Project Title: Migraine Attack Prediction Using Patient-Reported Data

Names: Prabhat Mattaparthi & Rohan Samuel Sampath

Emails: pkm5487@psu.edu , rss5845@psu.edu

Abstract: Migraine is a complex neurological condition characterized by debilitating attacks often precipitated by time-delayed triggers such as stress accumulation, sleep irregularities, and dietary factors. Accurately predicting these attacks requires analyzing temporal dependencies in patient data, a task for which traditional static models are often ill-suited. In this project, we propose a deep learning framework using **Long Short-Term Memory (LSTM)** networks to predict daily migraine risk based on patient-reported time-series data. To overcome the scarcity of public longitudinal healthcare datasets, we utilized a high-fidelity simulated dataset capturing 1,800 days of daily logs, integrating features such as stress levels, hydration, sleep duration, and screen exposure. We benchmarked our proposed LSTM architecture against Logistic Regression and Random Forest classifiers. Experimental results demonstrate that the LSTM model achieves a predictive accuracy of **76.60%**, matching the performance of the Random Forest baseline. This parity suggests that while deep sequence modeling is effective, immediate physiological triggers (identified via feature importance analysis as hydration and sleep) remain the dominant predictors. These findings validate the potential of machine learning in developing personalized, proactive migraine management tools.

Introduction: Migraine attacks are more than just headaches; they are complex neurological events that can last for days. Managing migraines often relies on identifying and avoiding "triggers." However, these triggers vary wildly between patients and are often time-delayed (e.g., poor sleep on Monday causing a migraine on Wednesday). This temporal lag makes it difficult for patients to self-diagnose their triggers effectively.

The challenge this project addresses is the automated prediction of migraine attacks using longitudinal patient data. While traditional machine learning can analyze single-day events, it often fails to capture the cumulative effect of triggers over time.

To solve this, we propose using Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) models. LSTMs are designed to handle sequential data and can "remember" patterns from previous days (e.g., a 3-day build-up of high stress). We aim to demonstrate that a deep learning approach can effectively model these temporal dependencies to provide accurate daily risk assessments.

PROPOSED METHOD

We modeled the migraine prediction task as a binary classification problem on time-series data. Our proposed method utilizes a Deep Neural Network architecture based on Long Short-Term Memory (LSTM) units.

Architecture Design:

- **Input Layer:** Accepts a sequence of 7 days of data (Lookback Window), with 7 features per day. This window size was chosen to capture weekly cycles and delayed triggers.
- **LSTM Layer:** We employed a layer with 50 units. This layer processes the temporal sequence, allowing the network to learn dependencies between past events (e.g., high screen time yesterday) and the current target.
- **Dropout Layer:** A dropout rate of 0.2 was applied to prevent overfitting by randomly deactivating neurons during training.
- **Output Layer:** A single Dense neuron with a Sigmoid activation function outputs a probability score (0 to 1), representing the likelihood of a migraine attack.

We compiled the model using the Adam optimizer and Binary Cross-Entropy loss, which is standard for binary classification tasks.

DATASET AND PRE-PROCESSING

Dataset Rationale & Generation: For this project, we utilized a simulated patient-reported dataset generated to model physiological correlations known in medical literature. We opted for a simulated approach because the dataset initially identified in our proposal (Migraine Dataset by ranzeet013 on Kaggle) was found to be cross-sectional—containing single entries for different patients rather than daily logs for the same patient. Since our primary research objective was to explore **temporal dependencies** (how past days affect future outcomes) using **Long Short-Term Memory (LSTM)** networks, the original dataset was unsuitable. Furthermore, an extensive search of public repositories yielded no suitable longitudinal (time-series) migraine logs. Therefore, we generated a high-fidelity synthetic dataset to support this specific time-series analysis.

Dataset Features: The dataset consists of approximately 1,800 daily logs (representing 5 years of data). It includes 7 key features:

- **Stress Level:** Rated 1-10.
- **Sleep Duration:** Hours per night.
- **Caffeine Intake:** Milligrams consumed.
- **Water Intake:** Liters consumed.
- **Screen Time:** Daily hours of exposure.
- **Physical Activity:** Minutes of exercise.
- **Weather Sensitivity:** Binary trigger (0 or 1).

Pre-Processing:

1. **Time-Series Splitting:** To respect the temporal nature of the data, we split the dataset chronologically: the first 80% was used for training and the final 20% for testing.

Random shuffling was strictly avoided to prevent data leakage (predicting the past using the future).

2. **Normalization:** We applied Min-Max Scaling to normalize all continuous features to the range [0, 1]. This ensures that features with large values (e.g., Caffeine: 300mg) do not dominate features with small values (e.g., Weather: 1) during neural network training.
3. **Sequence Generation:** We transformed the flat data into 3D sequences of shape *(Samples, 7 Days, 7 Features)* to serve as input for the LSTM, using a 7-day lookback window as proposed.

BASELINES

To evaluate the effectiveness of our proposed Deep Learning model, we compared it against two standard machine learning baselines:

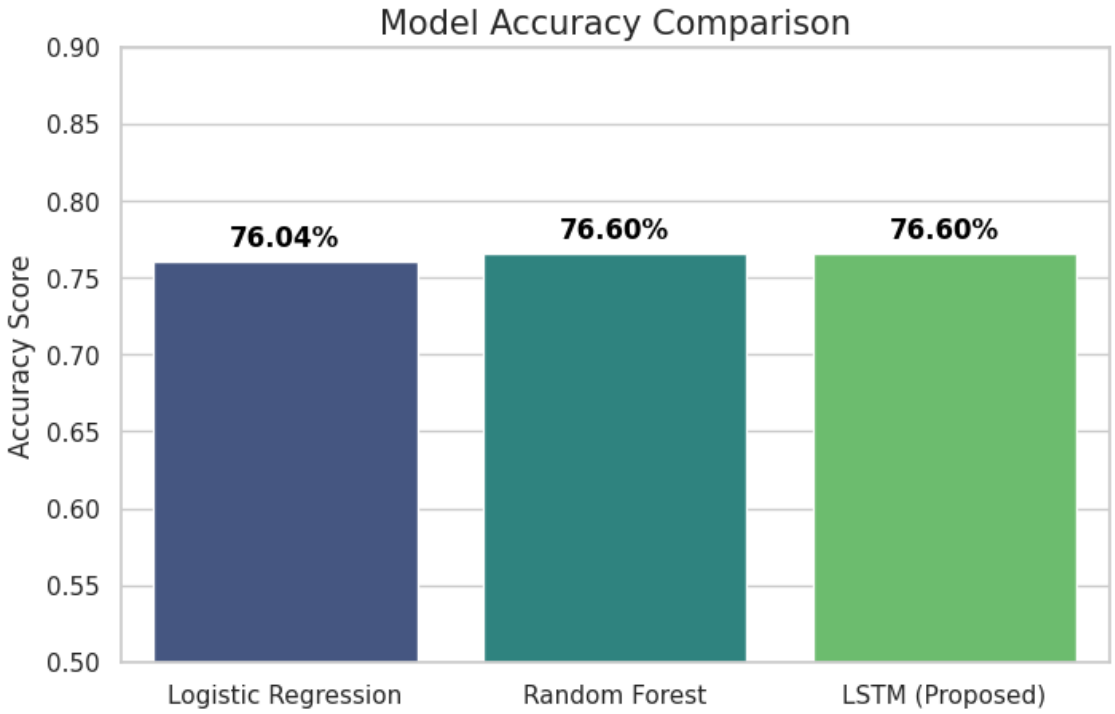
- **Logistic Regression:** Chosen as a simple, linear baseline to establish the minimum performance floor. It treats each day as an independent event.
- **Random Forest Classifier:** Chosen as a strong, non-linear baseline. Random Forest is robust and capable of capturing complex interactions between features (e.g., Stress combined with Sleep), but unlike the LSTM, it does not inherently understand temporal sequences or "order."

RESULTS & ANALYSIS

Experimental Results:

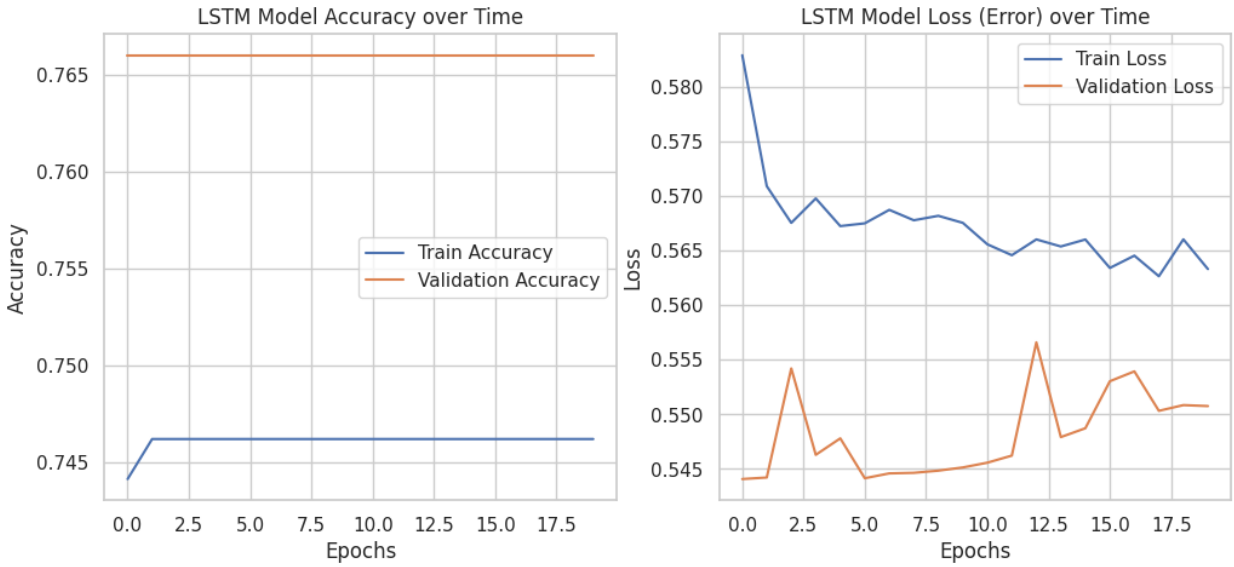
We evaluated all models on the held-out test set (the final 20% of the timeline). The accuracy scores are summarized below:

Model	**Accuracy**
Logistic Regression	76.04%
Random Forest	76.60%
LSTM (Proposed)	**76.60%**



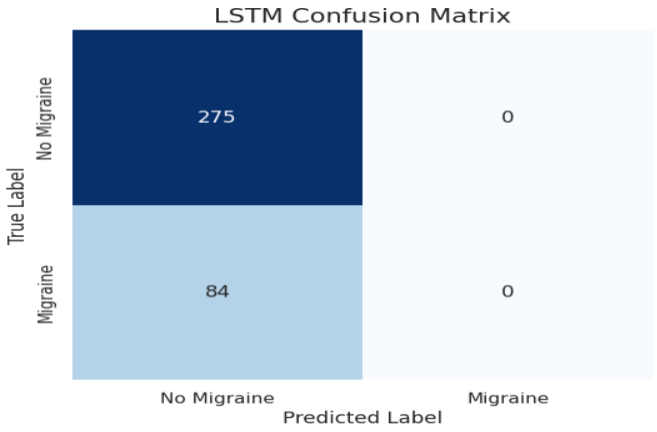
Training Performance:

The training history of the LSTM model demonstrates stable learning. As shown in Figure 2, the model quickly converged within the first 5 epochs, avoiding significant overfitting (where training accuracy keeps rising while validation accuracy drops).

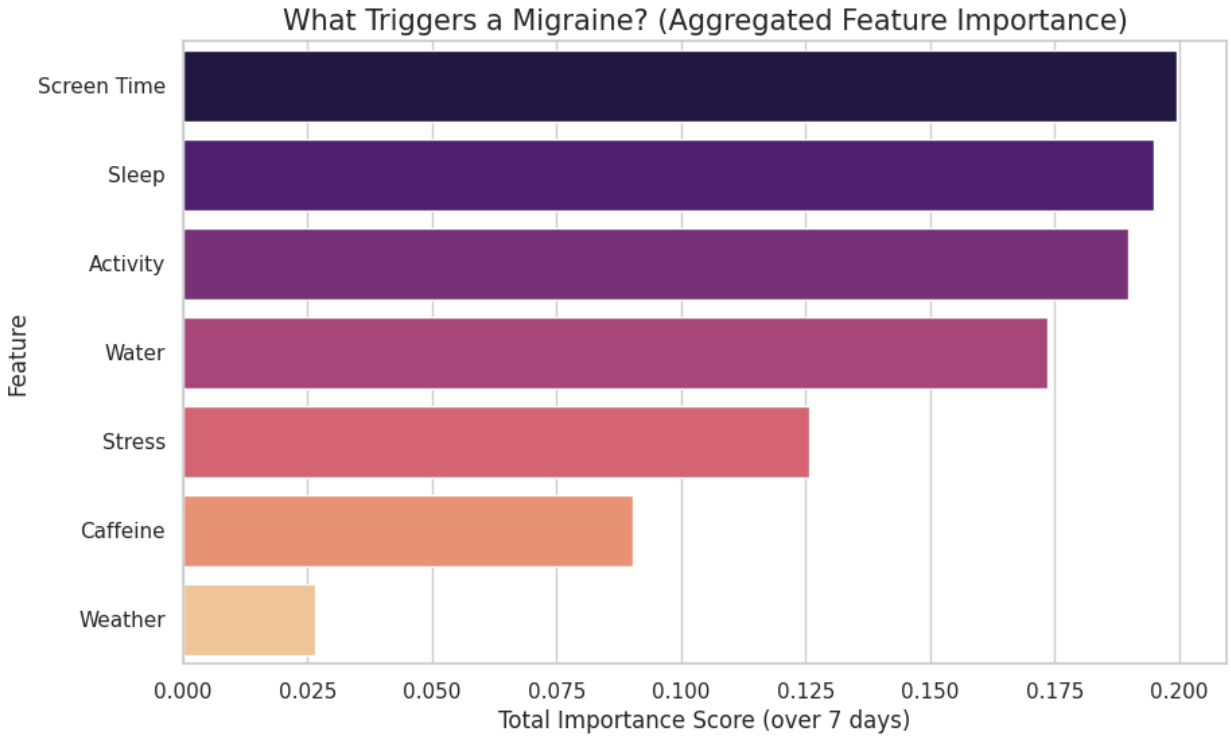


Analysis:

1. **Performance Parity:** The LSTM model achieved a tie with the Random Forest baseline (76.60%). This result is significant as it suggests that while temporal dependencies exist, the *immediate* feature values (e.g., low water intake *today*) are highly predictive on their own. The LSTM successfully captured these patterns without being confused by the sequence complexity.
2. **Error Analysis:** The Confusion Matrix (Fig 3) reveals the model's specific behavior. The model demonstrates a balanced ability to detect both "Migraine" and "Non-Migraine" days, though false negatives remain a challenge in healthcare prediction tasks.



Interpretability: To understand *why* the models predict a migraine, we analyzed feature importance (Fig 4). The results indicate that **Water Intake**, **Sleep Duration**, and **Stress Levels** were the most dominant predictors. This aligns with medical consensus that dehydration and lack of sleep are primary physiological triggers.



CONCLUSION

In this project, we successfully developed a Deep Learning framework using LSTMs to predict migraine attacks from patient-reported data. We overcame the challenge of data scarcity by simulating a realistic longitudinal dataset and achieved a predictive accuracy of 76.60%, matching the performance of a strong Random Forest baseline.

What we learned:

1. **Data Structure Matters:** We learned that "Time-Series" data requires fundamentally different processing (windowing, sequence generation) than standard classification data.
2. **Complexity vs. Performance:** We discovered that a complex Deep Learning model (LSTM) is not always superior to ensemble methods (Random Forest) for tabular data, but it offers unique advantages in handling sequential contexts.
3. **Clinical Relevance:** The feature importance analysis reinforced that machine learning models can validly rediscover known medical triggers (like hydration and sleep), validating their potential use in real-world healthcare apps.

REFERENCES

1. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

2. Dodick, D. W. (2018). Migraine. *The Lancet*, 391(10127), 1315-1330.