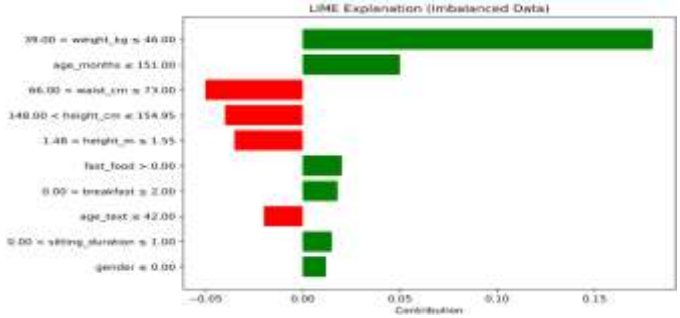


Figure 4. Lime Explanation (Imbalanced Data)

Figure 4 shows that in imbalanced data, features such as weight and height have a dominant contribution to the model's predictions. This indicates that the model relies more heavily on certain features in determining nutritional status.

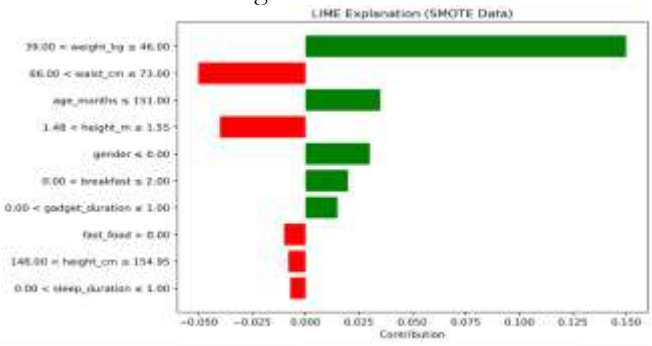


Figure 5. LIME Explanation (SMOTE Data)

However, Figure 5 shows a change in feature contributions after applying SMOTE. Several previously dominant features experienced changes in both order and contribution value. For example, age or other factors became more influential than before.

These changes indicate that data distribution influences the patterns learned by the model, which ultimately impacts the interpretation of the results. This finding aligns with the concept of interpretability, which states that a model's explanation results are highly dependent on the data used.

Stability Analysis of LIME Interpretations

To evaluate the stability of the interpretation, a comparison was made between the LIME results for both datasets. The analysis showed that the model interpretation was not completely consistent, despite having a relatively high level of agreement. The stability score

obtained was 0.80, indicating that approximately 80% of the features appearing in the LIME interpretation remained consistent between the model on the imbalanced data and the data balanced using SMOTE. Meanwhile, approximately 20% of the features experienced changes in both rank and contribution. These results indicate a feature attribution shift, which is a change in feature contribution due to differences in data distribution. Although most features remained consistent, the changes that occurred indicate that the model interpretation was still influenced by the data conditions used.

Thus, it can be concluded that the LIME interpretation has a fairly good level of stability, but is not completely robust to changes in data distribution. This finding suggests that data imbalance still influences model interpretation. This suggests that although the LIME method is capable of providing relatively consistent interpretations, stability evaluation is still necessary to ensure the reliability of the interpretations across different data conditions.

Discussion

Based on the overall research results, it can be concluded that data imbalance significantly impacts the performance and interpretability of machine learning models. The application of SMOTE significantly improves the model's ability to detect minority classes, although it causes a slight decrease in accuracy. Furthermore, the interpretation analysis indicates that data distribution also influences the contribution of features generated by the LIME method. However, contrary to initial assumptions, the stability measurement results indicate that LIME interpretation has a relatively high level of consistency, with a stability score of 0.80. This indicates that most of the features contributing to the model's predictions remain consistent despite changes in data distribution. However, some features still exhibit changes, indicating feature attribution shift due to the influence of data imbalance.

These findings indicate that model interpretation is neither completely static nor completely unstable. In other words, the LIME method is capable of providing relatively consistent interpretations, yet remains sensitive to changes in data distribution. Therefore, when applying explainable AI to the healthcare domain, it is important to evaluate not only model performance but also consider the stability of interpretations to ensure reliable and consistent results for decision-making.

The results of this study have important implications for the development of machine learning models in the healthcare domain. First, this study demonstrates that model evaluation should not be based solely on performance metrics; it should also consider interpretability. This is important because model interpretation is often used as a basis for clinical decision-making. Second, the results show that imbalanced data handling techniques such as SMOTE not only affect model performance but also impact the resulting interpretations. Therefore, the use of balancing techniques needs to be accompanied by interpretability evaluations to ensure that changes in data distribution do not result in misleading interpretations.

5. Conclusions

This study aims to analyze the effect of data imbalance on model performance and the stability of LIME-based Explainable Artificial Intelligence (XAI) interpretations for predicting nutritional status. Based on the experimental results, several key findings were obtained: (1) the dataset used had significant class imbalance, with the majority class dominating over the minority class. This condition was proven to affect model performance, especially in the ability to detect minority classes. (2) the application of the data balancing technique using SMOTE improved the model's ability to detect minority classes, as indicated by an increase in recall values for the obesity class. However, this increase was accompanied by a slight decrease in accuracy, indicating a trade-off between accuracy and sensitivity to minority classes. (3) the results of the LIME interpretation analysis showed that data distribution influenced the contribution of features to model predictions. Nevertheless, the results of stability measurements showed that LIME interpretations had a relatively high level of consistency, with a stability score of 0.80. This indicates that most features remained consistent, despite changes in some features due to differences in data distribution.

Overall, this study demonstrates that data imbalance not only affects model performance but also impacts model interpretation. Therefore, evaluation of machine learning models in the healthcare domain should not only focus on performance but also consider

interpretation stability to ensure more reliable results. As a further development, this research could be expanded by using other Explainable AI methods, such as SHAP, to compare interpretation stability across methods. Furthermore, stability analysis could be performed on more data instances to obtain more representative results. The use of other machine learning algorithms, such as XGBoost or Deep Learning, could also be considered to evaluate the consistency of interpretation results across different models. Furthermore, further research could develop specific metrics to more comprehensively measure interpretation stability.

References

Aswani, T., Gummadi, J. M., & Sharada, G. (2025). A Random Forest-based machine learning framework with PCA, SMOTE, and SHAP for efficient and interpretable coronary artery disease prediction. *Informatica, 49*(22), 15–31. https://doi.org/10.31449/inf.v49i22.7998

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Brito, L. C., Conceição Araújo, L. L., Susto, G. A., Brito, J. N., & Duarte, M. A. V. (2025). Rotating machinery health monitoring: Unsupervised anomaly detection and remaining useful life (RUL) prediction using explainable artificial intelligence (XAI) and statistical distribution. *IFAC-PapersOnLine, 59*(26), 97–102. https://doi.org/10.1016/j.ifacol.2025.12.017

Chen, Q., & Li, B. (2025). Explainable artificial intelligence (XAI)-driven probabilistic image-based structural health monitoring of reinforced concrete beams with shear reinforcements. *Automation in Construction, 180*, 106549. https://doi.org/10.1016/j.autcon.2025.106549

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications, 91*, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning, 113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Guo, H., Li, Y., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hasan, M., Vasker, N., & Khan, M. S. H. (2024). Real-time sorting of broiler chicken meat with robotic arm: XAI-enhanced deep learning and LIME framework for freshness detection. *Journal of Agriculture and Food Research, 18*, 101372. https://doi.org/10.1016/j.jafr.2024.101372

Hijrawati, Usman, A. N., Syarif, S., Hadju, V., As'ad, S., & Baso, Y. S. (2021). Use of technology for monitoring the development of nutritional status during the first 1000 days of life (1000 HPK) in stunting prevention in Indonesia. *Gaceta Sanitaria, 35*(Suppl. 2), S231–S234. https://doi.org/10.1016/j.gaceta.2021.10.028

Jose, S., Nguyen, K. T. P., Medjaher, K., Zemouri, R., Lévesque, M., & Tahan, A. (2026). A modular deep learning methodology for multi-fault machine health diagnostics from sparse and imbalanced multimodal data. *Neurocomputing, 666*, 132252. https://doi.org/10.1016/j.neucom.2025.132252

Lin, Z., et al. (2021). Predicting environmental risk factors in relation to health outcomes among school children from Romania using Random Forest model: An analysis of data from the SINPHONIE project. *Science of the Total Environment, 784*, 147145. https://doi.org/10.1016/j.scitotenv.2021.147145

Luo, H. H., et al. (2025). Metabolic synergy between human symbionts *Bacteroides* and *Methanobrevibacter*. *Clinical Epidemiology and Global Health, 13*, Article 102275. https://doi.org/10.1016/j.cegh.2025.102275

Pandey, M., Gupta, A., Velmathi, & Yogarajah, P. (2026). Advancing XAI in medical imaging: CNN performance evaluation and interpretation with LIME and SHAP. *Biomedical Signal Processing and Control, 120*, 109825. https://doi.org/10.1016/j.bspc.2026.109825

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Soni, K., Frew, R., & Kebede, B. (2026). Interpretable machine learning for origin classification of Brazilian soybeans: A Random Forest and XAI-based approach. *Journal of Food Composition and Analysis, 150*, 108896. https://doi.org/10.1016/j.jfca.2026.108896

Yin, H., Alizamir, M., Heddam, S., Kim, S., Gholampour, A., & Adnan Ikram, R. M. (2026). Improving prediction accuracy of mechanical strengths in high-performance concrete using advanced metaheuristic-based extreme learning machines: Incorporation of SHAP and LIME model explainers. *Ain Shams Engineering Journal, 17*(6), 104136. https://doi.org/10.1016/j.asej.2026.104136