Predicting migraine attacks based on weather data using recurrent neural networks

Roidos, Antonios

Predicting migraine attacks based on weather data using recurrent neural networks

Antonios Roidos

Thesis advisor: dr. Matthijs van Leeuwen
Thesis advisor: Marieke Vinkenoog
Second supervisor: Dr. Peter Grunwald

Defended on 24/04/2023

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Contents

1 Abstract

Objectives: To develop and compare models that make personalised predictions on day-to-day migraine attacks based on weather data and to identify which weather triggers are the most common for weather-sensitive patients.
Methods: The variables that we used were: barometric pressure, cloud cover, precipitation, sunlight, wind, temperature and humidity. Apart from these, our predictions also considered the influence of changes of the weather. The weather data were obtained from the KNMI (Dutch national weather service),while the migraine data were derived from patients' e-diaries from the LUMC (Leiden University Medical Center).

The two types of Recurrent Neural Networks (RNNs) that we used and compared were: Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). For the tuning of the hyperparameters of the two networks we used Bayesian optimization, an AutoML method. Finally, we utilized explanatory AI to explore which weather variables are correlated with an increased probability of migraine attacks.
Results: The LSTM network outperformed the GRU in the test set by a small margin (AUC PR LSTM: 0.177, GRU: 0.171) and both models yielded better results than the baseline: 0.157 (=positive class/number of observations). The most predominant and common trigger was found to be temperature, while the rest of the variables had similar effects. The importance of the triggers for weather-sensitive patients varied greatly from one to another.

2 Introduction

Migraine is a neurological brain disorder that affects a large number of people around the world. It is estimated that around 15% of the total population suffers from regular migraine attacks. In this percentage the number of women is around 3 times larger than the number of men [1]. The most likely period of the onset of migraines is before the age of 35 and the period that they get most intense is between ages 35 and 39.[1] Migraines mostly affect people at the most productive years of their lives and can have great consequences on their performance and their jobs. The main symptoms are headache, nausea, photophobia, vomiting and phonophobia.[1] [2] According to the World Health Organization, migraine is ranked sixth in the list of most disabling disorders [3] and according to the Global Burden of Disease Study 2016, the disability caused by migraines is even greater than the sum of the rest of neurologic disorders. [1] On a big scale, migraines are a problem that costs countries millions of dollars each year. For example, the annual costs in the European Union are estimated to be 27 billion euros. [3]

During the past decades, scientists have tried to get a better understanding of migraines and what triggers them. An example of a factor that is very frequently described as a trigger for migraines by both patients and the literature is stress.[4] [5] [6] However, the total number of possible triggers is believed to be very big and still unknown.[6] Also, their relationship with migraine incidences has not been fully understood yet. Investigating these triggers and how they affect patients is a step towards better clinical management of migraines. Being able to forecast migraine attacks based on triggers, could lead to better prevention through early medication use. [7] This would increase the chances of a successful treatment of the migraine. Furthermore, if patients knew what factors are associated with increased migraine incidence, they could adjust their everyday lives in a way that migraines would not be as common or as disabling as they are now.

Each person seems to have a different perception on what possible causes for the onset of a migraine are. Around half of migraine patients believe that they are weather-sensitive, which is defined as having at least one weather associated migraine trigger. [6][8] Sunny days, thundery weather [8] and temperatures changes [9] are just some of the perceived triggers that patients report. Yang et al. [9] showed that patients who reported sensitivity to temperature, were more likely to have more migraine attacks in winter. The perception of patients about their triggers is not always correct. However, it is an indication that migraine triggers are not the same for everybody. If patients are split in homogeneous groups regarding to their triggers, the predictions of migraine attacks could be more precise.[10]

Prince et al. showed that around half of the patients on their study were weather-sensitive. [6] Each weather factor was examined separately for each patient, and if its effect on migraine incidence was statistically significant, it was considered a trigger. It was also shown that statistically significant triggers were not the same for every patient and that not all patients had triggers associated with the weather. Surprisingly, in the same research, patients who considered themselves weather-sensitive, did not have higher chances of actually being weather-sensitive. [2] This indicates that the connection between weather and migraine attacks is not easily understood in experience and patients' beliefs are not reliable for our models.

Table 6 summarizes the results of 13 different studies that investigated the connection between weather and migraines. There are evident differences between their results and their findings do not all point to the same direction. Oftentimes, triggers that are found significant in one study, are found non-significant in others. Barometric pressure, temperature and weather changes are regularly found to be associated with migraine incidences, but in some studies these connections were not observed. However, all this research is helping us to get a better understanding of the weather related triggers and could be useful for the development of personalized models.

According to the modular headache theory, each symptom that is experienced by a patient during a migraine, is associated with the activation of a group of neurons. [11] The symptoms and the triggers that cause the neurological process that lead to a migraine, can be similar for groups of patients. [12][10] Making predictions based on groups of people with common triggers is a method that was proposed by Holsteen et al. and is believed to be the future of migraine predictions. [13]

2.1 Research design and contributions

This thesis project focuses on predicting migraines that were induced by weather triggers. To this end, we used two datasets. The first one is the migraine e-diaries from patients of the LUMC who followed the study for varying time periods. During these periods migraine onsets were monitored. For this study, as migraine days we considered only the first day of a migraine attack. In order to explore the causality between migraines and weather, it was necessary that we also had accurate records of the weather conditions for the area where patients lived. For this reason, the second dataset that we used was from from KNMI (Dutch national weather service) that included the weather conditions as collected by the closest weather station to where patients lived. This dataset covered all the days of the first dataset.

Many studies overlook the importance of the continuity of weather data, by processing each day independently. This way they are not taking into consideration the changes in the weather. However, as can be observed in Table 5, the weather changes have been found to be a trigger for many studies.

Also, previous studies on the field have not been successful in addressing the heterogeneity of the migraine triggers between patients. The triggers between patients can differ greatly, so dealing with the data of every patient individually can be a great asset for a predicting model.

For the reasons above, we have chosen to use Recurrent Neural Networks (RNN) models. RNNs are very effective in handling sequential data, such as time series data. Their key advantage is that they are able to catch dependencies in each sequence that they process. In this way, it is possible to develop a model that can extract information from the data of multiple patients, while dealing with every patient individually.

2.2 Objectives

The primary objective of this thesis project is to develop a temporal predictive model that can make personalised predictions for migraine patients based on the weather. The model must take into consideration the unique characteristics and patterns of each patient's records.

The secondary objective of our research is to test different architectures for RNNs including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). The focus here is to investigate how the use of different model structures can affect the predictive power of the model.

The final goal is the exploration of the connection between weather and migraine incidences. In section 4.8 we investigate which weather triggers are more heavily connected with an increased probability of a migraine with the use of explanatory AI methods.

3 Preliminaries

3.1 Structure of RNN models

Recurrent Neural Networks (RNNs) are a type of neural network that are particularly effective in processing sequential data as time series. To do that RNNs utilise hidden states that are interlinked. States are matrices that are produced in cells, shown in the Figure 1, and preserve valuable information from previous time steps. The outcome of each hidden state is affected by the hidden state of the previous observation. Therefore, each new prediction will be affected by the data that will have already been processed. [14] In our case, this structure has the ability to keep and use information about weather conditions of previous days and how these influence the results. This could also help us with the predictions of the migraines that are caused by the changes in the weather between days, which is a very commonly mentioned trigger.

The graphical representation of this process can be seen in Figure 1. The formula of update rule of the $(i+1)^{th}$ hidden state is

$$h_{i+1} = f(W_{(i+1)}x_{(i+1)} + U_i h_i), \tag{1}$$

where x_i is the input data of the i^{th} observation, U transition matrix, W weights matrix, and f a transition function of our choice.

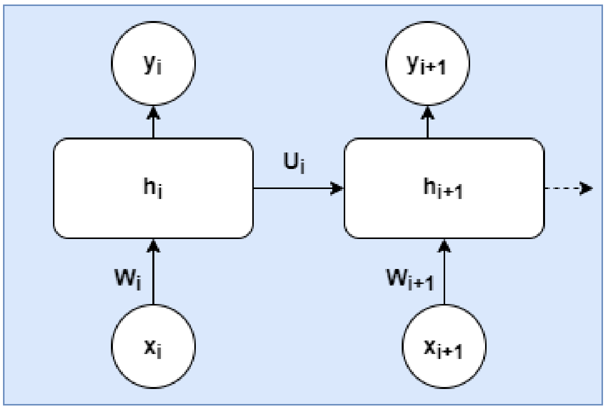
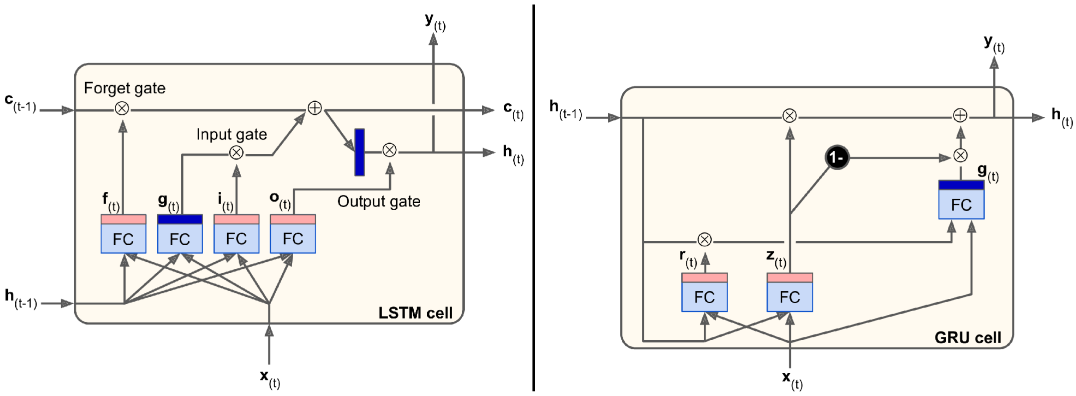


Figure 1: Schematic representation of the connection between the hidden states of an RNN network. The x_i is the input data for the i^{th} observation, y_i is its outcome variable, U transition matrix, W weights matrix and f a transition function of our choice.

The inputs of each step and the corresponding outcome, are processed one by one chronologically. In each time step, the input data are multiplied with the weights matrix **W** and the output of the previous state is multiplied with the transition matrix **U**. The weights matrix and the transition matrix are trainable parameters of the model. To get the new state, the sum of these two is then inserted in a transition function.

We will be comparing two different types of RNN architectures: LSTM and GRU. Figure 2 illustrates the structural differences in how they pass information between hidden states. Both LSTM and GRU have been used for medical predictions in the past, obtaining very good results in comparison to older methods. [15] [16]



(a) LSTM cell *(b) GRU cell*

Figure 2: Schematic representation of the architecture of two different types of cells. The arrows indicate the transfer of the matrices within the cell. When lines split, copies of the same matrix are created. The symbol "⊗" stands for element-wise multiplication and "⊕" for the addition of two matrices. The coloured boxes indicate that the matrices are fed to fully connected layers. The activation functions are displayed as blue and pink boxes, with blue being the tanh and pink being the sigmoid. (Taken from [17])

3.1.1 Gated Recurrent Unit

The schematic representation of GRU in Figure 2 indicates that it has a simpler structure than LSTM. There is only one state h_t that passes information between different time steps. This is achieved with the use of two gates: the update gate and the reset gate.

The cell has to determine which part of the information that it carried from the previous time steps is no longer useful and should be forgotten. To do that, a linear combination of the hidden state h_{t-1} and the input data x_t are fed in a sigmoid function. The resulting matrix z_t is then multiplied with the $h_{(t-1)}$ so that parts of the state will be discarded. This part of the cell is called update gate.

$$z_t = \sigma(W_{xz}^T x_t + W_{hz}^T h_{(t-1)} + b_z) \tag{2}$$

For the reset gate, a copy of the $h_{(t-1)}$ and the input data are processed so that the important information is stored in the matrix $(r_t \otimes h_{t-1})$. This matrix determines which contents of the previous state, will influence the state. To update the old hidden state, the cell also uses $1 - z_t$. As discussed earlier z_t included values between 0 and 1 and its use is to discard parts of the state. So, $1 - z_t$ is open when z_t is closed. This mechanism dictates which values are important for the predictions and should pass the gate.

The formulas that describe this procedure are the following:

$$r_t = \sigma(W_{xr}^T x_t + W_{hr}^T h_{(t-1)} + b_r) \tag{3}$$
$$g_t = tanh(W_{xg}^T x_t + W_{hg}^T (r_t \otimes h_{(t-1)}) + b_g) \tag{4}$$

The prediction of the model is again equal to the hidden state and it is given by:

$$y_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t \tag{5}$$

3.1.2 Long Short-Term Memory

The schematic representation of the LSTM's cell in Figure 2 demonstrates that for each time step, apart from the input x_t, the short term state h_t and long term state c_t of the previous time step are also passed in the cell. The long term state (or cell state) is a matrix that carries useful information and connects all consecutive states with each other. At every time step, LSTMs can modify the long term state by adding and removing information from it with the use of gates. The two gates that affect the output of the long term state are the forget gate and the input gate. First, $h_{(t-1)}$ and x_t are each multiplied with a weight matrix. Then the outcomes are summed and passed to a sigmoid function to transform the values between 0 and 1. The resulting matrix f_t, is then multiplied (element wise) with the long term state of the previous step. In this way, the values of the long term memory that are not important are multiplied with values close to zero, so the cell "forgets" them.

The formula that expresses the update rule of the f_t matrix is given by:

$$f_t = \sigma(W_{xf}^T x_t + W_{hf}^T h_{t-1} + b_f), \tag{6}$$

where W are the weight matrices, and b_f is the bias.

The same idea is used in the other two gates of the cell. The input gate is responsible for the update of the cell state based on h_{t-1} and x_t and it consists of two parts. On the first part, a linear combination of h_{t-1} and x_t is again transformed with sigmoid function and the matrix i_t is produced. On the other part, a different activation function is used and its outcome g_t is candidate values for the new cell state. The multiplication of the two is then added to the cell state to complete its update.

The rest of formulas for the update of the cell are the following:

$$g_t = tanh(W_{xg}^T x_t + W_h g h_{t-1} + b_g) \tag{7}$$
$$i_t = \sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i) \tag{8}$$
$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{9}$$

After the cell state is updated, it is time to move to the output gate that is responsible for the prediction. First, a copy of the cell state is transformed by tanh. Then, with the use of h_{t-1} and x_t only the useful information o_t is extracted. Finally, the element wise multiplication of these two constitutes the prediction of our model y_t. It is important to mention that the short-term state for each time step h_t is a copy of this prediction.

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \tag{10}$$
$$y_t = tanh(c_t) \otimes o_t \tag{11}$$

4 Methods

4.1 Dataset

In this thesis project we worked primarily with data from the LUMINA (Leiden University MIgraiNe Analysis program) e-diaries. This dataset included electronic headache diaries (e-diaries) of 1428 clinically diagnosed migraine patients who were asked to fill in online diaries. The diaries had to be filled once per day and included questions on whether they experienced a headache, its characteristics and symptoms, use of medication and menstruation. The mean number of days for patients was 161.8 days and the standard deviation 152.

A major strength of this dataset was that migraine days were defined based on reported symptoms instead of self-reports.[18] [9] This avoids problems that previous studies faced that had to do with misinterpretation of the symptoms by the patient and depending on a recall bias due to the episodic nature of the disease.

To get the weather conditions on the same days that the e-diaries were filled out, we used openly available weather data from KNMI (Dutch national weather service). The weather station that was responsible for our data, was located in De Bilt and it was the closest to where the patients lived. Weather variables were selected based on a literature review summarized in Table 6. The selected weather variables were: barometric pressure, sunlight, wind cover, temperature, precipitation, cloud cover, and humidity.

Weather condition	**Measurement**	**Unit**	**Mean**	**Std**
Temperature	Average temperature of the day	0.1 ^{o}C	111.7	59.6
Sun percentage	Percentage of longest possible sunshine duration	0.1 hours	39.1	31
Precipitation	Total precipitation of the day	mm	23.4	44.7
Pressure	Average atmospheric pressure of the day	hPa	10186	96.6
Cloud cover	Average cloud cover	[0,9] okta	6.1	2.1
Wind	Average wind speed	m/s	33.5	14
Humidity	Average relative humidity of the day	percent	79.3	10.8

Table 1: Weather predictor variables of the KNMI dataset

As expected, the weather variables had some correlations with each other, see Figure 3. The most obvious correlations were between sunlight and cloudiness, and between sunlight and humidity. For some types of models, such as linear regression, correlations between predictive variables could be an issue for their predictive power. Also, collinearity can affect the explainability of these types of models. On the contrary, RNNs do not include the assumption of independent prediction variables and we do not expect it to be a problem in our analysis. In RNNs the input variables are weighted and combined in a way that the impact of colliniarity is greatly reduced.

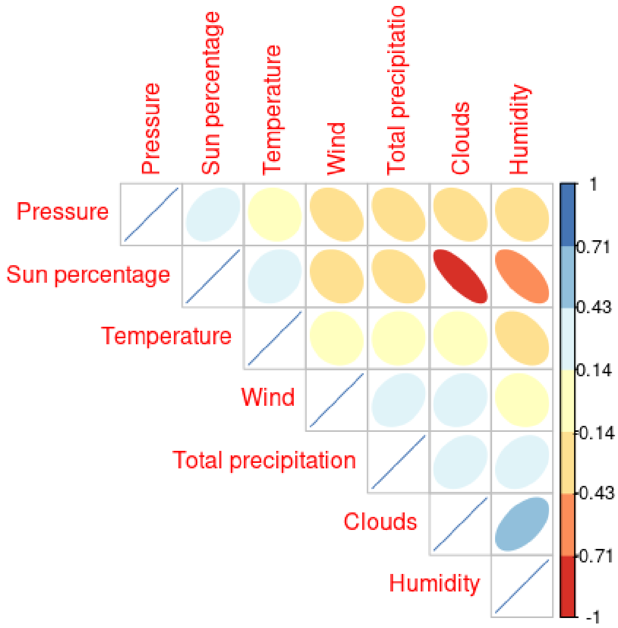


Figure 3: Correlation between predictor variables

4.2 Missing data

The ratio of missing values to diary inputs was approximately 70%. Every patient in this study has missing data and there are two reasons why. The first one is that patients did not fill in the diary every day of their followup period. Patients without at least 80% compliance for three consecutive months were left out of the study. The second reason is that, in the version of data for this thesis, diary inputs of the two days that follow the onset of every migraine were removed. The reason for this is that the duration of migraines is usually one to two days. In order to proceed with the training and the evaluation of our models, we first had to deal with these missing values.

It is observed that 344 out of the 1429 patients have missing data for periods greater than a month. As can be seen from Table 2 some patients even had multiple of these gaps in their data. These patients did not fill their diaries for many consecutive days, creating in this way big gaps of missing data in their sequences. In Figure 4 it is shown that within the last three months there was a great amount of missing data, which was caused by these big gaps. This phenomenon persisted throughout the whole duration of the study and could have led to big problems if not addressed properly.

Number of sequences	Number of patients
1	1084
2	235
3	73
4	26
5	8
6	1
7	1

Table 2: Number sequences of time data per patient for the dataset of our study. The right column indicates the number of patients and the left column the corresponding number of sequences. Patients with more than one sequence of data, had big gaps of missing data in their e-diaries.

Figure 4: Data of all patients with big gaps (over 30 days) of missing data in their diaries for the last two and a half months of the study.

In this study we regarded the available data as collections of independent sequences that represent the data of every patient. Big gaps within a sequence disturb the continuity, which is an important feature for the type of models that we have chosen. For this reason, when patients had periods of missing data greater than a month in their sequences, we chose to split their sequence and deal with it as many individual ones.

To deal with the small gaps (less than 30 days) of missing data we have chosen to classify all of the missing days, including the two days that follow the migraine onsets, as non-migraine days. The missing data were found to be uniformly distributed throughout the full period of the study. If the ratio of missing data to existing data was found to vary substantially for periods of the study, imputing the missing data in the way that we did would lead to imputation bias. The model would mistakenly associate the weather characteristics of periods consisting of a lot of missing data with a reduced probability of having a migraine attack. For example, if the amount of missing values was increased during winter, the model would assume that there is an inverse correlation between cold temperature and cloudy days with migraine incidences. In addition, imputing the two days that follow a migraine attack as non-migraine days is a very consistent and clear pattern, which we expected to be captured by the memory of the models. So, the predictions for these days would not be correlated with any weather conditions.

The imputed data was used for the training, but also the testing the model. However, in all sections of this project, where we assessed the models, the predictions for the imputed data were left out. In this way, we managed to deal with the problems that missing values were causing, and in the same time we avoided getting a biased evaluation of the performance of the models.

Although we have chosen to keep only patients with over 80% compliance, the exclusion of two days after a migraine attack caused many missing values, see Figure 5. During the procedure of the imputation of the missing data, the total number of data entries in the dataset was increased considerably, by a factor 1.7.

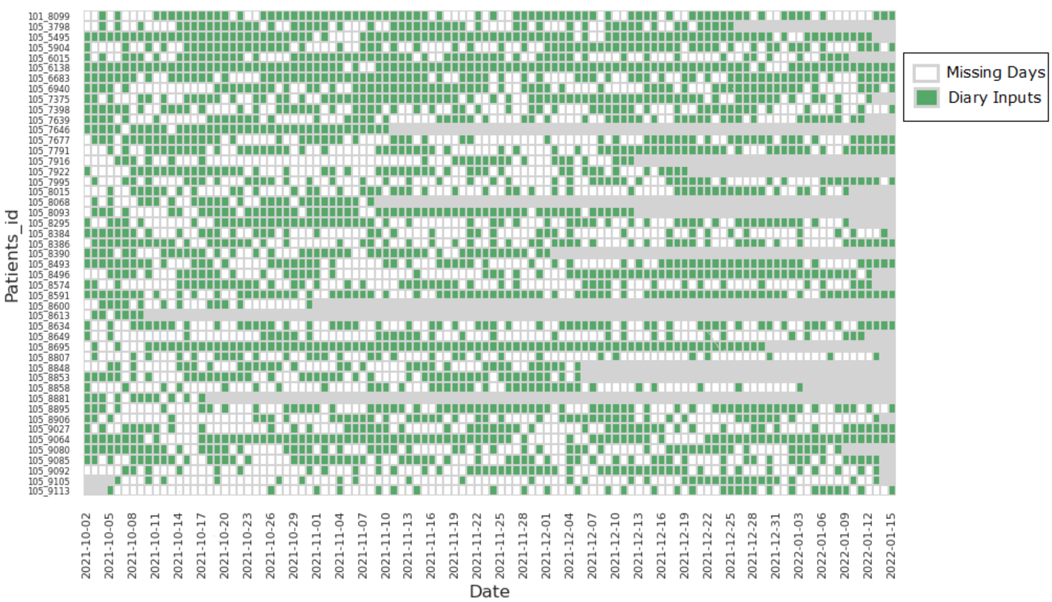


Figure 5: Data of 47 random patients without big gaps in their diaries for the last two and a half months of the study.

4.3 Train-test split

For the splitting of the dataset in train-test set we needed to take into consideration the following factors: the percentage of missing data, the train-test ratio and that weather conditions of every season are represented in both the test and the train sets.

Ultimately, the two sets had to cover different time periods and include a different set of patients to avoid auto-correlation. To achieve this, the dataset was first split into a train and a test set by time. However, in this way it was inevitable that some patients would have data in both the train and the test set. For each of these cases, we looked at how the patient's data were distributed between the two sets and calculated in which set each patient had the most data. Finally, we discarded the patient's data from the set with the smallest amount. During this procedure we would inevitably lose part of our initial data, so we had to carefully pick the dates that would be used as boundaries between the train-test sets. The distribution of the data for the span of our study can be seen in Figure 6

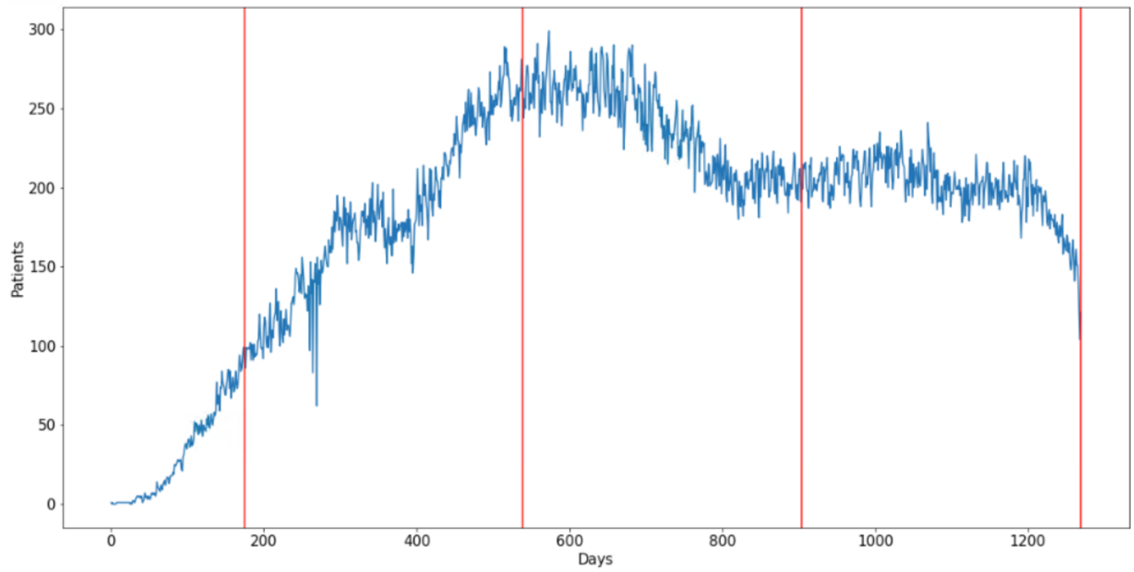


Figure 6: Number of patients as a function of the days of the study. Between two successive red lines are periods of a year.

From Figure 6 it is observed that the data in the last year of the study were almost uniformly distributed. So we picked the last year (16/01/2021-15/01/2022) as the test set. After going through the procedure that is described above, the data that were discarded were found to be 0.17% of the initial data. Also, the ratio between the test and the train set is 0.28. From Figure 7 it can be observed that this ratio was relatively stable for every day of the calendar year, and that both train and test set distributions were almost uniform throughout the year.

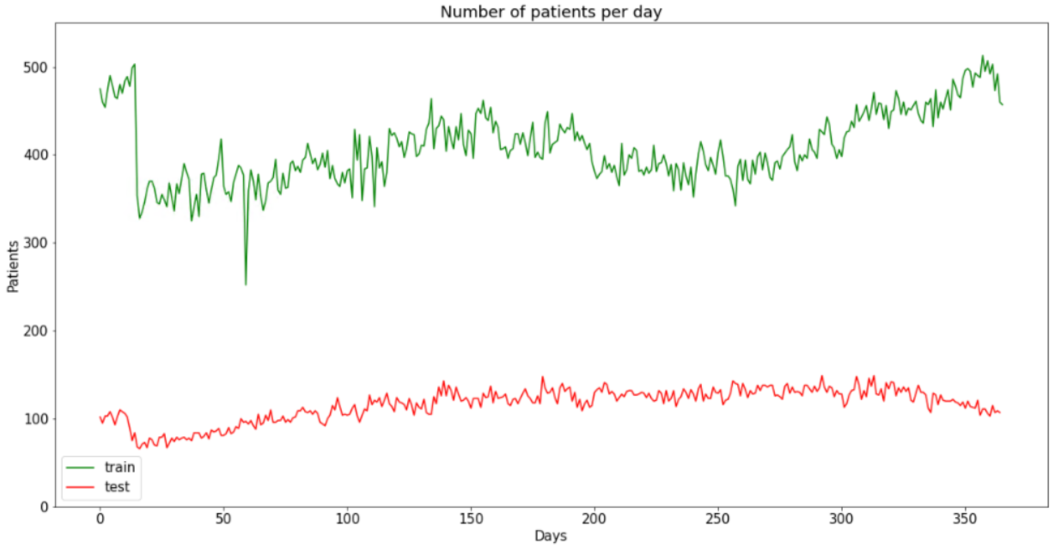


Figure 7: Graph of the total number of patients as a function of the days of one calendar year for the total of train-test sets.

4.4 RNNs for migraine prediction

In our case the time series data were not one big sequence, but numerous sequences that corresponded to numerous patients. These sequences had different lengths from each other and different dependencies between weather and migraines, so our models had to be adjustable. The reason for this is that each patient had different triggers and followed the diary for a different period than others.

To make our models flexible and able to deal with every sequence individually, we made use of the powerful feature of RNNs that is called memory. The structure of the model's memory is different for every type of RNN. For the simple RNN and the GRU, it has the form of the hidden states, while for the LSTM it also includes the cell state. Sequences were processed one by one, so that the memory of the model was able to identify the differences on every sequence independently. To do that, every time the model jumped to a new sequence, the memory that was created and used for the previous sequence was erased. Then, a new memory, which started from zero, was assigned to the model. In this way, the model could catch dependencies that existed for every patient individually. Also, separate sequences of the same patient were dealt independently, so dependencies that were found in one sequence could not be utilised for the rest sequences of the same patient.

In order to model non-linear and complex dependencies between input and output it is possible to create a network with stacked RNNs. To do that, multiple RNN layers get stacked on top of each other, and the output of every layer is used as input of the next layer. Along with the hidden state of previous layer, the RNN takes as input the hidden state (and the cell state for LSTMs) of the previous time step, as seen in Figure 8. In this way, the model manages to build some knowledge inside the hidden state (and cell state), which contributes to making better predictions. Finally, the output of the final layer of the network is passed in a sigmoid function so that our predictions are probabilistic.

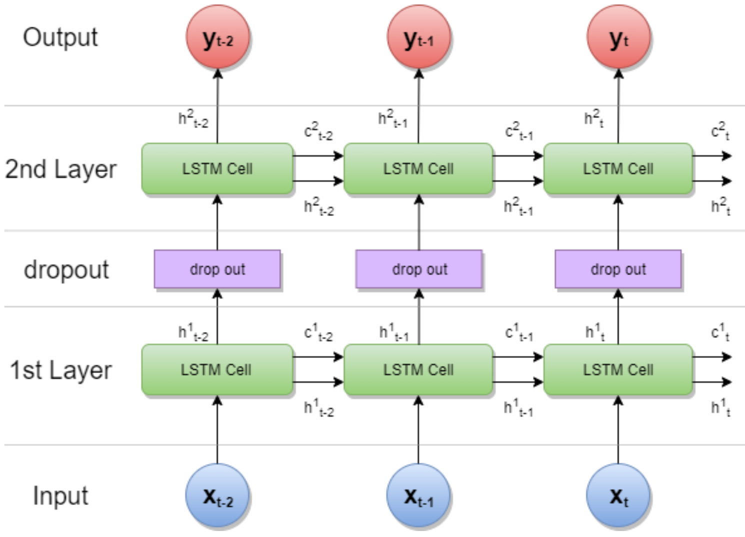


Figure 8: Example of a stacked LSTM model with 2 LSTM layers for three consecutive time steps. Between the two LSTM layers there is a dropout layer. The hidden state h_i of every cell is passed to the next layers of the model as input. Also, the hidden state along the cell state c_i are passed to the LSTM cells of the predictions for the time steps that follow.

The hidden states of every layer are matrices whose shape is determined by several factors, such as the number of the nodes of the layers and the size of the input. In a stacked network, the input of every RNN layer is determined by the hidden state of its previous RNN layer. Also, in case that there is a dropout layer in between the RNN layers, it may modify the input of the subsequent layer, resulting in a change to the shape of the hidden state. So, every small change in the network can make a very big difference on how the states are created, which translates to a different way of transferring information between different time steps.

Another important aspect of the model, which is determined by the size of the hidden states, are its trainable parameters. The bigger the hidden states, the bigger the weights matrices. The weights of the RNNs are the trainable parameters, and the number of trainable parameters can greatly influence the predictive power of the model. A large model is at risk of becoming overtrained and of lacking generalization. On the other hand, a model with a small number of parameters might turn out to be incapable of capturing complex relationships in the data.

To make predictions, we followed the same procedure that was described for the training. The predictions were made day by day and individually for each patient. The days that were initially missing or excluded during preprocessing were left out of the calculations of the accuracy.

4.5 Cross validation

This section examines how we used cross validation to tune the hyperparameters of the model. As when splitting in train-test set, avoiding autocorrelation between the train and validation sets of the cross validation was important. The two sets had to include completely different periods of time and different sets of patients.

The data of the training set covered the period between 26/07/2018 and 15/01/2021. This 905-day period could be split in five parts of 181 consecutive days each. We used these five parts to perform k fold cross-validation. In each of the five folds we trained our model in four parts and made predictions on the remaining part. For the predictions of each fold we then calculated the Area Under the Precision Recall Curve (AUC PR).

Similar with train-test split, when patients had data in both sets, we minimized the loss of information by discarding the period with less data. However, it was possible that patients had data both before and after the validation set. In such cases, if the total number of training data was higher than in the validation set, the data of the patient were split in two sequences, one before and one after the validation set.

During the first months of our study, we did not have many observations and the number of the patients was still increasing. As a result, the first 181 days, as a validation set of the first fold, had much less data than the other validation sets (see Figure 6). Therefore, in this fold, the estimation of the performance of the model could differ greatly from its actual value and this could lead in overestimating or underestimating it. Calculating the average of the performance metric would not give an accurate result of the performance of the model. Instead, taking into consideration the size of the validation set, we calculated a weighted average for the metric.

4.6 Bayesian Optimization

For the hyperparameter tuning of the model we used Bayesian Optimization. Testing sets of hyperparameters was computationally expensive and since the time for this project was limited, we could not create and test models with every possible combination of the configuration space. Bayesian Optimization is a very efficient method for choosing hyperparameters and it can help us achieve good results with much fewer tests than other methods, such as Grid Search and Random Search. With this method we did 50 tests by running Bayesian Optimization with 50 steps.

In Bayesian Optimization the goal is the optimization of a black box function using a set of parameters. This function, which is called objective function, is considered continuous and is usually difficult to evaluate. First, the model makes random observations of the objective function using the given parameter space. From these, it constructs a posterior which encapsulates all the knowledge of the model about the function that we want to optimise. The posterior is used for the creation of the acquisition function, which is then utilized to determine the next point in the parameter space. After every observation, the posterior gets updated and the algorithm gets a better estimation of the optimization function.

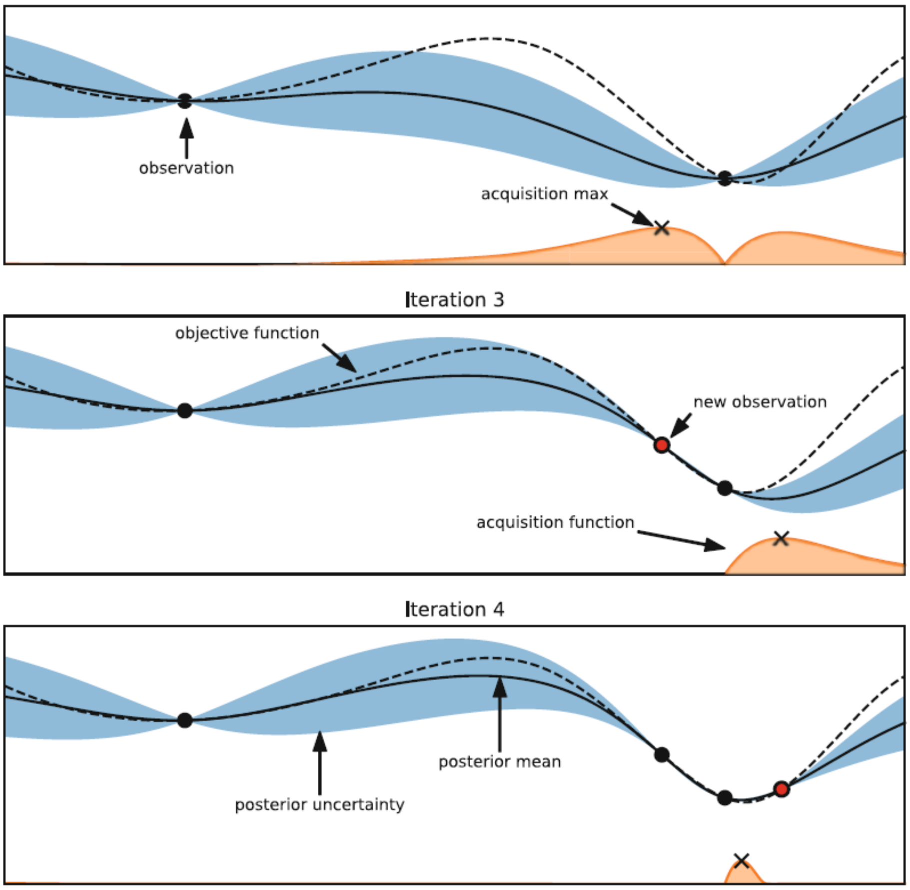


Figure 9: Example of Bayesian Optimization for one dimension. The dotted line is the objective function whose minimum we wish to find, and the black line is the predicted posterior mean based on the observations we have made. The maximum of the acquisition function is used to guide us to the next observation. Moving from two observations from the top picture to three and then four, it is shown how the posterior mean comes closer to the objective function. The more observations we have, the more we know about the function of interest. It is also shown that the closer we get to the observations, the more the uncertainty is reducing. (Taken from [19])

The procedure that is displayed in Figure 9 can be utilized for the search of hyperparameters. The chosen accuracy metric plays the role of the optimization function, and the parameter space consists of the possible values of our hyperparameters. In our case, the result of the cross validation was the objective function, which we were looking to maximize by searching the optimal settings for the hyperparameters. The configuration space of the hyperparameters can be seen in Table 3.

	Configuration Space
Nodes	n $\in Z : 10 \leq n \leq 100$
RNN Layers	[1, 2, 3, 4]
dropout rate	[0, 0.1, 0.2, 0.3, 0.4]

Table 3: Configuration space of hyperparameters

In the Bayesian optimization, the model accuracy was estimated using cross validation for every step of the process. After the completion of all 50 steps, we picked the hyperparameters that produced the model with the best Area Under the Precision–Recall Curve (AUC PR). It is important to note that each experiment was run only once because the cross validation was computationally expensive.

The hyperparameters with which we got the optimal performance were then used for the creation of the final models. These models were trained and tested with the corresponding train and test sets that we had created. The evaluation of the results of these models is explained in detail in the following section.

4.7 Model evaluation

Weather variables are often reported as a common trigger for migraines. However, every patient is different and it is very likely that a number of our patients are not weather-sensitive. The migraines that were recorded in the diaries might have been caused by numerous triggers that were not related to weather. So, a number of the migraines that occur in the diaries are impossible to be predicted in this study. Our goal was to identify days with increased probability of developing a migraine attack individually for every patient.

In this respect, looking at all the predictions of the model collectively, instead of individually for every patient, does not convey sufficient information on how well the model can perform for individuals. Moving in this direction we inevitably test the performance of the model on non-weather sensitive patients and on migraines that were induced by irrelevant triggers. However, there are a couple of advantages in looking at the results of many patients collectively. The inclusion of migraines that are not weather-related happens regardless of the different choices that were made for the model's creation. So, calculating the performance on all patients (overall performance) is a way of comparing models and choices that were made on the different aspects of the problem. From the comparison of the overall performance we could get insights on how to build a model and how our assumptions affect the results.

For this thesis project to evaluate our models' performance we selected the AUC PR. The baseline, with which we compared our models' performance, was the ratio of migraine days to the total number of days in the sequence. We chose this baseline and this performance metric due to the imbalance between the negative and the positive class of the outcome variable.

To evaluate our models' within-patient performance, we calculated the AUC PR for the predictions on each patient individually. In this way, we were able to have a close look on how our model's performance varies from one patient to another. It is important to note that the baseline performance for every patient is different due to the differences on the size and on the imbalance of their data.

In parts of the thesis that we compared the model's performance between different patients we used the values of (AUC PR - Baseline). Comparing the performance of the model between patients is not a trivial process. Each patient had different imbalances between the two classes of the outcome variable, so every patient's results were compared with a different baseline. Additionally, each patient followed the diary for a different number of days and for days with different weather conditions. For patients with a large number of diary inputs, we had the opportunity to test the model in a wider range of weather conditions and for a greater number of migraines. Thus, the more days there were in a patient's diary the better we could estimate the accuracy of our model.

4.8 Explainability

The interpretation of the predictions can be divided in two parts: all predictions for a group of patients as a whole and the predictions of each patient individually.

For the interpretation of the importance of each variable, it was essential to not use predictions for patients for which we get an AUC PR value close to the Baseline. By doing that we tried to avoid interpreting predictions on migraines that were not triggered by weather. In both parts, we made an arbitrary decision to use only the top 20% of the patients with the best (AUC PR-Baseline). In this way we excluded a number of patients that presumably were not weather-sensitive and weather-sensitive patients, whose migraines we did not manage to predict accurately. Finally, for both parts of the interpretation we only used the best performing model between LSTM and GRU. The choice of the model is explained in the Results section.

For the interpretation of the predictions we used the SHAP (SHapley Additive exPlanations) values. SHAP values are a method from coalitional game theory that can also be applied for interpretation of predictions and it is particularly useful for "black box" models. The SHAP values show the contribution of every predictor variable to each prediction comparing to the average prediction. [20] The total number of SHAP values that describe the contribution of predictor variables to a groups of predictions, can be illustrated in a graph. This graph helps us understand how each predictor variable influences our predictions.

To interpret the predictions for each patient individually, we had to identify the impact of each weather variable on our predictions separately for each patient. To do that, we have used the same subset of weather-sensitive patients as previously. For the predictions that we made for these patients, we calculated the SHAP values. By utilizing the SHAP values we then ranked the variables with respect to how influential they were to our predictions compared to the rest.

5 Results

The results of this thesis project are split in two sections. In the first section, we present the performance of the LSTM and the GRU models. The performance of the models was evaluated both in overall and with-patient. The second section examines the explainability of the predictions that were made with the LSTM model.

The hyperparameters settings that produced the best results in cross validation, can be seen in Table 4.

	Nodes	**Layers**	**dropout rate**
LSTM	39	4	4
GRU	10	4	2

Table 4: Resulting hyperparameters for each model type after Bayesian optimization

5.1 Overall Performance

	AUC Pr	
	Train set	**Test set**
Baseline	0.168	0.157
LSTM	0.185	0.177
GRU	0.183	0.171

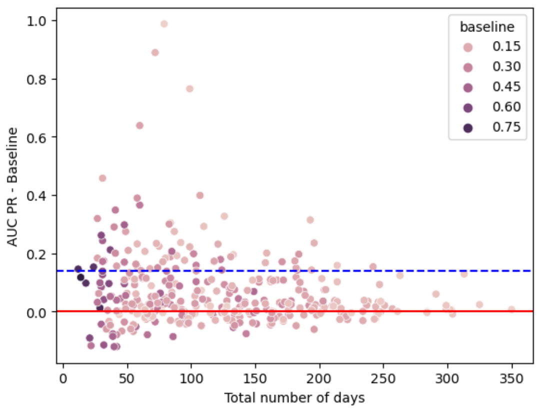
Table 5: Performance of the LSTM and GRU models using the optimal hyperparameters from the Bayesian optimization.

As we see in Table 5, the results of both the LSTM and GRU models are slightly better than the baseline in both the train set and the test set.

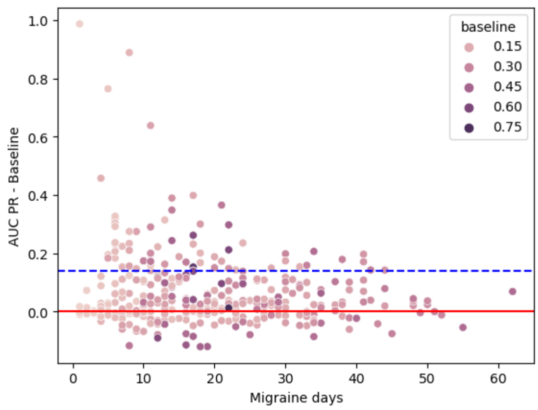
5.2 Within-Patient Performance

The within-patient results for the LSTM are displayed in the scatter plot of Figure 10a. The first thing that is evident here is that a big portion of the patients are around the red line, which indicates a performance equal to the baseline. The dot for approximately 30% of the patients in the test set is under the red line, which implies that for them we got worse predictions than the baseline.

However, in the same figure it can be seen that the results for some patients are much better than for others. There is a big number of patients whose dot is located well above the red line. For these patients we managed to get considerably better predictions than the baseline. This is an indication that our model might have some value for some patients.



(a) The total number of days that each patient had in their diaries (without counting the imputed days) is presented on the horizontal axis.



(b) The total number of migraines of every patient is represented on the horizontal axis.

Figure 10: Within-patient results of the LSTM model. The AUC PR - Baseline for every patient of this study is displayed with a dot. Both the baseline and the AUC PR are different from one patient to another. The colours of the dots indicate the baseline value of each patient. The red line corresponds to the AUC PR values that are equal to the Baseline value. The blue line indicates the (AUC PR - Baseline) value for the top 20% percentile.

5.3 Comparison of LSTM & GRU

As can be seen in Table5 the LSTM model is performing slightly better than the GRU model. This is observed in both the train and the test set.

To compare the within patients results we got for the LSTM and the GRU models, we have to choose patients that were found to have weather as a trigger in both models. For this reason, we identified the top 20% of all patients based on the (AUC PR- Baseline) for each model. From these two sets we found that 68% of the patients were shared between the two models, indicating that the two models achieved good predictions for similar sets of patients.

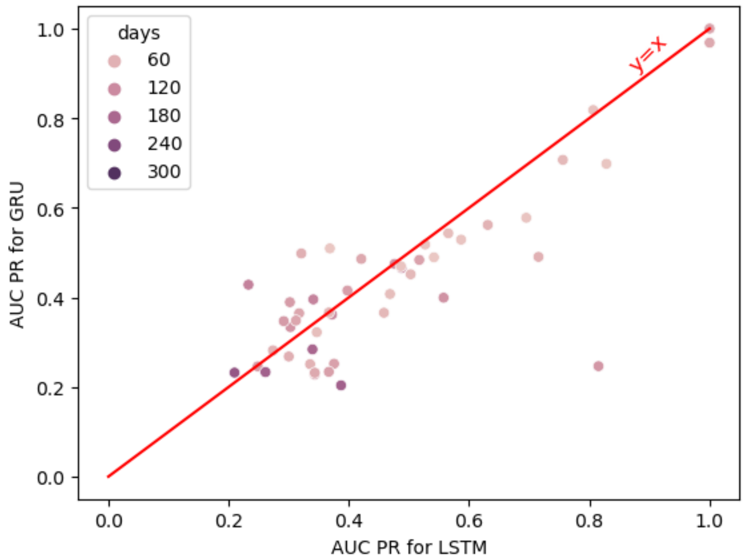


Figure 11: Scatter plot of patients with the AUC PR for LSTM and GRU. Patients of this graph belong in the top 20% of all patients regarding their value of (AUC PR - Baseline) for both the GRU and the LSTM models. The number of the dots displays the number of days each patient filled their diary. The red line represents the identity function x=y.

In Figure 11 it is shown that the predictive power of the two models were found similar for the majority of the patients. Also, the results from the LSTM tend to be better than the results of the GRU, as most patients are located under the diagonal of x=y. This provides further evidence of the superiority of the LSTMs in our case.

It is important to note that, the baseline and number of days change considerably between patients. In this respect, it is not easy to compare the model's performance between two patients. So, the patients that ended up to be in the top 20% are not necessarily the patients with the best results.

5.4 Explainability

For the explainability part of this project, we used the predictions made with the LSTM model. Although for some patients we managed better results with GRU, see Figure 11, the overall results of the LSTM model were found to be better than the GRU.

5.4.1 Collective interpretation of predictions

Figure 12 depicts the impact that each variable had on our predictions. The most clear finding from this graph is that for the patients in this study temperature is the most predominant trigger. The rest of the triggers have similar effects on the predictions and there are not big differences between them.

Another thing that can be observed from Figure 12 is that there is no apparent correlation between the values of each of the variables and predicted probability of a migraines. This means that it is not clear how each variable could lead to a migraine. Apparently, the connection between them and the possibility of getting a migraine is not a linear relationship. It is possible that it is the combination of weather conditions that causes an increased risk of getting a migraine.

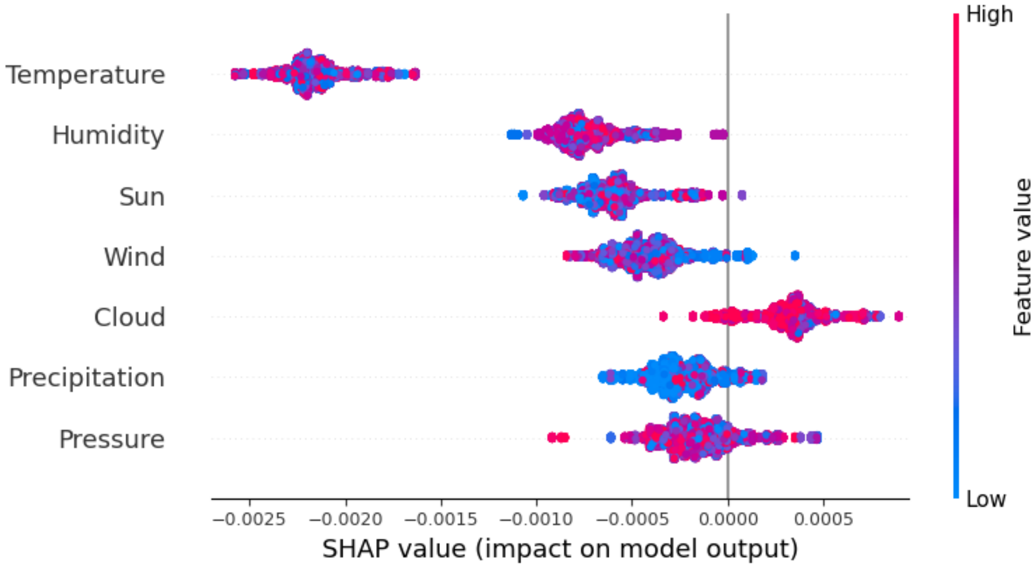


Figure 12: SHAP Values for all the predictions collectively of the LSTM on the 20% of patients. The colour indicates the value of each variable for each prediction.

5.4.2 Interpretation of individual predictions

The first thing that stands out from Figure 13 is that temperature is again the most predominant variable for most of the patients in this group. Following the results displayed in Figure 12, this is another indication of how important temperature is found to be for migraine predictions. In Figure 13 it is also illustrated that not all patients have the same triggers, and that the importance of every variable can change dramatically from patient to patient. This again stresses the need for models that can make personalised predictions and explore the patterns between triggers and migraines individually.

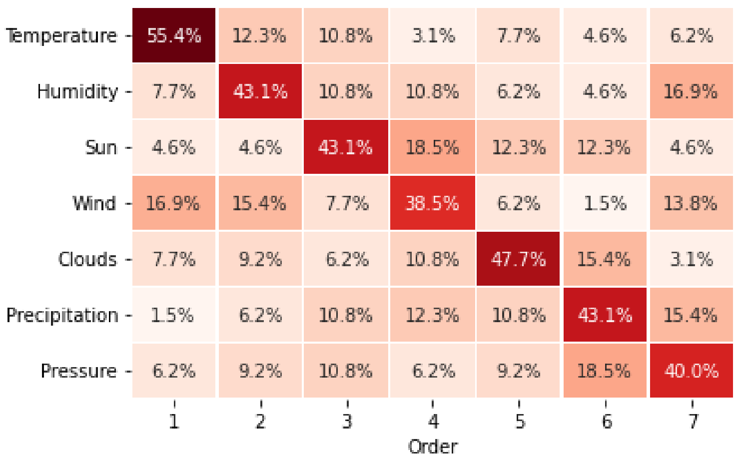


Figure 13: Percentage of patients, for which variable y was placed in position x of the order. This graph is based on the SHAP values for the predictions on a sub-group of weather-sensitive patients.

6 Discussion

Our study aimed to develop and compare models that make personalised predictions on migraines based on weather data, and to investigate which weather conditions are the most predominant triggers. Based on our findings, both LSTM and GRU models could be useful in making migraine predictions for some patients. The AUC PR that was calculated for the total of all the predictions was 0.185 for the LSTM and 0.183 for the GRU. Both performed only slightly better than the baseline.

It might have been anticipated that, because of the LSTM's long term memory, the results for patients with big number of days would be much better with the LSTM than with the GRU model. However, from Figure 11 it is not clear if the number of days affects which of the two types of models is performing better.

6.1 Conclusions

From the within-patient results, it was shown that the results between patients differed greatly. However, our two models performed very similarly for weather-sensitive patients. This is an indication that our models could provide some weather-sensitive patients with valuable information on when they would face an increased risk for weather-induced migraine. This could be done with the utilization of weather forecasts for the coming days.

The trigger that was the most important for the total of all predictions was also the trigger that was the most important for most patients individually. This trigger was temperature and it is consistent with previous studies. All of the variables that we included in this study were found to be important triggers for some of the weather-sensitive patients. These results demonstrate the need for personalised models and for the inclusion of a big number of potential triggers.

The models' performance for many of our patients was found to be close or below the baseline. This is something that was expected after the results we got for the overall AUC PR. For these patients, either weather is not a trigger or the dependency with migraine is not strong enough to cause an adequate number of migraines to give us the opportunity to capture it.

The models' performance was also affected by several limitations that were identified in this study. One of the prominent restrictions was the way that the weather was measured. It has been shown in many studies that migraines could be triggered because of certain weather conditions at a specific time during the day. In this study we only had one value per variable per day, and in most cases it was the average of the day. In our case it is impossible to capture such dependencies. Having at our disposal multiple values for the variables per day could help us make better predictions. Also, the averages we have for the day are affected by the weather conditions after the onset of migraine attacks.

An additional limitation is that the weather was measured by weather stations in the area close to where the patients lived. However, we do not know how close each patient was to the weather station. Thus, we can not be sure on how well the weather at the station resembles the weather where each patient lived. Additionally, it is expected that patients did not spend every day at their homes, which makes the estimation of the weather even less efficient.

The next limitation has to do with the hyperparameters selection that was done with Bayesian optimization. In order to get an accurate estimate of the accuracy of the model for each setting, it would be best if for every step of the optimization, we could re-run each experiment multiple times and then get their average. However, this would be computationally expensive and it would vastly increase the time that each experiment would need to finish.

As explained in section Missing Data, we encountered some problems with missing data, due to the nature of the models that we have chosen to use. A possible improvement would be to have stricter restrictions on the compliance of the patients. One could also try applying different imputation methods or a smaller threshold for differentiating "small" from "big" gaps of missing data.

6.2 Future work

There are several future directions that could be considered for potential improvements on our models.

The use of Bayesian Optimization for the tuning of the models helped us to search for the hyperparameters efficiently without the need of trying all the different combinations. However, one way to improve on the models of our study would be to run the Bayesian optimization for more steps and for a greater number of repetitions in each step. In this way, it would be easier for the algorithm to find the best values for the hyperparameters, and the results would be less affected by the randomness of the initialization.

In the way the models of this study were created, it would be easy to add new variables that are potential triggers. Adding levels of stress, sleep, caffeine and other triggers would enable us to predict more migraines, and thus vastly improve the performance of the models.

Also, improving the way the weather variables were measured would be a big step towards the direction of making our predictions more accurate. Utilization of smartphones and wearables could be beneficial to the solution of this problem. Smart devices could daily obtain accurate weather estimates from the closest located weather station to the patients' location.

Predicting migraines is a complicated task. The differences between patients and the big number of unidentified triggers are the two main difficulties that predicting models have to face. Personalised predictions with models that can incorporate a big number of parameters seem to be the future of migraine predictions. Technology is going to play a significant role, as variables will be more accurately measured and the computational power will grow. Hopefully, through this evolution, in the coming years migraines will become less disabling for migraine patients.

7 Appendix

7.1 Code

The Python and R codes used in this thesis can be found in the following GitHub repository: `https://github.com/AntonisRoid/Thesis.git`

The migraine dataset that was used is not in the GitHub repository for confidentiality reasons. The weather dataset can be found here: `https://www.daggegevens.knmi.nl/klimatologie/daggegevens`

7.2 Figures and Tables

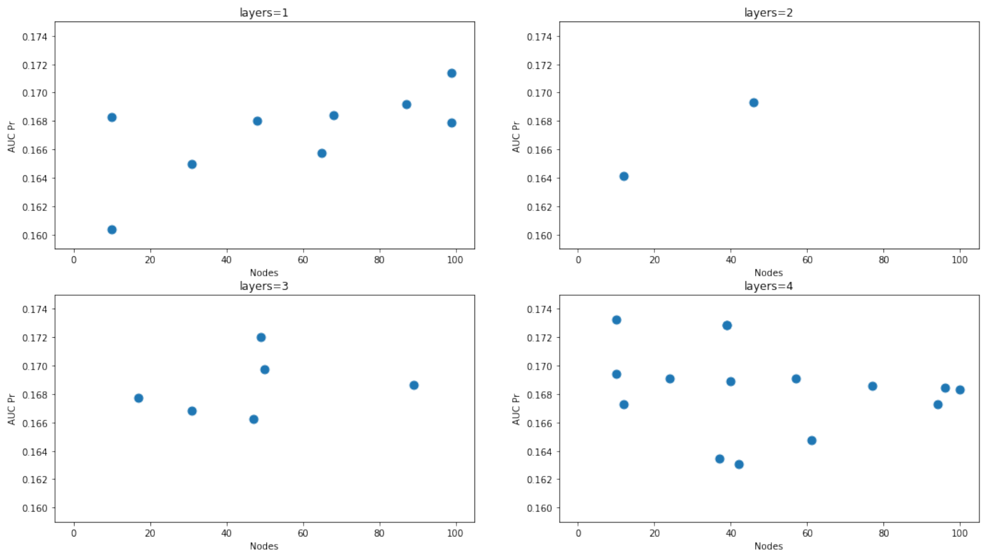


Figure 14: Results of Bayesian Optimization for the LSTM. The y axis displays the performance of the model and the x axis the number of Nodes per LSTM cell. Every one of the four graphs corresponds to a different value for the number of LSTM layers. The performance for combinations that were tested more than once are displayed here as averages.

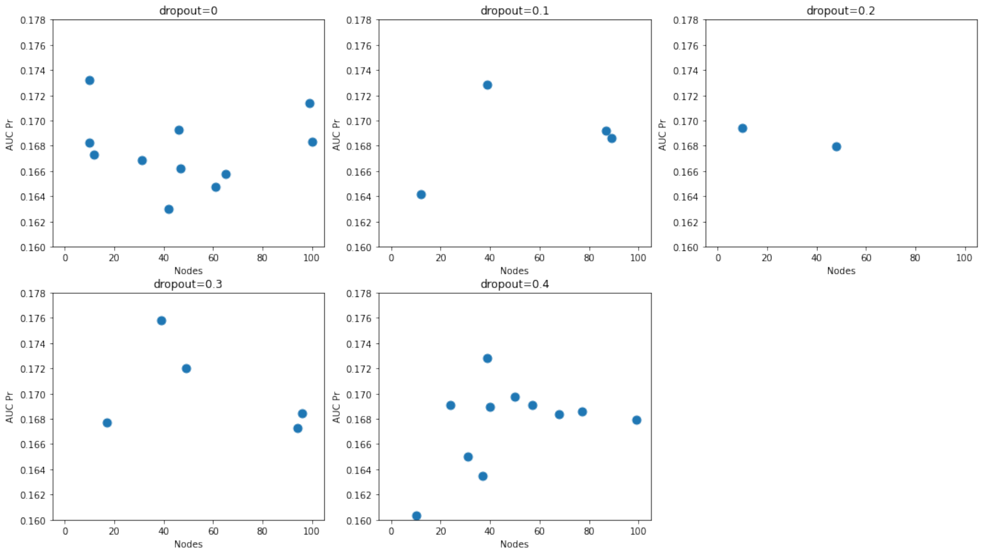


Figure 15: The Results of Bayesian Optimization for the combination of dropout rate and nodes are displayed similarly those on Figure 14.

Author	Year of publication	Number of Patients	Follow up period	Method	Used variables	Results
Wilkinson et. al [21]	1979	310	Each attack was monitored individually without any follow up	Variables were correlated to the frequency of migraine incidences	Wind Direction, Wind Velocity, Barometric Pressure, Humidity and Temperature Variables were measured at the time and three hours prior to the onset of the migraine attack	Atmospheric Changes were not correlated with the frequency of migraine attacks
Schulman et. al [22]	1980	75	One month	Log-Linear Model with use of Markov Chain	Barometric peressure	Little to no relationship between headache occurance and Barometric pressure
Cull et. al [8]	1980	44	Six months	Variables were correlated to the frequency of migraine incidences	Barometric pressure on the day and the day before the migraine attack. (Days, weeks and months of the year.)	Barometric pressure at 06.00 hrs bellow 1005mb were found to be associated with reduced numbers of migraine attacks Increase of 15 mb or more in pressure was associated with smaller migraine frequencies
Robbins et. al [23]	1993	494	Retrospective Study	Variables were correlated to the frequency of migraine incidences	Weather changes, Bright sunlight and Seasons of the year.	Weather changes were found highly significant as it was a trigger for 43% of patients. Sunlight was a trigger factor for 38% of the patients Small fluctuations between seasons with spring and fall being associated with more headaches
Prince et. al [6]	2004	77	Varying form 2 to 24 months	Variables were used to generate three factors which were later used in Linear Regression	Factor 1: Absolute temperature & humidity Factor 2: The changes of weather Factor 3: Barometric pressure	Most weather-sensitive patients were sensitive to high or low humidity combined with high or low temperature respectively. Many patients were influenced by the change in weather pattern and at the barometric pressure values.
Zebenholzer et. al [24]	2010	238	90 days	Stepwise Multivariate Cox Regression Analysis	28 weather related variables	Influence of weather on migraine was not found significant
Yang et. al [9]	2015	66	1 year	Empirical mode Decomposition and Regression Analysis	Temperature	During winter, when the temperatures are low, in some cases the migraine incidences are associated with temperature.
Pellegrino et. al [25]	2017	Meta-Analysis	Not applicable	Meta- Analysis	53 weather related variables	Weather/ Environment was in the top 4 of perceived triggers, with 33% of the patients.
Wenyuan et. al [2]	2019	98	Average of 45 days	Fixed Effects Logistic Regression Logistic Regression Models with GEE	Temperature, barometric pressure and relative humidity.	For the period from April to September, high humidity is associated with risen migraine incidences
Casteren et. al [26]	2021	6786	cross-sectional study	Logistic Regression	Bright (sun)light and Weather changes	Both were found significant. Bright (sun)light was perceived as a trigger for: 68% of females and 63.2% of males And Weather changes for: 45.9% of females and 38.7% of males
Hoffmann et. al [27]	2015	100	12 months	Logistic Regression	Atmospheric pressure, relative air humidity and ambient pressure	Weather sensitivity was detected for only subgroups of the patients.
Hoffmann et. al [28]	2011	20	12 months	Logistic Regression	Atmospheric pressure, temperature and relative air humidity	Low temperature and high relative humidity were found to be a trigger for a subgroup of patients
Holsteen et. al [10]	2020	178	90 days	Multivariable Multilevel Logistic Regression models	Max Barometric pressure, Max Temperature, humidity/ precipitation and wind speed	Could not predict migraine attacks with greater success than a random guess

Table 6: Summary of the findings on the association of Weather and Migraines

References

[1] M. Ashina. "Migraine". In: *N Engl J Med* 383.19 (Nov. 2020), pp. 1866–1876. DOI: `10.1056/NEJMra1915327` (cit. on p. 4).

[2] Li Wenyuan (ScD) et al. "Weather, ambient air pollution, and risk of migraine headache onset among patients with migraine". In: *Environment International* (2019). DOI: `https://doi.org/10.1016/j.envint.2019.105100` (cit. on pp. 4, 29).

[3] P. J. Goadsby et al. "Pathophysiology of Migraine: A Disorder of Sensory Processing". In: *Physiol Rev* 97.2 (Apr. 2017), pp. 553–622. DOI: `10.1152/physrev.00034.2015` (cit. on p. 4).

[4] T. T. Houle et al. "Forecasting Individual Headache Attacks Using Perceived Stress: Development of a Multivariable Prediction Model for Persons With Episodic Migraine". In: *Headache* 57.7 (July 2017), pp. 1041–1050. DOI: `10.1111/head.13137` (cit. on p. 4).

[5] S. J. Peroutka. "What turns on a migraine? A systematic review of migraine precipitating factors". In: *Curr Pain Headache Rep* 18.10 (Oct. 2014), p. 454. DOI: `10.1007/s11916-014-0454-z` (cit. on p. 4).

[6] P. B. Prince et al. "The effect of weather on headache". In: *Headache* 44.6 (June 2004), pp. 596–602. DOI: `10.1111/j.1526-4610.2004.446008.x` (cit. on pp. 4, 29).

[7] J. M. Pavlovic et al. "Trigger factors and premonitory features of migraine attacks: summary of studies". In: *Headache* 54.10 (2014), pp. 1670–1679. DOI: `10.1111/head.12468` (cit. on p. 4).

[8] R. E. Cull. "Barometric pressure and other factors in migraine". In: *Headache* 21.3 (May 1981), pp. 102–103. DOI: `https://doi.org/10.1111/j.1526-4610.1981.hed2103102.x` (cit. on pp. 4, 29).

[9] Albert C. Yang et al. "Patients with migraine are right about their perception of temperature as a trigger: time series analysis of headache diary data". In: *The Journal of Headache and Pain* 16.1 (May 2015), p. 49. DOI: `10.1186/s10194-015-0533-5` (cit. on pp. 4, 10, 29).

[10] K. K. Holsteen et al. "Development and Internal Validation of a Multivariable Prediction Model for Individual Episodic Migraine Attacks Based on Daily Trigger Exposures". In: *Headache* 60.10 (Nov. 2020), pp. 2364–2379. DOI: `https://doi.org/10.1111/head.13960` (cit. on pp. 4, 5, 29).

[11] W. B. Young, M. F. Peres, and T. D. Rozen. "Modular headache theory". In: *Cephalalgia* 21.8 (Oct. 2001), pp. 842–849. DOI: `10.1046/j.1468-2982.2001.218254.x` (cit. on p. 5).

[12] M. Schürks, J. E. Buring, and T. Kurth. "Migraine features, associated symptoms and triggers: a principal component analysis in the Women's Health Study". In: *Cephalalgia* 31.7 (May 2011), pp. 861–869. DOI: `https://doi.org/10.1177/0333102411401635` (cit. on p. 5).

[13] D. C. Buse, J. S. McGinley, and R. B. Lipton. "Predicting the Future of Migraine Attack Prediction". In: *Headache* 60.10 (Nov. 2020), pp. 2125–2128. DOI: `10.1111/head.14025` (cit. on p. 5).

[14] Samir Khan and Takehisa Yairi. “A review on the application of deep learning in system health management”. In: *Mechanical Systems and Signal Processing* 107 (2018), pp. 241–265. ISSN: 0888-3270. DOI: `https://doi.org/10.1016/j.ymssp.2017.11.024` (cit. on p. 7).

[15] Zachary Lipton et al. “Learning to Diagnose with LSTM Recurrent Neural Networks”. In: (Nov. 2015). DOI: `10.48550/ARXIV.1511.03677` (cit. on p. 7).

[16] E. Choi et al. “Using recurrent neural network models for early detection of heart failure onset”. In: *J Am Med Inform Assoc* 24.2 (Mar. 2017), pp. 361–370. DOI: `10.1093/jamia/ocw112` (cit. on p. 7).

[17] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition*. O’Reilly Media, Inc., 2019. ISBN: 9781492032649 (cit. on p. 8).

[18] Daphne S van Casteren et al. “Sex differences in prevalence of migraine trigger factors: A cross-sectional study”. In: *Cephalalgia* 41.6 (2021), pp. 643–648. DOI: `10.1177/0333102420974362` (cit. on p. 10).

[19] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, eds. *Automated Machine Learning - Methods, Systems, Challenges*. Springer, 2019 (cit. on p. 18).

[20] Christoph Molnar. *Interpretable machine learning: A guide for making Black Box models explainable*. Christoph Molnar, 2022 (cit. on p. 20).

[21] M. Wilkinson and J. Woodrow. “Migraine and weather”. In: *Headache* 19.7 (Nov. 1979), pp. 375–378. DOI: `10.1111/j.1526-4610.1979.hed1907375.x` (cit. on p. 29).

[22] J. Schulman et al. “The relationship of headache occurrence to barometric pressure”. In: *Int J Biometeorol* 24.3 (Sept. 1980), pp. 263–269. DOI: `10.1007/BF02249796` (cit. on p. 29).

[23] L Robbins. “Precipitating Factors in Migraine: A Retrospective Review of 494 Patients”. In: *Headache: The Journal of Head and Face Pain* 34 (1994). DOI: `https://doi.org/10.1111/j.1526-4610.1994.hed3404214.x` (cit. on p. 29).

[24] K. Zebenholzer et al. “Migraine and weather: a prospective diary-based analysis”. In: *Cephalalgia* 31.4 (Mar. 2011), pp. 391–400. DOI: `10.1177/0333102410385580` (cit. on p. 29).

[25] A. B. W. Pellegrino et al. “Perceived triggers of primary headache disorders: A meta-analysis”. In: *Cephalalgia* 38.6 (2018), pp. 1188–1198. DOI: `10.1177/0333102417727535` (cit. on p. 29).

[26] Daphne S van Casteren et al. “Sex differences in prevalence of migraine trigger factors: A cross-sectional study”. In: *Cephalalgia* 41.6 (2021), pp. 643–648. DOI: `10.1177/0333102420974362` (cit. on p. 29).

[27] J. Hoffmann et al. “The influence of weather on migraine - are migraine attacks predictable?” In: *Ann Clin Transl Neurol* 2.1 (Jan. 2015), pp. 22–28. DOI: `10.1002/acn3.139` (cit. on p. 29).

[28] J. Hoffmann et al. “Weather sensitivity in migraineurs”. In: *J Neurol* 258.4 (Apr. 2011), pp. 596–602. DOI: `10.1007/s00415-010-5798-7` (cit. on p. 29).