

Symptom-Based Classification of Migraine Severity Using Random Forest and LassoNet

Syehan Fariz Gustomo, Kemas Muslim Lhaksana*

School of Computing, Telkom University, Bandung, Indonesia

Email: ¹syehanfariz@student.telkomuniversity.ac.id, ^{2,*}kemasmuslim@telkomuniversity.ac.id

Submitted: 21/05/2026; Accepted: 29/06/2026; Published: 30/06/2026

Abstract—This study aims to develop and compare classification models for predicting symptom-based migraine intensity using a public dataset from Kaggle. The research was conducted on multiclass data with an imbalanced class distribution. Therefore, target creation and result interpretation must be carefully designed so that the resulting evaluation remains consistent with the characteristics of the data used. In this study, the Intensity variable was recoded into three operational classes: low, moderate, and high. The features used included symptoms, characteristics of migraine episodes, and ‘*symptom_count*’, which represents the number of symptoms in each sample. The two models compared were Random Forest and LassoNet, both of which were tested using Stratified 5-Fold Cross-Validation. Model performance was assessed using the Macro F1-score as the primary metric, Balanced Accuracy as the main supplementary metric, and Accuracy as a complementary metric. The test results showed that Random Forest performed better, with a Macro F1-score of 0.6366, Balanced Accuracy of 0.6181, and Accuracy of 0.6225. Meanwhile, LassoNet achieved a Macro F1-score of 0.2474, a Balanced Accuracy of 0.3333, and an Accuracy of 0.5900. These results indicate that symptom patterns in the dataset can still be utilized to distinguish migraine intensity within a computational classification framework, although the separation between closely related classes is not yet fully robust.

Keywords: Balanced Accuracy; LassoNet ; Migraine Intensity; Multiclass Classification; Random Forest

1. INTRODUCTION

Migraine is a neurological disorder characterized by recurrent headaches, often accompanied by other symptoms such as nausea, vomiting, photophobia, and phonophobia [1]. In such cases, it is important to consider not only the presence or absence of an attack but also how severe the attack is perceived by the patient. Differences in attack intensity can affect the patient’s experience during an attack, both in terms of the symptoms that appear and the impact on daily activities. In other words, intensity cannot be viewed merely as supplementary information, as differences in attack severity can indicate significant differences in the patient’s condition. If the available data includes symptoms and characteristics of the attacks, then this data can be used not only to describe the patient’s condition but also to identify patterns in migraine severity in a more targeted manner.

Advances in machine learning have spurred numerous studies that are beginning to utilize symptom data to aid the analysis process in the healthcare field. In the case of migraines, this type of approach is already widely used, both in the form of digital diagnostic systems and symptom-search systems. However, the results obtained still reveal several limitations. Previous research has shown that digital diagnostic systems for migraines have not yet demonstrated good performance or sufficient consistency in results [2]. A broader review of digital and online symptom checkers also reveals similar issues, particularly regarding diagnostic accuracy and triage [3]. In real-world use, unclear system outputs and users level of trust in the results remain challenges [4]. This situation indicates that the issues in migraine cases lie not only in model development but also in how the classification task is formulated to align with the problem at hand and how the results are subsequently interpreted accurately.

Previous research on migraines has also made significant progress, but has not yet focused extensively on the issue of attack intensity directly. Reddy et al. (2025) used a similar migraine dataset to classify triggers, phases, and types of migraines, and their results showed that symptoms and episode characteristics are indeed sufficiently informative for building meaningful classification models [5]. On the other hand, Kim et al. (2023) demonstrated that the most bothersome symptoms in migraine sufferers can vary from person to person [6], while other clinical studies have shown that pain intensity is associated with specific migraine characteristics as well as the impact of the headache as perceived by the patient [7], [8]. These findings demonstrate that symptoms can not only be used to identify migraine types but are also related to the severity of attacks experienced by patients. However, while machine learning-based research has advanced to the classification of migraine types, attack intensity has not yet been widely formulated as a separate classification target and this is where the gap in this study lies.

When intensity is chosen as the classification target, the problem becomes more specific than simply comparing algorithms. The target classes must be clearly defined, as blurred boundaries will make the multi-class task difficult to interpret from the outset. Feature selection also cannot be done haphazardly, as symptom data contain several attributes that may be too closely related to the target definition, posing a risk of data leakage if data splitting is not properly managed [9]. The same issue arises during the evaluation stage, as the accuracy often appears convincing in an imbalanced class distribution, even though this figure does not necessarily reflect the model’s ability to recognize all classes proportionally. Therefore, performance assessment must be supported by more appropriate metrics, such as Balanced Accuracy and Macro F1-score [10], [11]. Thus, what is being evaluated in intensity classification is not only the model used but also how the target is formulated, the features selected, and the accuracy in interpreting the results obtained.

This study focuses on developing a symptom-based migraine intensity classification model using a public dataset. Intensity is categorized into three operational classes, there are low, moderate, and high, so that the classification targets have clear boundaries and can be tested consistently. The two models used Random Forest and LassoNet were employed to examine how each performs on tabular symptom data. The aim of this study is not merely to add another classification approach to migraine research, but to address the limited attention previous studies have given to attack intensity as a standalone classification target. To that end, this study emphasizes the formulation of a well-defined task, attention to potential leakage, and the use of evaluation metrics better suited for imbalanced multiclass data. The contribution of this study lies in assessing the ability of symptom data to classify migraine intensity, as well as in comparing how the two modeling approaches perform on this task. Thus, this study is expected not only to produce classification results but also to provide a clearer picture of how symptom-based migraine intensity classification can be methodologically structured and interpreted. Consequently, the results of this study can serve as a basis for further research aiming to develop similar approaches on broader datasets or more diverse conditions.

2. RESEARCH METHODOLOGY

This study was conducted to build and compare symptom-based migraine intensity classification models using a public dataset. The primary focus of the study lies in the formulation of the intensity target, the selection of features used, the application of classification models, and the development of an appropriate evaluation scheme for multiclass data with an imbalanced distribution. In this study, the two models compared are Random Forest and LassoNet. The entire research process was organized in stages, ranging from data preparation, target formulation, preprocessing, modeling, to result evaluation. This sequence of stages was employed to ensure the analysis process was conducted in a more focused manner and that the results could be interpreted in line with the study's objectives. The complete flowchart of the research stages is shown in Figure 1.

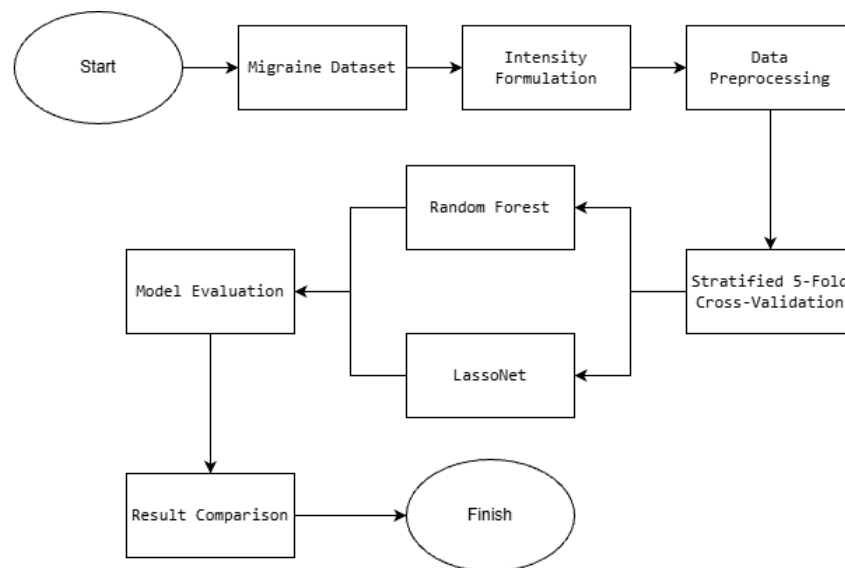


Figure 1. Research Stages

2.1 Dataset and Feature Engineering

This study uses a public migraine dataset obtained from Kaggle. The dataset consists of 400 data rows with 24 initial columns and contains two labels in the original data: Type and Intensity. In this study, Intensity is used as the basis for defining the classification target because the research focuses on grouping the severity levels of attacks as reflected by migraine intensity [7], [8]. Meanwhile, “Type” is not used as either a target or a predictor. The use of “Type” as a feature is also avoided because it risks causing target leakage [9]. The data used in this study has 15 final features. A summary of the dataset used can be seen in Table 1.

Table 1. Dataset Summary

Component	Description
Data Source	Public migraine dataset from Kaggle
Number of records	400 rows
Number of initial attributes	24 columns
Original labels in the data	Type, Intensity
Final number of features	15 features

The classification targets were divided into three operational classes: low, medium, and high. Recoding was performed by grouping Intensity values of 0 and 1 into the low class, values of 2 into the medium class, and values of 3 into the high class. This scheme was used to define a clearer multi-class target for modeling and evaluation, where the low, moderate, and high classes were treated as operational labels derived from the recoding of the Intensity variable [12], [13]. Therefore, this classification is used in the computational context of this research and is not intended to represent clinical categories in medical assessments.

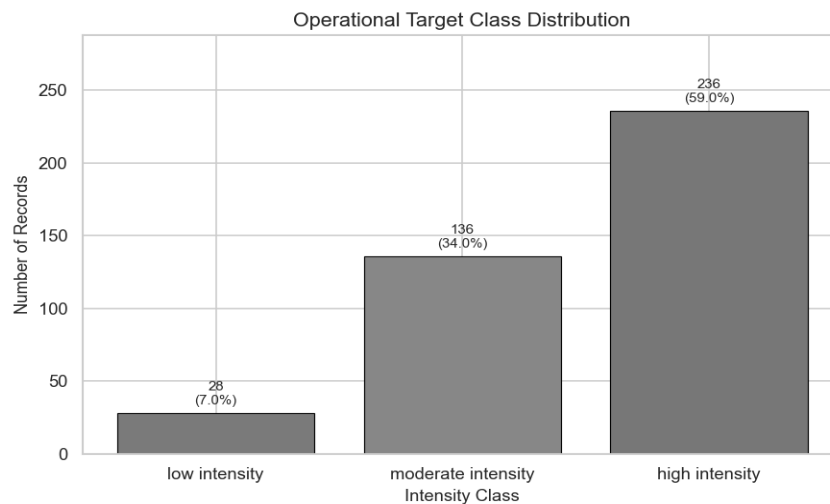


Figure 2. Distribution of Operational Target Class

The distribution resulting from the recoding process shows that the class proportions are not fully balanced, as illustrated in Figure 2. The high-intensity class contains the largest number of records, followed by the medium-intensity class, while the low-intensity class has a substantially smaller number of records. This condition provides important context for interpreting the experimental results, since the models were not evaluated on a perfectly balanced class distribution.

2.2 Feature Selection, Feature Engineering, and Preprocessing

The features used in this study are derived from two main groups: symptom features and episode characteristic features. The symptom group includes indicators of accompanying symptoms recorded in the dataset, such as nausea, vomiting, phonophobia, photophobia, and other sensory symptoms that occur alongside an attack. Meanwhile, the episode characteristics group includes attributes that describe the attack profile, such as frequency, duration, location, and pain characteristics. These features were selected to ensure that the data used accurately reflects the experience of the attack and the intensity of the migraine [6], [8]. Additionally, the “Type” attribute was excluded because it could potentially contain information too closely related to the classification label and cause leakage [9].

The final input attributes in this study consist of symptom variables, episode characteristics, and one simple engineered feature, namely *'symptom_count'*. The *'symptom_count'* feature is used as a summary of the number of symptoms appearing in each sample. With this addition, the data not only includes the types of symptoms that appear but also the number of symptoms in a single sample. This approach is used to capture the number of symptoms without excessively altering the data structure. After the feature selection and engineering processes were completed, the data were then organized into the same tabular format for all experiments. At this stage, the final features and the target labels resulting from recoding were separated, and the input format was prepared consistently for both Random Forest and LassoNet. With this arrangement, the input data remained uniform and suitable for the needs of migraine intensity classification.

2.3 Classification Models

This study uses two classification models, Random Forest and LassoNet, to examine how differences in model complexity affect symptom-based migraine intensity classification. Both models are run on the same features with operationally defined intensity targets, allowing for a more balanced comparison of results. In this study, Random Forest is used as the baseline model. This model is suitable for tabular data with a relatively limited sample size and remains capable of handling nonlinear relationships between features without requiring complex input processing [12], [13]. In this study, Random Forest was used to determine whether symptom patterns and episode characteristics could still be identified through a more classical classification approach. This choice is also consistent with the characteristics of the data used, as the dataset is not large, the features remain in tabular form, and the feature space has been simplified in a previous stage.

LassoNet is used as a comparison model to determine whether a more complex approach can offer advantages on the same data. This model remains relevant because it combines the learning process with regularization and

sparsity mechanisms, making it suitable for representing tabular data [14]. However, the literature also indicates that deep learning approaches for tabular data do not always yield better results than simpler models [15]. Therefore, in this study, LassoNet is not positioned as a model presumed to be superior, but rather as a methodological benchmark against Random Forest to examine the impact of model complexity on symptom-based migraine intensity classification.

2.4 Evaluation Scheme

This study uses Stratified 5-Fold Cross-Validation so that the evaluation does not depend on a single data split and class proportions are maintained in each fold. This scheme was chosen because the dataset used is relatively limited, while the final intensity labels consist of three classes with a distribution that is not entirely balanced. Under these conditions, stratification is necessary to ensure that each training and testing round consistently represents the class composition, particularly in multi-class classification settings with skewed distributions [10], [16].

Model evaluation focuses not only on overall prediction accuracy but also on the model's ability to distinguish between all intensity classes more evenly. Therefore, the Macro F1-score is used as the primary metric [17]. Balanced Accuracy is then used as the primary supplementary metric because the interpretation of results must still take into account performance across each class, especially when the data distribution is not entirely even [10], [11]. Meanwhile, Accuracy is still included as a supplementary metric to provide an overview of overall performance, but it is not used as the primary basis for drawing conclusions because its value may appear high even if the model does not perform equally well across all classes [11], [17]. With this structure, the interpretation of results does not stop at a single summary statistic but instead considers the roles of different metrics according to the evaluation objectives [18]. With this structure, the interpretation of results does not stop at a single summary figure but continues to consider the roles of different metrics according to the evaluation objectives [18], [19].

In addition to the composite score, a confusion matrix is used to trace patterns of misclassification across intensity classes [19]. This component is necessary because a single score does not indicate which classes are most frequently misclassified or the most dominant direction of error [11]. With the help of the confusion matrix, the analysis can focus on the classes that are most challenging to classify, emerging misclassification trends, and differences in model behavior that are not apparent from the composite metric alone.

3. RESULTS AND DISCUSSION

The evaluation results in this study show that Random Forest performs better than LassoNet on all the primary metrics used. A comparison of the two models' scores is shown in Table 2, while the patterns of difference are more easily discernible in Figure 3.

Table 2. Model Evaluation Results

Model	Macro F1-score	Balanced Accuracy	Accuracy
Random Forest	0.6366	0.6181	0.6225
LassoNet	0.2474	0.3333	0.5900

Table 2 shows that Random Forest yields better results than LassoNet across all metrics used in this study. This superiority is evident in the Macro F1-score, Balanced Accuracy, and Accuracy values. This indicates that Random Forest's advantage is not only evident in overall prediction accuracy but also in the model's ability to maintain performance across multiple intensity classes. Based on these results, Random Forest can be considered a more stable model for symptom-based migraine intensity classification using the tabular data employed in this study.

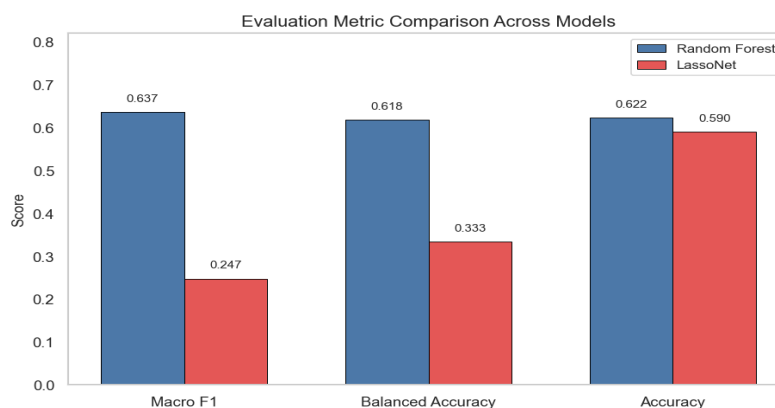


Figure 3. Comparison of Evaluation Metrics

When Table 2 is read in conjunction with Figure 3, the performance gap between the two models becomes more apparent. The most pronounced differences appear in the Macro F1-score and Balanced Accuracy, indicating

that the disparity in results between the models is not limited to a single evaluation metric. This means the study's findings go beyond simply concluding that one model achieved higher scores than the other, and they also suggest that Random Forest is better at maintaining prediction quality when the target classes are not perfectly balanced. Meanwhile, LassoNet still exhibits fairly clear limitations, particularly when results are evaluated using metrics that require a more evenly distributed across classes. With the experimental configuration used in this study, Random Forest emerges as the model with the best overall performance for migraine intensity classification.

3.1 Error Analysis and Interpretation of Results

Although Random Forest generally provided the best results, the evaluation results did not yet show a completely robust class separation. Classification errors still occurred in some classes, particularly in samples with similar symptom characteristics. This pattern can be observed in the confusion matrix in Figure 4.

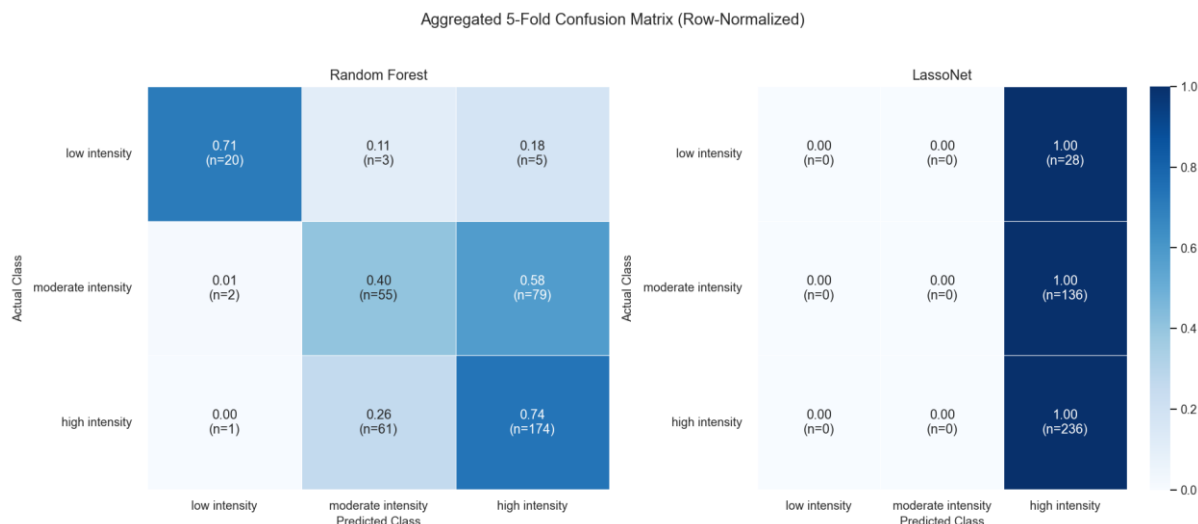


Figure 4. Confusion Matrix

Figure 4 shows that the model still struggles to consistently distinguish between the moderate and high classes. This point is important because the main issue in intensity classification lies not only in the model's ability to recognize symptoms but also in its ability to separate intensity classes that are closely spaced. In this study, that separation has not yet been fully achieved. The moderate and high classes still overlap, so some samples that should belong to one class are instead mapped to the adjacent class.

This overlap makes sense when considering the types of features used. The data in this study relies on symptoms and episode characteristics, so the model infers intensity through patterns emerging from this information. The problem is that samples in the moderate and high classes can still exhibit similar combinations of symptoms. An individual may experience nausea, phonophobia, photophobia, or pain in a specific location, but differences in intensity levels do not always serve as clear-cut distinctions within the available features. Under such conditions, the model tends to recognize pattern similarities but is not yet robust enough to consistently assign them to the correct class. Therefore, the errors that arise are not merely a sign that the model has failed to work, but also an indication that the boundaries between classes are not yet fully clear in the feature space used. At this stage, interpreting the results should not stop at the combined score. The evaluation metrics obtained need to be examined alongside the patterns of prediction errors to understand which classes remain difficult to distinguish and to what extent the model performs consistently across classes. In this study, moderate scores are associated with the model's difficulty in distinguishing between closely related classes, not with a total failure of the model to learn the data. Thus, the results indicate that the model is already capable of capturing some important patterns, but the separation between classes is not yet fully distinct.

In symptom-based health data, prediction quality is determined not only by the algorithm used but also by the target structure, data composition, and the information available for each sample [20]. In tasks centered on symptoms, two classes that appear conceptually distinct do not necessarily have a strong separator within the data. As a result, the model may perform reasonably well at a general level but still leaves gray areas when distinguishing closely related cases. This is evident in the results of this study. The same error pattern reappears when the analysis shifts from the confusion matrix to the error distribution per class in Figure 5. Here, it is evident that prediction errors are not evenly distributed. Some classes are relatively easier to classify, while others are more prone to being confused with neighboring classes.

Figure 5 helps illustrate that the main issue lies not merely in the number of errors, but in the direction of those errors. When a particular class is more frequently misclassified as a neighboring class, it means the model still tends to capture pattern proximity rather than clear separation boundaries. This directly affects how the results are interpreted. If the reader only looks at Accuracy, the impression might be overly optimistic, since some correct

predictions are still counted as overall success. However, when errors are analyzed per class, it becomes clear that the model's performance is not uniform. Classes that are harder to distinguish are holding back the overall improvement in the model's quality. From this, it is evident that the model's success in one class does not automatically imply the same success in other classes. Therefore, interpreting the results requires considering the error distribution, not just the summarized final values.

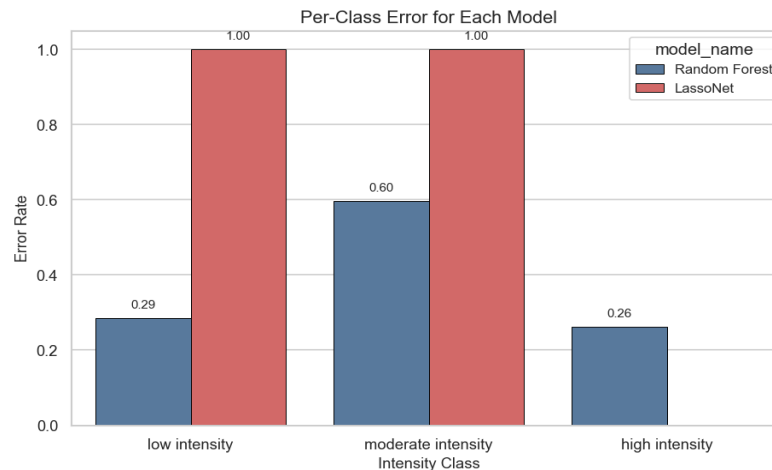


Figure 5. Error by Class

On the other hand, disparities in class distribution can still affect the model's learning process and result in uneven performance across all classes [21]. In the context of this study, classes with smaller distributions or less distinct boundaries are at risk of not achieving separation quality equivalent to that of other classes. Therefore, the results obtained should be understood as the outcome of the interaction between the model, the target structure, and the data distribution, rather than merely as a measure of a single algorithm's strength. Such conditions are also consistent with findings in medical benchmarks with long-tailed distributions, where metrics emphasizing inter-class balance may remain at moderate levels even though the model is already capable of capturing some important patterns [22]. A similar pattern is observed in more recent model validation studies, where seemingly good overall performance is not always accompanied by uniform results across classes [23]. These findings are not intended to equate the current research with those studies, but rather to place the obtained results within a more reasonable framework. Thus, scores that are not entirely high on class-sensitive metrics do not directly indicate that the design of this study is weak, but rather reflect that such results can indeed arise in difficult tasks, especially when the target is multi-class and the distribution is imbalanced.

Therefore, the results of this study are more accurately interpreted as a consequence of the difficulty level of the classification task at hand. Random Forest did emerge as the best model in this experiment, but this advantage does not negate the fact that separating closely spaced intensities remains a major challenge. From an analysis of these errors, it is evident that the research challenge lies not only in selecting the best model but also in addressing the limitations of feature representation when intensities must be distinguished across several closely spaced classes. This chapter thus not only demonstrates which model performs better but also highlights where the true difficulty lies in the symptom-based migraine intensity classification task.

3.2 Comparison with Previous Research

The results of this study should be interpreted with caution when considered alongside previous studies that used similar migraine datasets. In previous studies, the primary focus was on type classification [3]. This means that the models were trained to distinguish between migraine types, rather than to assess the intensity of attacks categorized into operational classes. Because the objectives of the studies differ, the results of the two studies cannot be accurately treated as a direct numerical comparison. Higher or lower scores do not automatically indicate which study is better, since the tasks the models were designed to perform from the outset were fundamentally different.

This difference directly affects how the results are interpreted. In this study, the focus is on operational intensity classification, so the task design is tailored to meet this need. The "Type" attribute is not used as a feature, and the results are not primarily evaluated based on Accuracy. Performance evaluation is directed more toward the Macro F1-score and Balanced Accuracy, as these two metrics are better suited to demonstrate whether the model is truly capable of distinguishing all intensity classes more proportionally [11], [17]. Thus, the significance of this study lies in the models' ability to handle a specifically designed intensity classification task, rather than in replicating studies that distinguish migraine types.

From this, it is evident that the results of this study not only demonstrate performance differences among models but also show that when the classification target shifts from type to intensity, the way the results are interpreted also changes. This shift means that target formulation, feature selection, and evaluation cannot be treated the same

way as in type classification studies. Therefore, the contribution of this study lies more appropriately in the formulation of the task and the interpretation of its evaluation, rather than in a mere comparison of numerical values. Thus, this study is not intended as a rebuttal to previous studies, but rather as an expansion of how to interpret the same migraine dataset. While previous studies have emphasized type classification, this research demonstrates that the dataset can also be used to assess operational intensity. The difference lies in how the task is formulated, how features are selected, and how evaluation results are interpreted more carefully when class distributions are not fully balanced [3], [11], [17]. Therefore, this study is best understood as a methodological advancement in interpreting the same dataset, rather than as a direct competition between research findings.

3.3 Research Limitations

This study was conducted on a relatively small dataset, so the classification results remain quite sensitive to the sample composition in each fold [10], [16]. In a multi-class setting, changes in the data distribution across classes can also affect the stability of the results obtained. Consequently, the model's performance cannot yet be considered fully stable, especially when class distributions are uneven and each class is not equally well-represented in the learning process.

Another limitation lies in the target structure and class distribution. The research target was not taken directly as the final label but was formed by recoding "Intensity" into three operational classes: low, medium, and high. This structure helps define the classification task more clearly but simultaneously makes the results highly dependent on the class boundaries used. On the other hand, the class distribution is not entirely balanced, making performance evaluation insufficient if based solely on overall accuracy, as a single composite score can mask the model's difficulties with specific classes [11]. This condition is one reason why the research results must be understood within the constraints of the target structure and data distribution used. Another limitation stems from the features available in the dataset. The model learns only from the symptoms and characteristics of episodes available in this public dataset. Therefore, the results of this study are best understood as computational findings applicable to the context of the data used, rather than as a direct representation of broader clinical conditions.

4. CONCLUSION

The results of this study show that symptom-based classification of migraine intensity can be performed computationally on the public dataset used, but the quality of the classification is still heavily influenced by the target format and the information available in the features. In the experimental design used, Random Forest yielded better results than LassoNet on all key metrics. This advantage is particularly evident in the Macro F1-score and Balanced Accuracy, indicating that the model's best performance is not only reflected in overall accuracy but also in its ability to distinguish between intensity classes more evenly. Given the data and feature conditions in this study, Random Forest is more suitable as a baseline model for symptom-based migraine intensity classification tasks. These findings also indicate that the research challenge extends beyond algorithm selection. The results in Chapter 3 show that the moderate and high classes still frequently overlap, meaning that the boundaries between classes are not yet fully distinct in the feature space used. This means that the available symptoms and episode characteristics are indeed sufficient to capture some intensity patterns, but they cannot yet be considered consistently strong enough to reliably distinguish adjacent classes. Therefore, the results of this study should be interpreted as evidence that migraine intensity classification is feasible but remains sensitive to how the target is defined and how features represent differences between classes. The limitations of this study lie in the relatively small dataset size, the class distribution that is not entirely balanced, and features that are still limited to publicly available data. These conditions make the model's results sensitive to the data composition in each fold and not yet fully stable for adjacent classes. Future research should focus on using more diverse and balanced data, adding more informative features to distinguish between closely related intensities, and conducting evaluations that prioritize class-specific performance rather than just overall accuracy. With this approach, symptom-based migraine intensity assessment can be developed on a stronger foundation.

REFERENCES

- [1] S. D. Silberstein, "Migraine," MSD Manual Professional Edition. Accessed: Jun. 13, 2026. [Online]. Available: <https://www.msdmanuals.com/professional/neurologic-disorders/headache/migraine>
- [2] Y. W. Woldeamanuel and R. P. Cowan, "Computerized migraine diagnostic tools: a systematic review," Jan. 24, 2022, doi: 10.1177/20406223211065235.
- [3] W. Wallace *et al.*, "The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review," Dec. 01, 2022, *Nature Research*. doi: 10.1038/s41746-022-00667-w.
- [4] Y. You, R. Ma, and X. Gui, "User Experience of Symptom Checkers: A Systematic Review," Aug. 2022. Accessed: Jun. 30, 2026. [Online]. Available: <https://arxiv.org/abs/2208.09100>
- [5] A. Reddy and A. Reddy, "Migraine triggers, phases, and classification using machine learning models," *Front. Neurol.*, vol. 16, 2025, doi: 10.3389/fneur.2025.1555215.
- [6] S. J. Kim, H. J. Lee, S. H. Lee, S. Cho, K. M. Kim, and M. K. Chu, "Most bothersome symptom in migraine and probable migraine: A population-based study," *PLoS One*, vol. 18, no. 11, Nov. 2023, doi: 10.1371/journal.pone.0289729.



- [7] A. K. Eigenbrodt *et al.*, “Diagnosis and management of migraine in ten steps,” *Nat. Rev. Neurol.*, vol. 17, no. 8, pp. 501–514, Aug. 2021, doi: 10.1038/s41582-021-00509-5.
- [8] N. Imai and Y. Matsumori, “Different effects of migraine associated features on headache impact, pain intensity, and psychiatric conditions in patients with migraine,” *Sci. Rep.*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-74253-3.
- [9] L. Sasse *et al.*, “Overview of leakage scenarios in supervised machine learning,” *J. Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-025-01193-8.
- [10] J. Yang, R. El-Bouri, O. O’Donoghue, A. S. Lachapelle, A. A. S. Soltan, and D. A. Clifton, “Deep Reinforcement Learning for Multi-class Imbalanced Training,” May 2022. [Online]. Available: <http://arxiv.org/abs/2205.12070>
- [11] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, “Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems,” *PLOS Digital Health*, vol. 2, no. 11 November, Nov. 2023, doi: 10.1371/journal.pdig.0000290.
- [12] S. Han, B. D. Williamson, and Y. Fong, “Improving random forest predictions in small datasets from two-phase sampling designs,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12911-021-01688-3.
- [13] Aurélien. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA: O’Reilly Media, 2022. Accessed: Jun. 13, 2026. [Online]. Available: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/>
- [14] I. Lemhadri, F. Ruan, L. Abraham, and R. Tibshirani, “LassoNet: A Neural Network with Feature Sparsity,” *Journal of Machine Learning Research*, vol. 22, no. 127, pp. 1–29, 2021, [Online]. Available: <http://jmlr.org/papers/v22/20-848.html>.
- [15] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, “Deep Neural Networks and Tabular Data: A Survey,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.
- [16] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, “Handling imbalanced medical datasets: review of a decade of research,” *Artif. Intell. Rev.*, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10884-2.
- [17] A. C. Muller and S. Guido, *Introduction to Machine Learning with Python*, 1st ed. Sebastopol, CA: O’Reilly Media, 2016. Accessed: Jun. 13, 2026. [Online]. Available: <https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/>
- [18] Sebastian. Raschka and Vahid. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*, 3rd ed. Packt Publishing, Limited, 2019. Accessed: Jun. 19, 2026. [Online]. Available: <https://www.packtpub.com/en-us/product/python-machine-learning-third-edition-9781789955750>
- [19] P. Bruce, Andrew. Bruce, and Peter. Gedeck, *Practical statistics for data scientists : 50+ essential concepts using R and Python*, 2nd ed. O’Reilly Media, 2020.
- [20] J. Pagán, J. L. Risco-Martín, J. M. Moya, and J. L. Ayala, “Modeling methodology for the accurate and prompt prediction of symptomatic events in chronic diseases,” *J. Biomed. Inform.*, vol. 62, pp. 136–147, Feb. 2024, doi: 10.1016/j.jbi.2016.05.008.
- [21] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, “The class imbalance problem in deep learning,” *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.
- [22] G. Holste *et al.*, “Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study,” in *Data Augmentation, Labelling, and Imperfections*, Cham: Springer Nature Switzerland, 2022, pp. 22–32. doi: 10.1007/978-3-031-17027-0_3.
- [23] V. M. Wagner *et al.*, “Real-World Benchmarking and Validation of Foundation Model Transformers for Endometrial Cancer Subtyping from Histopathology,” *Nature*, Apr., 2026. doi: 10.1101/2025.10.10.25337691.