University of Padova

Department of Mathematics "Tullio Levi-Civita"

Master Thesis in Computer Science

Predicting Migraine Episodes with Deep Learning - Enabling Patients to Anticipate Migraine Onset

Supervisor
Prof. Dr. Damiano Piovesan
University of Padova

Master Candidate
Syed Fahad Hassan

Co-supervisor
Dr. Marieke Vinkenoog & Ms. Nancy van Veelen
Leiden University & Leiden University Medical Center, Netherlands

Student ID
2044060

Academic Year
2023-2024

"This work is dedicated to my beloved parents, whose unwavering support, sacrifices, and love have been my foundation; to my dearest friends, for their companionship, encouragement, and countless moments of joy; and to my teachers, whose wisdom and guidance have illuminated the path of my academic journey."

Abstract

Migraine is a debilitating neurological condition that has shown a significant rise in prevalence globally, particularly among young to middle-aged females. Due to its profound impact on productivity, migraine has become a leading cause of missed deadlines and reduced efficiency, adversely affecting the education and work sectors. By 2024, numerous attempts have been made to treat and prevent migraines; however, its complex nature, particularly in terms of data variability and individual triggers, has limited the development of definitive solutions for prevention or cure.

This research aims to develop a migraine prevention model using artificial intelligence (AI) by leveraging deep learning techniques capable of processing high-dimensional data. By analyzing monthly and weekly e-diary data from patients, we sought to uncover hidden patterns and key triggers contributing to migraine episodes. SHAP (SHapley Additive exPlanations) analysis was employed to identify the most impactful features serving as predictors for migraines, providing crucial insights into the condition's triggers. Subsequently, we evaluated various deep learning models to determine their suitability for handling high-dimensional data. Among the models tested, BiLSTM demonstrated the best performance, achieving a recall of 0.6805, compared to the Transformer's recall of 0.5844.

Our findings underscore the significant potential of deep learning models, particularly BiLSTM, in aiding the medical field to analyze complex datasets. These models not only enable timely interventions but also pave the way for personalized healthcare solutions tailored to individual needs. This research highlights how AI-driven approaches can transform the understanding and management of complex chronic diseases like migraines, ultimately contributing to better patient outcomes.

Contents

Abstract v

List of figures x

List of tables xiii

Listing of acronyms xv

1 Introduction 1
- 1.1 Background . . . 1
  - 1.1.1 Impact on Quality of Life . . . 1
  - 1.1.2 Economic Impact and Loss in Productivity . . . 2
  - 1.1.3 Triggers and Challenges in Prediction . . . 3
  - 1.1.4 Machine Learning and Its Predictive Models . . . 4
- 1.2 Related Work . . . 5
  - 1.2.1 Evolution of Machine Learning in Healthcare . . . 5
  - 1.2.2 Deep Learning for Time-Series Data . . . 6
  - 1.2.3 Migraine Prediction Models . . . 6
- 1.3 Research Motivation . . . 8
- 1.4 Research Objectives . . . 8
- 1.5 Scope of the Study . . . 9
- 1.6 Structure of the Study . . . 9

2 Dataset 11
- 2.1 Data Sources and Ethics . . . 11
- 2.2 Population Selection Criteria . . . 12
- 2.3 Key Details of Dataset . . . 12
  - 2.3.1 Data Analysis . . . 13
- 2.4 Addressing Data Challenges: Imbalanced and Missing Data . . . 16
- 2.5 Data Preprocessing . . . 17
  - 2.5.1 Imbalanced Data . . . 17
  - 2.5.2 Missing Values . . . 17
  - 2.5.3 Feature Engineering . . . 18
- 2.6 Sequence Creation for Modeling . . . 19
  - 2.6.1 Sequence Creation Process . . . 19

3 Models 25
3.1 Models Overview . . . 25
3.1.1 Time-Series Forecasting with Deep Learning Models . . . 25
3.1.2 Bidirectional LSTM and Attention Mechanism Time-Series Prediction 26
3.1.3 Transformer Models for Time-Series Prediction . . . 26
3.2 Hyperparameter Optimization for Deep Learning Models . . . 27
3.3 Model Interpretability: The Role of SHAP Values . . . 28
3.4 Model Architectures . . . 28
3.4.1 BiLSTM Model . . . 29
3.4.2 Transformer Model . . . 32
3.4.3 Hybrid Approaches: BiLSTM + Attention and Use of Optimization Techniques . . . 36
3.5 Evaluation Metrics . . . 38
3.6 Experimental Setup . . . 39
3.6.1 Data Split . . . 39
3.6.2 Model Training Configurations . . . 40
3.7 Model's Misclassification . . . 40

4 Results and Analysis 41
4.1 Overall Performance . . . 41
4.1.1 BiLSTM Model . . . 42
4.2 Model Interpretability - Key Predictors for Migraine Episodes . . . 48
4.3 Error Analysis . . . 50
4.3.1 Insights from Misclassifications . . . 50
4.4 Model Performance in Healthcare Context . . . 50
4.4.1 The Importance of Recall in Healthcare . . . 50
4.4.2 Alignment with Real-World Needs . . . 51

5 Conclusion and Future Work 53
5.1 Conclusion . . . 53
5.2 Limitations . . . 54
5.2.1 Predicting Rare Events . . . 54
5.2.2 Generalizability and Variability . . . 54
5.2.3 Long Training Times . . . 54
5.3 Future Work . . . 54
5.3.1 Refining Validation Strategies . . . 54
5.3.2 Addressing Data Imbalance . . . 54
5.3.3 Integration of Multimodal Data . . . 55
5.3.4 Personalized Modeling . . . 55
5.4 Code Listings . . . 60

Code Listings 60

References 61

Acknowledgments 65

Listing of figures

2.1 Histogram of Data Completeness (Using averages) 15
2.2 Histogram of Data Completeness (Using total counts) 15
2.3 Histogram of Data Completeness (including All Patients) 16
2.4 Sequence Creation and Filtering process: it shows the creation of sequences from each patient records in the sliding window approach. Each sequence is 35 days, and the next sequence shifts one day. Sequences containing a migraine day in the last 48 hours are filtered out to get the final set of filtered sequences for model input. 19

3.1 Detailed architecture of a stacked Bidirectional LSTM (BiLSTM) network for temporal sequence modelling, illustrating forward and backward passes across layers with dropout and dense layers. 29
3.2 Our Implementation of BiLSTM model's Architecture 30
3.3 Transformer Architecture: Encoder-Decoder Framework. 32
3.4 Our Implementation of the Transformer Model's Architecture. 33
3.5 Integration of the Attention mechanism into the BiLSTM architecture. . . . 36

4.1 Training and Validation Loss and AUC-PR for BiLSTM 42
4.2 Training and Validation Loss and AUC-PR for BiLSTM with Attention Mechanism 43
4.3 Training and Validation Loss and AUC-PR for the Transformer Model . . . 44
4.4 ROC Curve and Precision-Recall Curve for all models. 46
4.5 Precision Comparison for each model. 46
4.6 Recall Comparisons for each model. 47
4.7 F1 Score Comparison for each model. 47
4.8 Key predictors identified for migraine occurrence using BiLSTM and Transformer models. 48

Listing of tables

2.1 Preliminary Analysis: Demographic Characteristics of Survey Respondents . 22
2.2 Preliminary Analysis: Preventive Medication Usage 23

3.1 BiLSTM Model Hyperparameters 31
3.2 Transformer's Configuration Space of Hyperparameters by Bayesian Optimization 35
3.3 Meaning of TP, FN, FP, TN. 38

4.1 Performance of BiLSTM 42
4.2 Performance of BiLSTM with Attention Mechanism 43
4.3 Performance of Transformer Model 44
4.4 Comparison of Models 45

Listing of acronyms

1 Introduction

1.1 Background

According to the Global Burden of Disease (GBD), around 1.04 billion people are affected by migraine worldwide, with an estimated prevalence of 14.7% in both genders [29, 30]. Migraine ranks as the third most common disease globally and is one of the leading causes of disability among people with primary headache disorders [29]. It is a primary headache disorder characterized by episodic attacks that can progress to chronic migraine [31], making it a complex neurological condition [32]. Among individuals aged 15–39, migraine is considered one of the most disabling neurological disorders [31, 30, 33]. Globally, it impacts millions of lives, causing recurring episodes of severe headaches often accompanied by symptoms such as vomiting, nausea, and sensitivity to light or sound [34, 35]. Migraine predominantly affects young to middle-aged women in a 3:1 ratio compared to men, often during their most productive years. This causes significant impacts on their personal and professional lives [29, 30, 34].

1.1.1 Impact on Quality of Life

Migraine's episodic nature poses significant challenges for patients, disrupting daily activities, impairing quality of life (QoL), and increasing migraine-related disability. This is often accompanied by psychiatric comorbidities such as anxiety and depression, especially in patients experiencing more frequent and prolonged episodes [29, 36]. Migraine is a paradoxical brain

disorder involving abnormal brain activity, affecting numerous bodily functions [32].

The burden of migraine extends beyond physical pain to significantly impact emotional and social well-being [37, 30]. Studies using tools such as the Migraine Disability Assessment Scale (MIDAS) and the Migraine-Specific Quality of Life (MSQoL) questionnaire indicate that patients with chronic migraines (15 or more headache days per month) experience substantial impairments in daily activities, work productivity, and interpersonal relationships [29]. These impacts are particularly detrimental during peak productive years, negatively affecting career prospects, financial stability, and social connections [37]. Additionally, psychiatric conditions like anxiety and depression are prevalent among migraine sufferers, compounding the overall disease burden [29, 38].

Despite advancements in the research of migraine and its high prevalence, predicting and preventing episodes remains challenging, and to be underdiagnosed and undertreated, with clinicians often failing to acknowledge the psychological and functional impact that led to shortcomings in a comprehensive, multifaceted approach [29, 30]. In this wide perspective of migraine impact on the quality of life, predictive models could play a valuable role in mitigating these impacts. By forecasting these episodes, such approaches could empower patients with preventive strategies that enable more effective management of symptoms and mitigate the emotional and social challenges associated with migraine.

1.1.2 Economic Impact and Loss in Productivity

Migraine is highly disabling not only on an individual level but also from a societal perspective, largely due to its significant impact on workplace productivity. It imposes a substantial economic burden on global economies. For instance, in the United States, migraine-related presenteeism (patients at work but with reduced effectiveness due to migraines) accounts for 89% of productivity losses, amounting to an estimated $19.6 billion annually [30]. Similarly, in Europe, around €27 billion per year is spent as a result of migraine-related absenteeism and presenteeism [30].

During a migraine attack, patients are reported to be adequately functional for less than 50% of the time, often experiencing severe pain, cognitive impairments, and increased sensitivity to light (photophobia) and sound (phonophobia) [30]. These productivity losses not only affect the individuals suffering from migraines but also employers and the overall economy [37].

The economic impact of migraine further underscores the value of predictive modeling. Accurate predictions of migraine episodes could enable preventive interventions and optimized

treatment schedules, reducing absenteeism and presenteeism. By mitigating productivity losses, predictive models could improve both patient well-being and economic stability.

1.1.3 Triggers and Challenges in Prediction

Migraine attacks are influenced by a variety of factors, or "triggers," which increase the probability of an attack over a relatively brief period and are of particular interest if they are avoidable or modifiable, creating a foundation for targeted management. Premonitory features, such as fatigue, change in mood or stiff neck, are initial manifestations that generally precede the onset of pain and can serve as early indicators in prediction models [36, 39]. Common triggers include sleep disruptions, menstruation, psychological stress, and environmental conditions like weather [36]. The response to such triggers is highly variable among patients, with some patients being sensitive to stress or weather changes while others may not respond to these factors in the same way [40, 39]. This variability complicates prediction, as models have to account for both common triggers and individualised responses. The wide range of triggers and individual sensitivities adds many features to the dataset, resulting in a high-dimensional and complex time-series structure. This increased dimensionality requires robust models capable of learning from a vast array of features and identifying those most relevant for the prediction of migraine onset. Building a model trained on a large dataset that consists of data from diverse patients' group would allow the identification of generalized patterns that may apply to broad individuals. For instance, one study using Recurrent Neural Networks (RNNs) has shown that weather conditions, like temperature and barometric pressure changes, were found to be major triggers for weather-sensitive patients, though with individual sensitivities varying a lot [40]. Identifying such broader patterns helps in the capture of the most influential factors associated with migraine onset in a broader patient population.

Self-reported data on migraine triggers, such as perceived stress, caffeine, and alcohol consumption, add further complexity due to potential recall biases and subjective interpretation. One of the studies using electronic diaries has shown that patients' beliefs about their triggers may not correspond to their observed patterns and thus it is desirable to validate self-reported data against objective measures where possible [36]. Additionally, in one more recent study, a multivariable prediction model has been implemented that incorporates self-reported and physiological data, which has shown modest success and points towards the necessity of more detailed data and advanced modelling techniques to capture patterns applicable across a broader population [36]

Given these challenges, effective models for migraine prediction would need to handle individual variability and high-dimensional data while combining different sources of subjective and objective data to improve their accuracy.

1.1.4 Machine Learning and Its Predictive Models

Given the complexity and burden of migraine, machine learning (ML) and deep learning (DL) have emerged as promising tools to predict and manage this condition in a more optimized manner. The recent ubiquity and advancements in ML and DL have created new opportunities in healthcare for improving the prediction and management of conditions like migraine [40]. By analyzing large high-dimensional datasets such as time-series data from electronic health records (EHRs) and patient self-reported diaries, predictive models can learn patterns that contribute to the onset of migraine attacks [41]. These models could facilitate early intervention to reduce the intensity and number of migraine episodes, improve patient care, and tailor treatment programs [42].

ML and DL models have been effectively used in different healthcare domains to predict health outcomes using complex datasets. For instance, in haematology, Vinkenoog et al. (2022) developed a support vector machine (SVM) model using SHAP values to predict haemoglobin deferral in blood donors. This model not only provided accurate predictions but also offered insights into feature importance, addressing issues related to donor deferrals that impact both donor retention and blood supply. Using explainable insights, the model suggested that prediction-guided donor invitations could potentially decrease the deferral rate by 60%, thereby enhancing blood collection while supporting donor motivation and health. Such explainable models show how ML can support clinical decision-making by providing actionable, transparent insights into complex physiological phenomena [43].

In oncology, ML and DL techniques have achieved phenomenal success in advancing diagnostic tools. For example, [44] demonstrated breast cancer diagnosis using ML models such as logistic regression, decision tree classifiers, and ensemble methods that include SHAP values for feature selection. Their models reached the highest accuracy of 99.82%, which was achieved by combining datasets and identifying impactful features through SHAP. This illustrates how such models can improve diagnostic accuracy and optimize feature selection to enhance methodologies used in breast cancer detection.

Focusing on migraine, ML and DL hold promise due to the complex and individualized triggers associated with the condition. For instance, [38] explored the use of ML in forecasting

migraine episodes by leveraging self-reported data and physiological measures collected using a mobile phone diary and wearable device. Random Forest classifiers were used to predict migraine episodes based on headache diary entries and physiological measures, including heart rate, skin temperature, and muscle tension. The model demonstrated moderate predictive performance, with an AUC of 0.62 on the test set, based on a sample of 18 migraine patients. This study highlights the potential integration of subjective and objective data for advancing migraine prediction and underscores the complexity of predicting migraine episodes due to the variability of triggers and symptoms across individuals [38].

Despite significant advancements in understanding migraine's pathophysiology and treatment options, predicting and managing migraine attacks remains a challenging task, particularly due to varied triggers and diversity among patients [30]. DL models offer promising solutions for discovering complex individual patterns and triggers for forecasting migraine episodes, enabling preventive measures and personalized care plans. However, further research is needed to enhance these models by improving predictive accuracy, handling complex time-series data, and making models adaptable to a wider range of patient-specific triggers. Such advancements are especially important for patients who may not exhibit predictable responses to common external factors, such as weather. These enhancements could significantly improve the clinical utility of migraine prediction models and further personalized patient care.

1.2 Related Work

1.2.1 Evolution of Machine Learning in Healthcare

The Healthcare ecosystem is widely adopting the new digital trends where all healthcare data is increasingly being recorded within electronic health records (EHRs) [41], however, medical practitioners are increasingly overwhelmed with massive amounts of recorded patient data, especially when they have relatively limited access to time, tools, and experience with this data daily [41], as a result, the application of machine learning in healthcare, particularly for time-series prediction, has significantly evolved in recent years [45]. Despite the quick success of traditional ML in healthcare, it faces challenges when dealing with the high-dimensional and temporal nature of healthcare data [41].

1.2.2 Deep Learning for Time-Series Data

Deep learning methods, especially those designed to handle sequential data such as BiLSTM and Transformer networks, outperform ML models and are proving to be more effective at capturing the temporal patterns in patient data [41, 45, 46, 47, 48]. These enhanced neural network techniques are capable of learning useful representations of key factors, such as esoteric medical concepts and their interactions, from high-dimensional to raw or minimally processed data [41, 45]. No longer dependent on experts to specify which manually crafted features to use, these end-to-end neural network learners can model data with rich temporal patterns and encode high-quality feature representations as nonlinear combinations of network parameters [41]. As a result, the recent popularity of DL methods has led to an increased number of associated publications in the healthcare domain. For instance, several studies have reviewed DL potential from different perspectives:

In the study [49], they applied DL-driven, transformer-based models to predict influenza prevalence using time-series data, which demonstrates the capability of the model to capture complex temporal dependencies through self-attention. This approach achieved state-of-the-art accuracy in forecasting influenza-like illness (ILI) prevalence highlighting the strength of DL to detect intricate disease prevalence patterns and enable timely public health interventions.

In another study [50], the author highlighted the transformative potential of AI in personalized medicine by developing deep neural networks that leverage patient-specific data to enhance predictions for disease outcomes. They also used model interpretability techniques to provide personalized, actionable insights, highlighting the importance of interpretability in clinical settings for informed decision-making [50]

Moreover, in the study [51], the authors use the EHR data and introduce the Bi-Attention model, which combines BiLSTM, a type of DL model, with an attention mechanism to improve disease prediction. It identifies and emphasizes critical time points, effectively processing health care data where not every moment holds equal significance. Such dynamic feature prioritization is particularly relevant for conditions like migraine, where trigger significance may vary across patients [51].

1.2.3 Migraine Prediction Models

Multivariable Predictive Modeling of Migraine Days

Holsteen et al. (2020) developed a multivariable logistic regression model aimed at predicting episodic migraine days by analyzing daily exposures to common triggers, such as stress, sleep patterns, caffeine intake, and menstruation. Using a prospective daily-diary approach, the model achieved modest predictive accuracy, with a within-person C-statistic of 0.56. This limited accuracy suggests that traditional models may struggle to capture individual daily variability and the complex interactions of migraine triggers. The author highlighted the need for models that incorporate a broader array of physiological and behavioural data, potentially collected via passive methods like wearable devices, to improve prediction. Furthermore, they identified the benefit of including additional common triggers and premonitory symptoms. These findings suggest that more advanced modelling techniques, such as deep learning, might offer a way forward, given DL's strengths in handling high-dimensional and temporal data, to capture dynamic interactions and improve individual-level predictions [36].

Disease Prediction using EEG signals

Ko et al. (2024) proposed a machine learning-based time series algorithm to predict migraine onset, focusing on the analysis of EEG data as input signals for early migraine identification. Their work underscores the promise of machine learning in recognizing patterns within high-dimensional EEG data, yet it also exposes significant challenges. One such limitation is the lack of standardization in migraine-related datasets, which affects model accuracy and generalizability across diverse patient populations. Additionally, due to the inherent complexity of EEG signals, characterized by high dimensionality, non-stationarity, and low signal-to-noise ratio, precise and reliable migraine predictions remain challenging. The study suggests that overcoming these limitations may require integrating advanced deep learning techniques that can handle complex, temporal data, thus allowing for more accurate, individualized predictions across patients with varying triggers and symptoms. This aligns with our project's focus on utilizing deep learning architectures, such as BiLSTM and Transformer models, to address similar challenges in migraine prediction using time-series diary data. **Forecasting migraine with machine learning based on mobile phone diary and wearable data**

As previously mentioned, Stubberud et al. (2023) explored the use of ML in forecasting migraine episodes by integrating data from mobile diaries and wearable sensors. The model achieved an AUC of 0.62, which illustrates the potential integration of subjective and objective data for migraine prediction. And study faced limitations, including a small sample size,

which restricted generalizability, and modest predictive accuracy, highlighting the need for larger datasets and more sophisticated, high-dimensional modelling approaches to capture the complexity of migraine triggers and premonitory signs. The authors emphasized that future research should incorporate additional data types and advanced machine learning models to improve predictive reliability and individual-level forecasting accuracy.

Predicting Response to Preventive Medications

Chiang et al. (2024) developed ML models to predict individual responses to various migraine preventive medications, including CGRP monoclonal antibodies, using detailed patient and migraine-specific features. Their approach, incorporating deep neural networks with SHAP for interpretability, demonstrated that features like baseline headache days, migraine duration, and migraine intensity were key predictors. With a high AUC of 0.825 for CGRP mAbs, this model represents a shift toward precision medicine in migraine treatment, highlighting the ability of DL models to refine treatment approaches based on individual characteristics [52].

Despite these advancements, further research is required to refine DL models for migraine prediction, especially in addressing individual variability and diverse trigger factors.

1.3 Research Motivation

Traditional approaches to treating and managing migraines often involve a significant trial-and-error in which patients along with their healthcare providers explore different medicines, diets or lifestyle modifications to reduce the number and severity of attacks [52]. Unfortunately, this approach is inefficient as it provides minimal predictability on when a migraine episode may or may not happen. Predicting the onset of a migraine attack can lead to timely intervention which will reduce the impact of the attack and improve patient well-being by lowering the disease burden [36]. Our study employs time-series data analysis to design optimised deep learning models, which include Bidirectional Long Short-Term Memory (BiLSTM) and a Transformer network. Utilizing SHapley Additive exPlanations (SHAP) values, these models forecast future migraine days and identify key predictors of migraine episodes.

1.4 Research Objectives

The key objectives of this thesis project are:

- To develop a model utilizing cutting-edge deep learning techniques that can predict the risks involving migraine attacks within the next 24 hours based on user data from filled diaries.
- To identify the top impacting predictors of migraine episodes using SHAP values.
- To enable timely intervention based on predictive insights to manage and reduce the impact of migraines on patients.

1.5 Scope of the Study

The thesis project explores the prediction of migraine attacks based on time series sequential data. The data was sourced from daily and monthly e-diaries of patients in the LUMC. This project primarily investigates deep learning techniques including BiLSTM and Transformer models to evaluate their ability to capture temporal dependencies in long sequential data. We have also used SHAP values to understand the relative importance of each predictor (variables), which could make model outcomes interpretable and explainable for all stakeholders including healthcare practitioners as well as patient and patient advocacy groups, enabling them to identify potential interventions for patients at a high risk of experiencing migraine episodes [50].

1.6 Structure of the Study

The remaining part of the thesis is organized as follows:

- **Chapter 2:** "Dataset", in which we provide an overview of the dataset used, including pre-processing steps, and feature engineering.
- **Chapter 3:** "Models", where we discuss the development of our models, selection choices, architecture, and training-validation setup.
- **Chapter 4:** "Results," shares the performance evaluation of the models, their comparison with each other, and the findings of key predictors for migraine episodes.
- **Chapter 5:** "Conclusion and Future Work", here we summarize our findings and highlight the future research directions.

2 Dataset

In this chapter, we discuss an overview of the dataset, and its preprocessing steps including data cleaning (missing value handling, feature engineering, and handling data imbalance) and sequence creation.

2.1 Data Sources and Ethics

The primary dataset we used in our study is of LUMC (Leiden University Medical Centre) who followed survey studies for varying periods, which is derived from electronic diaries filled out by migraine patients over the day before, capturing daily self-reported migraine symptoms, triggers, and medication use. The dataset has a variety of features (columns) that contain the key information including data on dietary habits, sleep, stress, and exposure to potential environmental triggers, supplemented by additional information from patients' medical records where applicable, including the target variable (output) which is the occurrence of a migraine attack, a binary outcome (Yes/No). This time-series data is collected over periods ranging from 3 to 12 months which allows the tracking of migraine patterns and analysis of daily variations in migraine risk.

As ethical considerations are integral to this study, given the sensitive nature of personal health data, the data is anonymized to protect patient identities, and only aggregated results are reported. Additionally, compliance with the General Data Protection Regulation (GDPR) and institutional guidelines ensures that all patient data is handled with strict confidentiality.

2.2 Population Selection Criteria

Participants were chosen based on the following selection criteria.

Inclusion criteria:

- Fulfil ICHD-3 criteria for migraine (verified diagnosis)
- Patient must experience active migraine, defined for this study as at least 2 days per month
- Willing to participate for at least 3 months and a maximum of 12 months
- ≥18 years of age;

Exclusion criteria:

- Unable/unwilling to use the headache E-diary application on a daily basis
- Diagnosed with other (chronic) neurological diseases such as Parkinson's disease, epilepsy etc. that may interfere with the results of this study;
- Chronic migraine as defined by the ICHD-3;
- Severe depression and/or panic disorders and/or schizophrenia and/or psychiatric disorders;
- Inability to differentiate between migraine and other headaches;
- Comorbidity with Cluster Headache or other TACs.
- Medication Overuse Headache as defined by the ICHD-3

2.3 Key Details of Dataset

This section includes the following details about the dataset:

- **Variables/features:** 153
- **Observations (days):** 26,855
- **Patients:** 206

- **Content:** Patient health conditions, medications, migraine triggers, daily inputs.
- **Key features include:**
  - **Symptoms:** Headache intensity, nausea, vomiting, photophobia, phonophobia, and aura.
  - **Triggers:** Stress, sleep quality, environmental conditions, and diet.
  - **Medication:** Whether or not medication was taken on a given day.
  - **Temporal Features:** Day of the week, time since the last migraine episode, etc.
- **Target variable:**
  - **MIGRDAGJN:** Binary label (1 for migraine day, 0 for non-migraine day) based on the occurrence of a migraine attack.

As mentioned, the dataset is based on daily and monthly e-diaries, which was a survey that each patient must fill out, Table 2.1, shows the details including the demographic characteristics of the respondents (patients) of the survey. Some of the respondents were found to be using some preventive medication for migraine, which medications are mentioned in Table 2.2.

2.3.1 Data Analysis

Given that dataset completeness is critical to our study, we evaluated the 'valid' (non-missing) and 'missing' observations for each patient to assess data availability and identify patterns of missing information. Our analysis employed two approaches: calculating the average number of valid and missing observations per feature and determining the total counts of valid and missing observations across all features. These methods provided insights into the variability in data availability and the completeness of each patient's dataset.

Averages

We calculated the average number of valid and missing entries per feature by dividing the total valid and missing values by the number of features (excluding RECORDID) in each patient record. This approach provides a normalized view, helping to compare patients with different observation lengths.

Formula:

For a given patient P:

$$\text{Average Valid Entries per Feature} = \frac{\text{Total Valid Entries for All Features (Excluding RECORDID)}}{\text{Number of Features (Excluding RECORDID)}}$$

$$\text{Average Missing Entries per Feature} = \frac{\text{Total Missing Entries for All Features (Excluding RECORDID)}}{\text{Number of Features (Excluding RECORDID)}}$$

We found that the patient with ID `'L118553'` has the most observations, 328 in total, among which the valid observations are 205.63 and missing observations are 122.38. While the patient with ID `'L120545'` has the least observations, 1 in total with 0.47 valid and 0.53 missing observations, which highlights that, on a per-feature basis, data completeness varies across patients, with some having consistently high levels of valid entries, while others have significant gaps.

Total Counts

In addition to averages, for each patient, we calculated the total number of valid and missing observations across all features, which provides a broader view of each patient's data, representing the full extent of valid and missing values.

Formula:

$$\text{Total Valid Observations} = \sum_{i=0}^{n} \text{Valid Entries in Feature}_i$$

$$\text{Total Missing Observations} = \sum_{i=0}^{n} \text{Missing Entries in Feature}_i$$

We observed similar results to those obtained through the average calculation: the patient with ID `'L118553'` had the highest number of observations, with 31,255 valid and 18,601 missing entries, while the patient with ID `'L120545'` had the fewest, with 71 valid and 81 missing values. This approach provides a comprehensive view of data availability per patient, which is crucial for understanding the potential impact of missing data on model performance.

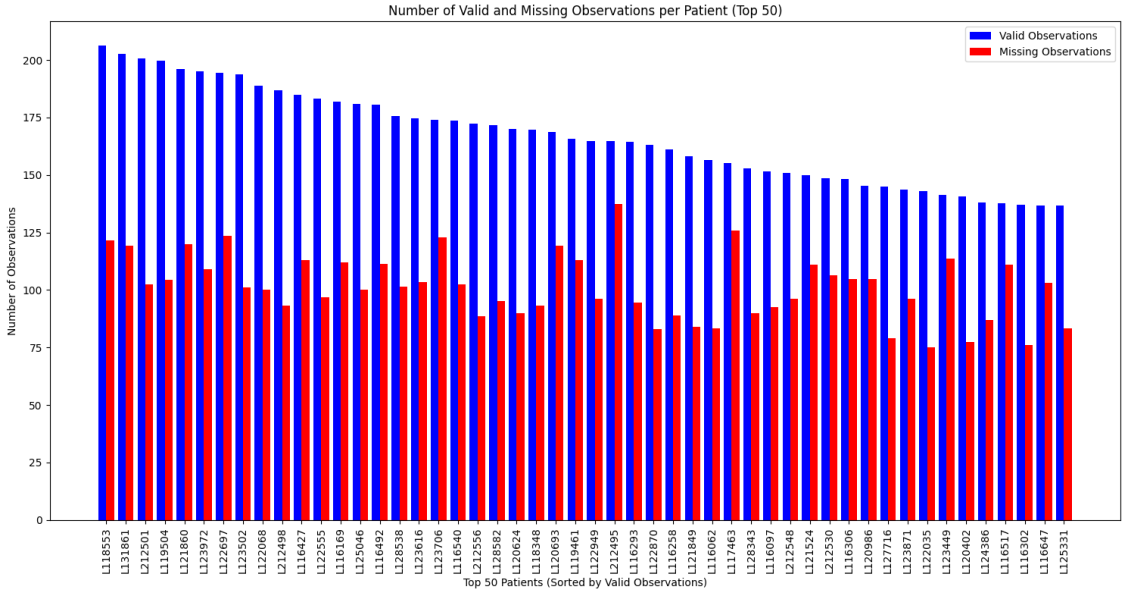


Figure 2.1: Histogram of Data Completeness (Using averages)

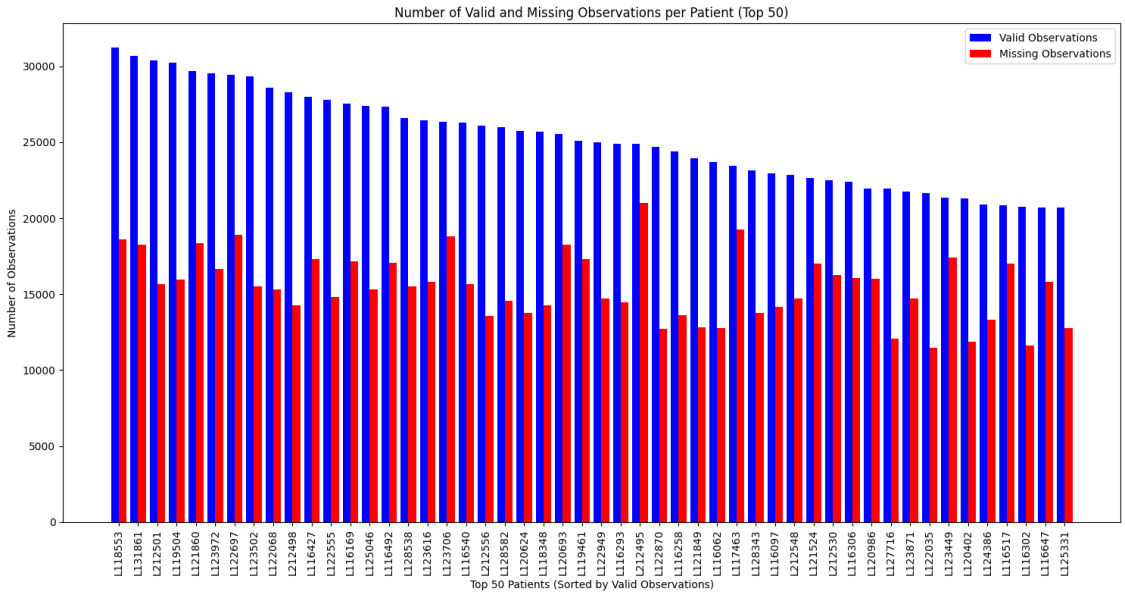


Figure 2.2: Histogram of Data Completeness (Using total counts)

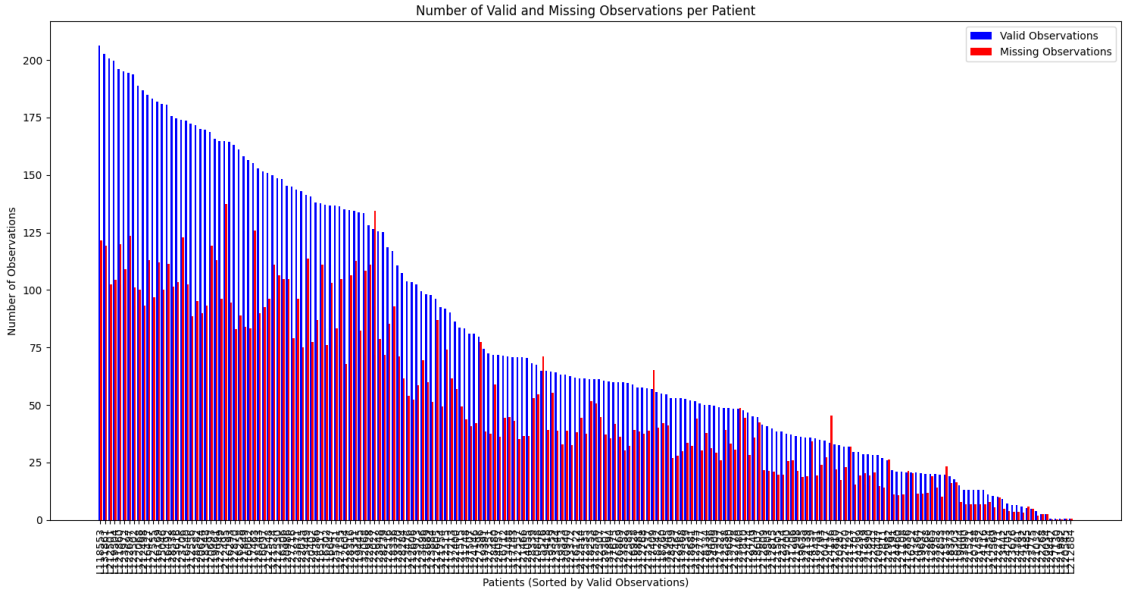


Figure 2.3: Histogram of Data Completeness (including All Patients)

All of the above histograms; Figures 2.1, 2.2, and 2.3 visualize the distribution of valid and missing observations across patients, showing the variability in data completeness, which helps to quickly assess how frequently patients have complete or partial entries, informing the strategies for handling missing data during preprocessing.

2.4 Addressing Data Challenges: Imbalanced and Missing Data

In the healthcare sector, datasets are often imbalanced in nature, and patients with diseases are less common than healthy patients [5, 18]. An imbalanced dataset refers to a classification of a dataset where the number of instances of a given class is much lower than for other classes [18]. The imbalanced data poses significant challenges which affect the performance and accuracy of the models. In particular, it is relevant in migraine prediction as well, where the number of non-migraine days is typically far more in number as compared to migraine days, which creates a class imbalance in the dataset ultimately making it a big challenge to deal with.

The conventional machine learning models favour the majority class due to their operating nature, which leads to poor performance in the prediction of minority class events, such as migraine days in our case. To address this concern, different approaches have been developed

and are in use. Like cost-sensitive learning that adjusts the loss function so that it gives more importance to the minority class and controls the model from being too biased towards just the majority class. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) and other resampling techniques have been applied in some studies [18] to balance the dataset. However, in this thesis, we addressed the imbalance problem mainly through class weighting which ensures that our models give main attention to the minority class (migraine days), not to the majority class. This approach makes models focus on learning from the less frequent migraine events while optimizing performance metrics such as F1 score and AUC-PR to ensure balanced evaluation in both classes.

When using healthcare datasets, handling missing data is another critical aspect, as missing values greatly affect and degrade model performance. , In our dataset, we encountered two distinct types of missing data: (1) completely missing rows for days when patients did not fill in their diaries, and (2) partially filled records within completed diary entries, often due to questions asked conditionally based on previous responses. For example, if one day a patient didn't experience any migraine symptoms, further questions about symptom severity or explicit trigger information wouldn't have been answered, making entries on that particular day partial.

2.5 Data Preprocessing

Data preprocessing was essential to prepare the e-diary entries for model development, addressing challenges such as class imbalance, missing values, and high-dimensional feature representation.

2.5.1 Imbalanced Data

The dataset is notably imbalanced, with fewer migraine days compared to non-migraine days. This imbalance could bias predictive models toward overestimating the likelihood of non-migraine days. To address this, we used the technique of implementing class weighting in the model training process, where higher weights were assigned to the migraine days.

2.5.2 Missing Values

Missing data points are common in self-reported data. For days when the e-diary was not completed, we managed it by filling missing entries with a placeholder (zero) to maintain continuity. This choice was informed by the observation that missing values might reflect the absence of

reported migraine-related symptoms or triggers for that day. Some missing entries are imputed where applicable. For example, we adjusted the "acute medication" variables as follows,

if `'AANVMEDPREJN = 0'`, all the variables `'AANVMEDPRE1 – AANVMEDPRE99'` should return `'0'` (in not NA). And if `'AANVMEDPREJN = 1'`, at least one of the `'AANVMEDPRE1 – AANVMEDPRE99'` should return `'1'`, and the rest of the `'AANVMEDPRE1 – AANVMEDPRE99'` should return `'0'` instead of NA.

Similarly, for the variables "Prophylactic medication," we implemented the following logic:

When the variable `'PROFYLAXEVERANDERD = 0'`, then the NA for all of the dependent variables is `'0'`. If `'PROFYLAXEVERANDERD = 0'`, and one of its dependent variables has a `'1'`, the rest of the NA can be set to `'0'`.

2.5.3 Feature Engineering

Feature engineering was essential to maximize the predictive capabilities of our models. The main steps included creating lag features, rolling averages, and patient-level sequences.

Lag and Rolling Features

Lagged and rolling window features were created for each numerical variable in the dataset to capture temporal trends and patterns. Specifically: Lag features: For each feature, a one-day lag feature was added to capture the previous day's data, enabling the model to recognize potential short-term dependencies.

$$\text{Lag}_k(t) = x_{t-k}$$

Rolling averages: A 7-day rolling mean was calculated for each feature, allowing the model to learn from recent trends and fluctuations over a week.

$$\text{RollingMean}_k(t) = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}$$

These features were concatenated with the main dataset and rows with missing values resulting from these operations were removed to avoid any interruptions in sequential data.

2.6 Sequence Creation for Modeling

As the dataset was large and each patient had a long series of data, so, we structured the prediction task as a time-series classification problem where we trained our model on sequences of data to predict whether a migraine attack will occur within the next 24 hours. Here sequences mean, for each patient creating a list of days that can contribute to the occurrence or non-occurrence of migraine episodes.

2.6.1 Sequence Creation Process

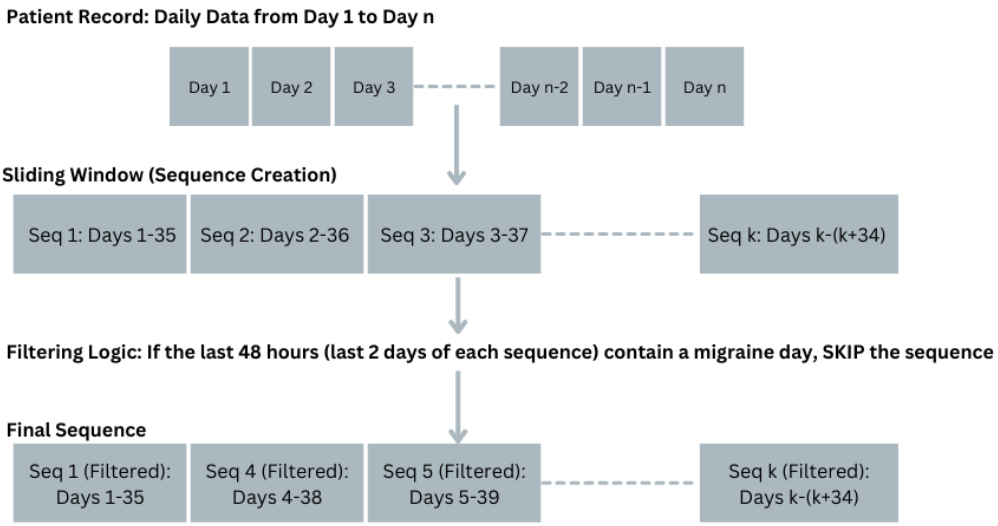


Figure 2.4: Sequence Creation and Filtering process: it shows the creation of sequences from each patient records in the sliding window approach. Each sequence is 35 days, and the next sequence shifts one day. Sequences containing a migraine day in the last 48 hours are filtered out to get the final set of filtered sequences for model input.

Sliding Window Approach

We created sequences of 35 consecutive days for each patient. The input to the model is a sequence of features from the first 35 days, and the output is whether a migraine occurred on the 36th day. Another input from day 2-36, predicts day 37, and it continues till the last entry.

Exclusion criteria: we set exclusion criteria for the sequences to avoid the model's focus on ongoing migraines, the criteria checks if a migraine day is present within the last 2 days `'(days`

`34-35 of the sequence)'`, if yes we excluded that sequence to keep the focus of the model on predicting new episodes rather than persistent attacks which also helps to avoid possible bias by ongoing migraines that are already being treated.

Sequence Creation Formula

Input Data:

- $X_{i,j}$: Feature value for patient i on day j, where $i \in \{1, 2, \ldots, N\}$ (number of patients).
- $y_{i,j}$: Target value (migraine day) for patient i on day j.
- S: Sequence length (here, $S = 35$).

Sequence Construction

$$\text{Seq}_{i,k} = \{X_{i,k}, X_{i,k+1}, \ldots, X_{i,k+S-1}\}, \quad \text{where } k \in \{1, 2, \ldots, T_i - S\}$$

Where:

- T_i: Total number of observations for patient i.

Target: For each sequence:

$$y_{i,k} = y_{i,k+S}$$

Sequence Filtering: Excludes sequence if:

$$\exists j \in \{k + S - 2, k + S - 1\} : y_{i,j} = 1$$

It ensures sequences ending with a migraine day in the last 48 hours are skipped.

Handling Temporal Data:

For handling temporality, we grouped our dataset by each patient ID (i.e., 'RECORDID') to ensure that temporal continuity is maintained within sequences. we segmented each patient's data into multiple overlapping sequences which captures both the short-term and long-term patterns leading up to a migraine attack.

Feature Selection and Standardization

Only numeric features were used for model input, and they were normalized to ensure consistency. Using only the training set, each numeric feature was standardized using a 'StandardScaler' to ensure that each feature has a mean of 0 and a standard deviation of 1, and this scaling was then applied to the test data to avoid information leakage. The StandardScaler computes the mean and standard deviation of each feature on the training set, then transforms each value as:

$$z = \frac{x - \mu}{\sigma}$$

Where:

- x: The original value of the feature.
- μ: The mean of the feature.
- σ: The standard deviation of the feature.

This Standardization was used to avoid model prioritizing features with larger magnitudes as features are in different scales (e.g., age in years vs. caffeine intake in units per day) The finalized dataset of sequences was then ready for input into our models, including both BiLSTM and Transformer models.

Table 2.1: Preliminary Analysis: Demographic Characteristics of Survey Respondents

Category	Total	MO	MA
Total, n	206	111	95
Sex (female), n (%)	174 (84.5)	93 (83.8)	81 (85.3)
Treated by general practitioner, n (%)	91 (44.2)	49 (44.1)	42 (44.2)
Treated by a neurologist, n (%)	91 (44.2)	50 (45.0)	41 (43.2)
Migraine Diagnosis			
Migraine with aura (MA), n (%)	37 (18.0)	-	37 (38.9)
Migraine without aura (MO), n (%)	72 (34.9)	72 (100)	-
MA & MO, n (%)	39 (18.9)	-	39 (41.1)
Age (years), median [IQR]	47.52 [47.37, 47.94]	47.52 [47.37, 47.94]	47.52 [47.37, 47.94]
Length (cm), mean (SD)	173.04 (8.09)	173.04 (8.09)	173.04 (8.09)
Weight (kg), mean (SD)	74.56 (16.12)	74.56 (16.12)	74.56 (16.12)
Birth Control			
Birth control, n (%)	37 (21.5)	19 (17.1)	18 (18.9)
Combination birth control, n (%)	14 (37.8)	7 (36.8)	7 (38.9)
Progesterone birth control, n (%)	19 (51.4)	10 (52.6)	9 (50.0)
Other, n (%)	4 (10.8)	2 (10.5)	2 (11.1)
Menstrual Cycle			
Menopausal, n (%)	75 (43.6)	42 (37.8)	33 (34.7)
Regular (21-35 days), n (%)	50 (82.0)	25 (22.5)	25 (26.3)
Irregular (<21 days >35 days), n (%)	8 (13.1)	-	4 (3.6)
Pregnant, n (%)	1 (0.5)	1 (0.9)	0 (0)
Breastfeeding, n (%)	1 (0.5)	1 (0.9)	0 (0)
Alcohol units/week, median [IQR]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]
Smoking, n (%)	7 (3.4)	3 (2.7)	4 (4.2)
Caffeine, units/day, median [IQR]	5.00 [3.00, 6.00]	5.00 [3.00, 6.00]	5.00 [3.00, 6.00]
OTC medication, days/month, median [IQR]	4.00 [2.00, 7.00]	4.00 [2.00, 7.00]	4.00 [2.00, 7.00]
Triptans, days/month, median [IQR]	4.00 [3.00, 6.00]	4.00 [3.00, 6.00]	4.00 [3.00, 6.00]
Acute medication days/month, median [IQR]	6.00 [4.00, 8.00]	6.00 [4.00, 8.00]	6.00 [4.00, 8.00]
Depression			
HADS total score, mean (SD)	8.30 (5.06)	8.30 (5.06)	8.30 (5.06)
HADS, anxiety score, median [IQR]	5.09 (2.98)	5.09 (2.98)	5.09 (2.98)
HADS, depression score, median [IQR]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]
CES-D, total score, median [IQR]	9.00 [4.00, 14.00]	9.00 [4.00, 14.00]	9.00 [4.00, 14.00]
Lifetime depression, n (%)	38 (100.0)	38 (100.0)	38 (100.0)

Table 2.2: Preliminary Analysis: Preventive Medication Usage

Current Preventive Medication	n (%)
Propranolol	71 (34.5)
Metoprolol	53 (25.7)
Sodium valproate	42 (20.4)
Flunarizine	5 (2.4)
Topiramate	49 (23.8)
Pizotifen	15 (7.3)
Candasartan	49 (23.8)
Lamotrigine	2 (1.0)
Acetazolamide	0 (0.0)
Verapamil	2 (1.0)
Amitriptyline	22 (10.7)
Mitrazapine	2 (1.0)
Botulinum	12 (5.8)
Galcanezumab	3 (1.5)
Fremanezumab	3 (1.5)
Other	15 (7.3)

3 Models

In this chapter, we discuss the development, training, and evaluation of our models in order to predict migraine episodes. It starts with an overview of the models developed, their selection criteria, architectures, evaluation metrics, and the experimental setup that we used to meet the research objectives.

3.1 Models Overview

3.1.1 Time-Series Forecasting with Deep Learning Models

To predict sequential events, the time-series forecasting models come with an importance, for instance, in events like migraine attacks where information from past patterns can offer insights for analysing future occurrences. Deep learning techniques, in particular BiLSTM and Transformer models, have shown commendable performance in this area. The author in [41] presented a comprehensive performance review of deep learning techniques used on healthcare time series data. They highlighted how deep learning methods have outperformed traditional machine learning models by specifically addressing the high-dimensional and temporal challenges of medical data. In migraine prediction, the deep learning models' ability to analyze both temporal and non-temporal factors such as patient behaviour, environmental triggers, and physiological data, have shown promise. BiLSTM networks have been widely used in the healthcare sector as they are able to capture both forward and backward sequences, which

makes them suitable for modelling complex temporal patterns. In [51] and [47], it is explained how BiLSTM can combined with attention mechanisms to significantly improve the accuracy of migraine prediction models, which has been achieved by more effectively learning the dependencies in time-series data that makes the model focus on the most crucial time steps and features.

3.1.2 Bidirectional LSTM and Attention Mechanism Time-Series Prediction

Bidirectional LSTM (BiLSTM) extends the unidirectional LSTM network by introducing a second layer in which the connections of the hidden layer to the hidden layer flow in backward timing [22]. It has emerged as a powerful model for time-series analysis as it is able to capture long-term dependencies in both forward and backward directions to preserve future and past information. In healthcare applications, where the past and future information about the events is important for predictions, the BiLSTM can be particularly useful [51, 32]. For instance, in migraine prediction, BiLSTM has been applied to sequential data, such as EEG signals data, for the early detection of migraine disease [25]. The attention mechanism in deep learning simulates the attention model of the human brain, it enhances the performance of BiLSTM by allowing the model to select more critical features from many features to achieve better prediction results, thereby improving interpretability and performance [22, 27]. Attention-based models have demonstrated their ability to weigh different input features dynamically which makes them particularly suitable for healthcare data, where not all time points carry equal importance [22]. By integrating attention mechanisms, BiLSTM models can prioritize critical moments leading up to a migraine attack, enabling more accurate predictions.

3.1.3 Transformer Models for Time-Series Prediction

The Transformer model emerged with the development of natural language processing (NLP), and was adapted for time-series forecasting due to its ability to handle long-range dependencies efficiently [21]. Transformer architecture leveraged the mechanism of self-attention to capture long-range dependencies in sequential data, which makes it well-suited for applications like the healthcare sector, where events may have high-dimensional and temporal dependencies [11]. Self-attention leverages mechanisms that assign different levels of importance to different parts of a sequence, which allows the model to capture complex dependencies in data effectively. This approach has been particularly beneficial in time series forecasting, as it enables the

model to process entire sequences simultaneously and identify both short- and long-range patterns, which is crucial for tasks with intricate temporal dynamics. Transformer models in some applications have outperformed the traditional RNN-based models, in terms of accuracy and scalability when combined with sophisticated optimization techniques such as Bayesian Optimization [27].

Transformer models are extensible to univariate and multivariate time-series data, as also demonstrated in a study [11], where the author uses the Transformer model for time-series forecasting by applying it to influenza-like illness (ILI) data, which employs self-attention mechanisms to capture complex dependencies within time series data, showcasing the model's capacity to handle both univariate and multivariate series effectively, particularly in health-related prediction tasks. As the transformer architecture benefits from self-attention, the model can efficiently learn complex temporal patterns from patient data, enabling more accurate predictions of migraine episodes.

3.2 Hyperparameter Optimization for Deep Learning Models

Hyperparameter tuning or hyperparameter optimization, is the process of determining which hyperparameters are optimal to utilize [26]. It is considered one of the greatest challenges in deploying DL models, as the model's performance is highly dependent on it [20]. The most crucial hyperparameters are learning rate, batch size, dropout rate, hidden units, number of hidden neurons, epochs, optimizer and activation functions. Manual hyperparameter tuning methods such as grid search are usually inefficient, in particular when using complex models like BiLSTM and Transformer networks. In [20], the author explores the use of Particle Swarm Optimization (PSO) as an alternative to traditional manual parameter setting and grid search approaches, while showing the potential of PSO in significantly reducing the time required to optimize hyperparameters that improve the model performance. In the context of healthcare applications like migraine prediction, where datasets are often large and heterogeneous, such optimization techniques are crucial for developing more efficient and accurate models, thereby overcoming the drawback of the trial-and-error approach. On the other hand, we have Bayesian optimization (BO), which is another technique that has been successfully applied to hyperparameter tuning in deep learning models [26]. BO helps in efficiently searching the hyperparameter space by balancing exploration and exploitation, thereby reducing the time re-

quired for model training. In our study, we have employed both PSO and BO to fine-tune the hyperparameters of BiLSTM model, respectively, to improve their performance on migraine prediction tasks and to do a comparison of both techniques. For Transformer, we have just used BO to fine-tune the hyperparameters as PSO is highly computationally expensive and as this project is of time-limited, both approaches cannot be used.

3.3 Model Interpretability: The Role of SHAP Values

Interpretability is one of the important aspects of deploying AI models in healthcare. While DL models are highly accurate, they are most difficult to interpret and often viewed as "black boxes" due to their complexity. The SHapley Additive exPlanations (SHAP) values have been designed specifically to tackle this perception. It offers a solution by making models explainable and interpretable, providing insights into the contribution of each variable to the prediction for each observation [16]. In the context of migraine prediction, SHAP values can help identify which factors (such as stress, sleep patterns or environmental triggers) have contributed the most to the risk of a migraine attack [15]. In our thesis project, we have used SHAP values to interpret the predictions made by our models, BiLSTM and Transformer. This has another advantage: it not only provides better explanations and transparency of the model's predictions but also enables healthcare professionals to make informed decisions about potential interventions to prevent or mitigate migraine attacks.

3.4 Model Architectures

In this section, we explain the architecture of our models, including Bidirectional LSTM, Transformer networks, and hybrid approaches used in this project.

3.4.1 BiLSTM Model

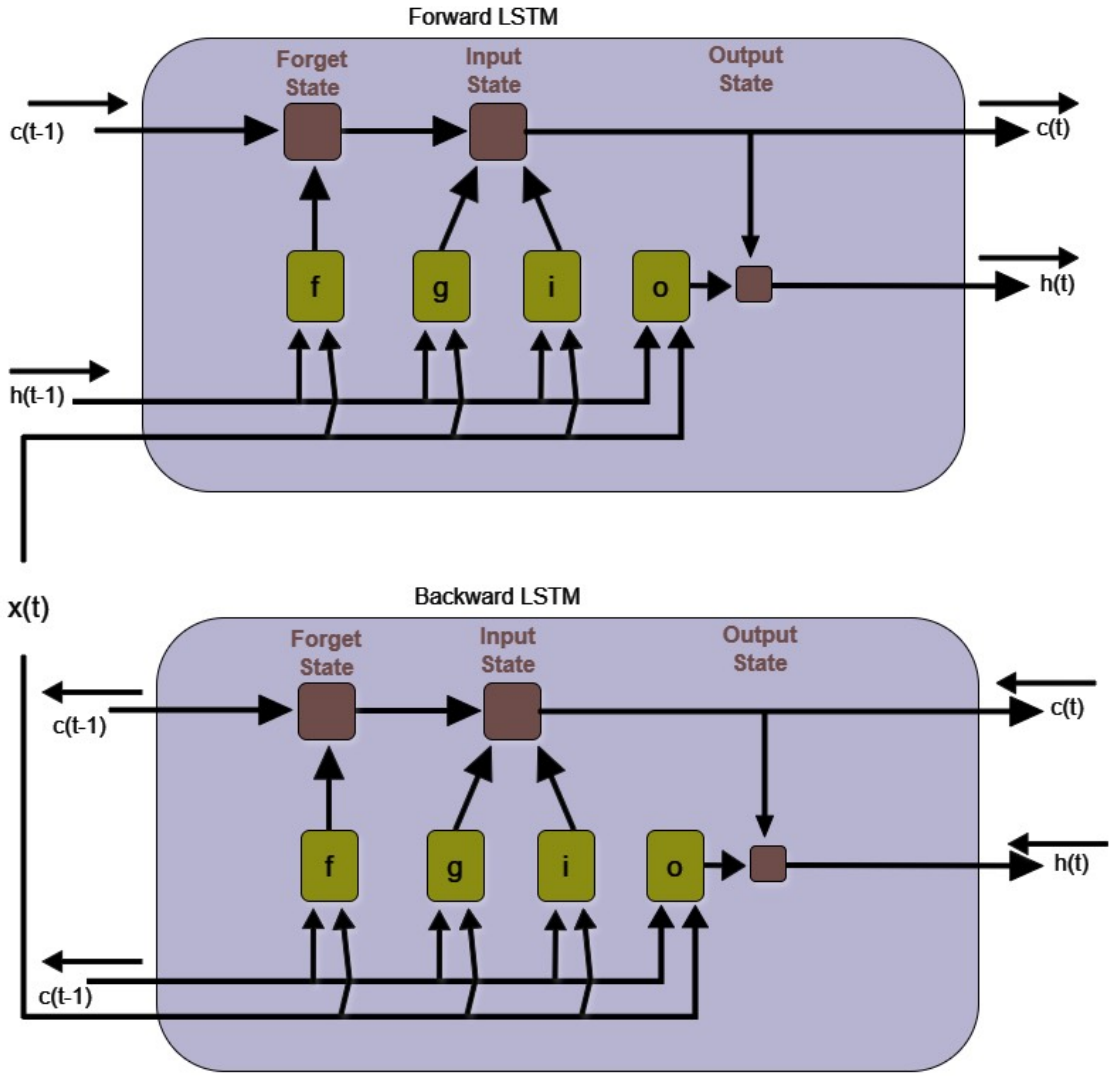


Figure 3.1: Detailed architecture of a stacked Bidirectional LSTM (BiLSTM) network for temporal sequence modelling, illustrating forward and backward passes across layers with dropout and dense layers.

Figure 3.2: Our Implementation of BiLSTM model's Architecture

As shown in the above Figure 3.2, we implemented our BiLSTM model consisting of 9 layers in total for sequence modelling. The input layer accepts patient diary sequences of 35 days (S = 35) with multiple numeric features (f) for each day. These inputs ($X \in \mathbb{R}^{T \times d}$) are passed

through each of three stacked BiLSTM layers to capture temporal dependencies both forward and backward:

$$h_t^{(l)} = \left[\overrightarrow{h}_t^{(l)} \; ; \; \overleftarrow{h}_t^{(l)} \right]$$

where l represents the layer index, $\overrightarrow{h}_t^{(l)}$ represents the forward hidden state, and $\overleftarrow{h}_t^{(l)}$ represents the backward hidden state for the t-th timestep.

Each BiLSTM layer includes 95 units (optimized through hyperparameter tuning) and outputs sequences to the next layer. A regularization is applied using dropout layers ($p = 0.2524$) after each BiLSTM layer to reduce overfitting. A dense layer with 128 units and L2 Regularization ($\lambda = 0.001$) uses ReLU activation, to refine the extracted features via L2-norm, defined as:

$$\text{Loss}_{\text{regularization}} = \lambda \sum_i w_i^2$$

where w_i represents the weights, and λ is the regularization strength.

Finally, the output layer with a softmax activation function is used to predict the probability of a migraine day ($y = 1$) or a non-migraine day ($y = 0$):

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

where z_i is the logit for class i. In our implementation, the logits z are calculated as:

$$z = W_{\text{out}} \cdot o + b_{\text{out}}$$

The hyperparameters, as shown in Table 3.1, including the number of LSTM units, learning rate, and batch size, were fine-tuned using Bayesian Optimization.

Table 3.1: BiLSTM Model Hyperparameters

Hyperparameters	Value
`'lstm_units'`	95
`'dropout_rate'`	0.2524
`'learning_rate'` (η)	1.48×10^{-5}
`'batch_size'`	64

Motivation for choosing BiLSTM

We considered BiLSTM different from other RNN structures because it can flow input in both directions, which allows it to learn both forward and backward dependencies, which is important when dealing with time-series data [47]. Given the temporal nature of migraine prediction, capturing both past and future data points improves the model's capacity to detect complex patterns and trends over time. [compact]titlesec

3.4.2 Transformer Model

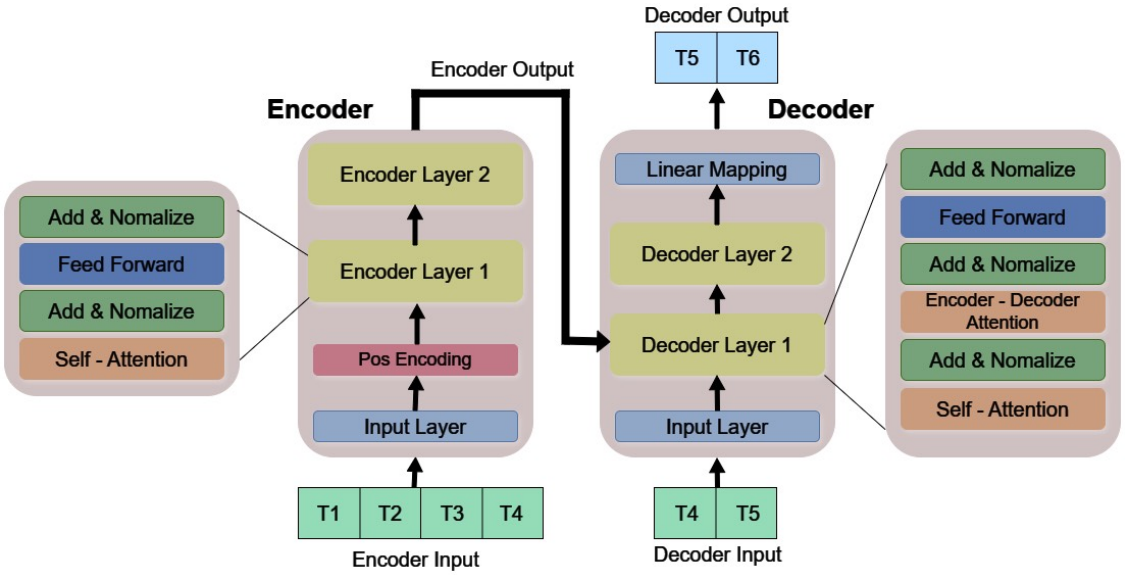


Figure 3.3: Transformer Architecture: Encoder-Decoder Framework.

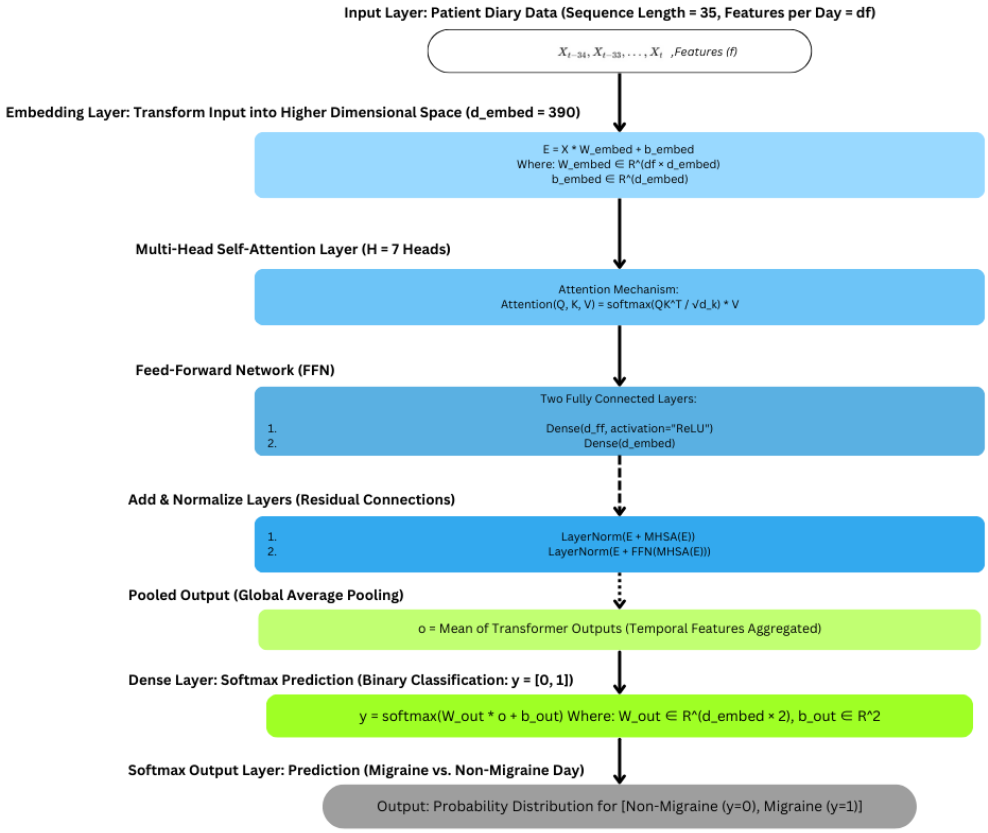


Figure 3.4: Our Implementation of the Transformer Model's Architecture.

We implemented our Transformer model consisting of 7 logical layers to capture temporal dependencies in multivariate time-series data. The input to the model is a sequence of 35 days, with df features for each day. Formally, the input sequence X can be represented as:

$X \in \mathbb{R}^{T \times d_f}$, where $T = 35$ is the sequence length, and d_f is the number of features per day.

The input is first transformed into a higher-dimensional space using an embedding layer, producing an embedding E:

$$E = XW_{\text{embed}} + b_{\text{embed}}$$

where $W_{\text{embed}} \in \mathbb{R}^{d_f \times d_{\text{embed}}}$ and $b_{\text{embed}} \in \mathbb{R}^{d_{\text{embed}}}$. Here d_{embed} is the embedding dimension, which is set to 390, allowing each day in the sequence to have a richer representation.

The core of the Transformer is the multi-head self-attention mechanism, which calculates attention for each head h as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where:

$$Q = EW_Q, \quad K = EW_K, \quad V = EW_V$$

$W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{embed}} \times d_k}$ are learnable weight matrices, and $d_k = \frac{d_{\text{embed}}}{H}$, where H is the number of attention heads.

The multiple attention heads here are used to focus on different parts of the sequence, capturing interactions between various migraine triggers and symptoms.

The output of all attention heads is concatenated and projected back to the embedding space:

$$\text{MHSA}(E) = \text{concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_H)W_O$$

where $W_O \in \mathbb{R}^{d_{\text{embed}} \times d_{\text{ff}}}$, $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{embed}}}$, and d_{ff} is the feed-forward dimension.

The final output of the Transformer is passed through a dense layer with a softmax activation to predict the probability of migraine or no migraine:

$$y = \text{softmax}(W_{\text{out}} \cdot o + b_{\text{out}})$$

where o is the pooled output of the model, $W_{\text{out}} \in \mathbb{R}^{d_{\text{embed}} \times 2}$, and $b_{\text{out}} \in \mathbb{R}^2$.

The categorical cross-entropy loss was used for optimization:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of training samples. Additionally, gradient clipping was applied to stabilize training, and a learning rate scheduler was used for optimization.

Hyperparameters for Transformer Model

The hyperparameters optimized using Bayesian Optimization are summarized in Table 3.2.

Table 3.2: Transformer's Configuration Space of Hyperparameters by Bayesian Optimization

Hyperparameter	**Value**
Number of Attention Heads (`num_heads`)	7
Dropout Rate (`dropout_rate`)	0.24627318773000064
Learning Rate (`learning_rate`)	0.0002652326485696 1323
Batch Size (`batch_size`)	32

Motivation for Choosing Transformer Networks

The Transformer model efficiently handles long-range dependencies, making it ideal for analyzing time-series data over extended periods (35 days). Its attention mechanism enables the model to focus on specific parts of the input, capturing the influence of symptoms or triggers leading up to a migraine.

3.4.3 Hybrid Approaches: BiLSTM + Attention and Use of Optimization Techniques

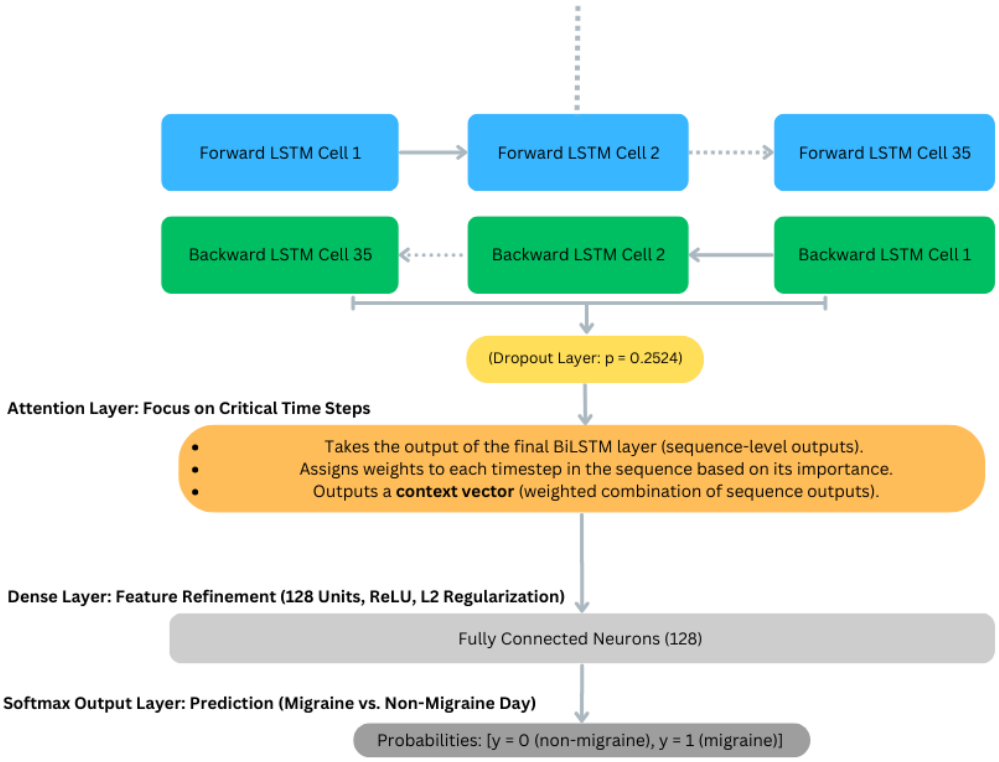


Figure 3.5: Integration of the Attention mechanism into the BiLSTM architecture.

Attention Layer

As shown in the Figure 3.5, the attention layer is added after the BiLSTM layers and before the Dense layer to focus on the most critical time steps such as key symptoms that precede a migraine, which enhances the ability of our model to assign different importance to each day in the sequence, boosts the interpretability and performance of the predictions.

Hyperparameter Optimization (PSO and BO)

PSO (Particle Swarm Optimization) was initially explored to optimize hyperparameters. It iteratively updates a population of particles, where each particle represents a potential solution

in the search space. The position and velocity of each particle are updated using the following equations:

Velocity Update:

$$v_i^{(t+1)} = \omega v_i^{(t)} + c_1 r_1 \left(p_i^{\text{best}} - x_i^{(t)} \right) + c_2 r_2 \left(g^{\text{best}} - x_i^{(t)} \right)$$

Position Update:

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)}$$

In the above equations:

- $v_i^{(t)}$: Velocity component of the i-th particle at time t.
- $x_i^{(t)}$: Position component of the i-th particle at time t.
- p_i^{best}: Personal best position of the i-th particle.
- g^{best}: Global best position among all particles.
- c_1, c_2: Acceleration coefficients representing the weights for approaching p_i^{best} and g^{best}.
- ω: Inertia coefficient that helps particles maintain momentum towards better positions.
- r_1, r_2: Uniform random values in the range $[0, 1]$.

BO (Bayesian Optimization) was eventually chosen due to its lower computational cost and effectiveness in fine-tuning models, which was critical given the project's time constraints.

BO models the objective function $f(x)$ using a probabilistic surrogate model, such as a Gaussian Process (GP). The optimization process involves:

- **Acquisition Function:** BO uses an acquisition function $\alpha(x)$ to determine the next point to evaluate x_{next}.
- **Gaussian Process Prior:** The surrogate model $f(x)$ is assumed to follow a GP prior.
- **Posterior Update:** After observing $f(x)$ at x_{next}, the posterior distribution is updated to refine the search for the optimal solution.

BO later replaced PSO, as it leveraged fewer evaluations by optimizing $\alpha(x)$, balancing exploration and exploitation of the hyperparameter search space while reducing computational cost.

3.5 Evaluation Metrics

The performance of the models is evaluated using a range of classification metrics, focusing on metrics relevant to healthcare, where minimizing false negatives (missed migraine attacks) is crucial [44, 53, 47, 52].

Precision: It is a measure of accuracy, representing the proportion of a positive example divided into positive examples [22]. In our case, it measures the proportion of true positives (correctly predicted migraine days) out of all predicted positives.

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

Where:

- TP: True Positives (correctly predicted migraine days)
- FP: False Positives (incorrectly predicted migraine days)

Recall (Sensitivity): It refers to the proportion of the positive example that is predicted to be positive [51]. Given the healthcare context, recall is of primary importance. It measures the ability of the model to correctly identify all actual migraine days, minimizing false negatives.

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

Where:

- TP: True Positives
- FN: False Negatives (missed migraine days)

Confusion Matrix:

Truth	**Forecast Result: Positive**	**Forecast Result: Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

Table 3.3: Meaning of TP, FN, FP, TN.

F1 Score (Weighted): It is the harmonic mean of precision and recall that provides a balance between both, which makes it useful for imbalanced datasets like ours, where there are more non-migraine days than migraine days.

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

This is the harmonic mean of precision and recall.

AUC-PR (Area Under the Precision-Recall Curve): It indicates the model's ability to distinguish between positive and negative instances, especially in imbalanced datasets [52]. It represents the area under the precision-recall curve. There is no direct formula for AUC-PR, but conceptually, it's the integral of the Precision-Recall curve:

$$\text{AUC-PR} = \int_0^1 \text{Precision}(r)\, dr$$

Where:

- r: Recall

Reason for focusing on Recall

In the healthcare context, predicting migraine episodes is critical to managing symptoms and improving patient quality of life. Focusing on recall ensures that the model catches as many migraine episodes as possible, even at the expense of some false positives [35]. This is essential to allow timely interventions.

3.6 Experimental Setup

3.6.1 Data Split

We used an 80-20 split for our training and test datasets to evaluate model performance. This setup does not specifically take individual patients into account. Rather, it randomly selects sequences from across the dataset which includes data from various patients for the training and test sets, which means that data from the same patient may be included in both the training

and test sets but anonymously, which allows us to assess the model's ability to generalize across diverse patient data, providing insight into the model's predictive performance on a broad range of patterns observed in the dataset.

3.6.2 Model Training Configurations

- **Learning rate:** We used an initial learning rate of 1.485×10^{-5} for BiLSTM model and 2.65×10^{-4} for the Transformer network model. ReduceLROnPlateau callback. It was adjusted dynamically using the `ReduceLROnPlateau` callback if the validation loss plateaued.
- **Early stopping:** We used an early stop technique to stop model training if the validation loss did not improve for 7 consecutive epochs to avoid overfitting.
- **Loss function:** We used categorical cross-entropy for all models as our target involves predicting the probability distribution over two classes (migraine and non-migraine).
- **Class weights:** We applied "class weights" to handle the class imbalance between migraine and non-migraine days which ensures that the model paid more attention to the minority class (migraine days).

This methodological framework serves as the foundation of our research that offers a strong data-driven approach for the prediction of migraine episodes using advanced DL techniques. The models and methods we explore are designed to capture temporality dependencies while balancing precision with recall.

3.7 Model's Misclassification

Even having an improvement in terms of increased recall and precision values, if the models struggle and occasionally fail to predict migraine episodes accurately. Such misclassifications fall into two categories:

- **False Negatives (Missed Migraine Episodes):** There are some sequences which failed to predict a migraine episode even though one occurred. It occurs when migraine symptoms are mild or inconsistent, making it difficult for the model to detect patterns. Such cases are particularly harmful because they represent missed opportunities for early intervention.
- **False Positives (False Alarms):** It is when model sometimes predicted a migraine episode that did not occur. Such false alarms are often due to noise in the input features or external (non-recorded) factors such as emotional stress that may not have been captured in the dataset.

4

Results and Analysis

Here we present the results and analysis of the models, including their performance evaluation and comparison. Additionally, in a healthcare context, the error analysis and a discussion on the model's performance are provided. The results are divided into two sections; the first section where we present the overall performance of our models; and The second section which shows the model interpretability, showcasing the most impactful predictors for migraine episodes.

4.1 Overall Performance

The models were evaluated using standard classification metrics, including Precision, Recall, F1 Score, and AUC-PR, which were selected as they have relevance in imbalanced healthcare datasets, where missing a positive (migraine day) is more harmful than predicting a false positive.

4.1.1 BiLSTM Model

Table 4.1: Performance of BiLSTM

Metric	Value
Precision	0.1502
Recall	0.6805
F1 Score (Weighted)	0.5009
AUC-PR	0.4124

The BiLSTM model demonstrated a strong recall of 0.6805 as shown in Table 4.1, which is important in the context of healthcare. Higher recall emphasises its ability to identify migraine episodes effectively, but the precision is relatively low which means that many false positives were also predicted. The weighted F1 score (0.5009) and AUC-PR (0.4124) are a little lower, which reflects its trade-off in overall balanced performance. Despite this, its strength in recall aligns well intending to maximize correct migraine predictions to enable timely interventions.

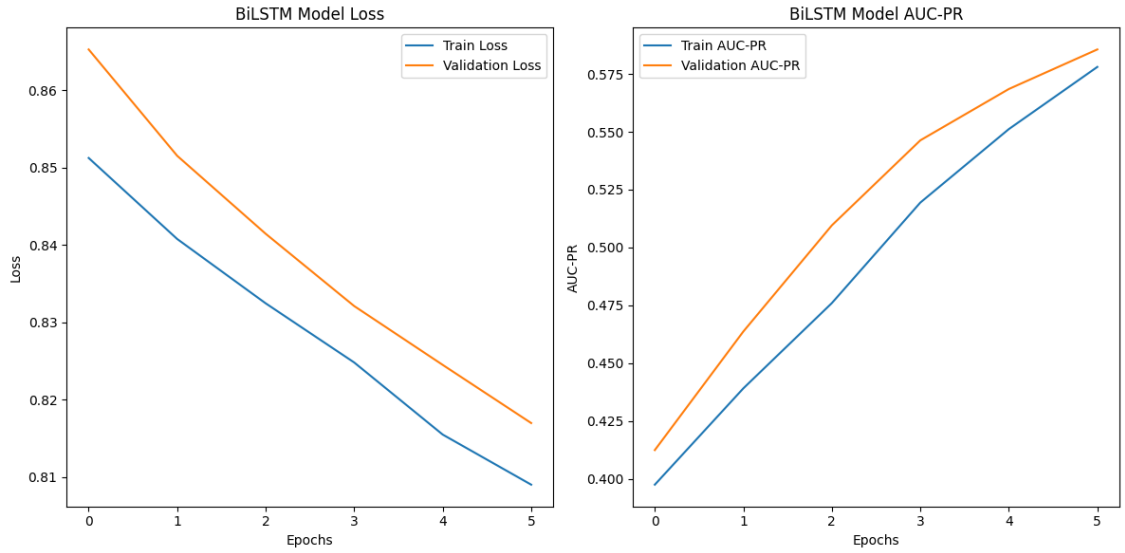


Figure 4.1: Training and Validation Loss and AUC-PR for BiLSTM

This Figure 4.1 illustrates the loss (left) and AUC-PR (right) curves for both training and validation sets during the training process of the BiLSTM model. The loss graph shows a steady decline, indicating effective learning, while the AUC-PR graph highlights improvements in precision-recall balance.

BiLSTM with Attention Mechanism

Table 4.2: Performance of BiLSTM with Attention Mechanism

Metric	Value
Precision	0.1860
Recall	0.5455
F1 Score (Weighted)	0.6716
AUC-PR	0.6788

The BiLSTM with an attention mechanism achieved improved performance, as shown in Table 4.2, with a recall of 0.5455, which is lower than the BiLSTM Model. However, it has achieved a higher weighted F1 score (0.6716) and AUC-PR (0.6788), hence more balanced across all metrics. This model is a strong contender for use cases where the primary objective is to maintain a balance between recall and precision.

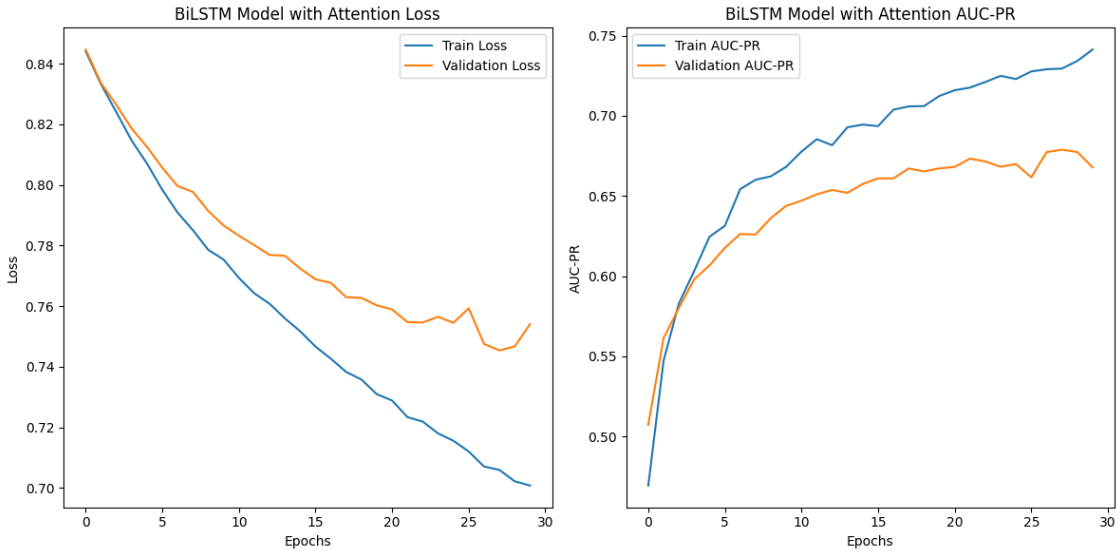


Figure 4.2: Training and Validation Loss and AUC-PR for BiLSTM with Attention Mechanism

This Figure 4.2 illustrates the loss (left) and AUC-PR (right) curves for both training and validation sets during the training process of the BiLSTM model with attention. The loss graph shows a steady decline, indicating convergence, while the AUC-PR graph highlights the model's performance in terms of precision-recall balance.

Table 4.3: Performance of Transformer Model

Metric	Value
Precision	0.1765
Recall	0.5844
F1 Score (Weighted)	0.6371
AUC-PR	0.6059

In comparison to the other two models, the Transformer model has achieved a recall of 0.5844, which is between the BiLSTM with Attention and the BiLSTM Model. Its weighted F1 score (0.6371) and AUC-PR (0.6059) highlight a well-balanced performance, meaning that it handles overall class imbalance better; hence, it is versatile on various evaluation metrics. However, the lower recall in comparison suggests that it missed more migraine days than the BiLSTM model, which is a limitation in healthcare contexts, where identifying all potential migraine days is important.

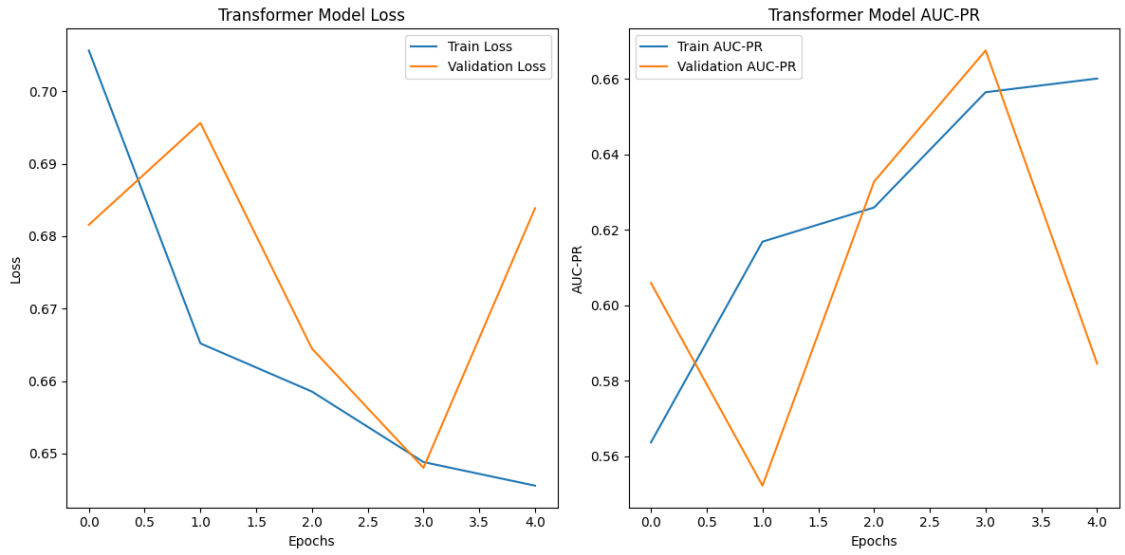


Figure 4.3: Training and Validation Loss and AUC-PR for the Transformer Model

Figure 4.3 illustrates the loss (left) and AUC-PR (right) curves for both training and validation sets during the training process of the Transformer model. The loss graph shows some

fluctuations in validation, while the AUC-PR graph reflects the model's ability to improve recall and precision over epochs.

Comparative Analysis

The performances of the BiLSTM, Transformer, and BiLSTM with Attention mechanism are summarized in Table 4.4 and are visualized in Figures 4.7–**??**, highlighting F1 score, recall, precision, and loss comparisons across epochs.

Table 4.4: Comparison of Models

Model	Precision	Recall	F1 Score (Weighted)	AUC-PR
BiLSTM	0.15	0.68	0.50	0.41
BiLSTM with Attention	0.19	0.55	0.67	0.68
Transformer	0.18	0.59	0.64	0.61

Based on our study's goal of emphasizing recall for identifying migraine episodes, the BiLSTM Model stands out as the best-performing model with the highest recall value, which makes it most suitable for our primary objective of prioritizing correct prediction of migraine episodes for timely interventions. Among these, the BiLSTM with the Attention mechanism runs slightly behind in terms of recall, though it provides superior overall balance, which will turn this into a competitive alternative when the scenarios require both precision and balanced metrics to be equally important. Additionally, the Transformer Model holds the second highest for recall and shows that this model would have consistent performance across all metrics, which reflects its robustness and adaptability to different prediction scenarios.

Accordingly, concerning the identification of migraine episodes, the BiLSTM Model is the most effective given its ability to maximize recall, which coincides with the intention to minimize missed episodes. Further confirmation of this can be seen in the ROC and PR curves from Figure 5, showing that the BiLSTM model with Attention has better capability in balancing recall and precision, as evidenced by its highest AUC-PR value.

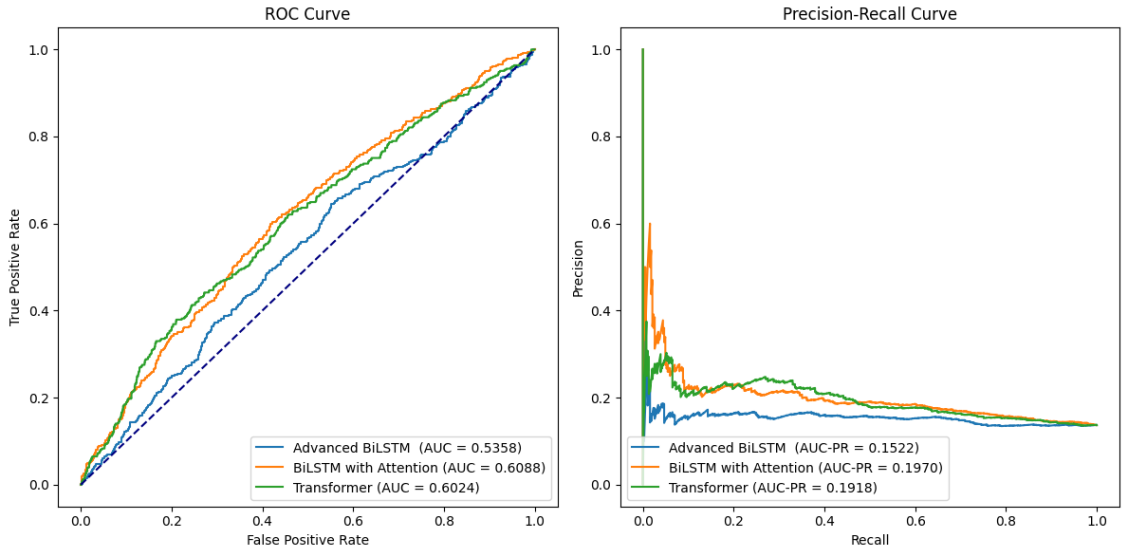


Figure 4.4: ROC Curve and Precision-Recall Curve for all models.

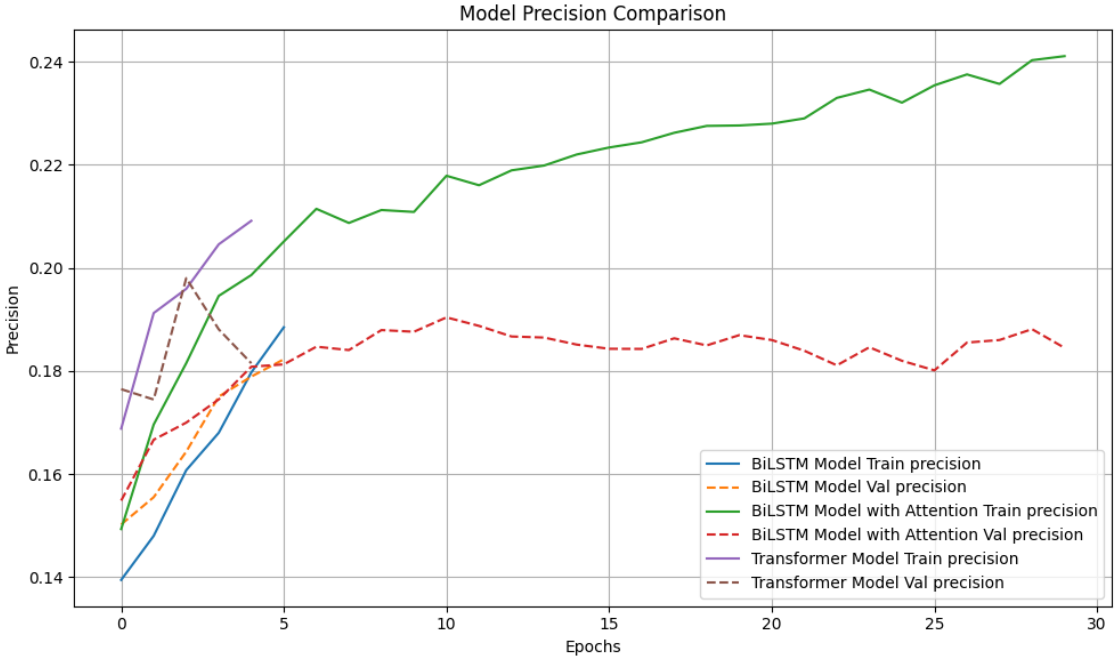


Figure 4.5: Precision Comparison for each model.

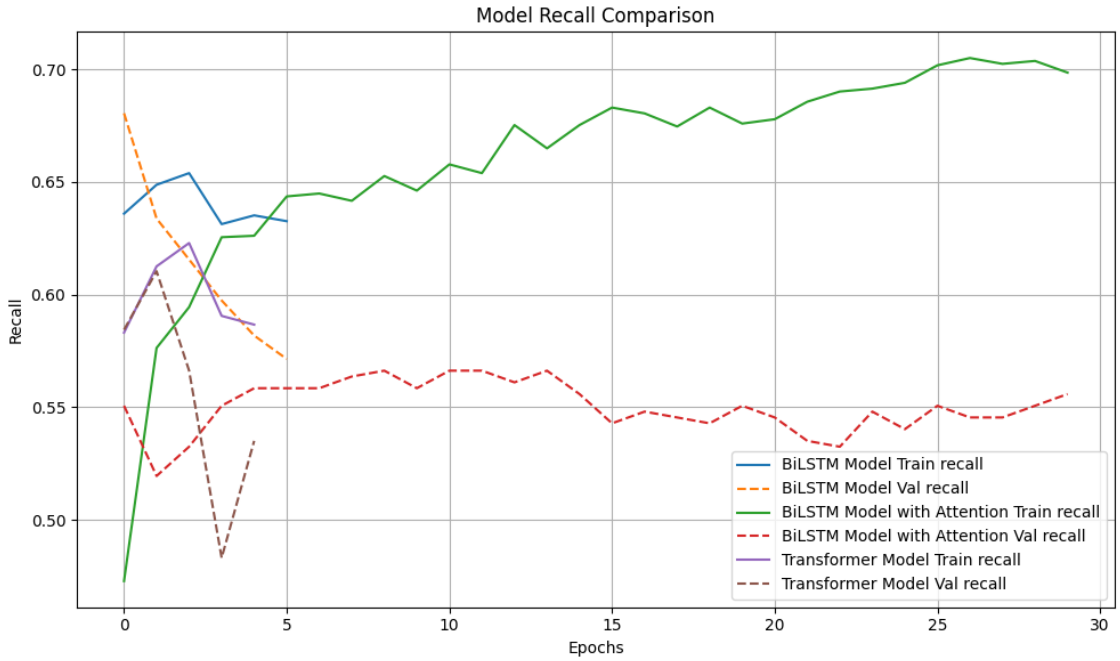


Figure 4.6: Recall Comparisons for each model.

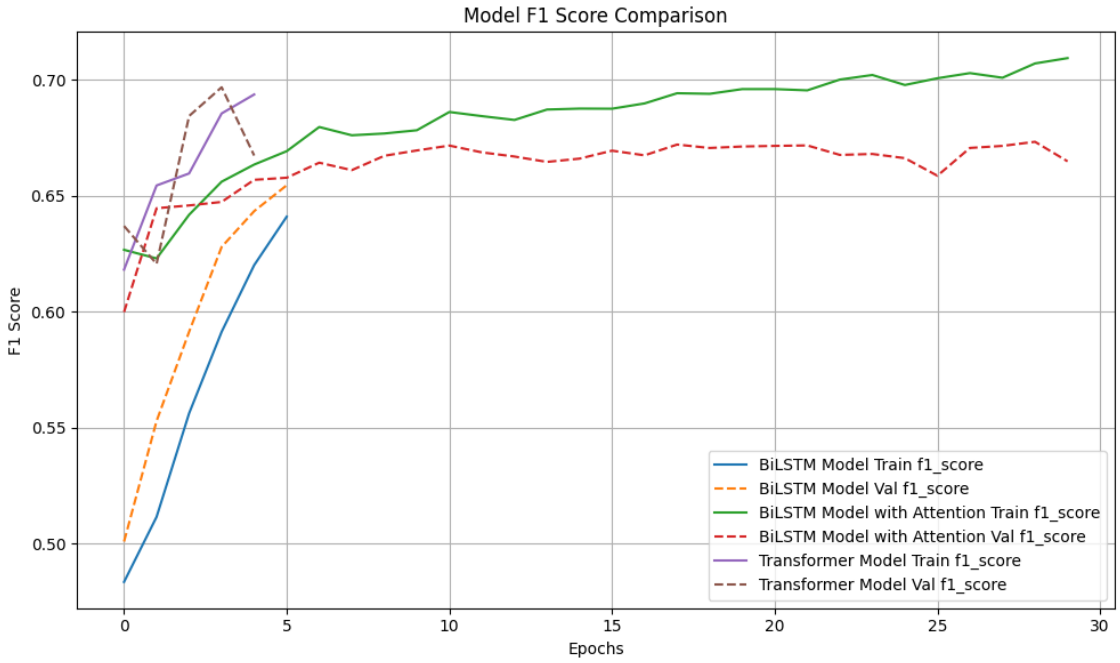
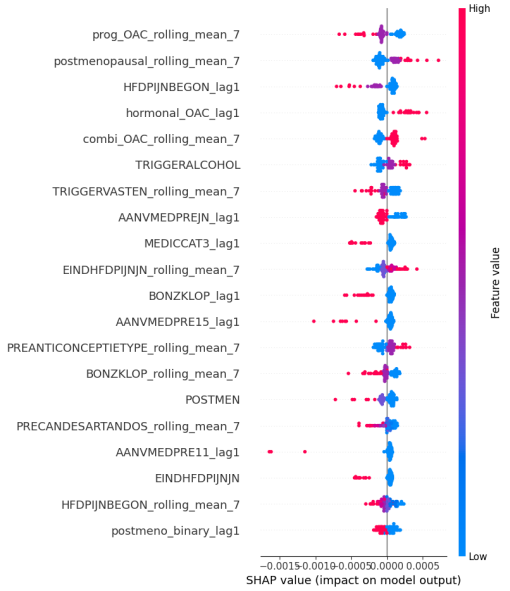
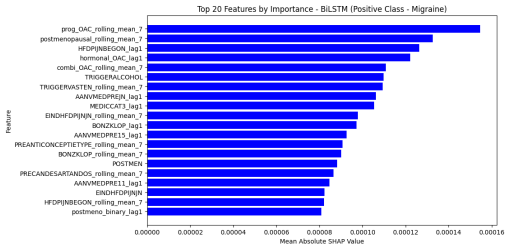


Figure 4.7: F1 Score Comparison for each model.

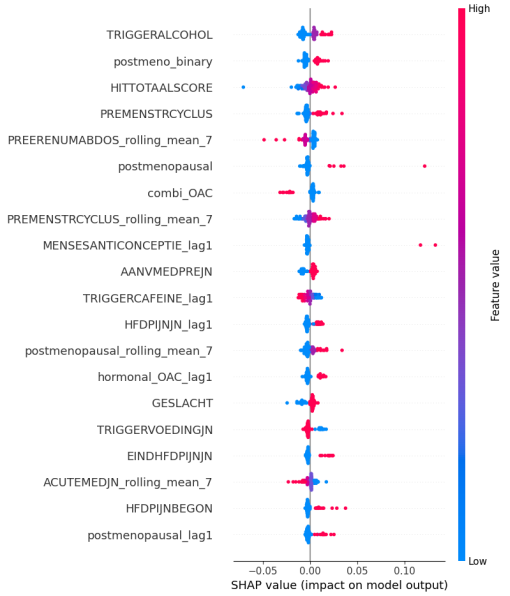
4.2 Model Interpretability - Key Predictors for Migraine Episodes



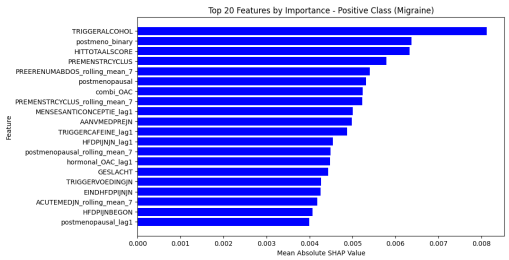
(a) BiLSTM Model: Key Predictors.



(b) BiLSTM Model: Key Predictors.



(c) Transformer Model: Key Predictors.



(d) Transformer Model: Key Predictors.

Figure 4.8: Key predictors identified for migraine occurrence using BiLSTM and Transformer models.

Figures 4.8a and 4.8c illustrate the top 20 features identified as most influential for the occurrence of migraine episodes for the BiLSTM and Transformer models, respectively.

In Figure 4.8a, the BiLSTM model highlights features such as "AANVMEDPREJN_lag1" (acute medication use lagged by one day), "prog_OAC_rolling_mean_7" (progesterone-based oral contraceptives averaged over 7 days), and "GESLACHT" (gender-specific identifier). It also emphasizes rolling mean and lagged features like "AANVMEDPRE6_rolling_mean_7" (medication use over 7 days) and "PREANTICONCEPTIEJN_lag1" (contraceptive use lagged by one day). These features align well with clinical understanding, demonstrating that migraines are often influenced by medication patterns, hormonal changes, and gender-specific factors. This indicates the model's capability to capture meaningful temporal patterns.

In Figure 4.8c, the Transformer model emphasizes complementary features such as "TRIGGERALCOHOL" (alcohol as a trigger), "postmeno_binary" (binary indicator for postmenopausal status), and "HITTOTAALSCORE" (a total headache intensity score). Additionally, it identifies lifestyle and hormonal factors, such as "PREERENUMABDOS_rolling_mean_7" (rolling mean of hormonal therapy) and "postmenopausal" (postmenopausal status), underscoring its ability to learn nuanced and long-term dependencies within the data.

Interestingly, both models show some overlap in significant features, such as the importance of hormonal factors ("AANVMEDPREJN") and lifestyle triggers, that could be considered the most generalized, influential triggers across the population. However, they also uniquely prioritize features based on their architectures. For instance, the BiLSTM model focuses more on lagged temporal features, while the Transformer model excels in identifying broader patterns related to hormonal and lifestyle influences. This ability may allow for the implementation of improved migraine management strategies serving larger groups while still retaining the personalized capabilities of these models. These SHAP-based visualizations provide valuable insights into the factors of influence driving migraine predictions, hence supporting the clinical relevance and interpretability of models. These models provide actionable insights that may improve migraine management strategies by highlighting common and unique predictors. Features aligned with established clinical knowledge further support the interpretability and reliability of models, ensuring their relevance in real-world healthcare applications.

4.3 Error Analysis

4.3.1 Insights from Misclassifications

A deep analysis of misclassifications across the models revealed several key insights. The BiLSTM model, achieving the highest recall, showed a tendency to overpredict migraine episodes, leading to a substantial number of false positives. This sensitivity was evident in features like headache frequency (`AANVMEDPREJN_lag1`) and lagged contraceptive use (`PREANTICONCEPTIEJN_lag1`). While these features align with known clinical indicators, their presence alone may not always signify migraine episodes. This highlights the trade-off between achieving high recall and reducing false positives.

The BiLSTM with Attention model, while achieving a more balanced performance with a weighted F1 score and an AUC-PR, exhibited lower recall. This suggests potential underrepresentation of certain temporal dependencies or overlooked subtle clinical patterns. Alternatively, the Transformer model, which demonstrated the second-highest recall and strong balanced metrics such as AUC-PR, exhibited distinct patterns of misclassification.

SHAP analysis revealed that features like `TRIGGERALCOHOL` (alcohol as a trigger) and `HITTOTAALSCORE` (headache intensity) significantly contributed to predictions, even in some false positives. Interestingly, false positives were also driven by features like `TRIGGERCAFEINE_lag1` (caffeine as a trigger) and `PREMENSTRYCYLUS_rolling_mean_7` (menstrual cycle patterns). These reflect the model's sensitivity to lifestyle and hormonal indicators. However, false negatives often involved overlooked long-term dependencies such as postmenopausal status and hormonal factors. This demonstrates the model's ability to capture nuanced patterns but also highlights the need for further refinement to handle inter-patient variability.

Shared key features across models underline the inherent difficulty in fully personalizing predictions, as their presence alone cannot account for inter-patient variability or individual symptomatology.

4.4 Model Performance in Healthcare Context

4.4.1 The Importance of Recall in Healthcare

In a healthcare setting, recall is considered the most critical evaluation metric as it reflects the model's ability to correctly identify all instances of a disease. Missing a critical event, such as a migraine episode, far outweighs the impact of a false positive prediction. High recall ensures

that most migraine episodes are flagged, enabling timely interventions that can significantly improve a patient's quality of life.

Our models prioritize recall to align with this objective. For instance, the Improved BiLSTM model demonstrated the highest recall of 0.6805, meaning it is very effective in capturing most migraine episodes. This aligns with our project goal of maximizing recall to provide proactive recommendations to patients, even at the cost of a few false alarms.

4.4.2 Alignment with Real-World Needs

These models were developed based on the actual clinical need for migraine prediction. In real-world scenarios, recall is more valuable than precision for both patients and healthcare providers because predicting most migraine episodes guarantees early preventive measures through medication or lifestyle adjustments.

While the Transformer model demonstrates a more balanced F1 score with higher precision, it was slightly less effective than the BiLSTM model in terms of recall. This reflects the real-world compromise between over-predicting migraine episodes and ensuring that critical cases are not missed. This underscores the need to fine-tune models for nuances in healthcare applications, where improving the quality of individual lives and outcomes is more critical than purely statistical metrics.

5

Conclusion and Future Work

5.1 Conclusion

This study demonstrates the potential of optimized deep learning models, including BiLSTM and Transformer architectures, for predicting migraine episodes from time-series data. By integrating attention mechanisms and hyperparameter optimization, the models improved predictive accuracy while enhancing their ability to identify key features contributing to migraine occurrence.

Furthermore, the application of interpretability techniques, such as SHAP values, bridged the gap between model predictions and clinical relevance. This provided actionable insights into influential predictors, thereby enhancing the clinical utility of the models. The findings represent significant progress toward leveraging machine learning to provide proactive healthcare solutions, enabling timely interventions to improve patient outcomes.

However, while these results are promising, this study also highlights the complexity of predicting chronic conditions like migraines due to inter-patient variability and diverse triggers. Future work is required to fully realize the potential of deep learning models in personalized healthcare.

5.2 Limitations

5.2.1 Predicting Rare Events

One of the primary challenges in this study was the inherent imbalance in the dataset, as migraine days were significantly fewer than non-migraine days. This imbalance posed difficulties in learning meaningful patterns for the minority class without compromising overall accuracy. While techniques such as class weighting were employed, achieving a balance between sensitivity to rare events and overall performance remains a challenge.

5.2.2 Generalizability and Variability

The variability in migraine triggers and symptoms across individuals further complicates the task of generalization. For example, while some models identified hormonal changes and lifestyle factors as significant predictors, their applicability across a diverse population remains limited. This restricts the models' ability to offer fully personalized predictions.

5.2.3 Long Training Times

One of the notable limitations of the proposed models is the significant computational resources and time required for training. Deep learning models, particularly BiLSTM and Transformer architectures, involve a large number of parameters, which necessitates extensive training on high-dimensional data. The inclusion of techniques such as Bayesian Optimization for hyperparameter tuning further increases the computational overhead. Also, our project was time-limited, further constrained the ability to experiment extensively with alternative optimization techniques or to fine-tune the models for even better performance.

5.3 Future Work

5.3.1 Refining Validation Strategies

Future work should focus on refining validation strategies, such as evaluating models on the last month of data for each participant. This individualized approach would better align with the goal of personalized predictions by accounting for unique patterns and predictors across patients.

5.3.2 Addressing Data Imbalance

Advanced techniques, such as synthetic data generation and transfer learning, could be explored to improve performance on rare events. These methods may help balance the dataset

and enhance the model's ability to learn from minority class examples.

5.3.3 Integration of Multimodal Data

Incorporating multimodal data, such as physiological signals (e.g., heart rate variability), environmental factors (e.g., weather conditions), and lifestyle metrics (e.g., physical activity), could provide a more comprehensive understanding of migraine triggers. This integration could lead to more accurate and actionable predictions.

5.3.4 Personalized Modeling

Future research should aim to develop models tailored to individual patients. Techniques like federated learning could allow models to learn from distributed datasets without compromising data privacy, paving the way for more personalized and ethical healthcare solutions.

Acknowledgments

The project was designed, organized, and supervised by Dr. Marieke Vinkenoog from LIACS and Ms. Nancy van Veelen, MD, from LUMC. Co-supervision was provided by Dr. Damiano Piovesan. Their guidance and expertise were invaluable in shaping this research.

References

[1] R. P, S. C. N, S. H, and K. R, "Migraine disability, quality of life, and its predictors," *Ann Neurosci*, vol. 27, no. 1, pp. 18–23, Jan 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7499825/

[2] P. J. Goadsby, P. R. Holland, M. Martins-Oliveira, J. Hoffmann, C. Schankin, and S. Akerman, "Pathophysiology of migraine: A disorder of sensory processing," *Physiological Reviews*, vol. 97, no. 2, pp. 553–622, 2017.

[3] M. Doane, S. Gupta, and J. e. a. Fang, "The humanistic and economic burden of migraine in europe: A cross-sectional survey in five countries," *Neurol Ther*, vol. 9, pp. 535–549, 2020. [Online]. Available: https://doi.org/10.1007/s40120-020-00196-2

[4] H.-K. Ko, "Design of disease prediction algorithm applying machine learning time series prediction," *International Journal of Internet, Broadcasting and Communication*, vol. 16, no. 3, pp. 321–328, Aug 2024. [Online]. Available: https://koreascience.kr/article/JAKO202425043323869.page

[5] C. ZF, K. XM, Y. CH, L. XY, G. H, and W. ZW, "Global, regional, and national burden and trends of migraine among youths and young adults aged 15-39 years from 1990 to 2021: findings from the global burden of disease study 2021," *J Headache Pain*, vol. 25, no. 1, p. 131, 2024.

[6] S. wei Cui, P. Pei, and W. ming Yang, "Application value of a machine learning model in predicting mild depression associated with migraine without aura," *British Journal of Hospital Medicine*, vol. 85, no. 9, pp. 1–12, 2024.

[7] D. Riskin, R. Cady, and A. e. a. Shroff, "Using artificial intelligence to identify patients with migraine and associated symptoms and conditions within electronic health records," *BMC Med Inform Decis Mak*, vol. 23, p. 121, 2023.

[8] H. M. B. M. Holsteen, K.K. and L. Nelson, "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*, vol. 60, pp. 2364–2379, 2020.

[9] O. B. de Dhaem and F. Sakai, "Migraine in the workplace," *eNeurologicalSci*, vol. 27, p. 100408, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240565022200017X

[10] S. A, I. SH, and B. E. et al., "Forecasting migraine with machine learning based on mobile phone diary and wearable data," *Cephalalgia*, vol. 43, no. 5, 2023.

[11] M. J. Buse, D.C. and R. Lipton, "Predicting the future of migraine attack prediction," *Headache: The Journal of Head and Face Pain*, vol. 60, pp. 2125–2128, 2020.

[12] A. Roides, "Predicting migraine attacks based on weather data using recurrent neural networks," Master Thesis, Universiteit Leiden, Statistics and Data Science, 2023.

[13] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. Article 2, 29 pages, March 2023.

[14] H. H and G. S, "Machine learning in healthcare," *Curr Genomics*, vol. 22, no. 4, pp. 291–300, Dec 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822225/

[15] V. M, van Leeuwen M, and J. MP, "Explainable haemoglobin deferral predictions using machine learning models: Interpretation and consequences for the blood supply," *Vox Sang*, vol. 117, pp. 1262–1270, 2022.

[16] M. S. H. Shaon, T. Karim, M. S. Shakil, and M. Z. Hasan, "A comparative study of machine learning models with lasso and shap feature selection for breast cancer prediction," *Healthcare Analytics*, vol. 6, p. 100353, 2024.

[17] D. Roy, A. Roy, and U. Roy, "Learning from imbalanced data in healthcare: State-of-the-art and research challenges," in *Computational Intelligence in Healthcare Informatics*, ser. Studies in Computational Intelligence, D. Acharjya and K. Ma, Eds. Singapore: Springer, 2024, vol. 1132.

[18] Y. Hu, W. An, R. Subramanian, N. Zhao, Y. Gu, and W. Wu, "Faster clinical time series classification with filter based feature engineering tree boosting methods," in *Explainable AI in Healthcare and Medicine*, ser. Studies in Computational Intelligence, A. Shaban-Nejad, M. Michalowski, and D. Buckeridge, Eds. Cham: Springer, 2021, vol. 914.

[19] H. Göker, "Automatic detection of migraine disease from eeg signals using bidirectional long-short term memory deep learning model," *SIViP*, vol. 17, pp. 1255–1263, 2023.

[20] A. Lew, "A brief survey of ml methods predicting molecular solubility: Towards lighter models via attention and hyperparameter optimization," *Preprints*, vol. 2024, no. 2024090849, 2024.

[21] G. B. B. X. . O. S. Wu, N., "Deep transformer models for time series forecasting: The influenza prevalence case," *ArXiv*, vol. abs/2001.08317, 2020.

[22] P. S. N. S. P. K. S. P. D. Mendhe, A. Dogra and S. B. G. T. Babu, "Ai-enabled data-driven approaches for personalized medicine and healthcare analytics," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pp. 1–5, 2024.

[23] Y. Yang, X. Zheng, and C. Ji, "Disease prediction model based on bilstm and attention mechanism," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 1141–1148.

[24] C. C-C, S. TJ, and D. G. et al., "Advancing toward precision migraine treatment: Predicting responses to preventive medications with machine learning models based on patient and migraine features," *Headache*, vol. 64, pp. 1094–1108, 2024.

[25] F. Yang, X. Wang, and H. e. a. Ma, "Transformers-sklearn: a toolkit for medical language understanding with transformer-based models," *BMC Med Inform Decis Mak*, vol. 21, no. Suppl 2, p. 90, 2021.

[26] F. G, S. A, A. A, K. I, B. E, K. P, K. Y, K. C, K. V, and V. VS, "Integrating shapley values into machine learning techniques for enhanced predictions of hospital admissions," *Applied Sciences*, vol. 14, no. 13, p. 5925, 2024.

[27] B. Qolomany, M. Maabreh, A. Al-Fuqaha, A. Gupta, and D. Benhaddou, "Parameters optimization of deep learning models using particle swarm optimization," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Valencia, Spain, 2017, pp. 1285–1290.

[28] S. Vedhanayaki and V. Indragandhi, "A bayesian optimized deep learning approach for accurate state of charge estimation of lithium ion batteries used for electric vehicle application," *IEEE Access*, vol. 12, pp. 43 308–43 327, 2024.

5.4 Code Listings

The complete codebase for this project is available and can be accessed through the following Google Drive link:

Code Listings on Google Drive

Feel free to explore the folder for implementation details, scripts, and data preprocessing routines.

References

[29] R. P, S. C. N, S. H, and K. R, "Migraine disability, quality of life, and its predictors," *Ann Neurosci*, vol. 27, no. 1, pp. 18–23, Jan 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7499825/

[30] P. J. Goadsby, P. R. Holland, M. Martins-Oliveira, J. Hoffmann, C. Schankin, and S. Akerman, "Pathophysiology of migraine: A disorder of sensory processing," *Physiological Reviews*, vol. 97, no. 2, pp. 553–622, 2017.

[31] M. Doane, S. Gupta, and J. e. a. Fang, "The humanistic and economic burden of migraine in europe: A cross-sectional survey in five countries," *Neurol Ther*, vol. 9, pp. 535–549, 2020. [Online]. Available: https://doi.org/10.1007/s40120-020-00196-2

[32] H.-K. Ko, "Design of disease prediction algorithm applying machine learning time series prediction," *International Journal of Internet, Broadcasting and Communication*, vol. 16, no. 3, pp. 321–328, Aug 2024. [Online]. Available: https://koreascience.kr/article/JAKO202425043323869.page

[33] C. ZF, K. XM, Y. CH, L. XY, G. H, and W. ZW, "Global, regional, and national burden and trends of migraine among youths and young adults aged 15-39 years from 1990 to 2021: findings from the global burden of disease study 2021," *J Headache Pain*, vol. 25, no. 1, p. 131, 2024.

[34] S. wei Cui, P. Pei, and W. ming Yang, "Application value of a machine learning model in predicting mild depression associated with migraine without aura," *British Journal of Hospital Medicine*, vol. 85, no. 9, pp. 1–12, 2024.

[35] D. Riskin, R. Cady, and A. e. a. Shroff, "Using artificial intelligence to identify patients with migraine and associated symptoms and conditions within electronic health records," *BMC Med Inform Decis Mak*, vol. 23, p. 121, 2023.

[36] H. M. B. M. Holsteen, K.K. and L. Nelson, "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*, vol. 60, pp. 2364–2379, 2020.

[37] O. B. de Dhaem and F. Sakai, "Migraine in the workplace," *eNeurologicalSci*, vol. 27, p. 100408, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240565022200017X

[38] S. A, I. SH, and B. E. et al., "Forecasting migraine with machine learning based on mobile phone diary and wearable data," *Cephalalgia*, vol. 43, no. 5, 2023.

[39] M. J. Buse, D.C. and R. Lipton, "Predicting the future of migraine attack prediction," *Headache: The Journal of Head and Face Pain*, vol. 60, pp. 2125–2128, 2020.

[40] A. Roides, "Predicting migraine attacks based on weather data using recurrent neural networks," Master Thesis, Universiteit Leiden, Statistics and Data Science, 2023.

[41] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. Article 2, 29 pages, March 2023.

[42] H. H and G. S, "Machine learning in healthcare," *Curr Genomics*, vol. 22, no. 4, pp. 291–300, Dec 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822225/

[43] V. M, van Leeuwen M, and J. MP, "Explainable haemoglobin deferral predictions using machine learning models: Interpretation and consequences for the blood supply," *Vox Sang*, vol. 117, pp. 1262–1270, 2022.

[44] M. S. H. Shaon, T. Karim, M. S. Shakil, and M. Z. Hasan, "A comparative study of machine learning models with lasso and shap feature selection for breast cancer prediction," *Healthcare Analytics*, vol. 6, p. 100353, 2024.

[45] D. Roy, A. Roy, and U. Roy, "Learning from imbalanced data in healthcare: State-of-the-art and research challenges," in *Computational Intelligence in Healthcare Informatics*, ser. Studies in Computational Intelligence, D. Acharjya and K. Ma, Eds. Singapore: Springer, 2024, vol. 1132.

[46] Y. Hu, W. An, R. Subramanian, N. Zhao, Y. Gu, and W. Wu, "Faster clinical time series classification with filter based feature engineering tree boosting methods," in *Explainable AI in Healthcare and Medicine*, ser. Studies in Computational Intelligence, A. Shaban-Nejad, M. Michalowski, and D. Buckeridge, Eds. Cham: Springer, 2021, vol. 914.

[47] H. Göker, "Automatic detection of migraine disease from eeg signals using bidirectional long-short term memory deep learning model," *SIViP*, vol. 17, pp. 1255–1263, 2023.

[48] A. Lew, "A brief survey of ml methods predicting molecular solubility: Towards lighter models via attention and hyperparameter optimization," *Preprints*, vol. 2024, no. 2024090849, 2024.

[49] G. B. B. X. . O. S. Wu, N., "Deep transformer models for time series forecasting: The influenza prevalence case," *ArXiv*, vol. abs/2001.08317, 2020.

[50] P. S. N. S. P. K. S. P. D. Mendhe, A. Dogra and S. B. G. T. Babu, "Ai-enabled data-driven approaches for personalized medicine and healthcare analytics," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pp. 1–5, 2024.

[51] Y. Yang, X. Zheng, and C. Ji, "Disease prediction model based on bilstm and attention mechanism," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 1141–1148.

[52] C. C-C, S. TJ, and D. G. et al., "Advancing toward precision migraine treatment: Predicting responses to preventive medications with machine learning models based on patient and migraine features," *Headache*, vol. 64, pp. 1094–1108, 2024.

[53] F. Yang, X. Wang, and H. e. a. Ma, "Transformers-sklearn: a toolkit for medical language understanding with transformer-based models," *BMC Med Inform Decis Mak*, vol. 21, no. Suppl 2, p. 90, 2021.

[54] F. G, S. A, A. A, K. I, B. E, K. P, K. Y, K. C, K. V, and V. VS, "Integrating shapley values into machine learning techniques for enhanced predictions of hospital admissions," *Applied Sciences*, vol. 14, no. 13, p. 5925, 2024.

[55] B. Qolomany, M. Maabreh, A. Al-Fuqaha, A. Gupta, and D. Benhaddou, "Parameters optimization of deep learning models using particle swarm optimization," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Valencia, Spain, 2017, pp. 1285–1290.

[56] S. Vedhanayaki and V. Indragandhi, "A bayesian optimized deep learning approach for accurate state of charge estimation of lithium ion batteries used for electric vehicle application," *IEEE Access*, vol. 12, pp. 43 308–43 327, 2024.

Acknowledgments