Universiteit Leiden

MyGraine: The Utility of Machine Learning Approaches for Personalized Migraine Attack Onset Prediction among Episodic Migraine Patients

Tome Jolevski (s3814300)

Supervisors:
Marieke Vinkenoog, PhD & Sandra van der Griend, MD

Bachelor Thesis – Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

August 26, 2026

Abstract

Background: Migraine is a severely debilitating neurological condition that is widely prevalent worldwide, representing one of the leading global causes of disability and significantly impairing patients' physical and social functioning. Its episodic nature, as well as the presence of triggers and premonitory symptoms that have been shown to indicate the onset of a migraine attack, jointly call for the development of approaches which can help patients accurately anticipate the onset of their next attack and take appropriate preventative measures in a timely manner, thereby minimizing the burden cast. ML and AI approaches have found their footing in this field, with such approaches commonly making use of self-reported patient-level electronic diary data.

Objective: To investigate two distinct ML approaches to migraine onset prediction, deployed on e-diary data: a generalized linear mixed-effects model (GLMM) and a long short-term memory model (LSTM), with the aim of evaluating their respective predictive and clinical utility.

Methodology: Both models were trained and evaluated on longitudinal electronic diary data collected at the Leiden University Medical Center. The dataset encompasses 78 individual candidate predictors (including various triggers, as well as health and lifestyle factors) across 381 episodic migraine patients whose observations were deemed viable. Both models aimed to predict following-day migraine onset, held to a minimal precision constraint of 0.20, with a decision threshold dynamically tuned to maximize recall within this constraint, thereby ensuring minimal missed attacks and maintaining the clinical viability of the results.

Results: The GLMM achieved an AUC score of 0.67 and a migraine recall score of 0.42 on the test set, while the LSTM achieved an AUC score of 0.65 and a migraine recall score of 0.53. Results collected from both models are consistent with results sourced across the literature. SHAP analysis carried out on the LSTM results isolated pain-coping ability, well-being and stress as the three most significant predictors of an impending migraine attack, with most binary trigger variables lagging behind in predictive value.

Conclusion: The relatively modest results achieved by both models seem to indicate that ML models trained on self-reported electronic diary measures are only moderately successful when it comes to reliable migraine prediction. Improvement upon these results would call for the collection of larger and more diverse volumes of data (integrating precise physiological measures alongside self-reported measures), as well as the development of precise and personalized models on the individual patient level, capable of capturing the sensitivity and attack frequency patterns that are unique to each patient, yet not clearly addressable using population-level prediction approaches.

Contents

1 Introduction

Migraine is a neurological disorder characterized by bouts (*attacks*) of throbbing, unilateral headache, which are accompanied by symptoms such as nausea, vomiting, and/or sensitivity to light and sound, with some patients also experiencing temporary episodes of sensory (usually audiovisual) disturbances known as *aura* [1], [2], [3], [4]. Aside from the headache itself, patients often experience symptoms such as fatigue, stiffness and increased irritability in the lead-up to a migraine attack, with patients often interpreting such symptoms as warning signals of an impending migraine attack [5], [6], [7]. Depending on the frequency of headache days, migraine can be classified as either episodic (less than 15 headache days per month) or chronic (more than 15 headache days per month, of which at least 8 display the characteristic features of a migraine headache), with the latter representing by far the most detrimental form of migraine disorder [1], [8]. Aside from physically incapacitating patients during an attack, migraine also imposes a substantial societal burden on affected patients, affecting their social participation capacity and workplace productivity, as well as their quality of life in general [9], [10], [11]. With around 1.04 billion people affected, migraine is the third most prevalent medical condition worldwide and (when expressed as years lived with disability) a leading cause of disability among young and middle-aged adults, positioning its treatment and prevention as a crucial issue of global importance [11], [12]. Research shows that the onset of a migraine attack may be motivated by an interplay of precipitating internal and external factors, commonly referred to as *triggers* [13]. Internal (endogenous) factors originate within the patient's physiology and affect their neurological sensitivity (examples being hormonal fluctuations, psychological stress, and baseline sensitivity to headache), whereas external (exogenous) factors are environmental and behavioral in nature, thereby making them a more suitable target for prediction-based attack prevention by allowing patients to avoid exposure to such triggers in a timely manner [5], [14], [15]. Commonly reported exogenous triggers of a migraine attack include skipping meals or otherwise not eating enough, disturbed sleep or sleep deprivation, smoking, alcohol consumption, exposure to strong odors, strenuous exercise, and many other examples [13]. A notable example of an extensively studied predictor of migraine onset is the menstrual cycle, which is highly clinically relevant given the fact that migraines are disproportionally more prevalent among female patients than male patients, with cases among female patients outnumbering those among male patients 3 to 1 [6], [13]. Given the well-studied nature of migraine triggers and their correlation to the onset of a migraine attack, many researchers have inquired into the deployment of various machine learning models in order to better predict the onset of a migraine attack, both on the patient level as well as on the population level [14], [16], [17]. The discrepancy between male and female patients previously discussed only further highlights the necessity of patient-level prediction models that would allow individual patients to accurately anticipate an impending migraine attack based on historical data of their individual predictors [14]. The accurate prediction of migraine attacks (especially among patients suffering from episodic migraines) would enable both patients and medical personnel to act in a timely manner, with the intention of mitigating the harm caused by the impending attack (or avoiding it entirely) through medical intervention or behavioral change [14], [18]. Patient-level attack forecasting efforts usually aim to estimate the likelihood of an attack unfolding within a pre-defined future time interval, usually the following day [14], [19]. As part of this paper, two approaches to migraine prediction were investigated: a generalized linear mixed-effects model (GLMM) [20], [21], as well as a long short-term memory model (LSTMM) [22], both of which had been previously investigated within the research group of

the Leiden Headache Center (LHC) and showed promise in the realm of migraine prediction. The parameters of both models, the patient dataset used for training and testing, as well as the results achieved, are discussed in the following sections.

1.1 Thesis Overview

This section of the paper includes the introduction. The following sections will elaborate on the background (section 2) of migraine, including its symptoms, triggers and pathophysiology, the burden it casts on its sufferers, as well as related work in the longstanding tradition of employing machine-learning-based approaches to the prediction of migraine attacks, all of which motivate decisions made in subsection 4.2. This is followed by the research question (section 3), which encapsulates the research gap that this thesis project attempts to address. As part of the methods section 4, subsection 4.1 elaborates on the dataset of trigger and health data used to answer the research question outlined in the previous section. Subsection 4.2 describes in detail the approaches taken to obtain an answer to the research question, including the specifications of both predictive models employed and how they are uniquely suited for answering the research question. The outcomes of these approaches are outlined in section 5. Section 6 provides an in-depth interpretation and discussion of the findings reported in section 5 and situates them within the context established by section 2. Section 7 elaborates on the limitations in the chosen approach and suggests directions for how to improve upon them, which may be beneficial for future researchers in the field. Finally, section 8 draws a conclusion from all information presented in the aforementioned sections.

1.2 Acknowledgments

Many thanks to my amazing supervisor Dr. Marieke Vinkenoog (of LIACS), whose insight, mentorship and encouragement have been of immeasurable value throughout the entirety of my work on this bachelor thesis project. Her willingness to take me under her supervision and assign me to a project of such social and medical importance has provided me with an opportunity that I am immensely grateful for. I also extend my gratitude to my co-supervisor, Drs. Sandra van der Griend (of LUMC), who has been nothing short of insightful, patient and understanding when faced with any inquiry on my part, as well as to Judith Essenburg, MSc, for her helpful direction and encouragement throughout the entire process. Special thanks as well to Prof. Dr. Gisela Terwindt for her expert advice on the direction of my research.

2 Background & Related Work

2.1 Migraine Prevalence & Diagnosis

Migraine is the third-most common neurological disorder worldwide, affecting approximately 12.7% of the world population, with lower estimates placing this proportion around 2.6% and higher estimates placing it at 21.7% [23]. In numerical terms, the Global Burden of Disease initiative (GBD) estimates that around 1.16 billion individuals around the world suffer from migraine, with its incidence rising to an astounding 90.34 million cases in 2021 [12]. Despite significant advancements in medical technology and a broadening array of treatments available to patients, the

health burden that migraine places on its sufferers remains quite large. As populations grow and urbanize and lifestyles continue to change, the importance of migraine as a pressing public health issue will continue to be exceptionally relevant as the future unfolds [12]. Migraine is a primary headache disorder, meaning that it arises independently (not as a result) of any underlying health conditions the patient may be facing, making its pathogenesis quite unclear in the majority of cases [24]. It is usually characterized by unilateral, pulsating headaches of moderate to severe intensity which can last anywhere from 4 to around 72 hours, which may be caused (or worsened) by any combination of environmental or sensory triggers. Headaches are accompanied by prolonged periods of vomiting, nausea, and hypersensitivity to sensory stimuli such as light (dubbed *photophobia*) and sound (dubbed *phonophobia*) [2]. During the *inter-ictal* period, which is the period between two subsequent migraine attacks, patients usually display minimal to no levels of headache pain or cognitive impairment. In spite of this, their overall brain functioning remains abnormal and more excitable than that of a non-migraine-affected person, with anxiety concerning an impending migraine attack only serving to further impede their well-being [2], [6]. Migraine attacks are broadly classified into one of two categories: migraine with aura and migraine without aura, with *aura* being defined as any combination of reversible neurological, audiovisual, sensory or speech symptoms that precede the attack, worsen as the attack progresses, and then gradually dissipate as the attack disappears, ranging over a relatively short 5-60 minute time period [1], [4]. Around 30% of all migraine patients suffer from migraine with aura, with the most widespread form of aura being *visual* aura, displayed in around 90% of aura-having patients [3], [4]. If attack onset frequency worsens, the condition can progress to *episodic* migraine, characterized by regular attacks at a rate of less than 15 headache days per month, as well as *chronic* migraine in the most severe of cases, characterized by more than 15 headache days per month, of which at least 8 fulfill the classification criteria for migraine attacks [1], [8]. Chronic migraine is by far the most debilitating form of migraine disorder, affecting around 8% of all migraine patients and taking an enormous toll on their physical, mental and social functioning [25]. Patients suffering from chronic or episodic migraine often transition from one condition onto the other, with around 2.5% of all episodic migraine cases progressing into chronic migraine, and around 26% of chronic migraine cases remitting into episodic migraine on a yearly basis [8].

2.2 Migraine & Disability

As common as it may be, migraine is also incredibly disabling, positioning itself among the leading causes of years lived with disability among both sexes, as well as among diverse age groups [6]. Migraine is routinely classified as an incapacitating condition, one that severely degrades the quality of life of those who suffer from it by placing a strain on their physical health, outlook on life, interpersonal relationships and workplace productivity, among many other medical and social dimensions [8], [11], [26]. Namely, migraine headache elements such as pain intensity and attack frequency have been shown to directly contribute to a worsened health-related quality of life score among migraine sufferers [9]. Further illuminating the state of migraine as a pressing public health issue, global disability rates due to migraine are on a steady rise, shifting from 27.41 million disability-adjusted life years (DALYs) in 1990 to a worrying 43.38 million DALYs in 2021, amounting to an increase of over 58% [12]. In addition to this, migraine shows strong comorbidity with a variety of physiological and psychiatric conditions, which only further increases its toll. Conditions such as anxiety, depression, panic disorders and chronic fatigue are strongly comorbid with migraine:

migraine patients are both at a higher risk of developing the aforementioned conditions, as well as at a higher risk of worsening symptoms during a period of heightened attack frequency [27], [28]. Migraine attacks also heavily restrict workplace productivity, with migraine-prone workers exhibiting a substantial loss in productivity and working capacity whilst they are experiencing a migraine attack [10]. This is in part due to the actual, physical pain of a headache, as well as due to the cognitive impairment and heightened sensory sensitivity that are characteristic of most migraine attacks. When untreated, attacks in the workplace serve to further destabilize the day-to-day lives and job security of migraine patients, which may increase levels of stress and anxiety, resulting in a vicious cycle [29].

2.3 Sex & Migraine Burden

Migraine burden is not equally distributed among the sexes: in fact, women are over three times as likely to be affected by migraine than men [6]. This discrepancy is usually (not fully, however) attributed to ovarian hormone fluctuation, heightened pain sensitivity and socio-environmental factors, but it is not the only aspect of migraine burden that is unequally shared among the sexes. In addition to the skewed prevalence, women also report longer attack durations, higher likelihoods of recurrence, heightened severity of symptoms and longer post-attack recovery periods [26], [30]. It's interesting to note, however, that this imbalance in prevalence is not present from birth: prepubescent children exhibit similar rates of migraine incidence across both sexes, and it is only during and after puberty that migraine incidence becomes noticeably skewed among women [30]. It comes as no surprise, then, that migraine is the leading cause of disability among young women, totaling to the most years of healthy life lost due to disability out of any other medical condition at around 4.9 disability-adjusted life years [26]. Common patterns of migraine incidence also differ, as men are, on average, more likely to develop migraine earlier but exhibit longer remission periods, whereas women have been shown to carry a far larger cumulative lifetime incidence in spite of later migraine development on average [30]. Regardless of sex, however, migraine is at its most prevalent within the 30-39 age interval (peaking around age 40), with both attack frequency and attack severity steadily decreasing as the patient progresses through life [31]. The skewness of migraine incidence has also led to the unfortunate portrayal of migraine as a "feminine" condition within the public consciousness, a reputation that is further aided by the fact that male patients have historically been underdiagnosed and have sought help for migraine at much lower rates than female patients, as well as being having been chronically underrepresented in clinical trials related to migraine disorders [30].

2.4 Migraine Triggers

As mentioned in subsection 2.1, even in the inter-ictal period (when no migraine attack is taking place), migraine patients display abnormal brain activity and a heightened predisposition to the onset of another migraine attack [2], [5], [13]. Furthermore, the brains of migraine patients have (in general) been shown to be hyperexcitable when compared to the brains of healthy patients, resulting in a hypersensitivity to sensory stimuli that heightens during the inter-ictal period and reaches its peak in the *ictal* period (the headache attack itself), with the attack being triggered when a hypothesized "sensory threshold" is reached [5], [7], [32]. This sensitivity threshold is determined in part from a spectrum of internal neurological and genetic factors (sex being the most

evident exemplar, as discussed in subsection 2.3), and in part from a complex interplay of external factors, such as the patient's lifestyle and environment [14]. These factors are commonly referred to as *triggers*, with a trigger encompassing any factor that a patient believes may precipitate, increase the probability of or temporally precede the onset of a migraine attack [5], [14]. Different patients report diverse sets of triggers, with commonly cited triggers including lack of sufficient sleep, stress and worsening well-being, smoking, exposure to strong odors, bright lights or loud sounds, consumption of alcohol, caffeine, dairy, nuts or any other problematic dietary group, strenuous physical exercise, and many others, with many of these triggers acting in combination with one another [13], [14]. However, the role of triggers in the onset of a migraine attack is subject to discussion. It is important to note that not everything that a patient may perceive as a migraine trigger is necessarily causally related to the onset of a migraine attack, with many patients conflating well-documented premonitory symptoms of a migraine attack (electrophysiological changes in the brain, tiredness, irritability and stiffness, to name a few) with environmental triggers that are experienced simultaneously, leading to a false conclusion that the trigger is responsible for the attack taking place [33], [34]. Furthermore, the cross-sectional nature of most trigger studies only serves to proliferate this error further, with patients filling out questionnaires retrospectively over long time periods and thus making findings prone to response bias and recall bias [5].

2.5 E-Diaries & Longitudinal Data Collection

In order to mitigate the risk of recall bias present in cross-sectional studies, many migraine forecasting studies using trigger data employ an *e-diary* approach to collect the required longitudinal data, with such an approach allowing for consistent tracking of personal patient-specific migraine days and triggers and significantly reducing patient recall error [35], [36]. This approach has been experimentally validated and is actively utilized by the Leiden Headache Center for the purposes of monitoring migraine data among patients at the Leiden University Medical Center (LUMC), which is how the data used for the purposes of this paper was collected [36]. As part of this approach, patients receive an electronic headache diary and are tasked with filling it out on a daily basis, reporting on whether they experienced a migraine the previous day (or the day before), which symptoms they displayed, the intensity of their headache, which medication they took, as well as questions about their well-being [36]. After each patient report is saved, the e-diary carries out an automatic algorithm which classifies the day in question as migraine day or a non-migraine day based on established criteria sourced from the International Classification of Headache Disorders (ICHD-3), which serves a double purpose of eliminating patient subjectivity and ensuring that migraine attacks are differentiated from regular headaches based on their associated symptoms [1], [37]. The e-diary approach has been shown to not only reduce recall bias, but also to reduce variability in self-reported migraine frequency among patients, as patients tended to both underestimate the number of migraine attacks during lower-frequency periods and, conversely, to overestimate said attack number during higher-frequency periods [35].

2.6 Attack Forecasting

Subsections 2.4 and 2.5 establish that the timely identification of triggers, premonitory symptoms, and electrophysiological changes in the brain alike could allow migraine patients to more accurately estimate the onset of a future migraine attack and take precautions accordingly [14], [38]. This

assumption lays the foundation of the field of migraine attack forecasting, in which researchers attempt to build predictive models on the patient level, using longitudinal data on triggers, symptoms, health factors and physiological changes in order to predict the onset of a migraine event [16]. With this type of longitudinal data in mind, migraine forecasting efforts functionally boil down to a supervised learning problem, with most studies in the field attempting to predict a future migraine day from a vector of previously collected patient data, with timely and accurate prediction allowing patients to take appropriate preventative measures [16]. Of particular interest to this paper are approaches utilizing machine learning (ML) paradigms. The body of literature on this topic is recent and quickly growing, with many researchers acknowledging the potential of ML and artificial intelligence (AI) approaches in the effort to predict migraine attack onset [16], [39]. Most studies in the field focus on one of three data types: 1) a patient's physiological metrics captured by wearable sensors, 2) constant external environmental factors, such as weather changes and, most importantly for this paper, 3) longitudinal self-reported data on triggers and premonitory symptoms that may foreshadow a migraine attack [19], [40]. Many of these studies show that ML and AI approaches are indeed valid when it comes to forecasting migraine attacks from trigger data (especially when combined with metrics such as heart rate variability and muscle tension measured through wearable devices), although the results shown remain somewhat inconclusive on their effectiveness [40]. Similar studies that focused on combining insights from trigger data and meteorological data through a smartphone app interface also showed promising results, with significant relationships between internal and external triggers and migraine attack onset being reported [19], [41]. Recent research also underscores the importance of predictions being made on the patient-level as opposed to the population-level, with the person-specific nature of migraine triggers being reflected in the very low generalizability of migraine prediction models to test sets consisting of previously unseen patients [42]. These findings are further complicated by the relatively small cohort sizes typically employed in migraine studies [19], [42].

2.6.1 Relevant External Work

In one of the first studies of this kind, Holsteen et al. developed a multivariate logistic regression model trained on trigger data in order to investigate and quantify the predictive potential of migraine triggers [43]. Participants were recruited online and were tasked with filling in a daily questionnaire (facilitated through an iPhone app) about the occurrence of a migraine attack and their exposure to common triggers such as sleep deprivation, menstruation and alcohol and/or caffeine consumption, with data collection spanning a 90-day period. In addition to the self-reported measures, the research team also employed the Google Maps and Dark Sky APIs in order to assemble data on daily weather conditions specific to each participant's home postal code during the aforementioned 90-day interval. Combining both trigger and weather data, Holsteen et al. started work on a multi-level logistic regression model with varying intercepts per patient, testing multiple lag configurations for the predictors (one day prior, two days prior and cumulative averages) as well as using raw values versus day-to-day changes in values for predictions; the best-fitting configuration was used to assemble the final model. Despite these efforts, the final model failed to exhibit meaningful predictive power, with a cross-validated between-patient C-statistic of 0.56 (which is in itself equivalent to the AUC for binary predictions [44]) indicating that predictions were only slightly better than chance. Additionally, the distribution of predicted probabilities generated by the model was narrowly clustered around the sample-level migraine onset rate of

17%, which Holsteen et al. interpreted as the model being unable to differentiate between high-risk days and low-risk days for most patients. Holsteen et al. conclude that the model is not clinically useful, owing to the inherent subjectivity of self-reported metrics, the high variability in triggers and symptoms between each individual patient, the lack of directly measured physiological data (through e.g. wearable sensors), as well as the relatively small pool of triggers included in the study. They noted that premonitory symptoms had been excluded as predictors (as they have been proven to highly correlate with the initial stages of a migraine attack, and therefore obscure the comparatively weaker effects of triggers) and called on future researchers to broaden the scope of both the predictors as well as the model architectures used for addressing the migraine prediction problem.

Following up on these findings, a noteworthy 2023 paper by Stubberud et al. explored the predictive utility of a suite of ML models when it comes to forecasting migraine attacks, using both self-reported e-diary data (including the presence and intensity of headache, quality of sleep, exercise, type of medication used etc.), as well as biofeedback data (such as heart rate, skin temperature and muscle tension) gathered on a daily basis through wearable sensors [45]. After the data was collected and processed (with 18 patients contributing 388 complete feedback days), Stubberud et al. proceeded to construct six predictive model architectures that are commonplace in ML: logistic regression, support vector machines (SVMs) and random forest classifiers, as well as gradient-, adaptive- and extreme gradient boosting machines. The target variable for prediction was defined as the presence or absence of headache the following day. Notably, to prevent data leakage, the train, test and validation sets were comprised of disjoint subsets of patients, with each patient's data being included in at most two of the three sets. Conducting a SHAP analysis revealed that the most important predictors were premonitory swelling and feeling of coldness, sleep duration, peak and average headache intensity, and impaired daily functioning, among others. Upon evaluation, the best-performing model was random forest, achieving a test set AUC of 0.62 and a test set accuracy of 0.56 (with a validation set AUC of 0.56), results comparable to those reported by Holsteen et al. [43]. Crucially, the random forest model exhibited a sensitivity of 0.0 and a specificity of 1.0, meaning that the model essentially exclusively predicted "no headache" for all cases. Despite achieving a minimally lower validation set AUC of 0.55, the gradient boosting model displayed a test set AUC equal to that of random forest (AUC=0.62), whilst displaying more balanced sensitivity and specificity values at 0.21 and 0.95 respectively. Interpreting these results, Stubberud et al. reflected on how although their findings somewhat succeeded in proving the utility of ML approaches to the problem of migraine forecasting, their results are called into question due to the extremely small sample size (18 participants overall and only 3 in the test set), as well as the lack of differentiation between migraine and non-migraine headaches.

Taking an alternative approach and focusing on migraine detection instead of migraine forecasting, Göker found that employing an LSTM-based approach to automate and aid the migraine detection process using EEG resting-state data gathered during the inter-ictal phase yielded promising results [46]. To address this problem, a dataset consisting of EEG recordings from 21 healthy (control) patients and 18 migraine patients was used. After filtering, 128 meaningful feature frequencies were extracted and two model training configurations were carried out: one which used all 128 channels, and one which used the 14 channels deemed most clinically significant (i.e. ones that correspond‘ to brain areas known to be sensitive to migraine). Comparing four different classifiers (random forest, SVM, linear discriminant analysis and bidirectional LSTM - BiLSTM),

BiLSTM proved by far to be the most effective, reaching accuracy, sensitivity, specificity, and F1 scores of 0.96, 0.96, 0.95 and 0.96 respectively on the full 128 channels. As previously mentioned, the problem formulation of Göker's study relates to migraine detection (by means of the differing neurological profiles of migraine patients when compared to healthy ones) and not migraine forecasting, therefore making Göker's work less directly comparable to the aims of this thesis paper than the other studies mentioned in this subsection. Nevertheless, Göker succeeds in showing how uniquely suited LSTMs are to working with biomedical data (which is in many cases inherently noisy and multifaceted), as well as how promising LSTM-based approaches are for the migraine realm in particular.

Building on top of Stubberud et al.'s work and working within the same research group, Faisal et al. made use of the combined e-diary and bio-feedback data approach originally employed by Stubberud et al. [45], albeit on a much larger scale [47]. Namely, Faisal et al. analyzed data originating from a randomized clinical trial encompassing 146 episodic migraine patients, with each patient contributing three months worth of daily e-diary entries (notably without gathering any trigger data, instead focusing purely on headache intensity, duration, medication use, and menstruation data when applicable) and biofeedback sessions, both targeting the same measures and facilitated through the same means as in Stubberud et al., with the final day count totaling to 21,550 rows, a significant increase when compared to Stubberud et al. [45]. The patient data was strictly separated on the patient level between the training, validation and test sets, and the binary target value was defined as the presence of a moderate-to-severe headache (a headache rated at or above a 4 on a scale from 1 to 10) both the following day, as well as within the following three days. Following data collection, three broad model categories were evaluated on the dataset: a total of 15 traditional ML classifier models (including random forest, logistic regression, (extreme) gradient boosting etc.), a sequence of time-series models (such as ARIMA, SARIMA, RNN, LSTM and GRU-D, the latest of which notably utilised a sliding-window approach wherein six days of observations were used to predict the seventh), as well as a strong transformer-based model suited to tabular datasets (TabPFN). When interpreting and comparing the model results with AUC as the guiding metric, both the standard ML models and the transformer model showed lukewarm results, with the best ML model (decision tree classifier) reporting a cross-validated AUC of 0.61 and the TabPFN model reporting a cross-validated AUC of 0.56. The GRU-D time-series model outperformed the rest by far, achieving a mean cross-validated AUC of 0.85 for following-day prediction, and 0.76 for the 72-hour prediction window. Despite this strikingly high AUC score, though, its sensitivity peaked at 0.39 (with a specificity of 0.95) for following-day prediction, indicating a subpar performance on identifying headache days when contrasted with headache-free days. These results were evened out when evaluating over the 72-hour period, with the model achieving sensitivity and specificity values of 0.58 and 0.83, respectively. Headache intensity, headache duration and maximal heart rate were identified as the most important features through permutation importance (PI), which works on the principle of permuting feature values randomly, and subsequently looking into how severely this disrupts the model output. Despite these seemingly impressive results and the usage of a time-series model serving as an important baseline to the LUMC's own migraine research, the lack of trigger and premonitory symptom data within the e-diary framework meant that the potentially crucial predictive value of these factors was not taken into account as part of Faisal et al.'s study.

In a yet unpublished study, Tsai et al. illustrated the importance of patient-level (as opposed to

population-level) predictions [48], employing a cohort of 25 patients and tracking features such as headache duration and onset time, intensity, aura, medication use, menstruation, etc. Instead of focusing on migraine attack onset, however, Tsai et al. decided instead to focus on following-day migraine attack recurrence using present-day data, thereby predicting the continuation of a migraine day from the present day into the following day. Four model types were investigated as part of the study (random forest, SVM, KNN and XGBoost), with each model being trained for each patient individually and the results being averaged across all patients. Before deciding on this patient-level prediction framework, Tsai et al. tested generalized versions of the given models, all of which achieved lukewarm predictive performances, all of which were later superceded by their patient-level counterparts. KNN proved to be the best-performing model when evaluating on mean recall (with a mean recall of 0.91), whilst random forest achieved the highest specificity (with a mean specificity of 0.77). Through an ablation study (in which predictors are removed in order to gauge how their absence impacts model output), pain intensity was isolated as the most significant predictor, followed by menstrual cycle, effectiveness of medication, and the location of the pain itself. It is important to mention, however, that the sequential nature of the given task (predicting attack recurrence / continuation, as opposed to onset) significantly simplifies the migraine prediction problem, thereby decreasing the significance of Tsai et al.'s results to the ones described in this paper. The relatively small sample consisting of only 25 patients, combined with the shuffling of patient data across the train and test sets (thereby destroying the sequential nature of patients' attack and symptom histories) are additional noteworthy limitations that may skew the relevance of the final results.

2.6.2 Relevant Work within LUMC

Within the research group of the Leiden Headache Center, two approaches have been investigated in particular: a generalized linear mixed-effects model, and several variants of a long short-term memory model, with each architecture being uniquely suited to tackling the problem at hand. As part of an LSTM-based approach, Hassan [17] used a dataset of e-diary data gathered from LUMC migraine patients and creating sliding-window sequences of 35-day migraine data for each patient, in order to predict the onset of an attack on day 36. Hassan's approach only tackles the emergence of a new migraine attack (i.e., the first migraine day in a potential sequence of migraine days), ensuring that migraine attacks that last several days are treated as a singular onset event, with the definition of a migraine day being a built-in feature of the e-diary used for data collection, and therefore compliant with the findings laid out by Van Der Arend et al. [37]. After using class weighing to handle the class imbalance inherent to the dataset and SHAP to isolate and report on the most significant predictors [49], Hassan then compared three LSTM architectures on the given dataset: BiLSTM (bidirectional LSTM), BiLSTM with a built-in attention mechanism, and a transformer model. BiLSTM exhibited the highest recall score (totaling to 0.68), with recall being identified and the most important and informative metric for the issue at hand. Interestingly, despite their lower comparative recall scores (0.55 for BiLSTM+A and 0.59 for the transformer model), both models exhibited comparatively higher F1 scores, with 0.67 for BiLSTM+A and 0.64 for transformer to the BiLSTM's original F1 score of 0.50 [17]. Alternatively, in an (as of yet) unpublished paper within the same research group, Van Veelen et al. utilize a generalized linear mixed-effects model (GLMM) for migraine prediction, using the same e-diary configuration for data

selection as Hassan, which is elaborated on in Van Veelen et al. [36]. Using at least three months of continuous trigger, lifestyle and health data for each patient, Van Veelen et al. proceeded to apply the Least Absolute Shrinkage and Selection Operator (LASSO) [50] to all candidate predictors in order to isolate the most relevant predictors, which were then used to fit a simple GLM model. The model in question was able to isolate menstruation, sleep deprivation, and migraine attacks in the recent past (among others) as the most indicative predictors of an upcoming migraine attack, with the LASSO-adjusted GLMM correctly identifying 74% of migraine days with a specificity (the proportion of true negatives identified as such) [51] of 49%, indicating middling results. The utilities of using a GLM model, as well as an LSTM model for migraine attack prediction are further elaborated upon in section 4.2.

3 Research Question

Having touched on the background in section 2, and with the guiding principle of accurate prediction enabling efficient prevention of impending attacks among episodic migraine patients being discussed in subsection 2.6, the research question (RQ) addressed by this paper is formulated as follows: **Can machine learning models of different classes (namely, a linear model - a GLMM, and an RNN-based model - an LSTM) be effectively utilized in order to predict the onset of a migraine attack on the patient level, from a combination of longitudinally measured medical and lifestyle factors?** The approach in question is inspired by the two approaches already undertaken within the research group (as discussed in subsection 2.6.2) and will employ patient-level predictions, with the aim of discovering whether such approaches are relevant to the field of migraine prediction, as well as whether they can effectively be utilized to prevent future migraine attack onset. As part of this paper, the two approaches and their results are compared and contrasted, and feature importance of relevant predictors (gathered through SHAP analysis) is presented for the LSTM results. The focus of the project lies heavily on the interpretability and applicability of the findings within the broader medical context, thereby adding new insight into the limited literature on the topic of migraine onset prediction and hopefully aiding patients suffering from episodic migraine in the alleviation of their symptoms.

4 Methodology

4.1 Dataset

The dataset utilized for the purposes of this project was collected by the Leiden Headache Center, from patients enrolled in the outpatient clinic of the Department of Neurology at the Leiden University Medical Center (LUMC), as well as from patients enrolled through web-based enrollment methods such as social media. It is important to note that a significant subset of patient variables pertaining to of migraine data collected for the purposes of the study were not used for training the prediction models, as they were either deemed irrelevant to the prediction task at hand (e.g. information about the time of day the attack took place, or whether it recurred after being treated with medication) or omitted at the supervisors' discretion (e.g. information about prophylactic medication prescribed to the patient, or acute medication taken to treat an ongoing migraine attack).

4.1.1 Participant Cohort

To gather data, participation in the study was offered on a voluntary basis to adult, Dutch-speaking patients. In order to be included, patients had to meet the ICHD-3 diagnostic criteria for episodic migraine either with aura or without aura (with episodic status being defined as a monthly headache frequency of less than 15 headache days per month), with patients diagnosed with chronic migraine or any secondary headache disorder being excluded [1]. Said episodic migraine patients were either presently undertaking treatment and monitoring as part of the outpatient clinic of the neurology department of the LUMC, or had applied to participate in the research through the headache center's website. Prospective participants received an initial screening for migraine diagnosis and were checked against the inclusion and exclusion criteria over telephone by a member of the research team, with the diagnosis being confirmed by a medical doctor and through consultations with headache specialist Prof. Dr. G. M. Terwindt. Given a positive screening result, participants were invited to complete a screening questionnaire in order to confirm their diagnosis. Post-questionnaire, participants were instructed with completing a much more extensive questionnaire inquiring into more detailed aspects of their migraine status and symptoms (such as treatment they are already receiving, or their menstrual status), after which they were cleared to join the cohort. Patients within the cohort received an electronic headache diary [36] and were instructed to complete its associated questionnaire on a daily basis, a task which included answering questions about symptoms they had experienced in the previous 24 hours, whether they had or had not experienced a migraine attack, their stress and wellbeing levels, the severity of said migraine attack, how well they were able to cope with the pain, any exposure to common triggers such as bright lights, loud sounds or strong odors, whether they were currently menstruating, to name a few. A full overview of all 78 predictors used can be found in Table 1. Employing the headache day classification algorithm previously utilized by Van Der Arend et al. and Van Veelen et al. [36], [37], a headache day is only classified as a migraine day if the patient exhibited both phonophobia and photophobia, or either nausea or vomiting within the duration of their headache attack, as outlined in the definition of a migraine day by Van Der Arend et al. [37]. Alongside the daily e-diary entries, participants were also tasked with completing a monthly *pre-diary*, in which they provided more detailed information about their prescribed prophylactic and acute medications (including type and dosage), their menstrual cycle (or lack thereof) and its regularity, any contraception they might be using, their pregnancy status, the average intensity of headaches they experience, among other questions. Data from the daily e-diary, as well the pre-diary columns related to the patient's menstrual cycle were then temporally organized and sorted by patient ID (with the pre-diary values being filled in for a month's worth of daily e-diary rows, as they are only sourced on a monthly basis). Per LUMC standards, migraine prediction studies set their inclusion criterion for patient data at a baseline of 60 sequential observations per patient (with patients being allotted the opportunity to go back and fill in days that they might have skipped), in order to ensure that about three months of viable migraine data are gathered. In accordance with this, the inclusion criterion for the model train and test sets (see subsection 4.2) was set at a minimum of 60 rows per patient.

Variable	Description	N	%	Mean (SD)	Median (Q1–Q3)
Patient cohort and dataset structure summary					
Total patients	Number of unique patients in the dataset	381			
Total rows	Number of diary day rows in the cleaned dataset	138,555			
Total columns	Number of predictor columns in the dataset	85			
Total migraine days	Number of rows indicating a migraine day (`MIGRDAGJN` = 1)	26,460			
Total migraine onset events	Number of rows indicating the first day of a new migraine episode	16,454			
Migraine onset rate	Proportion of all rows that correspond to a migraine onset day		11.9		
Female patients	Proportion of patients who indicated their sex as female (vs. male)		85.3		
Patients with aura	Proportion of patients diagnosed with migraine with aura (vs. without aura)		21.0		
Rows per patient	Number of diary day rows recorded per patient			363.66 (329.44)	354.00 (190.00–364.00)
Missing rows per patient	Rows per patient with at least one missing feature value (before imputation)			113.11 (317.87)	0.00 (0.00–3.00)
Migraine days per patient	Number of migraine days per patient				41.00 (21.00–79.00)
Migraine onset events per patient	Number of distinct migraine episodes per patient				30.00 (15.00–48.00)
Mean monthly migraine frequency	Average migraine days per 28-day period, per patient			4.74 (2.82)	
Mean monthly migraine onset frequency	Average migraine onset events per 28-day period, per patient			3.06 (1.51)	
Continuous e-diary variables and counts					
`PIJNCOPING`	Ability to cope with pain, rated from 0–10			1.61 (2.76)	0.00 (0.00–3.00)
`WELBEVINDENX`	Well-being score, rated from 0–10			6.89 (1.72)	7.00 (6.00–8.00)
`SRSS`	Level of stress, rated from 0–10			2.48 (2.08)	2.00 (1.00–4.00)
`KATERERNST`	Hangover severity, rated from 0–10			0.02 (0.31)	0.00 (0.00–0.00)
`TRIGGERCAFEINE`	Caffeine consumption, measured in cups of coffee consumed			1.71 (2.08)	1.00 (0.00–3.00)
`beer`	Beer consumption, measured in number of beers consumed			0.07 (0.47)	0.00 (0.00–0.00)
`redwine`	Red wine consumption, measured in number of glasses consumed			0.04 (0.29)	0.00 (0.00–0.00)

continued on next page

Table 1 (continued)

Variable	Description	*N*	%	Mean (SD)	Median (Q1–Q3)
whitewine	White wine consumption, measured in number of glasses consumed			0.08 (0.43)	0.00 (0.00–0.00)
liquor	Liquor consumption, measured in number of shots consumed			0.04 (0.30)	0.00 (0.00–0.00)
TRIGGERTIJDZONEVERSCHIL	Experience of a time zone difference, measured in hours			0.48 (12.94)	0.00 (0.00–0.00)
Binary e-diary variables					
TRIGGERVOEDINGJN	Consumption of any trigger foods	348	91.3		
TRIGGERVOEDING0	Consumption of cheese	341	89.5		
TRIGGERVOEDING1	Consumption of chocolate	339	89.0		
TRIGGERVOEDING2	Consumption of nuts	337	88.5		
TRIGGERVOEDING3	Consumption of citrus fruits	332	87.1		
TRIGGERVOEDING4	Consumption of sweeteners	245	64.3		
TRIGGERCAFEINEJN	Consumption of caffeine	344	90.3		
alcoholjn	Consumption of alcohol	279	73.2		
TRIGGERKATER	Experience of a hangover	144	37.8		
TRIGGERHOOFDTRAUMA	Experience of a head injury	229	60.1		
TRIGGERVASTEN	Fasting status	329	86.4		
physicalactivityJN	Strenuous physical activity	338	88.7		
physicalactivity_short	Strenuous physical activity of up to an hour	323	84.8		
physicalactivity_long	Strenuous physical activity lasting longer than an hour	301	79.0		
tijdzone_oostwest	Time zone change east/west	79	20.7		
SLAAP_depriv	Sleep deprivation, defined as less than 30% of usual sleep duration	332	87.1		
MENSESNATCYCLUS	Natural menstrual cycle	155	40.7		
menses_dag_calc	Menstruation day	186	48.8		
menses_AC_stop	Menstruation day due to stop in contraceptive use	88	23.1		
menses_AC_doorbraak	Menstruation despite contraceptive use	39	10.2		
change_profylaxe	Change in prophylactic medication	67	17.6		
regular_pre_menstruation	Regular menstruation, indicated within the pre-diary (same as pre_menstruation_1)	71	18.6		
irregular_pre_menstruation	Irregular menstruation, indicated within the pre-diary	232	60.9		
Sensory exposure category indicators					
TRIGGERLICHT_0	No exposure to bright light	349	91.6		

continued on next page

Table 1 (continued)

Variable	Description	*N*	%	Mean (SD)	Median (Q1–Q3)
TRIGGERLICHT_1	Exposure to bright light for less than 30 minutes	315	82.7		
TRIGGERLICHT_2	Exposure to bright light for 30–60 minutes	329	86.4		
TRIGGERLICHT_3	Exposure to bright light for over 60 minutes	297	78.0		
TRIGGERGELUID_0	No exposure to loud sounds	349	91.6		
TRIGGERGELUID_1	Exposure to loud sounds for less than 30 minutes	297	78.0		
TRIGGERGELUID_2	Exposure to loud sounds for 30–60 minutes	290	76.1		
TRIGGERGELUID_3	Exposure to loud sounds for over 60 minutes	280	73.5		
TRIGGERGEUR_0	No exposure to strong odors	349	91.6		
TRIGGERGEUR_1	Exposure to strong odors for less than 30 minutes	290	76.1		
TRIGGERGEUR_2	Exposure to strong odors for 30–60 minutes	216	56.7		
TRIGGERGEUR_3	Exposure to strong odors for over 60 minutes	160	42.0		
Temporal variables encoded as binary, batched to reduce redundancy					
migraine1–migraine7	Days elapsed since the most recent migraine attack, ranging from 1 to 7 days	381	100.0		
SLEEPDEPR1–SLEEPDEPR7	Days elapsed since the most recent sleep deprivation event, indicated from 1 to 7 days	331–334	86.9–87.7		
perimenstrual1–perimenstrual10	Days remaining until the onset of menstruation, ranging from 1 to 10 day lead-up periods	67–68	17.6–17.8		
menstrual1, menstrual2	Days elapsed since the first day of menstruation, accounting for lags of 1 and 2 days	186	48.8		
Pre-diary menstruation status variables					
pre_menstruation_1	Regular menstruation, indicated within the pre-diary (same as regular_pre_menstruation)	71	18.6		
pre_menstruation_2	Irregular menstruation, indicated within the pre-diary	28	7.3		
pre_menstruation_3	Irregular menstruation due to pregnancy	7	1.8		
pre_menstruation_4	Irregular menstruation due to breastfeeding	2	0.5		
pre_menstruation_5	Irregular menstruation due to postmenopausal status	90	23.6		
pre_menstruation_6	Irregular menstruation due to the use of oral contraception	55	14.4		

pre_menstruation_7	Irregular menstruation due to the use of hormonal contraception	63	16.5

Table 1: Descriptive statistics of the study population and the electronic diary predictor variables. For the continuous variables, summary statistics are computed across all rows. For all remaining variables, N and the percentage indicate the number and proportion of patients who reported the given trigger within at least one diary observation. Cells corresponding to statistics that are inapplicable to the given variable's type have been left blank. All binary diary variables are reported as yes/no (1/0). For the temporal variables, values are reported as a range where they differ across the constituent indicators.

4.1.2 Data Preprocessing

After the single participant with less than the minimal amount of 60 rows had been filtered, columns which were to be excluded as predictors (namely: sex, final aura / non-aura diagnosis, increases / decreases in stress levels, stress increase and decline days, exposure to height, physical activity variables measured in duration, all pre-diary values excluding post-menopausal status and menstruation regularity, headache start- and end-time values, headache intensity scores, migraine recurrence information, and prophylaxis dosages) were dropped from the dataset, with columns showing large gaps in data (such as the columns corresponding to night shifts and large timezone shifts, for example) and ordinal columns (for which dummy variables were ultimately created) also being omitted. Missing well-being and stress scores were imputed as a mean of the well-being scores of the given patient on the day before and the day after the missing observation. Variables representing time intervals (such as amount of intense exercise, for example) were binned as part of the supervisors' preprocessing, and only binned columns were used for the purposes of the final prediction model (given in Table 1). Menstruation data was encoded within the pre-diary questionnaire as one of 7 values, with 1 corresponding to regular menstruation and 2-7 corresponding to various types of irregular menstruation, elaborated on in the final section of Table 1. For prediction purposes, this variable was treated in two stages: firstly, dummy variables corresponding to each category were created, and secondly, categories 2-7 were collapsed into a new column labeled "irregular", thereby ensuring that the effects of individual type of irregular menstruation are taken into account whilst also leaving room to compare and contrast the effects of regular vs. irregular menstruation. Missing values for all other predictors were then defaulted to 0. Individual patient rows were then sorted chronologically, in order to account for any potential inconsistencies in their initial ordering. Finally, a new column corresponding to the onset of a migraine event was added, which assumes a value of 1 if the given migraine day is the first in a sequence of migraine days, and 0 if it directly succeeds a migraine day. All data preprocessing was carried out using the `pandas` statistical analysis package for Python [52].

4.2 Approach

The following subsections will elaborate on the two genres of models used to address the RQ outlined in section 3. Subsection 4.2.1 will introduce the generalized mixed-effects model and explain how it is uniquely suited to address both the within-patient and inter-patient variability inherent to the dataset, tying into the framework implemented by Van Veelen et al., as discussed in subsection 2.6. Fixes, tuning and other optimizations made as part of the research for this paper will be touched on as well. Subsection 4.2.2 will introduce the long short-term memory model and explain how it addresses the issues faced by the GLM model, especially as they pertain to the GLM model's inability to be dynamically trained. The LSTM model makes up for this by introducing a growing prediction window with a pre-set amount of "warm-up" rows used for scaling and training, after which each row is predicted and subsequently added back to the training set, with prediction moving on to the following row, an approach inspired by that of Hassan, as outlined in subsection 2.6. As with the GLMM, fixes and improvements will be discussed, as well as limitation encountered in the implementation phase.

4.2.1 Generalized Linear Mixed-Effects Model

Broadly, generalized linear models (GLMs, an example being logistic regression) are widely used in machine learning for problems which require the outcome variable to follow a non-Gaussian distribution, such as a binary distribution in the case of this paper. This is achieved through employing a link function that maps each predictor to the scale of the outcome. Generalized linear *mixed-effects* models expand on the GLM structure by accounting for both fixed effects (effects whose impact applies equally across all input cases) and random effects (effects that may fluctuate across individual inputs or input groups). To be more precise, fixed effects encapsulate the average relationship between a certain predictor and the final outcome across the entirety of the dataset, whereas random effects account for the fact that a sequence of observations of the same individual are not mutually independent and are therefore more similar to each other than those of a separate individual [20], [21]. This approach is well-suited to the e-diary migraine onset prediction problem due to the repeated daily observations on the patient level, thereby allowing both the correlation of broader effects across patients (such as the average effect of stress on migraine onset, for example) as well as the inherent diversity of patient-specific effects (such as a specific patient's sensitivity to stress) to be properly accounted for. As GLMs are comparably simpler to implement in R rather than Python, the `rpy2` package was used for the purposes of the GLMM implementation, enabling efficient interfacing between Python and R [53]. This interfacing further allowed for the use of the `lme4` R package, with the package being by far the most efficient method of computationally implementing and fitting linear mixed-effects models [54].

In order to facilitate following-day attack onset prediction based on current-day patient data, the previously processed data was loaded and checked for chronological ordering within each patient, ensuring that predictors are sequentially ordered before lagging. Non-predictor columns (such as the target column, set as migraine onset, as well as patient ID) were dropped, and the values of all remaining predictors were shifted downwards by one row (being mindful of dropping rows that do not have a previous observation) and concatenated back to the original data. An additional check is employed in order to locate and remove any potential columns consisting of a single constant value in all cells. As doing a standard shuffled train / validation / test split would destroy the sequential ordering of the patient data (and thus cause, for example, past patient rows to be predicted using future rows), the train / validation / test split was carried out chronologically, such that the first 67% of each patient's entries were placed in the training set, the following 17% in the validation set, and the remaining 16% were placed in the test set. Aside from ensuring that the data is sequentially sound, this approach provides the added benefit of the same patients appearing in both the train, validation test sets, meaning that all patients in the test set are previously made known to the model. Most columns within the dataset are either binary variables or dummy variables standing in for ordinal variables, which means that the non-binary numeric columns, (such as well-being and stress) have to be scaled in order to avoid skewness. The scaler used for this task was fitted exclusively on the training data in order to prevent validation and test ("future") data from influencing the scaling process.

After the data was prepped, what remained was to build the GLMM model using the `lme4` package. The leftmost side of the underlying GLMM formula contains the target variable (the onset of a migraine attack), encoded as a binary value. The right side is composed of two separate

parts: firstly the sum of the fixed effects of all predictors, as well as the random intercept (`(1 | LUMRECID)`). The sum assigns a lagged predictor term to each individual feature in order to ensure that results are temporally consistent and present-day data doesn't leak into the same day's prediction. The intercept instructs the model that baseline migraine risk values are to be assigned independently to each patient, as patients do not share a common baseline odds of developing migraine. Patients that are more frequently affected by migraine attacks obtain a higher intercept value and are therefore treated as more likely to experience a migraine attack overall, whereas those experiencing attacks more rarely obtain a lower value to reflect their lower perceived risk. These two terms combined account for both the daily variability of attack risk induced by triggers or environmental factors, as well as the intrinsic differences between patients when it comes to propensity to migraine attacks. Due to the binary nature of the outcome variable, the GLMM itself employs a logit link function in order to express the fixed-effects coefficients on the log-odds scale. After the model was fitted on the training data using `lme4`'s `glmer()` method, the final model (containing both the fixed-effect coefficients and the estimated per-patient intercepts) was used to generate predictive probabilities on the test set through inverse logit transformation. This results in a predicted value between 0 and 1 for each day, which represents the estimated probability of that patient experiencing the onset of a migraine attack on that day (e.g., an output of 0.64 indicating a 64% chance of a migraine attack that day). Predicted values are then stored for further processing and analysis.

After the initial predictive probabilities are generated, a cutoff point must be selected to convert said probabilities into binary values, such that if the probability is equal to or greater than the threshold, the model would predict *migraine*. The most important metrics for evaluating this threshold are the *recall* ($\frac{\text{true positives}}{(\text{true positives + false negatives})}$), which indicates the rate of true migraines predicted as such; the *precision* ($\frac{\text{true positives}}{(\text{true positives + false positives})}$), which indicates the rate of correctly predicted migraines; and the *F1* score, which is defined as the harmonic mean of precision and recall and indicates the overall quality of the model fit. A very low threshold may lead to very high recall, entailing many correctly identified migraines in addition to many false alarms, whereas a very high threshold may lead to very low recall, meaning that substantial amounts of migraine attacks would be missed in exchange for a lower false alarm rate. As recall is the most clinically significant metric within this context (a few false alarms are acceptable, as long as as many migraines are correctly anticipated as possible; a false alarm in this context is merely a minor inconveniece for the patient). The minimum precision was set at 0.2 in order to ensure that false alarms are kept in check, with the threshold tuning algorithm aiming to select a threshold with maximal recall, whilst adhering to the aforementioned minimal precision constraint. After the optimal threshold is selected, it is then applied to the array of predicted probabilities, converting them into binary predictions and allowing for the calculation of a confusion matrix, classification report, and visual plots. These results are outlined in section 5.

4.2.2 Long Short-Term Memory Model

An LSTM is a specialized subtype of a recurrent neural network (RNN) which parses an input sequence step by step, whilst maintaining an ongoing history and summary of everything it has evaluated thus far. As opposed to a regular RNN, in which older information (encountered earlier in the stream of input sequences) progressively decays with each timestep, the LSTM

architecture mitigates this issue with the introduction of a *cell* that is continuously updated in a non-overwriting, additive fashion. The cell enables the LSTM to recover more distant information and ensures that the effects of such information are not ignored, as well as enabling the training signal to backpropagate through the cell. The cell itself interacts with three different *gates*, with a binary value (based on the current and previous output) corresponding to each of them and indicating which of the three gates an input sequence will pass through. The *forget* gate serves as a discounting factor for the cell content by selectively removing ("forgetting") information that is deemed no longer relevant. The *input* gate scales newly read values before incorporating them into the cell, thereby ensuring that seemingly irrelevant inputs don't destabilize the stored content. The *output* gate scales the cell content that is shared with the remainder of the network and ensures that circulation of information currently deemed irrelevant is halted, and circulation of relevant information is facilitated [22]. The utility of LSTM architectures in the realm of migraine prediction rests on the assumption that small patterns of triggers and symptoms that accumulate over days can be picked up by the model over time, with the relevance of each individual predictor being subject to variability. When contrasted with the predefined nature of the GLMM's lagged predictors, the LSTM allows for more robust and longitudinal learning, being able to detect relationships between predictors and the target that may span a much longer time frame.

The LSTM implementation for the purposes of this paper was carried out using `PyTorch` (the standard library for ML / deep learning problems in Python) [55], with model interpretability and explainability being handled by `SHAP` [56]. After loading the preprocessed data and ensuring that the temporal ordering of patient rows was intact, migraine onset was defined as the target variable, and all other feature columns (excluding patient ID and date) were set as predictors. The data was then grouped on the patient level (by patient ID) and patient-level dictionaries were created, each of them containing: a 2D array of all feature values (labeled $\mathbf{x}$, where rows correspond to timesteps and columns correspond to features), a 1D array of labels (labeled $\mathbf{y}$, where 1 corresponds to migraine and 0 corresponds to no migraine), the timestamp for each row (labeled **date**), as well as each patient's ID. After these sequences were assembled, the first 56 timesteps of each patient's data were isolated into a **warm-up** set that is used to fit and calibrate a standard feature scaler (through which all feature values assume a mean of 0 and a standard deviation of 1), which is then deployed on all subsequent patient sequences. The warm-up window also serves an additional purpose as the absolute-minimum row history requirement for prediction, ensuring that the model exclusively generates predictions based on at least 56 rows of data, corresponding to two 28-day months. It's important to note that this window is distinct from the actual growing context window used for prediction (discussed below), as the warm-up window is used mainly for scaler fitting. We proceed by assembling *prefix samples*, wherein a sample is created at each timestep $\mathbf{t}$ from 1 all the way to 243 (the absolute training cutoff calculated as the 67% point of the mean length of all patient sequences), after which a growing context window approach is implemented, with the model reading $\mathbf{t}$ feature rows in order to predict the label corresponding to row $\mathbf{t+1}$. After the scaler is fit on the set of warm-up rows, it is applied to all sequences (including those in the validation and test sets) in order to standardize normalization across data. The train / validation / test split was set to absolute numbers of rows, respectively corresponding to 67% / 17% / 16% of the mean number of rows per patient. This results in each patient having their initial 243 rows allocated to their training set and their subsequent 61 rows being allocated to the validation set, with all subsequent rows being allocated to the test set. Patients that did not note enough rows

to qualify for this split were subsequently excluded. The train and validation rows were then converted to prefix samples whilst the test rows remained full sequences for growing predictions later on. Sequences in a batch will necessarily have different lengths, which means that they can't be effectively stacked. This issue is dismantled through the addition of a collation function which separates sequences from their associated labels and converts them to tensors, all while recording and memorizing the original length of each patient sequence. The function then pads the shorter sequences with an amount of zeros corresponding to the longest sequence in the batch, with the stored original sequence lengths helping the LSTM model to discriminate between real and padded datapoints in sequences.

After the sequences are assembled and scaled and the training and validation loaders are created, the LSTM model itself is assembled. This model (consisting of 64 internal states or *hidden units*) takes in sequences of differing lengths and learns longitudinal, temporal patterns embedded within them. The model's *linear head* transforms the model's final hidden state of 64 dimensions down into a singular binary output value, through a logit function. Each forward pass (movement of input data through the network) consists of the same sequence of steps, including packing padded sequences to remove padding before they are processed by the LSTM, running the model itself and obtaining the final hidden state, and passing said final state through the linear head in order to obtain an output logit (this logit is ultimately converted into a probability value using the sigmoid function during the inference phase). The loss function implemented for training the model was selected as binary cross-entropy (BCE) with logit loss, which heavily penalizes incorrect predictions and as such is suited for binary classification problems based on heavily imbalanced datasets, such as the one described in section 4.1 [57]. As part of the BCE process, the minority class (migraine) is assigned a positive weight in order to mitigate the class imbalance within the dataset. This positive weight value was derived by dividing the amount of negative class instances (non-migraine days) by the amount of positive class instances (migraine days). This landed the positive weight value at 7.6, meaning that misidentified migraine rows are penalized 7.6 times as heavily as non-migraine rows. As for the optimizer, ADAM with a fixed learning rate of 0.001 was selected, due to its stochastic nature making it uniquely suitable to LSTM problem formulations [58]. The main model loop runs for up to 15 epochs, with the training phase for each batch consisting of the forward pass, loss computation, backpropagation and weight adjustment stages, with the training loss being averaged across all batches. Gradient updating stops in the validation phase, during which the validation loss using the current model configuration is evaluated and tracked. In order to prevent model overfitting, an early stopping mechanism was also implemented, wherein the best performing model state (lowest validation loss) is dynamically saved. If the validation loss value has not improved for 5 consecutive epochs, training is ceased and the best model configuration is fetched and restored. The LSTM hyperparameters were pre-set as follows: 56 warm-up rows, 15 maximum training epochs, 64 hidden dimensions, 32 samples per batch, learning rate of 0.001, minimal precision of 0.20. The only hyperparameters that were dynamically calculated for the purposes of the model were the positive weight (calculated as a quotient of the amounts of negative and positive cases) and the decision threshold, the tuning of which is described in the following paragraph.

Post-training, the model takes in a history window associated with a given patient and generates a predictive probability for migraine in the row directly succeeding the history window, with said predictive probability ranging between 0 and 1. As the model purely outputs predictive probabilities,

a decision threshold needs to be decided on. This is done through assembling predictions on the entirety of the validation set and sweeping all possible thresholds from 0.01 to 0.70 in increments of 0.01. Recall and precision values are calculated for each threshold evaluated as part of the sweep, with the aim of selecting a threshold that maximizes recall while keeping the precision above or equal to 20%. If no such threshold exists, then the previously obtained threshold with the highest recall is utilized. In order to efficiently deploy the trained model, evaluate its performance, and check whether said performance degrades as predictions move forward in time and away from the warm-up period, a growing prediction approach was implemented. As part of this approach, for each patient's evaluation on the validation set, their initial history is comprised of their entire set of training rows. Similarly, for evaluation on the test set, each patient's history is comprised of both their training and their validation rows, thereby making sure that the model receives as much context as possible in order to make predictions on the test set. For each subsequent row after the initial patient history, a prediction is made using said history, artifacts concerning said prediction are recorded (patient ID, date, true migraine / no migraine label, predicted probability, length of history, and prediction step number), the true outcome value is appended to the history as ground truth feedback, and the model expands onto the next row and repeats the process. The learned decision threshold is then applied to convert the achieved predicted probabilities into binary predictions, after which metrics are calculated and the confusion matrix is assembled. To finalize everything, SHAP values are computed and illustrated in order to give a helpful visual aid on how and to what extent certain predictors influence the final model outcome [56]. SHAP works on the basis of individual marginal predictor contributions, investigating how severely the final prediction changes when a specific feature is used as a predictor, whilst accounting for said feature's interplay with all other features. In practice, this is done using a collection of around 50 patients with around 5 sampled predictions each (for a total of 250), such that the values of each predictor at the last timestep are extracted as a snapshot for each patient. A SHAP explainer is built on top of this snapshot, which then approximates which predictors most heavily influenced predictions on the sequence in question. This is depicted in a *beeswarm plot*, as part of which every feature is assigned a separate row with the x-axis depicting the magnitude of model impact. Points shown in red raise predictions higher, whereas points shown in blue push predictions even lower.

5 Results

This section elaborates on the model results achieved through the configurations discussed in section 4.2, including tables of evaluation metrics, confusion matrices, and plots. These are interpreted further in section 6. All in all, 382 patients took part in the study, totaling to a final amount of 138,613 rows of data. A singular patient with a total of 58 rows of data was filtered out of the dataset, rounding out the final amount of participants to 381, with 138,555 total rows of data. Descriptive statistics about the patient dataset are given in Table 1, with the monthly migraine onset frequency per patient being depicted visually in histogram form in Figure 1.

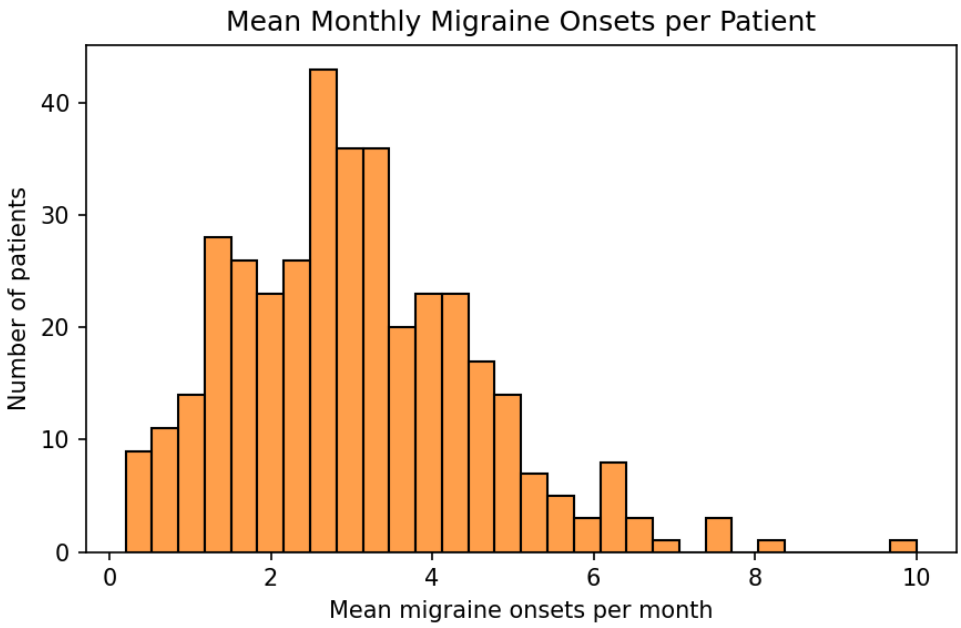

Figure 1: Histogram of mean monthly migraine onset frequency per patient, with a month being defined as a 28-day period and an onset day being defined as the first in a series of migraine days.

5.1 GLMM Results

Using a minimum precision of 0.20, the threshold tuning process for the GLM model settled on a decision threshold of 0.15 (achieving a precision of 0.205 and a recall of 0.478 on the validation set). Using this decision threshold and all predictors (with R automatically dropping mutually explainable columns), the GLMM achieved modest recall values of 0.42 and 0.80 for migraine and no-migraine respectively, thereby achieving respective precision values of 0.21 and 0.91, as can be seen in the full summary in Table 2. The model achieved an overall ROC-AUC score of 0.67. The confusion matrix can be seen in Figure 2, achieving a true negative count of 15691, false positive count of 3870, false negative count of 1479 and true positive count of 1055. The histogram of predicted probabilities by true class (including the tuned threshold) can be seen in Figure 5. The ROC-AUC curve, as well as the precision-recall (PR) plot are given in Figure 3. The patient recall distribution is given in Figure 4, with over 175 patients noting a migraine recall value of 0.

	precision	recall	F1-score	support
no migraine	0.91	0.80	0.85	19,561
migraine	0.21	0.42	0.28	2,543
accuracy			0.76	22,095
macro avg	0.56	0.61	0.57	22,095
weighted avg	0.83	0.76	0.79	22,095

ROC-AUC: 0.67

Table 2: GLMM classification performance, using a decision threshold of 0.15 and a minimal precision of 0.20.

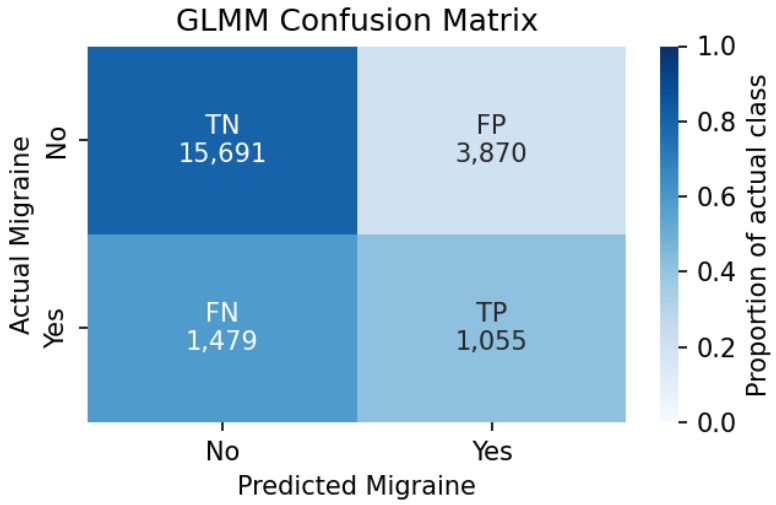

Figure 2: GLMM confusion matrix, using a decision threshold of 0.15 and a minimal precision of 0.20. Rows are normalized to sum up to 1, such that each quadrant represents the proportion of the given actual class assigned to the given predicted class. Darker quadrants on the main diagonal indicate larger proportions of correct predictions, while darker quadrants off the main diagonal indicate larger proportions of errors.

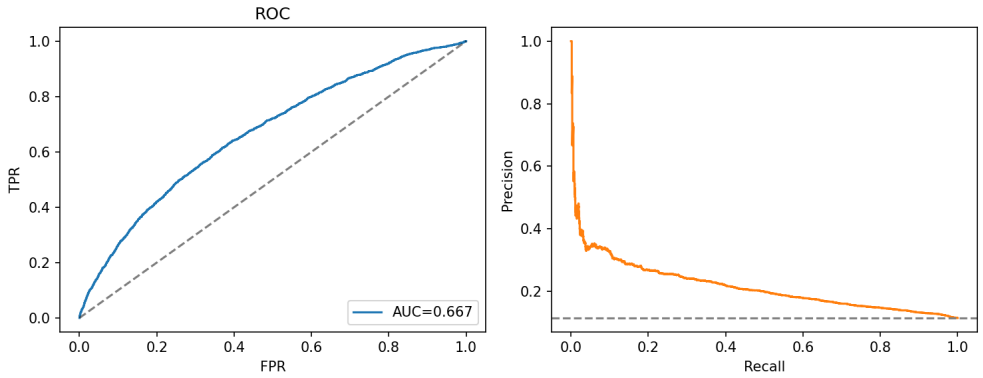

Figure 3: GLMM ROC-AUC (left) and precision-recall (right) curves.

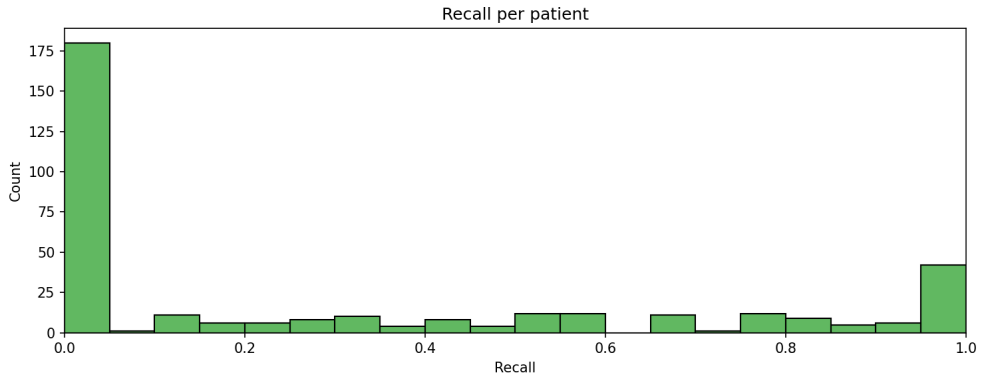

Figure 4: GLMM patient-level recall distribution, pertaining to patients that experienced at least one migraine onset event throughout the study.

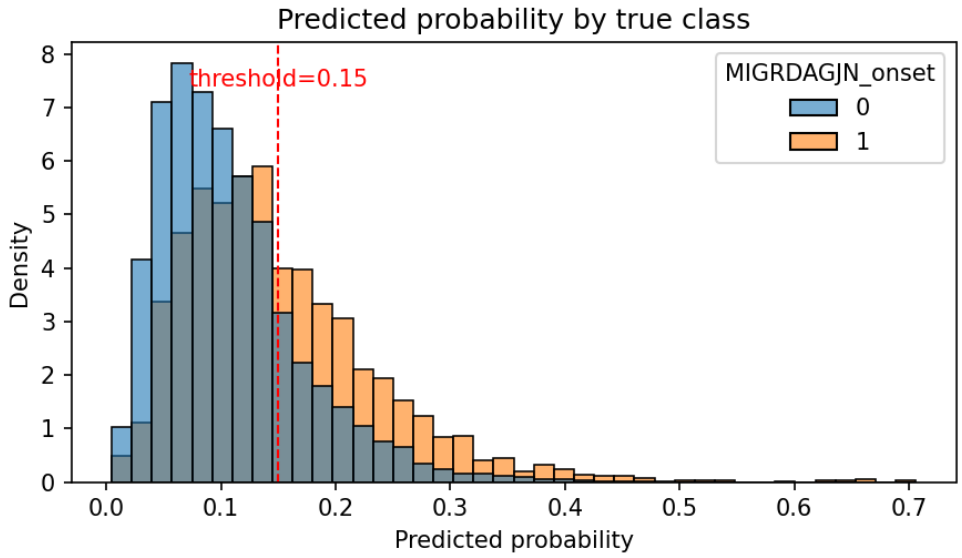

Figure 5: Histogram of GLMM predicted probability by true class, using a decision threshold of 0.15.

5.2 LSTM Results

LSTM evaluation halted at epoch 7 out of 15 due to the early stopping mechanism, as the validation loss had not improved in 5 consecutive epochs. The threshold tuning process (aiming to maximize recall, while keeping a minimal precision of 0.20) settled on a decision threshold of 0.58, thereby achieving a precision score of 0.201, a recall score of 0.411 and an F1 score of 0.270 on the validation set. The final evaluation metrics following test-set evaluation can be found in Table 3, with the model scoring an ROC-AUC of 0.65. The confusion matrix can be seen in Figure 6, achieving a true negative count of 26243, false positive count of 11875, false negative count of 2766 and true positive count of 3116. The histogram of predicted probabilities by true class (including the tuned threshold) can be seen in Figure 7, exhibiting a slight negative skewness. Finally, the SHAP beeswarm plot (Figure 8) visualizes the most significant features on the model output, with pain-coping, well-being and stress taking the top positions.

	precision	recall	F1-score	support
no migraine	0.90	0.69	0.78	38,118
migraine	0.21	0.53	0.30	5,882
accuracy			0.67	44,000
macro avg	0.56	0.61	0.54	44,000
weighted avg	0.81	0.67	0.72	44,000

ROC-AUC: 0.65

Table 3: LSTM classification performance, using a decision threshold of 0.58 and a minimal precision of 0.20.

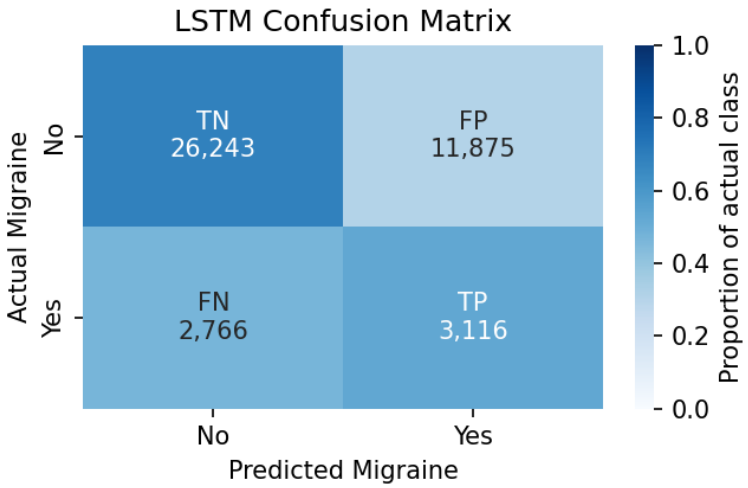

Figure 6: LSTM confusion matrix, using a decision threshold of 0.58 and a minimal precision of 0.20. Rows are normalized to sum up to 1, such that each quadrant represents the proportion of the given actual class assigned to the given predicted class. Darker quadrants on the main diagonal indicate larger proportions of correct predictions, while darker quadrants off the main diagonal indicate larger proportions of errors.

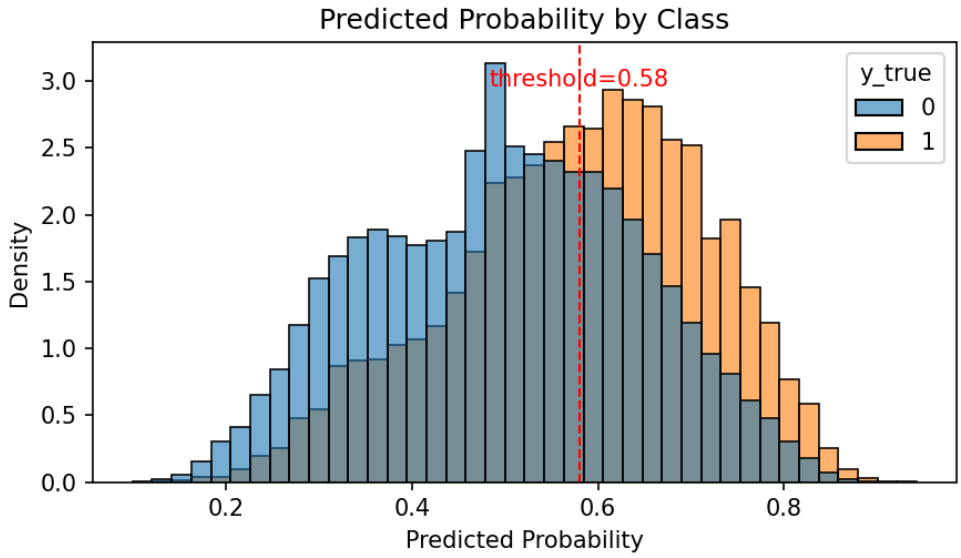

Figure 7: LSTM predicted probability by true class, using a decision threshold of 0.58.

6 Discussion

6.1 Comparison with Previous Literature

Upon situating the results outlined in section 5 within the literature discussed in subsection 2.6, the results prove to be above par. Both the GLMM's AUC of 0.67 and the LSTM's AUC of 0.65 exceed Holsteen et al.'s reported C-statistic of 0.56 [43] (which is equivalent to the AUC for binary predictions [44]), and are comparable to Stubberud et al.'s best test-set AUC value of 0.62 [45]. It is important to note that Stubberud et al. additionally benefited from the use of wearable biofeedback data in tandem with self-reported e-diary data. The lack of biofeedback data somewhat limits the predictive potential of our models, since biofeedback data has proven to be immensely useful when it comes to RNN-based migraine prediction (as evidenced by Faisal et al.'s GRU-D model achieving an AUC of 0.84 on a dataset consisting of both e-diary and biofeedback data [47]). Despite its prevalence and seemingly universal status as the preferred evaluation metric within the literature on migraine forecasting, AUC offers quite a limited view into what a model is actually capable of. AUC simply gauges the model's ability to correctly discriminate positive cases from negative ones, meaning that in cases concerning imbalanced datasets, using AUC as a metric alone would result in a skewed representation of model performance due to the over-representation of true-negative cases (the majority class) within the underlying dataset. Different research papers differ in the base attack rate present in their respective datasets (Holsteen et al.'s of around 17%, Faisal et al.'s of about 19%, and the LUMC's of about 11.9%), with models facing a more imbalanced dataset being tasked with a more difficult classification problem that a simple AUC score does not give meaningful information on. Furthermore, predicting migraine attack onset (as opposed to a headache day in general, which is what Faisal et al. [47] and Stubberud et al. [45] target) is more challenging by default, as the model cannot make use of the autocorrelation present between days comprising a

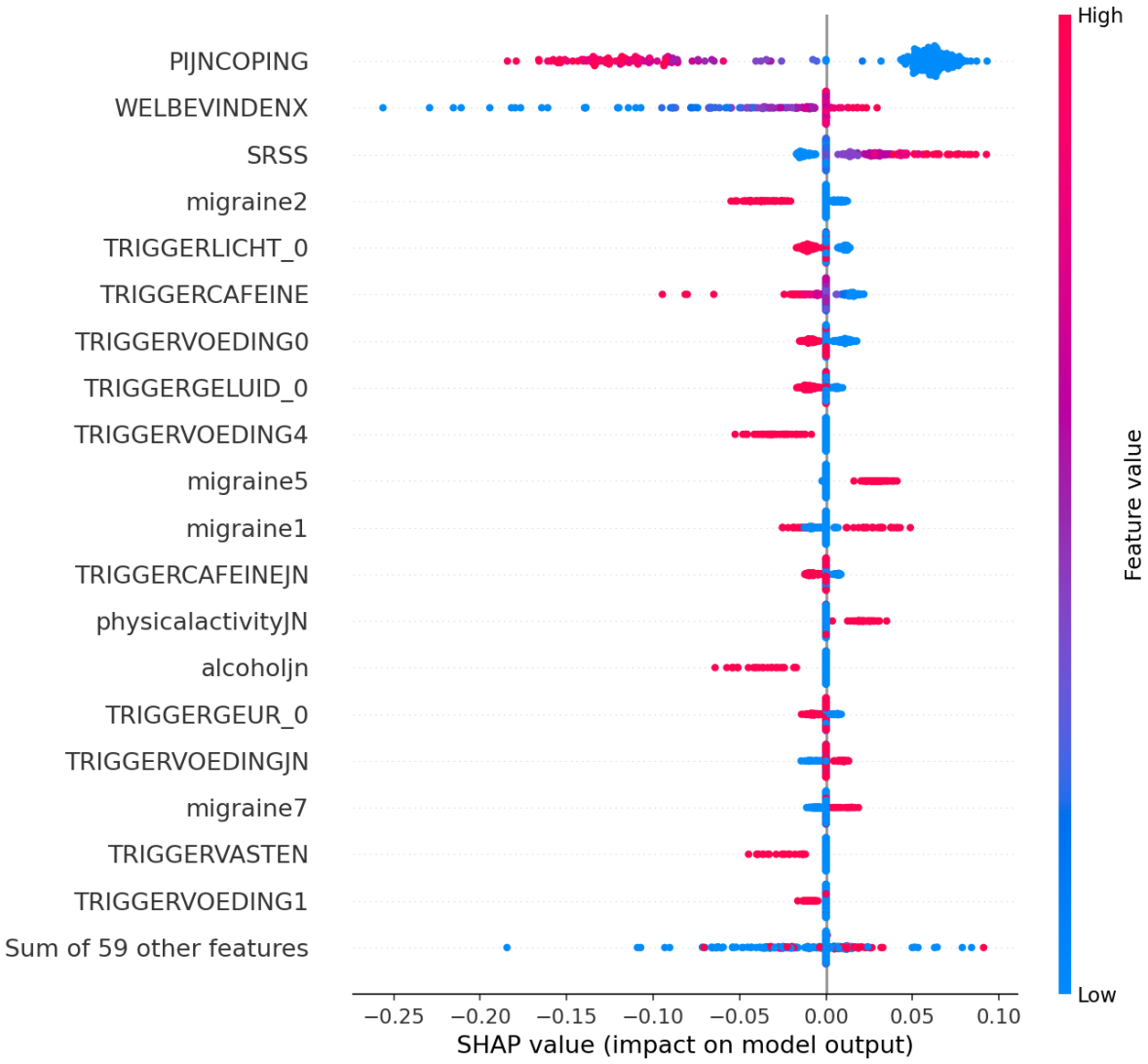

Figure 8: SHAP feature importance beeswarm plot, identifying the most impactful features (as well as their direction of impact on final predictions) included as part of the LSTM.

multiple-day migraine attack. Casting aside AUC, recall on the positive (migraine) class (when held against a strict precision constraint in order to avoid all cases uniformly being classified as migraine) proves to be the most viable metric on which to judge model performance. Recall is defined as the proportion of true migraine onset days that are correctly identified as such by the model, owing its clinical significance to the fact that the burden inflicted on a patient due to a false alarm is comparatively much lower than the burden inflicted by a missed migraine attack. A missed attack gives the patient in question no opportunity to anticipate and take timely preventative action in the lead-up to an impending migraine attack, which, depending on the severity of the attack and the sensitivity of the patient, may result in debilitating consequences. The consequences of a false alarm, on the other hand, are much milder, merely amounting to unnecessary precautions taken and some additional stress. At a precision level of 0.21, both the GLMM's achieved recall of 0.42 and the LSTM's achieved recall of 0.53 provide modest but solid results, showing the promise of ML models in the realm of migraine prediction, subject to further improvements and more extensive data.

6.2 GLMM vs. LSTM Performance Comparison

That being said, the GLMM and LSTM achieving comparable recall results, with recall being established as the most significant metric from a clinical standpoint, is quite an unexpected finding (said findings are reported in section 5). Given the fact that the LSTM is significantly more architecturally complex than the GLMM, as well as LSTM models specifically being designed in order to carefully detect and learn even small-scale temporal patterns in longitudinal data, the natural expectation would've been for LSTM to outperform the GLMM by a significant margin. Nevertheless, this architectural superiority was not shown to translate into a significant clinical advantage. The simplest explanation for this result is that the scope and content of the dataset in question is simply insufficient when it comes to utilizing the full predictive power of the LSTM architecture, the utility of which only becomes apparent when deployed on much larger datasets with patient observations spanning longer temporal scales. Additionally, said observations ideally should not consist purely of self-reported values (which are inherently noisy), but should rather incorporate these self-reported values alongside directly measurable ones. Bearing in mind that the median number of rows per patient totals to 354, 243 of which are dedicated to training, the model is only presented with roughly 8.6 months of e-diary data to learn from. This is an improvement over other studies discussed in section 2.6.1, most of which employ much more modest magnitudes of patient data [40], [43], but is nevertheless relatively short when accounting for the vast daily variability in trigger and diary data. Practically speaking, however, managing to gather (on average) almost 9 months of consistent patient data for training purposes is already a non-negligible accomplishment, given that previous work within the LUMC utilized comparatively modest amounts of 2-3 months of data. Any future work which aims to expand the temporal scope of the data collected may run into practical issues when getting patients to agree to track their data for such prolonged periods of time in a consistent manner, and even though any potential inconsistencies may be bridged through, for example, retroactively filling in observations, doing so would introduce additional levels of recall bias within the reported data and may even hinder predictive performance. Furthermore, several factors work to introduce additional noise into the dataset, including but not limited to: the self-reported and personally interpretable status of some diary measures (especially as it pertains to variables such as well-being and stress), patients

reserving the ability to fill diary entries in retrospectively, as well as the patient-specific nature of the relationship between a particular trigger and migraine onset. With all of these factors at play, the temporal scale of the data may simply be too small for the LSTM to effectively discriminate said noise from meaningful information. This is further complicated by the train / validation / test split being based on patient row counts, which inevitably means that patients with shorter histories contribute comparatively fewer training samples to the LSTM, further complicating the learning task by introducing additional randomness that is not temporally extensive enough to be effectively learned (randomness which the GLMM can handle much better, owing to its random patient-level intercepts). This explanation is further aided by findings from Faisal et al., who attribute the success of their RNN-based GRU-D model on the fact that it was trained on around 180 observations per patient on a daily basis, claiming further that the advantages of time-series models for attack forecasting can only fully shine through when presented with dense, regular and temporally long observations [47]. Additionally, the manner in which both models approach between-patient randomness is fundamentally different. The usage of random intercepts within the GLMM architecture provides an easy and logical way for baseline differences in attack propensity between patients to be encapsulated, with each patient being assigned an intercept that is derived from their own historical data. The LSTM is not explicitly equipped with any such mechanism, and so must learn to represent patient-level randomness throughout its architecture in a much more abstract manner. Finally, the LSTM evaluation halting on epoch 7 out of 15 due to a lack of improvement in validation loss, and still managing to achieve a solid recall value of 0.58 seems to indicate that the architecture is broadly suitable for this type of problem formulation, and that the issue does in fact lie with the scope of the data. From the perspective of clinical utility, both models testify to the potential of e-diary data when it comes to migraine attack forecasting, even though results may be more modest than anticipated. Despite being too underdeveloped to be used as a standalone forecasting technique, their predictions (specifically in the format of predicted probabilities, which could be perceived as proxies for migraine risk) could aid neurologists in the assessment of an individual patient's migraine risk from their respective e-diary data, thereby offering additional insight into said patient's sensitivity to triggers and propensity for migraine.

6.3 Recall and Threshold Tuning

As mentioned in the previous subsection, the clinical consequences of a false alarm vs. a missed attack are incomparable. An attack that is not preceded by a valid warning renders patients unable to take preventative action in a timely manner, whether that be avoiding triggers or taking preventative medication. A false alarm, however, constitutes a minor inconvenience in most cases, aside from the undue stress it might cause. The asymmetrical nature of this clinical burden is precisely why recall should be prioritized as a metric within episodic migraine forecasting efforts, especially when attacks are infrequent yet damaging. Hence, recall was chosen as the metric to focus on when tuning the binary decision threshold. In order to ensure that the models don't maximize recall by simply classifying every single instance as a migraine day (which would result in a precision of 11.9%, corresponding with the proportion of migraine rows in the dataset), a minimal precision constraint of 0.20 was introduced, motivating the model to be more selective whilst still prioritizing recall. The results show that, even at low precision levels, both models can manage to identify a meaningful subset of onset days, which lends them some level of clinical credibility.

Taking a look at the GLMM patient-level recall distribution depicted in Figure 4, however, shows the true variability in patient-level recall values, as well as how misleading the value of 0.42 might be. Namely, a large number of patients had every single instance of migraine onset be missed by the model, despite the threshold having been tuned specifically to maximize recall on the population level. This indicates that the threshold which correctly identifies most migraine cases overall may not be the most universally useful one. With the current formulation, patients whose predicted probabilities for migraine onset tend to fall below the 0.15 threshold (either due to a low attack frequency or their low sensitivity to triggers) will never have their attacks correctly predicted without generating a significant number of false alarms. The peak on the far right side of the histogram indicates that this approach works exceedingly well for a different subset of the patient population, meaning that a viable solution for future research could be to tune the decision threshold for each individual patient and their historical attack pattern, rather than applying a uniform threshold to an otherwise very heterogeneous population when it comes to triggers. It is important to note that this limitation could partially be attributable to the GLMM's random intercept element, as the decision threshold is not adapted to the historical baseline attack propensity assigned to each patient. Namely, a patient who is assigned a low intercept (and is therefore treated as low-risk by the model) might have their predicted probabilities pushed down below the population-wide threshold even on days when the fixed effects elements point toward high migraine probability, which might push patients with inherently low baseline intercepts deeper into zero-recall territory. The need for patient-level decision thresholds becomes even more apparent when taking a look at the predicted probability histograms. The GLMM predicted probability histogram (depicted in Figure 5) shows that probabilities clearly trend towards 0, with a large degree of overlap between the migraine and no-migraine distributions. This level of overlap directly informs the low decision threshold of 0.15, as the model rarely ever predicts probabilities above 0.3 (even on actual migraine days) and stays very close to the base rate of migraine prevalence in the dataset, making differentiation between the population-level migraine onset rate and individual low-risk migraine days more difficult. The heavy positive skewness of the histogram itself (with some high-probability observations still being noted, presented at the far right end of the histogram) aids this interpretation, additionally indicating that the model is capable of predicting high-risk migraine onset days, but struggles more the closer to the base rate the predictions get. Due to the positive class weighting applied during the training phase significantly shifting model outputs toward identifying true migraine onset days, the LSTM predicted probability histogram (depicted in Figure 7) is much more evenly distributed, with values ranging from 0.2 to 0.9, centered near 0.5. The migraine class is noticeably shifted further to the right and more negatively skewed when compared to the no-migraine class, which is situated more heavily around the high-probability area above 0.7, seemingly indicating that the LSTM has, in fact, learned to separate the classes from each other to a certain extent and is more likely to assign higher probabilities to observations than the GLMM. The no-migraine class is mostly clustered between 0.3 and 0.6, with the decision threshold of 0.58 being located close to the crossing point of both distributions. Compared to the GLMM, this wider spread of predicted probabilities seems to indicate that the LSTM investigates a wider array of potential probabilities, rather than simply clustering around the base rate of migraine onset as the GLMM has been shown to do, which makes the LSTM more sensitive to data deviating from the population mean. Practically, this might give the LSTM a slight clinical edge in more ambiguous cases, as its wider spectrum of

probabilities renders it more likely to interpret middle-of-the-road cases as potential migraine cases and call for action accordingly, rather than to risk missing out and potentially inflicting undue harm.

6.4 Validation and Hyperparameters

As the same validation set is utilized for both early stopping and decision threshold selection, a small but meaningful dependency is introduced within the LSTM model. Namely, the early stopping loop selects the best model state based on stagnating validation loss, and the threshold tuning loop checks through and selects the most optimal decision cutoff based on precision and recall on that same held-out validation set. This results in the selected decision threshold not being fully independent of the model that it is then applied to, causing the metrics reported on the validation set to be a bit more optimistic than they would be otherwise. As there is still no data leakage between the validation and test sets and the test set remains fully held out, the test set metrics remain valid and untouched by either of the two processes. In the future, for a more robust and less optimistic implementation, separate subsets of the validation set should be used to carry out threshold tuning and early stopping independently of each other. Another issue arises from the low amount of hyperparameters that were tuned. The positive class weight of 7.6 was derived mathematically from the quotient of the number of exemplars in both classes and was applied as-is, without any actual tuning procedure. Even though this is common practice, the positive class weight is, in fact, a very important hyperparameter when it comes to controlling the precision-recall trade-off both during the training phase and during the test phase. Too high of a weight could lead to even higher recall prioritization at the cost of precision, whereas too low of a weight could make the final predictions overly selective. The optimal weight can then only be isolated through optimization, which should be addressed in future work on the topic. Several other LSTM parameters (namely, the number of hidden dimensions, batch size, learning rate, number of layers and number of warm-up rows) were pre-set as well, and future results could certainly benefit from a thorough hyperparameter search, especially when considering the possibility of potential clinical deployment.

6.5 SHAP Analysis

According to the SHAP beeswarm plot generated from the LSTM predictions depicted in Figure 8 (with the methodology of SHAP being elaborated on in subsection 4.2.2), the most influential features on the final model predictions were pain-coping, well-being and stress levels, which are interesting for a variety of reasons. As the most significant predictor, the SHAP analysis showed that lower levels of pain-coping ability are closely correlated with a higher risk of migraine onset. Low pain-coping ability might appear as a premonitory symptom due to a patient experiencing heightened sensitivity in the lead-up to a migraine attack [7], thereby foreshadowing an imminent migraine day. Well-being ranks as the second most significant predictor, with low well-being scores serving as indication of an imminent migraine onset event, and stress ranks third, with heightened levels of stress increasing the predicted probability of an attack. Both of these observations are consistent with the literature on migraine triggers and premonitory symptoms, where these relationships have been extensively cataloged [7], [13]. It is interesting to note that the top three

highest ranking variables are all continuous in nature, as the presence of an interval of values rather than a yes / no binary indicator allows for a much more nuanced relationship between each feature level and the final outcome. Keeping in mind that all of the aforementioned variables are self-reported (and therefore equally prone to recall bias) this phenomenon may reflect the known tendency of feature attribution methods (such as SHAP) to assign greater importance to predictors measured on scales with more levels (here, continuous, as opposed to the binary nature of most other predictors used), overblowing their predictive utility as a result [59]. The following variables are all differing in type, with both lagged migraine days and exposures to lights, sounds and problematic foods being amongst the most relevant predictors. The inclusion of the migraine lag variables may indicate that the LSTM has learned to exploit some level of autocorrelation between migraine attacks, estimating future onset based on recent attack history. Another interesting observation is the high proportion of values lying on the 0-line, indicating that they only negligibly contribute to the final model output, if at all. This appears to be a consequence of both their binary nature, as well as the low proportion of non-zero values for these features across the dataset. Since most binary triggers given are only reported by patients in a small subset of observations, these features take on a value of 0 for the majority of the rows in the dataset. The LSTM subsequently learns that a value of 0 carries little extra information beyond the baseline level, and assigns it a SHAP score that approximates 0. The minority of observations where a given trigger is actually present contribute to the prediction in more meaningful ways and are thus assigned non-zero SHAP values, the predictive value of which may or may not be diluted based on the prevalence of zero values across that same feature. This could partially also explain why the continuously measured features display such a predictive predominance, since they are reported much more consistently across many more levels. The wide array of different triggers and their differing frequencies of appearance (as well as different patients being affected by them differently) significantly complicate the prediction task, resulting in the LSTM sourcing most of its predictive power from the values that have shown to be the most consistent: pain-coping, well-being and stress.

7 Limitations & Directions for Future Work

As only the decision threshold was actually tuned for both models (with the LSTM positive class weight being chosen dynamically from the data distribution, and therefore not truly tuned), future research could greatly benefit from more robust hyperparameter tuning efforts, in order to strike the perfect balance between model performance and clinical applicability. For the LSTM in particular, the hyperparameter space should be searched in order to optimize all of the model's hyperparameters, including the number of hidden dimensions (drawing comparisons between 32, 64 and 128 units), number of stacked layers, the learning rate (which can be explored through a thorough grid search), the batch size (16, 32, or 64), as well as the warm-up row count (starting at 28, 56 or 84 rows corresponding to one, two or three months respectively). Given the intricate ways in which all of these hyperparameters interact, a robust solution can only be reached if all of them are jointly optimized, which could be possible through a randomized search over the entire hyperparameter space, followed by more localized grid searches in regions identified as promising. Said optimizations would undoubtedly lead to an improvement in model output when compared to fixed, pre-set values. The positive class weight and decision threshold also ought to be tuned jointly rather than sequentially, as both play a role in the precision-recall trade off at different steps of

the model's evaluation (namely, the positive class weight informs the initial learned probability distribution during the training phase, whereas the decision threshold transforms said distribution into individual binary predictions at the end of the pipeline). This could be addressed through training a set of models independently and employing one of several potential positive weight values on each, with a decision threshold being tuned separately for each model on a validation set distinct from the one used for early stopping. Under these conditions, the selected positive weight would be certain to maximize migraine recall (of course, subjected to a minimal precision value).

Future work would also unquestionably benefit from the introduction of patient-specific (rather than population-wide) decision thresholds, as the population-wide decision thresholds employed within this paper were shown to fail certain subsets of patients whose attack patterns might have been too weak or inconsistent to properly predict. A patient-specific threshold would mitigate this issue through careful calibration (even for patients whose attack pattern may be more sparse of difficult to learn), allowing for potential onset events to be more effectively flagged for both high- and low-risk patients. This threshold could be selected on top of the population-wide threshold based on each individual patient's predictions on the validation set, with the individual threshold being derived through either increasing or decreasing the global threshold in accordance with each patient's personal migraine attack probability distribution.

The final major limitation of the approach outlined in the preceding sections is the limited sample size, both in regards to the raw number of patients enrolled within the study cohort, as well as in regards to the rows of data obtained from each patient. Firstly, the dataset is quite severely imbalanced with respect to patient sex, as only 56 out of the 381 patients indicated their sex as male, which necessarily biases predictions in favor of patients who have indicated their sex as female. Future research should ideally employ a larger sample size in order to improve predictions, more effectively filter out noise and be able to verify that accurate predictions can be made for highly diverse patient cohorts, both demographically and geographically. Expanding recruitment to multiple headache research groups, ideally including ones abroad, would help with the generalizability of the findings presented above beyond the LUMC (and web-recruited, yet still exclusively Dutch-speaking) patient population. The relatively small temporal scale additionally meant that the LSTM model's predictive capabilities were not utilized to their full potential. The LSTM's key advantage over the GLMM lies in its ability to identify underlying patterns across extended time periods, an dvantage which can only make itself apparent when the training data used is long-form and consistent enough to allow for such patterns to emerge. In the future, in order to avoid subpar results originating from an otherwise extremely promising model architecture, researchers should ideally look at maximizing the temporal scale and obtaining patient-level data in the longer term, such that more migraine onset events can be observed, more accurate predictions can be made and small-level noise in the dataset can be more efficiently ignored. An increase in dataset size would come with the added benefit of ensuring that the chronological train / validation / test split within the patient data is accurate and representative of each patient's true attack frequency. This point of inquiry could also be concretely explored in the future, looking into how exactly model performance changes with increasing levels of patient-level data and where an optimal convergence point (after which all predictions can be deemed useful) would lie, if anywhere at all. The importance of integrating measurable physiological data alongside e-diary data within future migraine forecasting endeavors cannot be overstated, as doing so would aid both to alleviate

some of the recall bias inherent to self-reported metrics, as well as to contribute larger volumes of data for training and prediction.

8 Conclusion

The promise of automated migraine attack onset prediction has become an issue of increasing interest within the realm of neurology, motivated by the principle that timely and accurate anticipation of attacks facilitates preventative intervention and therefore alleviates the burden of a migraine attack on patients' quality of life. This paper has investigated and evaluated two separate, yet complementary machine learning approaches to this problem, both of which had historically been employed in the literature: a generalized linear mixed-effects model, as well as a long short-term memory model, with both models being trained and evaluated on a dataset gathered from 381 episodic migraine patients at the Leiden Headache Center of the LUMC. The dataset in question was comprised of 138,555 daily observations spanning a wide spectrum of triggers, symptoms, lifestyle and health factors. Post-evaluation (held under a strict minimal precision value of 0.20 to retain clinical significance), the GLMM noted an AUC score of 0.67 and a migraine recall score of 0.42, whilst the LSTM noted an AUC score of 0.65 and a migraine recall score of 0.53. The similarity of results between the two models (in spite of the LSTM's significantly superior model architecture when it comes to learning from time-series data) is quite an interesting finding, as it seemingly indicates that the main issues hindering progress on this migraine attack forecasting problem may hinge on the self-reported nature and low volume of the data itself, rather than on the particular model architectures. A valuable avenue to be explored in this regard would be the integration of physiological biofeedback data that is directly measured on the patient level alongside self-reported diary data, with such approaches in the literature showing promising results [45], [47]. Furthermore, as evidenced by the GLMM's modest population-wide recall score of 0.42 (despite accounting for patient-level variability using random intercepts), population-level modeling approaches ultimately prove inadequate toward approaching the true patient-level variability inherent to clinical data. This calls for a shift towards more personalized modeling approaches that estimate and apply feature weights and decision thresholds individually for each patient, based exclusively off of said patient's attack history. Future research and results may greatly benefit from such an approach, doubly so if additional metrics (such as the positive class weight) are treated as primary model hyperparameters and actively tuned. Both approaches investigated as part of this paper show that longitudinally collected e-diary data, in combination with ML approaches, can yield clinically promising results, and are thus an avenue worth exploring in future neurological research. Effectively exploring this avenue would mean gathering more objective data on a longer temporal scale, careful modeling to account for patient-specific variability, and thorough hyperparameter tuning in order to arrive at an optimal model configuration, all of which would aid in increasing the robustness and clinical reliability of the models developed.

References

[1] "Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders, 3rd edition," en, *Cephalalgia*, vol. 38, no. 1, pp. 1–211, Jan. 2018, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/0333102417738202. Accessed: Aug. 25, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0333102417738202.

[2] T. N. Ward, "Migraine Diagnosis and Pathophysiology," en, *Continuum*, vol. 18, no. 4, pp. 753–763, Aug. 2012, ISSN: 1080-2371, 1538-6899. DOI: 10.1212/01.CON.0000418640.07405.31. Accessed: Jun. 2, 2026. [Online]. Available: https://continuum.aan.com/doi/10.1212/01.CON.0000418640.07405.31.

[3] C. Lucas, "Migraine with aura," en, *Revue Neurologique*, vol. 177, no. 7, pp. 779–784, Sep. 2021, ISSN: 00353787. DOI: 10.1016/j.neurol.2021.07.010. Accessed: Jun. 2, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0035378721006214.

[4] J. Lai and E. Dilli, "Migraine Aura: Updates in Pathophysiology and Management," en, *Current Neurology and Neuroscience Reports*, vol. 20, no. 6, p. 17, Jun. 2020, ISSN: 1528-4042, 1534-6293. DOI: 10.1007/s11910-020-01037-3. Accessed: Jun. 3, 2026. [Online]. Available: https://link.springer.com/10.1007/s11910-020-01037-3.

[5] R. B. Lipton, J. M. Pavlovic, S. R. Haut, B. M. Grosberg, and D. C. Buse, "Methodological Issues in Studying Trigger Factors and Premonitory Features of Migraine," en, *Headache: The Journal of Head and Face Pain*, vol. 54, no. 10, pp. 1661–1669, Nov. 2014, ISSN: 0017-8748, 1526-4610. DOI: 10.1111/head.12464. Accessed: Jun. 5, 2026. [Online]. Available: https://headachejournal.onlinelibrary.wiley.com/doi/10.1111/head.12464.

[6] D. C. Buse, M. F. Rupnow, and R. B. Lipton, "Assessing and Managing All Aspects of Migraine: Migraine Attacks, Migraine-Related Functional Impairment, Common Comorbidities, and Quality of Life," en, *Mayo Clinic Proceedings*, vol. 84, no. 5, pp. 422–435, May 2009, ISSN: 00256196. DOI: 10.4065/84.5.422. Accessed: May 25, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0025619611605612.

[7] G. Sebastianelli et al., "Insights from triggers and prodromal symptoms on how migraine attacks start: The threshold hypothesis," en, *Cephalalgia*, vol. 44, no. 10, p. 03 331 024 241 287 224, Oct. 2024, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024241287224. Accessed: Jun. 5, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024241287224.

[8] Z. Katsarava, D. C. Buse, A. N. Manack, and R. B. Lipton, "Defining the Differences Between Episodic Migraine and Chronic Migraine," en, *Current Pain and Headache Reports*, vol. 16, no. 1, pp. 86–92, Feb. 2012, ISSN: 1531-3433, 1534-3081. DOI: 10.1007/s11916-011-0233-z. Accessed: May 25, 2026. [Online]. Available: http://link.springer.com/10.1007/s11916-011-0233-z.

[9] G. M. Terwindt, M. D. Ferrari, and L. J. Launer, "The impact of headache on quality of life," en, *The Journal of Headache and Pain*, vol. 4, no. S1, s35–s41, Mar. 2003, ISSN: 1129-2369. DOI: 10.1007/s101940300007. Accessed: May 31, 2026. [Online]. Available: https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1007/s101940300007.

[10] T. Shimizu et al., "Disability, quality of life, productivity impairment and employer costs of migraine in the workplace," en, *The Journal of Headache and Pain*, vol. 22, no. 1, p. 29, Dec. 2021, ISSN: 1129-2369, 1129-2377. DOI: 10.1186/s10194-021-01243-5. Accessed: Jun. 2, 2026. [Online]. Available: https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1186/s10194-021-01243-5.

[11] P. R., N. S. C., H. S., and R. K., "Migraine Disability, Quality of Life, and Its Predictors," en, *Annals of Neurosciences*, vol. 27, no. 1, pp. 18–23, Jan. 2020, ISSN: 0972-7531, 0976-3260. DOI: 10.1177/0972753120929563. Accessed: May 25, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0972753120929563.

[12] L. Dong, W. Dong, Y. Jin, Y. Jiang, Z. Li, and D. Yu, "The Global Burden of Migraine: A 30-Year Trend Review and Future Projections by Age, Sex, Country, and Region," en, *Pain and Therapy*, vol. 14, no. 1, pp. 297–315, Feb. 2025, ISSN: 2193-8237, 2193-651X. DOI: 10.1007/s40122-024-00690-7. Accessed: Jun. 2, 2026. [Online]. Available: https://link.springer.com/10.1007/s40122-024-00690-7.

[13] L. Kelman, "The Triggers or Precipitants of the Acute Migraine Attack," en, *Cephalalgia*, vol. 27, no. 5, pp. 394–402, May 2007, ISSN: 0333-1024, 1468-2982. DOI: 10.1111/j.1468-2982.2007.01303.x. Accessed: May 25, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1111/j.1468-2982.2007.01303.x.

[14] D. P. Turner, A. D. Lebowitz, I. Chtay, and T. T. Houle, "Forecasting Migraine Attacks and the Utility of Identifying Triggers," en, *Current Pain and Headache Reports*, vol. 22, no. 9, p. 62, Sep. 2018, ISSN: 1531-3433, 1534-3081. DOI: 10.1007/s11916-018-0715-3. Accessed: May 25, 2026. [Online]. Available: http://link.springer.com/10.1007/s11916-018-0715-3.

[15] S. J. Peroutka, "What Turns on a Migraine? A Systematic Review of Migraine Precipitating Factors," en, *Current Pain and Headache Reports*, vol. 18, no. 10, p. 454, Oct. 2014, ISSN: 1531-3433, 1534-3081. DOI: 10.1007/s11916-014-0454-z. Accessed: Aug. 25, 2026. [Online]. Available: http://link.springer.com/10.1007/s11916-014-0454-z.

[16] G. Dumkrieger, "The promise of machine learning in predicting migraine attacks," en, *Cephalalgia*, vol. 45, no. 11, p. 03 331 024 251 391 207, Nov. 2025, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024251391207. Accessed: May 25, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024251391207.

[17] S. F. Hassan, "Predicting Migraine Episodes with Deep Learning: Enabling Patients to Anticipate and Prevent Migraine Onset," M.S. thesis, University of Pisa, Pisa, 2024.

[18] S. K. Parikh and S. D. Silberstein, "Preventive Treatment for Episodic Migraine," en, *Neurologic Clinics*, vol. 37, no. 4, pp. 753–770, Nov. 2019, ISSN: 07338619. DOI: 10.1016/j.ncl.2019.07.004. Accessed: May 25, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0733861919300647.

[19] W. Lee and M. K. Chu, "The Current Role of Artificial Intelligence in the Field of Headache Disorders, with a Focus on Migraine: A Systemic Review," en, *Headache and Pain Research*, vol. 26, no. 1, pp. 48–65, Feb. 2025, ISSN: 3022-9057, 3022-4764. DOI: 10.62087/hpr.2024.0024. Accessed: Jun. 8, 2026. [Online]. Available: http://e-hpr.org/journal/view.php?doi=10.62087/hpr.2024.0024.

[20] W. W. Stroup, M. Ptukhina, and J. Garai, *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, en, 2nd ed. New York: Chapman and Hall/CRC, Mar. 2024, ISBN: 978-0-429-09206-0. DOI: 10.1201/9780429092060. Accessed: May 25, 2026. [Online]. Available: https://www.taylorfrancis.com/books/9780429092060.

[21] UCLA Statistical Consulting Group, *Introduction to Generalized Linear Mixed Models*, 2026. [Online]. Available: https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/.

[22] R. C. Staudemeyer and E. R. Morris, *Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks*, Version Number: 1, 2019. DOI: 10.48550/ARXIV.1909.09586. Accessed: May 31, 2026. [Online]. Available: https://arxiv.org/abs/1909.09586.

[23] W. Z. Yeh, L. Blizzard, and B. V. Taylor, "What is the actual prevalence of migraine?" en, *Brain and Behavior*, vol. 8, no. 6, e00950, Jun. 2018, ISSN: 2162-3279, 2162-3279. DOI: 10.1002/brb3.950. Accessed: May 31, 2026. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/brb3.950.

[24] Z. Wang, X. Yang, B. Zhao, and W. Li, "Primary headache disorders: From pathophysiology to neurostimulation therapies," en, *Heliyon*, vol. 9, no. 4, e14786, Apr. 2023, ISSN: 24058440. DOI: 10.1016/j.heliyon.2023.e14786. Accessed: May 31, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S240584402301993X.

[25] A. May and L. H. Schulte, "Chronic migraine: Risk factors, mechanisms and treatment," en, *Nature Reviews Neurology*, vol. 12, no. 8, pp. 455–464, Aug. 2016, ISSN: 1759-4758, 1759-4766. DOI: 10.1038/nrneurol.2016.93. Accessed: Jun. 3, 2026. [Online]. Available: https://www.nature.com/articles/nrneurol.2016.93.

[26] on behalf of Lifting The Burden: the Global Campaign against Headache, T. J. Steiner, L. J. Stovner, R. Jensen, D. Uluduz, and Z. Katsarava, "Migraine remains second among the world's causes of disability, and first among young women: Findings from GBD2019," en, *The Journal of Headache and Pain*, vol. 21, no. 1, 137, s10194–020–01208–, Dec. 2020, ISSN: 1129-2369, 1129-2377. DOI: 10.1186/s10194-020-01208-0. Accessed: Jun. 2, 2026. [Online]. Available: https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1186/s10194-020-01208-0.

[27] F. Radat and J. Swendsen, "Psychiatric Comorbidity in Migraine: A Review," en, *Cephalalgia*, vol. 25, no. 3, pp. 165–178, Mar. 2005, ISSN: 0333-1024, 1468-2982. DOI: 10.1111/j.1468-2982.2004.00839.x. Accessed: Jun. 2, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1111/j.1468-2982.2004.00839.x.

[28] G. E. Tietjen, N. A. Herial, J. Hardgrove, C. Utley, and L. White, "Migraine Comorbidity Constellations," en, *Headache: The Journal of Head and Face Pain*, vol. 47, no. 6, pp. 857–865, Jun. 2007, ISSN: 0017-8748, 1526-4610. DOI: 10.1111/j.1526-4610.2007.00814.x. Accessed: Jun. 2, 2026. [Online]. Available: https://headachejournal.onlinelibrary.wiley.com/doi/10.1111/j.1526-4610.2007.00814.x.

[29] W. N. Burton, S. H. Landy, K. E. Downs, and M. C. Runken, "The Impact of Migraine and the Effect of Migraine Treatment on Workplace Productivity in the United States and Suggestions for Future Research," en, *Mayo Clinic Proceedings*, vol. 84, no. 5, pp. 436–445, May 2009, ISSN: 00256196. DOI: 10.4065/84.5.436. Accessed: Jun. 2, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0025619611605624.

[30] K. G. Vetvik and E. A. MacGregor, "Sex differences in the epidemiology, clinical features, and pathophysiology of migraine," en, *The Lancet Neurology*, vol. 16, no. 1, pp. 76–87, Jan. 2017, ISSN: 14744422. DOI: 10.1016/S1474-4422(16)30293-9. Accessed: May 31, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1474442216302939.

[31] L. Kelman, "Migraine Changes with Age: IMPACT on Migraine Classification," en, *Headache: The Journal of Head and Face Pain*, vol. 46, no. 7, pp. 1161–1171, Jul. 2006, ISSN: 0017-8748, 1526-4610. DOI: 10.1111/j.1526-4610.2006.00444.x. Accessed: May 31, 2026. [Online]. Available: https://headachejournal.onlinelibrary.wiley.com/doi/10.1111/j.1526-4610.2006.00444.x.

[32] K.-P. Peng and A. May, "Migraine understood as a sensory threshold disease," en, *Pain*, vol. 160, no. 7, pp. 1494–1501, Jul. 2019, ISSN: 0304-3959, 1872-6623. DOI: 10.1097/j.pain.0000000000001531. Accessed: Jun. 5, 2026. [Online]. Available: https://journals.lww.com/00006396-201907000-00003.

[33] M. J. Marmura, "Triggers, Protectors, and Predictors in Episodic Migraine," en, *Current Pain and Headache Reports*, vol. 22, no. 12, p. 81, Dec. 2018, ISSN: 1531-3433, 1534-3081. DOI: 10.1007/s11916-018-0734-0. Accessed: Jun. 6, 2026. [Online]. Available: http://link.springer.com/10.1007/s11916-018-0734-0.

[34] N. Giffin et al., "Premonitory symptoms in migraine: An electronic diary study," en, *Neurology*, vol. 60, no. 6, pp. 935–940, Mar. 2003, ISSN: 0028-3878, 1526-632X. DOI: 10.1212/01.WNL.0000052998.58526.A9. Accessed: Jun. 6, 2026. [Online]. Available: https://www.neurology.org/doi/10.1212/01.WNL.0000052998.58526.A9.

[35] D. S. Van Casteren et al., "E-diary use in clinical headache practice: A prospective observational study," en, *Cephalalgia*, vol. 41, no. 11-12, pp. 1161–1171, Oct. 2021, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024211010306. Accessed: Jun. 5, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024211010306.

[36] N. Van Veelen et al., "The psychometric properties of an e-headache diary in migraine," en, *Cephalalgia*, vol. 45, no. 8, p. 03 331 024 251 359 222, Aug. 2025, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024251359222. Accessed: Jun. 5, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024251359222.

[37] B. W. H. Van Der Arend, I. E. Verhagen, M. Van Leeuwen, M. Q. T. P. Van Der Arend, D. S. Van Casteren, and G. M. Terwindt, "Defining migraine days, based on longitudinal E-diary data," en, *Cephalalgia*, vol. 43, no. 5, p. 033 310 242 311 666, May 2023, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024231166625. Accessed: Jun. 5, 2026. [Online]. Available: http://journals.sagepub.com/doi/10.1177/03331024231166625.

[38] I. P. Martins, M. Westerfield, M. Lopes, C. Maruta, and R. Gil-da-Costa, "Brain state monitoring for the future prediction of migraine attacks," en, *Cephalalgia*, vol. 40, no. 3, pp. 255–265, Mar. 2020, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/0333102419877660. Accessed: Jun. 7, 2026. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0333102419877660.

[39] K. Pardo, T. J. Schwedt, F. M. Cutrer, and C.-C. Chiang, "The promise of artificial intelligence and machine learning for migraine treatment outcome prediction: A narrative review," en, *Cephalalgia*, vol. 45, no. 11, p. 03 331 024 251 395 541, Nov. 2025, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024251395541. Accessed: Jun. 7, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024251395541.

[40] A. Stubberud, H. Langseth, P. Nachev, M. S. Matharu, and E. Tronvik, "Artificial intelligence and headache," en, *Cephalalgia*, vol. 44, no. 8, p. 03 331 024 241 268 290, Aug. 2024, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024241268290. Accessed: Jun. 8, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03331024241268290.

[41] A. Torrente et al., "The Clinical Relevance of Artificial Intelligence in Migraine," en, *Brain Sciences*, vol. 14, no. 1, p. 85, Jan. 2024, ISSN: 2076-3425. DOI: 10.3390/brainsci14010085. Accessed: Jun. 8, 2026. [Online]. Available: https://www.mdpi.com/2076-3425/14/1/85.

[42] V. Kapustynska, V. Abromavičius, A. Serackis, S. Andruškevičius, K. Ryliškienė, and Š. Paulikas, "Advancing a generalizable model for migraine prediction: Analysis of filtering techniques on physiological signals," en, *Technology and Health Care*, vol. 33, no. 5, pp. 2184–2193, Sep. 2025, ISSN: 0928-7329, 1878-7401. DOI: 10.1177/09287329251332415. Accessed: Jun. 8, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/09287329251332415.

[43] K. K. Holsteen, M. Hittle, M. Barad, and L. M. Nelson, "Development and Internal Validation of a Multivariable Prediction Model for Individual Episodic Migraine Attacks Based on Daily Trigger Exposures," en, *Headache: The Journal of Head and Face Pain*, vol. 60, no. 10, pp. 2364–2379, Nov. 2020, ISSN: 0017-8748, 1526-4610. DOI: 10.1111/head.13960. Accessed: Jun. 21, 2026. [Online]. Available: https://headachejournal.onlinelibrary.wiley.com/doi/10.1111/head.13960.

[44] A. M. Carrington et al., "A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms," en, *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 4, Dec. 2020, ISSN: 1472-6947. DOI: 10.1186/s12911-019-1014-6. Accessed: Aug. 22, 2026. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1014-6.

[45] A. Stubberud et al., "Forecasting migraine with machine learning based on mobile phone diary and wearable data," en, *Cephalalgia*, vol. 43, no. 5, p. 033 310 242 311 692, May 2023, ISSN: 0333-1024, 1468-2982. DOI: 10.1177/03331024231169244. Accessed: Jun. 19, 2026. [Online]. Available: http://journals.sagepub.com/doi/10.1177/03331024231169244.

[46] H. Göker, "Automatic detection of migraine disease from EEG signals using bidirectional long-short term memory deep learning model," en, *Signal, Image and Video Processing*, vol. 17, no. 4, pp. 1255–1263, Jun. 2023, ISSN: 1863-1703, 1863-1711. DOI: 10.1007/s11760-

022-02333-w. Accessed: Jun. 21, 2026. [Online]. Available: https://link.springer.com/10.1007/s11760-022-02333-w.

[47] F. Faisal et al., "Forecasting migraine with time-series machine learning from mobile health data," en, *The Journal of Headache and Pain*, vol. 27, no. 1, p. 91, Mar. 2026, ISSN: 1129-2377. DOI: 10.1186/s10194-026-02346-7. Accessed: Jun. 19, 2026. [Online]. Available: https://link.springer.com/10.1186/s10194-026-02346-7.

[48] Y.-H. Tsai et al., *Personalized prediction of the following-day migraine attacks: A machine learning approach based on digital headache diary records*, Jul. 2025. DOI: 10.21203/rs.3.rs-7226289/v1. Accessed: Jun. 20, 2026. [Online]. Available: https://www.researchsquare.com/article/rs-7226289/v1.

[49] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to ¡span style="font-variant:small-caps;"¿SHAP¡/span¿ analysis: Explaining supervised machine learning model predictions in drug development," en, *Clinical and Translational Science*, vol. 17, no. 11, e70056, Nov. 2024, ISSN: 1752-8054, 1752-8062. DOI: 10.1111/cts.70056. Accessed: Jun. 8, 2026. [Online]. Available: https://ascpt.onlinelibrary.wiley.com/doi/10.1111/cts.70056.

[50] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," en, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, Jan. 1996, ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1996.tb02080.x. Accessed: Jun. 8, 2026. [Online]. Available: https://academic.oup.com/jrsssb/article/58/1/267/7027929.

[51] A. Swift, R. Heale, and A. Twycross, "What are sensitivity and specificity?" en, *Evidence Based Nursing*, vol. 23, no. 1, pp. 2–4, Jan. 2020, ISSN: 1367-6539, 1468-9618. DOI: 10.1136/ebnurs-2019-103225. Accessed: Jun. 8, 2026. [Online]. Available: https://ebn.bmj.com/lookup/doi/10.1136/ebnurs-2019-103225.

[52] T. pandas development team, *Pandas-dev/pandas: Pandas*, Jul. 2026. DOI: 10.5281/ZENODO.3509134. Accessed: Jul. 23, 2026. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.3509134.

[53] L. Gautier, *Rpy2: Python interface to the R language (embedded R)*, 2026. Accessed: Jul. 23, 2026. [Online]. Available: https://rpy2.github.io.

[54] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using **lme4**," en, *Journal of Statistical Software*, vol. 67, no. 1, 2015, ISSN: 1548-7660. DOI: 10.18637/jss.v067.i01. Accessed: Jul. 23, 2026. [Online]. Available: http://www.jstatsoft.org/v67/i01/.

[55] A. Paszke et al., *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, Version Number: 1, 2019. DOI: 10.48550/ARXIV.1912.01703. Accessed: Jul. 23, 2026. [Online]. Available: https://arxiv.org/abs/1912.01703.

[56] S. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, Version Number: 2, 2017. DOI: 10.48550/ARXIV.1705.07874. Accessed: Jul. 23, 2026. [Online]. Available: https://arxiv.org/abs/1705.07874.

[57] . R. Dr.A, "Binary cross entropy with deep learning technique for Image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5393–5397, Aug. 2020, ISSN: 22783091. DOI: 10.30534/ijatcse/2020/175942020. Accessed: Jul. 23, 2026. [Online]. Available: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse175942020.pdf.

[58] Z. Chang, Y. Zhang, and W. Chen, "Effective Adam-Optimized LSTM Neural Network for Electricity Price Forecasting," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China: IEEE, Nov. 2018, pp. 245–248, ISBN: 978-1-5386-6565-7. DOI: 10.1109/ICSESS.2018.8663710. Accessed: Jul. 23, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/8663710/.

[59] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," en, *BMC Bioinformatics*, vol. 8, no. 1, p. 25, Dec. 2007, ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-25. Accessed: Aug. 23, 2026. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25.