

Artificial Intelligence in Neurological Healthcare: Improving Migraine Diagnosis

Hanan M. Shukur^{1,2*}, Ismael Abdulrahman¹

¹Department of Information Systems Engineering, Technical Engineering College, Erbil Polytechnic University, Erbil, Kurdistan Region, Iraq; ²Department of Computer Techniques Engineering, Technical Engineering College, Al-Kitab University, Kirkuk 36015, Iraq

ABSTRACT

Migraines are a complicated neurological illness that affects millions worldwide, its diagnosis is problematic due to dependence on subjective symptomatology and the absence of reliable biomarkers. This study aims to utilize Artificial Intelligence (AI) to improve migraine diagnosis based on tabular clinical data. Two scenarios were adopted: in the first scenario, a single-stage, multi-class Deep Neural Network (DNN) model was used along with several ML algorithms to classify migraine into four classes. In the second scenario, a two-stage hierarchical framework based on DNN was developed. The first stage (S1) is binary classification to classify (Migraine vs non-migraine), while the second stage (S2) classifies the migraine cases into their three common subtypes, providing an innovative AI-based diagnostic tool that mimics clinical diagnostic processes. The results revealed that the DNN model in the first scenario achieved the best performance with an accuracy of 99.55%. In the second scenario, S1 and S2 tests achieved high performance with accuracy $\approx$ 99.55, 93.75 respectively. In order to assess generalizability, the hierarchical model was evaluated using an external dataset. The generalization accuracy of S1 reached 96.5% while S2 reached 96.61%. In summary, although the first case outperforms well in terms of the classification performance using the source dataset, our second proposed hierarchical framework not only achieves high classification accuracy but also have a strong ability of generalization for more realistic clinical applications. These findings may lead to a more accurate, faster, and cost-effective diagnosis of migraine in clinical practice.

ARTICLE HISTORY

Received September 23, 2025

Accepted December 28, 2025

KEYWORDS

Migraine Diagnosis,
DNN,
Machine Learning,
Artificial Intelligence,
Neurological Healthcare,
Hierarchical Classification

1. Introduction

As the third most prevalent illness in the world, migraine is a common neurological disorder characterized by episodes of moderate-to-severe unilateral throbbing and headaches, in some cases, accompanied by phonophobia, photophobia, nausea, and other neurological disturbances that last hours to days (Ashina et al., 2021). It is one of the most widespread neurological diseases, with an estimated frequency nearing 15% of the worldwide population (Steiner & Stovner, 2023), and the second most disabling condition worldwide, resulting in 45.1 million years of disability (Viana et al., 2020). All forms of migraine diagnosis are clinical challenges due to the lack of objective neuroimaging or laboratory findings. Current diagnosis procedures are based on symptom reporting and clinical criteria illustrated by ICHD in 2013, and ICHD-3 in 2018 (Yang et al., 2023). It is perfectly suited for building automated diagnoses and deploying artificial intelligence. However, this method results in delay and erroneous diagnosis due to different mimics, which can be a challenge for neurologists and prevent immediate actions, thus increasing patient suffering (Jain et al., 2024). For this purpose, our primary objective is to leverage different ML algorithms to effectively detect and classify migraine into subtypes like migraine without aura, migraine with aura, and chronic migraine.

Artificial intelligence (AI) is defined as an intelligent system or computer program that can imitate the cognitive function of humans (Adegboro et al., 2022). Nowadays, AI has influenced almost all areas of human life, including cybersecurity, Industry, and the medical field (Basu et al., 2020; Cacciamani et al., 2023; Dave & Patel, 2023; Pashkov et al., 2020). In healthcare, new frontiers have been opened, aligning with the advancement of AI and its technologies, especially ML and DL, where powerful tools have appeared to analyze high-dimensional, complex datasets and uncover patterns outperforming the traditional statistical approaches (Johnson et al., 2020; Poalelungi et al., 2023). In the case of migraine,

AI can play a vital role by aiding patients and physicians in several ways. For instance, AI-based models have the potential to improve diagnostic accuracy, particularly for non-headache experts, and could aid in classifying various groups of patients correctly (Woldeamanuel & Cowan, 2022). AI models can explore details in medical images that human professionals cannot detect and could make a quantitative assessment automatically (Burgos et al., 2021). In the context of migraine management, AI applications illustrated efficiency in identifying outcome metrics, therapy reaction prediction, and suitable treatment selection (Petrušić et al., 2024).

Current research focuses on predicting and classifying migraines, often relying on the analysis of electroencephalography (EEG) or magnetic resonance imaging (MRI) data. However, access to these technologies is expensive and sometimes difficult, which leads us to use data based on patient-reported symptoms. This work aims to develop a reliable classification system based on ML techniques to predict and classify migraine into its subtypes, aiming to accurate diagnostic with cost-effective manner. By leveraging the capabilities of ML and DL, we aim to advance the field toward more precise, timely, and individualized diagnostic frameworks for migraine patients. This paper targeted migraine because it is considered a highly studied headache diagnosis. In clinical practice, migraine is diagnosed using a hierarchical, step-by-step approach. The clinician begins by determining whether the condition falls under the category of migraine disorders or another type of headache, then proceeds to precisely classify the subtype. Based on this realistic diagnostic sequence, a two-stage, hierarchical deep learning framework has been proposed to embody the logic of clinical reasoning and systematically and methodically mimic its steps.

1.1 Problem Statement

Migraine is a complex disorder characterized by a variety of symptoms and multiple subtypes, making accurate diagnosis a real challenge. Furthermore, a large number of sufferers do not exhibit early warning signs, making timely diagnosis difficult and sometimes leading to overuse of medication or worsening of the condition. Therefore, accurate and early diagnosis of migraine is essential for developing a personalized treatment plan that helps reduce the severity and frequency of attacks. Despite the increasing use of AI in migraine diagnosis, current models often rely on expensive and not always readily available to all patients due to their cost or limited accessibility such as EEG or MRI. Hence, there is a need to employ artificial intelligence techniques to develop automated classification systems capable of accurately and reliably diagnosing migraine types based on the clinical symptoms reported by patients, achieving greater efficiency and reducing costs. This study focuses on tabulated clinical data, making the proposed approach more accessible and realistic for large-scale application

1.2 Scientific and Methodological Value of the Study

This study presents an advanced methodological framework for enhancing the reliability of AI systems in diagnosing migraine disorders through the following:

- 1- Proposing a two-stage hierarchical deep learning model that mimics the realistic clinical diagnostic pathway for migraine assessment.
- 2- Conducting a direct empirical comparison between flat multi-class classification and hierarchical classification using the same data to ensure fair assessment.
- 3- Implementing a systematic protocol for feature alignment across heterogeneous datasets to enable external validation.
- 4- Evaluating the robustness of the model under distribution variation across an independent external dataset.
- 5- Providing a reproducible framework for migraine diagnosis using tabular clinical data.

Most studies rely on single-stage models or basal systems (such as fuzzy logic) that do not utilize pattern learning from data. Despite the existence of such models, this study is the first to present a multi-stage hierarchical model using data-driven ML for migraine diagnosis based on tabular data. This model combines a comprehensive main task followed by classification.

2. Related Works

Different ML and DL methods are used for addressing migraine diagnosis problems. ML and DL models collect data from different sources and imitate decision-making logically to get the desired result. These AI models aid clinical, radiological, and neurophysiological diagnosis of migraine. The most popular models used were Support Vector Machine (SVM), Logistic Regression (LR), and Random Forests (RF), where Holsteen et al. (2020) utilized multivariable logistic regression to develop a prediction model for episodic migraine attacks according to migraineurs' self-prediction and regular trigger conditions like stress levels, menstruation, sleep quality, consumption of alcohol and caffeine, and weather conditions. Kwon et al. (2020) used an ML-based stacked classifier model to classify the primary headache diseases,

including migraine, tension-type headache (TTH), trigeminal autonomic cephalgia (TAC), epicranial, thunderclap, and other headaches. [Simić et al. \(2021\)](#) proposed a hybrid fuzzy clustering to classify various types of primary headache diseases, including migraine (with aura and without aura), tension-type headache (TTH), and other primary headaches. [Kapustynska et al. \(2024\)](#) focused on the applications of ML algorithms as migraine classifiers using statistical learning and functional learning. [Stubberud et al. \(2023\)](#) used ML to forecast migraine attacks by analyzing data from physiological measurements and preictal headache diaries collected by wearable devices. [Ashina et al. \(2021\)](#) used ML, the Least Absolute Shrinkage and Selection Operator (LASSO), and Random Forest algorithms to discover the determinants most closely associated with the quest for healthcare for migraines.

DL is a subset of AI that replicates the human brain's neural structures through complex neural network algorithms. Where features are automatically learned in DL, DL methods tend to be more efficient than ML ([Aslan & Akin, 2022](#); [Joshi et al., 2024](#)). [Katsuki et al. \(2020\)](#) developed a DL-based model utilizing an Artificial Neural Network (ANN) to diagnose basic headache, including migraine, trigeminal autonomic cephalgias, and tension-type headaches. [Sanchez-Sanchez et al. \(2020\)](#) designed a system for early migraine classification through ANNs to classify seven different types of migraine based on patient symptoms. [Khan et al. \(2024\)](#) used different ML algorithms with data augmentation to classify seven different types of migraine. The authors used data augmentation to increase the data size during the data pre-processing stage.

3. Materials and Methods

3.1 Study Design Framework

This study presents two different approaches/scenario to diagnosing migraine using AI techniques and clinical symptom analysis. The first scenario is based on a straightforward, single-stage classification model, where ML and DL algorithms are trained to classify migraine type to one of four classes: Chronic, With Aura, Without Aura, and Other for atypical cases.

The second scenario employs a two-stage hierarchical structure using a DNN, mimicking the diagnostic pathway used in medical practice. Where S1 distinguishes between migraine and non-migraine cases through binary classification, and S2 classifies migraine cases into its three common types: Chronic, With Aura, and Without Aura. To ensure an objective assessment of model reliability, the analysis was not limited to internal testing. The proposed framework was also applied to an independent external dataset to measure performance consistency and generalizability beyond the training dataset.

3.2 Description of Data Used in the Study

3.2.1 Source Data

The source dataset used in this research was obtained from Azadi Teaching Hospital in Kirkuk Governorate, Iraq. Data were collected from 1100 patients who have been diagnosed with various types of migraines and other headache types similar to migraine. The dataset is collected via a questionnaire that included 30 questions that represent features describing each patient's clinical and demographic characteristics, as well as the type of migraine they were experiencing. These features included a wide range of variables, such as age, gender, duration and frequency of headache attacks, location, and intensity, as well as associated symptoms such as nausea, vomiting, phonophobia, photophobia, dizziness, speech disturbances, and sensory and motor disturbances. Additional factors such as food or sleep habits, menstruation relation, at what point of menstruation, family history, blood pressure, and heart disease were also included, in addition to migraine type (Chronic, With Aura, Without Aura, and Other) for each individual. It is worthy to mention, that "Other" category represents cases where patients have symptoms similar to migraine, such as tension that do not fit the standard migraine types, or cases of unclear definition.

3.2.2 External Data for Generalization Testing

The external dataset was obtained from the Kaggle repository and comprises 400 documented migraine cases with 24 clinical variables ([Khan et al., 2024](#)). These variables include demographic characteristics and detailed symptoms associated with headache attacks, such as age, attack duration, etc., as well as associated symptoms like nausea, vomiting, photophobia, etc., in addition to the diagnostic type variable. The cases are distributed across seven main diagnostic categories, including: Migraine without aura, Typical aura with migraine, Typical aura without migraine, Basilar-type aura, Sporadic hemiplegic migraine, Familial hemiplegic migraine, and Other. For generalization, class mapping is performed by mapping 'Migraine without aura' to 'Without Aura', and mapping 'Typical aura with migraine', 'Typical aura without migraine',

'Familial hemiplegic migraine', 'Sporadic hemiplegic migraine', 'Basilar-type aura', to 'With Aura' where these classes are subtypes of With Aura according to ICHD-3, while 'other' represent non-migraine cases. Since "Chronic" migraine cases were not available in the external data, this type was not included in the generalization analysis in the second stage.

3.3 Data Preprocessing

Data preprocessing is an important step before supplying the data to ML models, as it contributes to improving the accuracy and quality of results. During this stage, missing or illogical values were detected, corrected, or excluded, while duplicate records and irregular patterns were removed to ensure that all samples are accurate and representative of the clinical landscape. The dataset contains two types of data: numerical data, such as age and pain intensity, and categorical data, such as gender, pain location, and associated symptoms, which are converted into a numerical representation using the One-Hot Encoding technique to make them processable by statistical algorithms. After applying one-hot encoding, the dataset contained 71 dummy columns. To reduce dimensionality, a filter-based feature selection method was used. Where it calculates the mutual information score (MI) between each attribute and the target variable (i.e. measures the strength of linear and non-linear relationship). After that the attributes are ranked in descending order, and the 50 most informative attributes were selected for classification (for scenario 1) while for scenario 2 only common features between both source and external dataset were used, which is about 16 features. The target "migraine type" column was encoded using label encoding to transform the four migraine types to numerical values to be appropriate for model training. The dataset splitting ratio is 80:20, where 80% represents the training set and the rest 20% for test set for scenario1 while this 20% represents validation data in scenario2 because in scenario2 the external dataset considered as test data.

The target distribution was analyzed, where imbalance class distribution was revealed, especially the minor class like "Chronic" for scenario 1 and second stage of scenario 2. For this purpose, data augmentation techniques were used to expand the training dataset size and improve its predictive power. Among these techniques, the Synthetic Minority Oversampling Technique (SMOTE) was adopted, an effective method used to address class imbalance. The SMOTE was applied to expand training data only to 1000 instead of 880 (80% of the dataset), targeting the minor classes to achieve balanced distribution.

3.4 Scenario One: Direct Multi-Class Classification

In this scenario, the problem is treated as a four-class classification task performed in a single step. After preparing the dataset, it is applied to several ML algorithms such as SVM, K-Nearest Neighbors (KNN), LR, RF, Decision Trees (DT), GaussianNB, and DNN models for migraine classification. The basic architecture of a DNN consists of an input layer representing 50 selected attributes, followed by four hidden layers, and then an output layer that used the Softmax function to generate the probabilities of belonging to the four classes. The categorical cross-entropy loss function is used with the Adam optimizer, and the hyperparameters for instance, batch size, learning rate, and number of training epochs were experimentally fine-tuned, with an early stop mechanism employed to avoid overfitting.

3.5 Scenario Two: A Two-Stage Hierarchical Classification Model

To simulate the mechanism of real clinical diagnosis workflow, a deep learning framework was developed based on two interconnected stages. In the first stage, a binary classifier based on a deep neural network was trained to differentiate between migraine and non-migraine cases, using a binary cross-entropy loss function. This stage focused on simplifying the decision by answering the fundamental question: Is the case migraine or not?

In S1, a fully connected deep neural network was used to classify samples as migraine and non-migraine. It consists of an input layer with the number of features shared between the training (source dataset) and test data (external dataset), followed by two hidden layers 64,32. The final layer is a linear output layer consisting of two binary classification units. The CrossEntropyLoss loss function and the Adam optimizer with a learning rate of 0.001 were used, and the network was trained for 80 training periods (epochs). Batch Normalization and a Dropout were also used to minimize overfitting, and a random seed was fixed to ensure reproducible results.

The S2 only accepted samples that were classified in S1 as migraine and reclassified into three migraine subtypes: Chronic, With Aura, and Without Aura, using a multi-class neural network and categorical cross-entropy loss. This approach limits overlap between classes and reflect the logical sequence of medical assessment, where the general probability is first ruled out before moving to specifics. To ensure fair assessment in the absence of class imbalances, overall means of the measures were reported.

In S2, a multi-layered fully connected network was designed, it consists of input layer with four hidden layers (128, 64, 32, and 16 neural units) followed by a three-class output layer. The hidden layers are equipped with the ReLU activation function to enable the model to learn nonlinear relationships between variables. Batch normalization was applied after each hidden layer to improve stability and accelerate the learning process, along with dropout at decreasing rates (0.4, 0.3, 0.2,

and 0.1) to limit overlearning. The benefit of these deep, graded layers lies in enabling the model to learn complex patterns and nonlinear relationships between clinical features. The first layers extract general features, while the deeper layers capture the subtle differences between migraine subtypes. Batch normalization helps stabilize and accelerate the learning process, dropout reduces the risk of overfitting, and SMOTE technology addresses class imbalances and improves classification accuracy. The CrossEntropyLoss loss function and the Adam optimizer with a learning rate of 0.0015 were used. The model was trained for 120 training periods, with SMOTE applied to the training data only to address class imbalances, StandardScaler used to standardize features, and random seeding adjusted to ensure reproducibility of the results.

3.6 Validation and External Generalization

To ensure a rigorous and systematic evaluation of the generalizability of the proposed models, two complementary evaluation scenarios were adopted, strictly separating the training, validation, and test data. In the first scenario, the source dataset was divided into training data (80%) and test data (20%). Within the training data only, a validation mechanism was used to fine-tune the hyperparameters and select the optimal model. The test set was never used in training, weighting, or parameter adjustments; its use was limited to the final evaluation after the training process was complete. This provided an unbiased assessment of performance within the same data distribution. The second scenario followed a two-stage protocol: First, in-distribution evaluation using the test set from the source dataset, following the same mechanism as the first scenario. Validation data was used only within the training set to fine-tune the model, and then it was evaluated on the test set, which was not used in training. Second, external validation was performed using a completely independent dataset (external dataset) that did not participate in any stage of training, internal split, or parameter adjustment. In this stage, the trained models were directly applied to the external dataset without retraining or parameter refitting, to simulate the actual clinical usage scenario after model deployment.

3.7 Performance Evaluation Criteria

The models were evaluated using a set of statistical indicators, including accuracy, precision, recall, and the F1 score. Confusion matrices were also used to analyze error patterns and understand the model's behavior toward each category. Macro averages were reported to ensure fair representation of the different classes, especially in cases of imbalance. This methodology enables an objective comparison between the single-stage and hierarchical models, as well as an assessment of the generalization's power on independent external data. Thus, while the internal test in both scenarios measures the model's ability to generalize within the same data distribution, the external validation in the second scenario measures its ability to generalize under a data distribution shift. Because the number of features differed between the two datasets, only common features (16 features) were precisely aligned, maintaining the same order and structure as in the training dataset. The same coding methods and measurement parameters, extracted exclusively from the training dataset, were applied to the external dataset without refitting to prevent information leakage and ensure the integrity of the evaluation. The common features that have been used for generalization were: age, ataxia, character, duration, dysphasia, frequency, hypoacusis, intensity, location, nausea, phonophobia, photophobia, sensory, tinnitus, vertigo, and visual.

The external validation maintained the inference hierarchy; only cases predicted as migraine by the first-stage model were transferred to the second stage, reflecting the realistic clinical decision-making process. Performance was measured using accuracy, precision, recall, and the F1-score, in addition to confusion matrix analysis. This multi-level framework provides a comprehensive assessment of the model's stability and reliability in different clinical settings.

3.8 Implementation Details

All experiments were performed using Python (version 3.9.23) with PyTorch (version 2.5.1) for deep learning model development and Scikit-learn (version 1.2.2) for data processing and statistical analysis. The experiments were conducted on a machine with an Intel Core i7 processor and 16 GB of RAM. To ensure reproducibility, the random seed value was fixed at 42 across all experimental phases. The source code will be made available upon request to support transparency and enable replication of the study.

4. Results and Model Evaluation

4.1 The Performance of Scenario 1: Direct Multi-Class Classification

In Scenario 1, various models were trained to directly classify migraine type into four classes: Chronic, With aura, Without Aura, and Other. The performance of the classifier models was analyzed and evaluated using different performance metrics, including the classification report and confusion matrix, to reveal the detailed accuracy in classifying migraine



subtypes on the test dataset. These matrices show the number of correct and incorrect predictions for each class, allowing for a deeper understanding of the strengths and weaknesses of each model's performance. Key performance indicators, including recall, precision, accuracy, and F1-score, were derived from the classification report of all used models to provide a comprehensive assessment of the models' ability to accurately discriminate between different types of migraine, as shown in table 1. Evaluation results on the test group showed that the DNN model performed best compared to traditional models, achieving 99.55% accuracy. This was followed by the SVM model with 99.09% accuracy, while the LR and RF models achieved 97.27% accuracy, the DT model 96.82%, the KNN model 96.36%, and the GaussianNB model the lowest accuracy at 86.36%. These results highlighted the power of combining feature selection with preprocessing and data augmentation techniques to improve model performance. The confusion matrices, classification reports, and learning & loss curves of DNN in scenario 1 are shown in figures 1,2, and 3, respectively. The training and validation accuracy and loss curves of DNN across epochs showed consistent learning behavior and stable convergence throughout the training process, meaning the absence of overfitting the model.

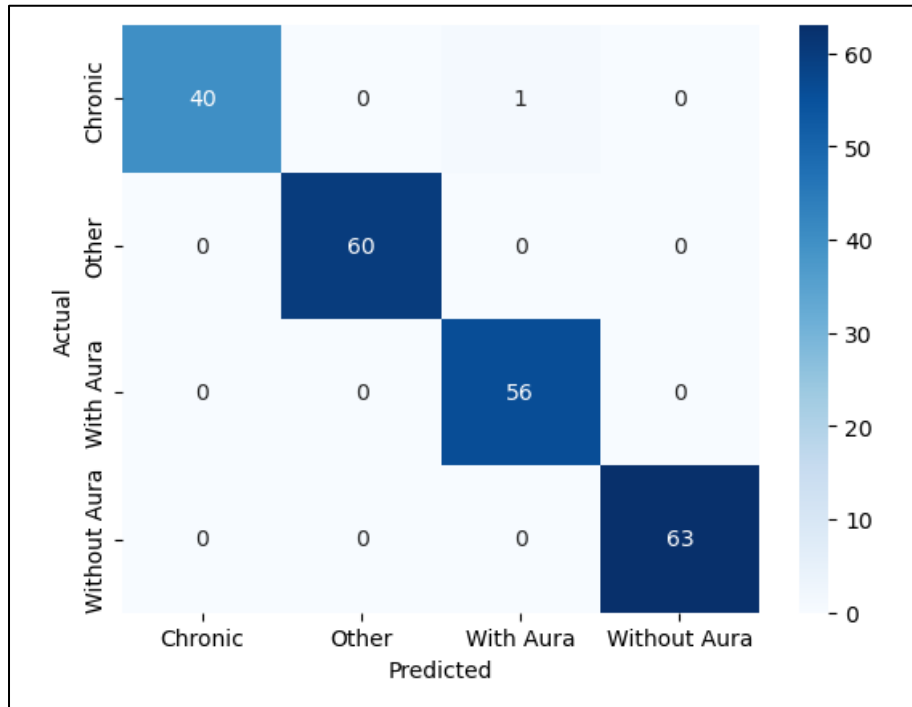


Figure 1: Confusion Matrix of DNN for Scenario1

Classification Report:				
	precision	recall	f1-score	support
Chronic	1.0000	0.9756	0.9877	41
Other	1.0000	1.0000	1.0000	60
With Aura	0.9825	1.0000	0.9912	56
Without Aura	1.0000	1.0000	1.0000	63
accuracy			0.9955	220
macro avg	0.9956	0.9939	0.9947	220
weighted avg	0.9955	0.9955	0.9954	220

Figure 2: Classification Report of DNN for Scenario1

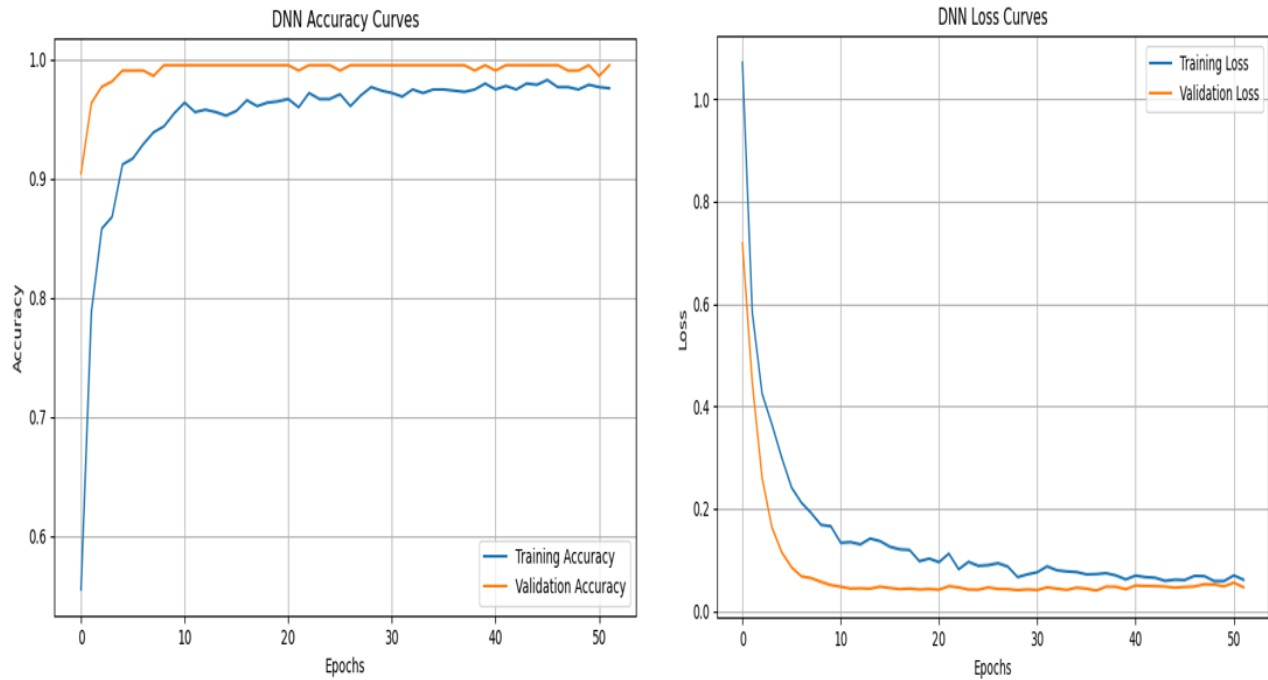


Figure 3: Training vs Validation Accuracy and Loss Curves of DNN for Scenario1

Table 1: Classification Report of all Used Classifiers of Scenario 1

Model	Precision	Recall	F1-Score	Accuracy
DNN	99.55	99.39	99.47	99.55
SVM	99.11	99.09	99.09	99.09
Logistic Regression	97.38	97.27	97.28	97.27
Random Forest	97.39	97.27	97.26	97.27
Decision Tree	96.90	96.82	96.78	96.82
KNN	96.42	96.36	96.35	96.36
GaussianNB	87.05	86.36	86.44	86.36

4.2 The Performance of Scenario 2: Two-Stage Hierarchical Classification

4.2.1 Results of Stage 1: Classification of Migraine vs. Non-Migraine

The results of S1 on the internal test group demonstrated the model's high ability to distinguish between migraine and non-migraine patients, with an accuracy of 99.55%. Detailed report analysis showed that the model accurately classified the non-migraine group with 100% accuracy, 98.33% recall, and 99.16% F1-score, on the other hand, accuracy, recall, and F1-score for migraine class were 99.38%, 100%, and 99.69% respectively. This performance indicates the model's effectiveness in minimizing errors between the main classes.

4.2.2 Results of Stage 2: Classification of Migraine Subtypes

In S2, samples classified as migraine in S1 were further classified into the following subtypes: Chronic, With Aura, And Without Aura. The results on the internal test set showed an overall accuracy of 93.75%. Specifically, the Chronic subtype achieved 81.63% accuracy and 97.56% recall, the With Aura subtype achieved 97.92% accuracy and 83.93% recall, while the Without Aura subtype performed flawlessly with 100% accuracy and recall.

These results demonstrate the model's ability to distinguish between different subtypes with high accuracy, although there were some challenges in the Chronic subtype due to the smaller sample size compared to the other subtypes. The confusion matrices, classification report, and learning & loss curves of train/test for S1 and S2 are illustrated in figures 4, 5, 6, and 7 respectively.

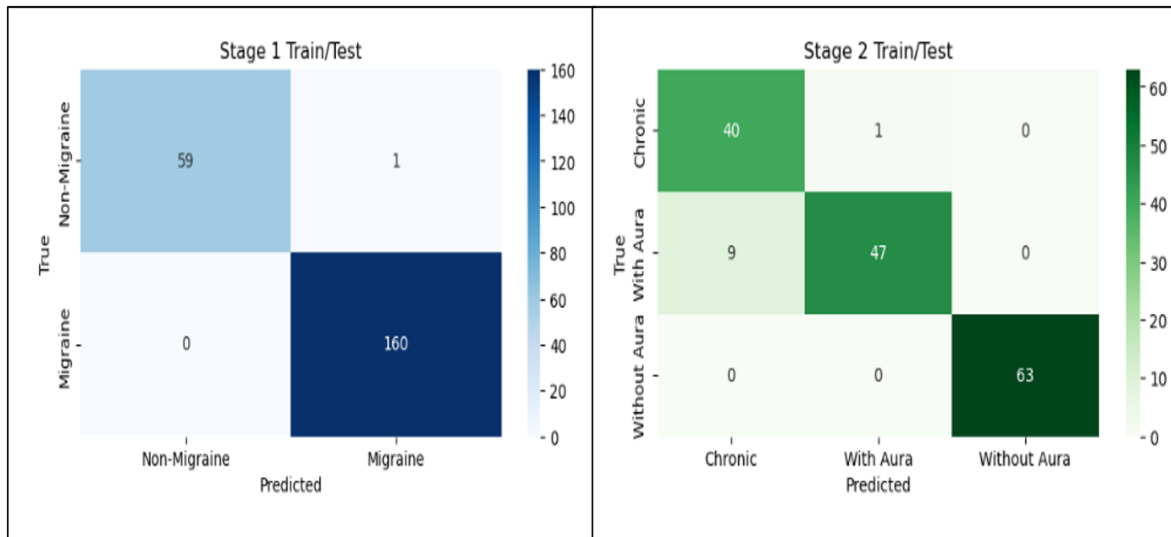


Figure 4: Confusion Matrices of Second Scenario-Both Stages

===== STAGE 1 REPORT (Train/Test) =====				
	precision	recall	f1-score	support
Non-Migraine	1.0000	0.9833	0.9916	60
Migraine	0.9938	1.0000	0.9969	160
accuracy			0.9955	220
macro avg	0.9969	0.9917	0.9942	220
weighted avg	0.9955	0.9955	0.9954	220

===== STAGE 2 REPORT (Train/Test) =====				
	precision	recall	f1-score	support
Chronic	0.8163	0.9756	0.8889	41
With Aura	0.9792	0.8393	0.9038	56
Without Aura	1.0000	1.0000	1.0000	63
accuracy			0.9375	160
macro avg	0.9318	0.9383	0.9309	160
weighted avg	0.9456	0.9375	0.9379	160

Figure 5: Classification Report of Second Scenario-Both Stages

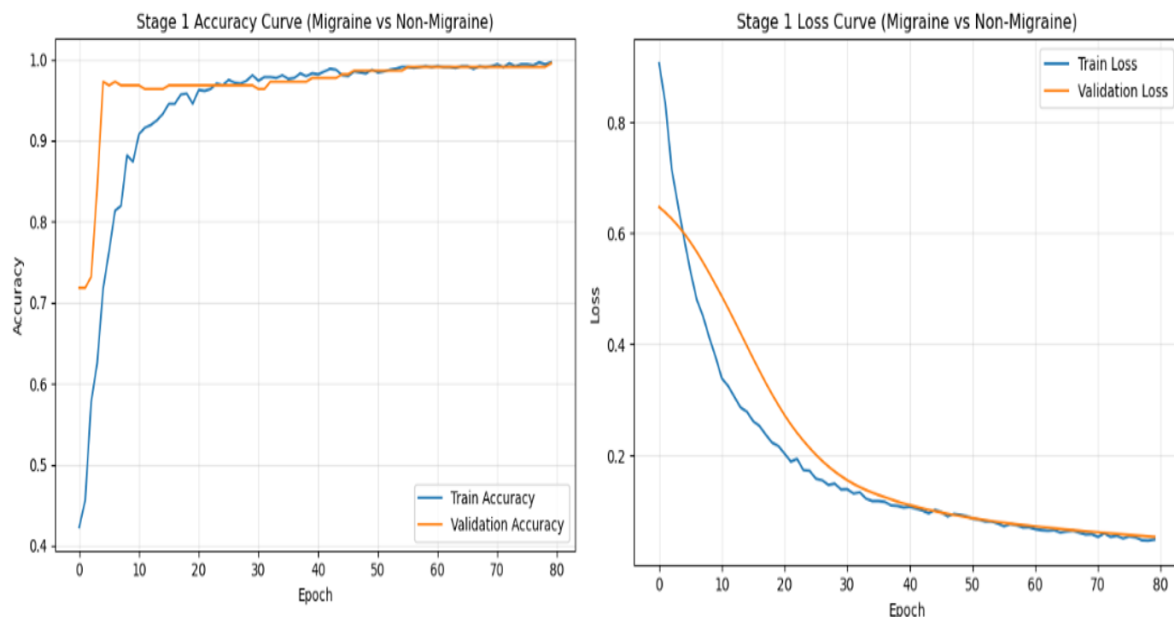


Figure 6: Training vs Validation Accuracy and Loss Curves of Second Scenario-S1

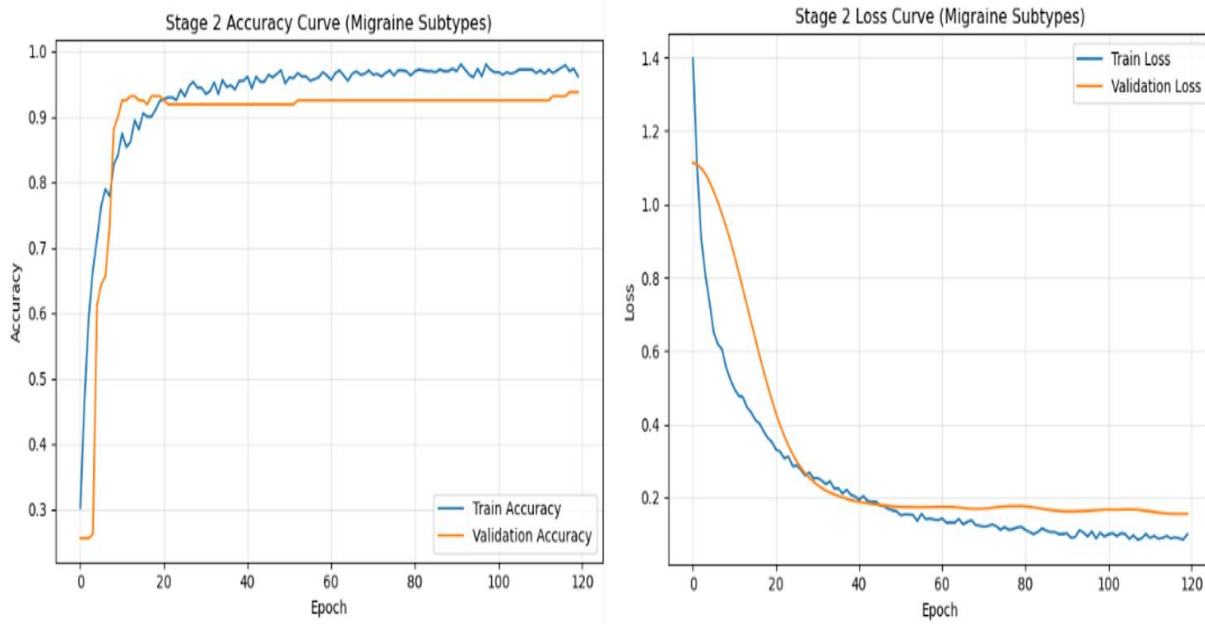


Figure 7: Training vs Validation Accuracy and Loss Curves of the Second Scenario-S2

4.2.3 Results of Generalization on External Dataset

To measure the model's ability to generalize, it was tested on an external dataset (unseen data) not used in the training, in order to assess its stability and performance in an environment different from the training data. The generalization is applied only for scenario2.

When the model was evaluated on the external dataset, S1 accuracy reached 96.50%, with a significant difference in recall for the non-migraine group (only 17 samples were classified with an accuracy of 0.1765), while the recall for the migraine group was 100%. This difference reflects the effect of data imbalance in the external generalization set, but it does not prevent the overall performance of the model from being good, as the weighted accuracy rate reached 96.62% and the weighted F1 rate reached 95.31%. In addition to the impact of data imbalance between the migraine and non-migraine classes, another factor contributing to the low recall of the Non-Migraine category in the outgoing generalization set is the data class labeled "Other." This class contains multiple types of headaches, it might be including Chronic migraine, and its content is not precisely defined. The model struggles to classify these samples correctly, resulting in lower recall accuracy for this category despite the model's overall high performance. However, the potential for future performance improvement can be done by expanding the "Other" class by collecting more specific data. The confusion matrix and classification report for generalization of the second scenario result are shown in figures 8 and 9.

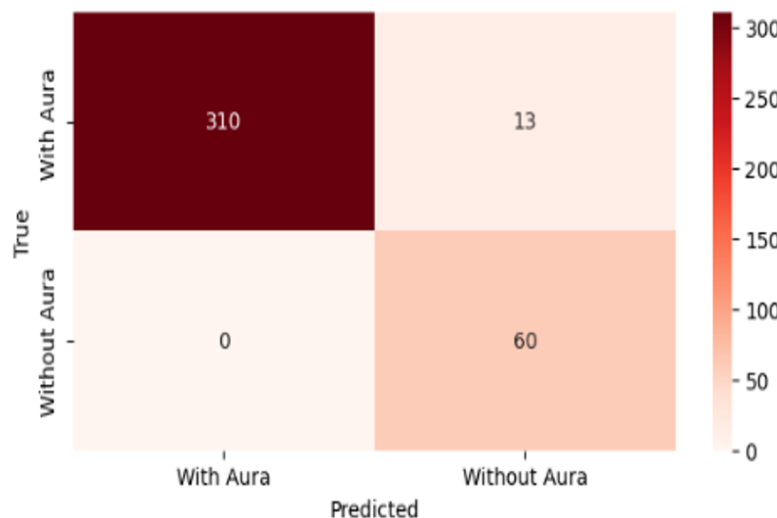


Figure 8: Confusion Matrix of Generalization for the Second Scenario.

===== GENERALIZATION REPORT (With / Without Aura) =====				
	precision	recall	f1-score	support
With Aura	1.0000	0.9598	0.9795	323
Without Aura	0.8219	1.0000	0.9023	60
accuracy			0.9661	383
macro avg	0.9110	0.9799	0.9409	383
weighted avg	0.9721	0.9661	0.9674	383

Figure 9: Classification Report of Generalization for the Second Scenario

On the other side, when S2 was applied to the external dataset, the samples were classified into only two subtypes (With Aura and Without Aura) due to the absence of Chronic migraine class in the external dataset. The results showed an overall accuracy is 96.61%, With Aura class achieving 100% accuracy and 95.98% recall, while Without Aura class achieved 82.19% accuracy and 100% recall. The results revealed that the model is capable of generalizing well while maintaining high performance for the most representative classes. Furthermore, it confirms that the hierarchical (two-stage) approach provides a framework more compatible with medical practice and enhances the model's reliability when applied in different clinical settings.

5. Discussion

The results of this study demonstrate the vital role of AI in accurately diagnosing and classifying migraine, compared to the previous studies listed in Table 2. Comparing the results of the current study with previous studies reveals that while some studies achieved high accuracy, such as [Sanchez-Sanchez et al. \(2020\)](#) and [Khan et al. \(2024\)](#) many relied on small sample sizes and data from a single center, or only internal validation without independent testing ([Stubberud et al., 2023](#)), ([Ashina et al., 2024](#)) which may affect clinical generalizability. In contrast, the current study used a relatively large sample size (1100 patients) and 30 clinical features, comparing several traditional ML algorithms alongside a DNN model. A two-stage classification framework was also applied, mimicking the clinical diagnostic process (first identifying migraine and then classifying its subtypes), with performance evaluated on non-visual data. This enhances the reliability and stability of the results compared to studies that relied solely on cross-validation. Many studies still suffer from limited dataset size or a lack of external validation. In this context, the current study contributes to strengthening this research approach by incorporating a more realistic phased classification methodology, using a relatively larger sample, and an evaluation based on non-visual data, thus enhancing the maturity of the proposed models and bringing them closer to practical clinical application.

Despite the high performance of both models, several points should be considered. First, the generalization report in the first phase showed a significant decrease in recall for the "non-migraine" class, indicating the model's high sensitivity in detecting migraine cases versus challenges in distinguishing some non-affected cases when the data distribution changes. This reflects the complex nature of differentiating between migraine and conditions with similar symptoms and underscores the importance of improving data balance or employing strategies to address imbalances in the future.

Second, the external dataset did not include Chronic migraine cases, limiting the ability to assess the generalization of the S2 across all subtypes. Therefore, future studies should expand the external dataset to include all clinical classes. Accordingly, although the first scenario achieves higher accuracy within the same dataset, the second scenario is more methodologically robust and clinically realistic, given its external testing and reliance on a graded diagnostic structure that simulates the actual medical decision. Based on these findings, it can be concluded that the use of AI, particularly hierarchical approaches and DNN, provides an effective framework for developing accurate and reliable diagnostic systems for migraine. This system can also be applied to support medical decision-making, improve the quality of healthcare, reduce misdiagnosis, and decrease medication overuse, thus contributing to personalized and effective patient care.

Finally, the study recommends continuing to collect larger and more diverse datasets, including clinical data and medical images, to enhance the models' ability to learn and generalize, as well as validating the model's effectiveness in different centers and hospitals to ensure its practical applicability. Furthermore, for future research, it is recommended to

explore the possibility of incorporating additional features, such as genetic data or other biomarkers, to further enhance the model's performance and increase classification accuracy.

6. Conclusion

I. In summary, migraine is a widespread neurological disease, it is marked by unilateral, pulsating, and pulsing headaches. Unfortunately, a large percentage of migraineurs are not correctly diagnosed when typical diagnostic criteria are employed following the International Headache Society's (IHS) guidelines. In this context, there is considerable interest in the advancement of automated methods that can assist in the diagnosis of migraine. For this purpose, this study proposed two different approaches to classify migraine to its subtypes; a direct multi-class classification and two-stage hierarchical classification. The results of this study demonstrate that the application of AI techniques, particularly DNN, can significantly improve the accuracy of migraine diagnosis and subclassification. In the first scenario, the DNN model was able to classify patients into four categories (Chronic, With Aura, Without Aura, and Other) with 99.55% accuracy on the internal test set, outperforming all conventional models. These findings approve the effectiveness of the direct approach in rapidly identifying all classes, but it may struggle to differentiate between underrepresented class or non-migraine cases.

II. In contrast, the second scenario performed more closely in line with clinical diagnostic methods. The first stage successfully distinguished between migraine and non-migraine patients with 99.55% accuracy on internal data and 96.50% when generalized to external data. The second stage demonstrated the ability to classify subtypes (with aura, without aura, and chronic) with 93.75% accuracy on internal data and 96.61% on external data after excluding the chronic category due to its absence in the generalization set. This scenario demonstrates clear advantages in improving category discrimination and reducing errors between unaffected cases and underrepresented subtypes.

Furthermore, generalization results on external data showed that the models maintain high performance even with variations in feature size and sample distribution, reflecting the model's adaptability and reliability in diverse clinical settings.

Table 2 (a): Comparison of the Proposed Migraine Classification Models with Previous Studies Used Tabular Dataset.

#	Ref.	Model	Sample Size	Variables /Features	Best Model with Accuracy/AUC	Results	Limitation
1	Holsteen et al. (2020)	Multivariable LR	178 participants	Stress levels, self-predicted probability, menstruation, sleep quality, consumption of alcohol and caffeine, and weather conditions.	-----	The model's performance was considered insufficient for practical application.	It is essential to improve both the accuracy and breadth of data acquisition to develop migraine prediction models that are practically helpful.
2	Kwon et al. (2020)	Stacked XGBoost, KNN, SVM, and RF in each layer.	2162 participants	For the first layer (migraine classifier), 32 features.	XGBoost achieved the highest performance for migraine classification (Acc. 81%)	The study underlined the important nature of particular features in the classification procedure; for instance, the nature of pain, mode of onset, and related symptoms were significant in the differentiation of various headache types. This knowledge may help clinicians in focusing on essential symptoms during the diagnostic process.	1. A single center of data that affects the validity for other populations. 2. Excluding secondary headaches. 3. The study applied conventional ML methods that might not leverage the potential of advanced DL.
3	Simić et al. (2021)	Hybrid fuzzy clustering "FMLE" (combination of fuzzy partition and Maximum Likelihood Estimation (MLE) Clustering)	579 instances for the first dataset.	9 features	Hybrid fuzzy clustering "FMLE" (combination of fuzzy partition and Maximum Likelihood Estimation (MLE) Clustering) achieved (Acc. 83%, and 86%) using the first and second datasets, respectively.	The result was evaluated using two datasets, where it achieved (Acc. 83%, and 86%) using the first and second datasets, respectively.	-----

Table 2 (b): Comparison of the Proposed Migraine Classification Models with Previous Studies Used Tabular Dataset.

#	Ref.	Model	Sample Size	Variables /Features	Best Model with Accuracy/AUC	Results	Limitation
4	Kapustynska et al. (2024)	1.Statistical learning: Bayes Net, Naïve Bayes, Naïve Bayes Multinomial Updateable, Naïve Bayes Updateable, Naïve Bayes Multinomial 2. Functional learning: Latent Dirichlet allocation, Sequential minimal optimization, Quadratic Discriminant Analysis. Multinomial LR, MLP.	400 records from the Kaggle repository.	24 variables	Naïve Bayes and Naïve Bayes Updateable (Acc. 93.50% both) in classifying migraine.	Naïve Bayes is an effective tool to predict migraine characteristics that can provide beneficial knowledge for clinical applications.	1. Limited dataset size. 2. Lack of model validation on independent datasets, which is important for real-world applications.
5	Stubberud et al. (2023)	RF, GB, LR, SVM, AdaBoost, and XGBoost	18 participants	438 headache diaries and 391 daily biofeedback sessions collected for 388 days	RF (Acc. 56%) and GB (Acc. 62%) classifiers showed the best performance.	Models with high dimensions can forecast with high accuracy.	1. Small sample size. 2. To enhance the predictive accuracy, further research is needed.
6	Ashina et al. (2024)	RF and LASSO	61,826 participants' responses	54 features	RF (AUC 0.7)	The modeling methodology was highly useful for identifying and validating the factors determining the decision to seek medical care for migraines.	1. Reliability may be affected due to possible inaccuracies in reporting symptoms. 2. The findings may not be generalized to all populations.
7	Katsuki et al. (2020)	NLP using ANN	848 patients (46%, migraineurs, 47%, tension-type headaches, 5% trigeminal autonomic cephalalgias, and 2% other primary headache disorders)	Four features: Sex, age, consultation date, and unstructured description of the Japanese patients	ANN (Acc. 77.59%)	After eliminating secondary headaches, the study determined that the model could be useful for effective medical practice, possibly improving the quality of life and reducing the load on patients as well as doctors.	1. Limited Data Variables. 2. Limited sample size relatively which may limit the model's generalization. 3. Only internal cross-validation was performed without external validation, which means applying the model to other hospitals is uncertain. 4. Poorly detailed classification where the study did not separate non-chronic migraine (less than 15 days per month) and chronic migraine and TTH (15 days or more per month).
8	Sanchez-Sanchez et al. (2020)	ANN (MLP), LR, SVM, NN, DT	400 participants	23 for the Complete Model, 18 for the Reduced Model	Complete Model & Reduced Model: ANN (1 hidden layer & 10, 20 hidden neurons respectively) (Acc. 97.5%, 97.5%, Precision: 97%, 98% respectively).	Automatic classification of 7 types of migraines	Lack of generalizability of the results to other populations (conducted at a single institution in Colombia), limited dataset, absence of independent validation: separate test data for final model validation isn't used (only cross-validation has been used), and absence of comparison to clinical practice.

Table 2 (c): Comparison of the Proposed Migraine Classification Models with Previous Studies Used Tabular Dataset.

#	Ref.	Model	Sample Size	Variables /Features	Best Model with Accuracy/AUC	Results	Limitation
9	Khan et al. (2024)	SVM, KNN, RF, DT, and deep neural networks (DNN)	400 patients, and after data augmentation data size increased to 1447 patient records.	24 attributes.	DNN with data augmentation outperformed ML algorithms (Acc. 99.66%)	1. The classification model with data augmentation could classify seven types of migraine. 2. The study investigated the data augmentation impact on these algorithms' performance.	1. Dataset size limitation. 2. Dependence on data augmentation because it may limit the effectiveness of the classification model in different clinical settings. 3. Lack of model generalization on different datasets. 4. Migraine diagnosis complexity can affect the classification process, particularly in undeveloped countries due to a shortage of trained professionals.
10	Proposed Study	Scenario 1: A direct, multi-class DNN model with a comparison to ML algorithms (SVM, KNN, RF, DT, GaussianNB). Scenario 2: A two-stage hierarchical classification framework that simulates clinical decision-making (migraine detection followed by subtype classification).	1100 patients (the largest compared to most previous studies).	30 features in the first scenario, and 16 common features in the second scenario after aligning the features with an external dataset.	In the first scenario, the DNN model achieved the highest performance (Acc. 99.55%). In the hierarchical scenario, the first stage achieved 99.55% accuracy, while the second stage achieved 93.75% accuracy.	The results showed that the phased framework reflects true clinical diagnostic logic and reduces confusion between similar categories. Furthermore, the model's reliability was enhanced through external evaluation, addressing a common shortcoming of previous studies.	1- Data from a single center limits generalization. 2- The absence of "chronic" migraine-type in external data restricted overall assessment. 3- Alignment of features between the two datasets limited some predictive variables.

7. Institutional Review Board Statement

This study was approved by the Kirkuk Health Directorate – Iraq, in its capacity as the ethical oversight body for Azadi Teaching Hospital/ Kirkuk Governorate, Iraq, in accordance with the approved ethical standards and regulations for handling clinical data. Official approval to collect the dataset was granted under reference number (817) on October 31, 2024.

8. References

- Adegboro, C. O., Choudhury, A., Asan, O., & Kelly, M. M. (2022). Artificial intelligence to improve health outcomes in the NICU and PICU: a systematic review. *Hospital Pediatrics*, 12(1), 93-110.
- Ashina, M., Katsarava, Z., Do, T. P., Buse, D. C., Pozo-Rosich, P., Özge, A., Krymchantowski, A. V., Lebedeva, E. R., Ravishankar, K., & Yu, S. (2021). Migraine: epidemiology and systems of care. *The Lancet*, 397(10283), 1485-1495.
- Ashina, S., Muenzel, E. J., Nicholson, R. A., Zagar, A. J., Buse, D. C., Reed, M. L., Shapiro, R. E., Hutchinson, S., Pearlman, E. M., & Lipton, R. B. (2024). Machine learning identifies factors most associated with seeking medical care for migraine: results of the OVERCOME (US) study. *Headache: The Journal of Head and Face Pain*, 64(8), 1027-1039.
- Aslan, Z., & Akin, M. (2022). A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals. *Physical and Engineering Sciences in Medicine*, 45(1), 83-96.
- Basu, K., Sinha, R., Ong, A., & Basu, T. (2020). Artificial intelligence: how is it changing medical sciences and its future? *Indian J Dermatol*, 65(5), 365-370.
- Burgos, N., Bottani, S., Faouzi, J., Thibeau-Sutre, E., & Colliot, O. (2021). Deep learning for brain disorders: from data processing to disease treatment. *Briefings in Bioinformatics*, 22(2), 1560-1576.
- Cacciamani, G. E., Chu, T. N., Sanford, D. I., Abreu, A., Duddalwar, V., Oberai, A., Kuo, C.-C. J., Liu, X., Denniston, A. K., &

- Vasey, B. (2023). PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nature Medicine*, 29(1), 14-15.
- Dave, M., & Patel, N. (2023). Artificial intelligence in healthcare and education. *Br Dent J*, 234(10), 761-764.
- Holsteen, K. K., Hittle, M., Barad, M., & Nelson, L. M. (2020). Development and internal validation of a multivariable prediction model for individual episodic migraine attacks based on daily trigger exposures. *Headache: The Journal of Head and Face Pain*, 60(10), 2364-2379.
- Jain, A., Bali, O., Ritchie, N., & Singla, S. (2024). Artificial Intelligence in the Diagnosis of Migraine (with or without Aura) in Adults: A Systematic Review. *Journal, Indian Academy of Clinical Medicine*, 25(3), 87-99.
- Johnson, J., Simms-Ellis, R., Janes, G., Mills, T., Budworth, L., Atkinson, L., & Harrison, R. (2020). Can we prepare healthcare professionals and students for involvement in stressful healthcare events? A mixed-methods evaluation of a resilience training intervention. *BMC Health Serv Res*, 20(1), 1094.
- Joshi, G., Jain, A., Araveeti, S. R., Adhikari, S., Garg, H., & Bhandari, M. (2024). FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics*, 13(3), 498.
- Kapustynska, V., Abromavičius, V., Serackis, A., Paulikas, Š., Ryliskenė, K., & Andruskevicius, S. (2024). Machine learning and wearable technology: Monitoring changes in biomedical signal patterns during pre-migraine nights. *Healthcare*, 12(1), 1-15.
- Katsuki, M., Narita, N., Matsumori, Y., Ishida, N., Watanabe, O., Cai, S., & Tominaga, T. (2020). Preliminary development of a deep learning-based automated primary headache diagnosis model using Japanese natural language processing of medical questionnaire. *Surg Neurol Int*, 11, 475.
- Khan, L., Shahreen, M., Qazi, A., Jamil Ahmed Shah, S., Hussain, S., & Chang, H.-T. (2024). Migraine headache (MH) classification using machine learning methods with data augmentation. *Scientific Reports*, 14(1), 5180.
- Kwon, J., Lee, H., Cho, S., Chung, C.-S., Lee, M. J., & Park, H. (2020). Machine learning-based automated classification of headache disorders using patient-reported questionnaires. *Scientific Reports*, 10(1), 14062.
- Pashkov, V. M., Harkusha, A. O., & Harkusha, Y. O. (2020). Artificial intelligence in medical practice: regulative issues and perspectives. *Healthcare*, 10(1), 1-15.
- Petrušić, I., Ha, W.-S., Labastida-Ramirez, A., Messina, R., Onan, D., Tana, C., & Wang, W. (2024). Influence of next-generation artificial intelligence on headache research, diagnosis and treatment: the junior editorial board members' vision-part 1. *The Journal of Headache and Pain*, 25(1), 151.
- Poalelungi, D. G., Musat, C. L., Fulga, A., Neagu, M., Neagu, A. I., Piraianu, A. I., & Fulga, I. (2023). Advancing patient care: how artificial intelligence is transforming healthcare. *Journal of personalized medicine*, 13(8), 1214.
- Sanchez-Sanchez, P. A., García-González, J. R., & Ascar, J. M. R. (2020). Automatic migraine classification using artificial neural networks. *F1000Research*, 9, 618.
- Simić, S., Banković, Z., Villar, J. R., Simić, D., & Simić, S. D. (2021). A hybrid fuzzy clustering approach for diagnosing primary headache disorder. *Logic Journal of the IGPL*, 29(2), 220-235.
- Steiner, T. J., & Stovner, L. J. (2023). Global epidemiology of migraine and its implications for public health and health policy. *Nature Reviews Neurology*, 19(2), 109-117.
- Stubberud, A., Ingvaldsen, S. H., Brenner, E., Winnberg, I., Olsen, A., Gravdahl, G. B., Matharu, M. S., Nachev, P., & Tronvik, E. (2023). Forecasting migraine with machine learning based on mobile phone diary and wearable data. *Cephalgia*, 43(5), 03331024231169244.
- Viana, M., Khaliq, F., Zecca, C., Figuerola, M. D. L., Sances, G., Di Piero, V., Petolicchio, B., Alessiani, M., Geppetti, P., & Lupi, C. (2020). Poor patient awareness and frequent misdiagnosis of migraine: findings from a large transcontinental cohort. *European journal of neurology*, 27(3), 536-541.
- Woldeamanuel, Y. W., & Cowan, R. P. (2022). Computerized migraine diagnostic tools: a systematic review. *Therapeutic Advances in Chronic Disease*, 13, 20406223211065235.
- Yang, Y., Peng, X., & Chen, Y. (2023). A case of migraine misdiagnosed as epilepsy. *Acta Epileptologica*, 5(1), 3.