Computer Science and Economics

Weather as a Trigger for Migraines:

Predictive Modelling on Migraine E-Diaries using Machine Learning

Milou Schamhart

Supervisors:
Dr. M. van Leeuwen & H.A.J. Spaink & M. Vinkenoog

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University Medical Centre (LUMC)

04/08/2021

Abstract

Background: Migraine is a prevalent, multifactorial brain disease, characterized by an intense form of headache, typically causing pain in one half of the head, nausea and visual changes. It affects approximately 1 out of 7 adults annually. There is a lot known about the pathology of migraine, however there still is a lot of speculation about the triggers. 53% of migraineurs indicate weather to be a trigger for attacks.
Related work: Research into the correlation between weather and migraine is contradictory. This is explained by the existence of a subgroup of migraine patients with weather dependent migraine. However, there is little research done into identifying this subgroup. This research will expand the current research area of the relation between migraine and the weather by using machine learning to predict the start of migraine attacks and thereby identify migraineurs with weather dependent migraine.
Data: For the migraine data, LUMC LUMINA e-diaries are used. From these diaries, the age of patients, the patients' indication on whether they have weather dependent migraine and the start dates of migraine attacks were used. The weather data was publicly available from the KNMI. The temperature, sunshine, precipitation, air pressure, cloudiness and relative humidity were taken into account.
Methods: First, the data was preprocessed, selecting the usable patients. The 860 patients selected were linked with the weather data using zip codes. These patients were divided into 5 folds using stratified train test split. The stratification was based on the age, number of months in the diary and whether the patient indicated to have weather dependent migraine. Then the data sets were applied on random forest and support vector machine using 5-fold cross validation.
Results: The prediction on the dataset using random forest and support vector machine have a mean accuracy over all the folds of 60.5% for random forest and 59.7% for support vector machine. Also, a subgroup of patients with a sensitivity > 0.5 and a specificity > 0.5 was analysed. This subgroup contains approximately 10% of the patients. Random forest and support vector machine show a linear correlation when plotting the accuracy per patient of both models.
Discussion & Limitations: Random forest is better at finding a pattern in the subgroup leading to 71 more patients with a sensitivity > 0.5 and specificity > 0.5 than support vector machine. No distinctive pattern could be found in the baseline characteristics of the subgroup.
Conclusion & further research Using classification models we found that for approximately 10% of the patients we were able to make predictions with a sensitivity > 0.5 and a specificity > 0.5 solely based on weather and weather changes. This shows that there are patterns in weather conditions which could be a trigger for patients with weather dependent migraine.

Contents

1 Introduction

1.1 The situation

Migraine is a prevalent, multifactorial brain disease, characterized by an intense form of headache, typically causing pain in one half of the head, nausea and visual changes. It is a neurological abnormality affecting approximately 1 out of 7 adults annually. It predominantly affects women: 1 in 3 experience migraine symptoms in their life, making it the third most common condition for adults [BSLS15] [Rop20].
Globally, migraine is ranked number two for highest burden diseases [VAA+17]. This ranking is based on the number of people with the disease and the number of days the patients are unable to function due to the disease. However in the Netherlands, migraine does not appear in the top 10 highest burden diseases. This difference might be caused by the fact that, compared to the global average of migraine patients in the population (1 out of 7), the Netherlands has relatively few migraine patients (approximately 1 out of 70) [Mig].
The quantity of migraine attacks per patient differs drastically. According to headache frequency migraine patients can be divided into two groups: chronic (CM) and episodic migraine (EM). Here we define CM according to ICHD-3 criteria of having more than 15 headache days per month of which at least 8 were migraine days for at least 3 consecutive months [IHS21]. EM is when the patient does not fulfil the criteria for CM [IHS21]. Due to overuse of acute medication a patient can develop medication overuse headache (MOH). MOH is defined according to the ICHD-3 criteria as 15 or more simple analgesics days, 10 or more triptan days, or 10 or more acute medication days including at least 1 triptan day during 3 subsequent months [IHS21].
Because migraine is not a deadly disease, less focus is placed on migraine prevention when compared to migraine treatment. More research into migraine prevention would be beneficial not only for patients, but also economically. Approximately 37.4 million euros is spent per year on migraine healthcare costs in the Netherlands, even though there are relatively few migraine patients. Of the 37.4 million euros only 1.3% is spent on prevention, the remainder is spent on treatment after the start of an attack [Mig]. On top of the healthcare cost, it is also costly for companies, as migraine attacks are generally too severe for patients to work during an attack [LR13]. A start to preventing migraine attacks is to know more about what triggers them. Although a lot is known about the pathology of migraine [PM13], there is still a lot of speculation about what triggers migraine.
Weather is one of the top five most named triggers by migraineurs, indicated by approximately 53% of patients [Kel07]. However, Hoffmann et al. state that the scientific evidence for weather being a trigger factor for migraine attacks is inconclusive [HHLMR11]. Studies performed to find a correlation between weather and migraine took place in different climatological zones, using different sample sizes and different methods [ZRF+11] [GS73] [LBM+19] [YFH+15]. This could influence the results of the studies. If all migraine patients with weather dependent migraine are sensitive to similar weather conditions or weather changes, the climate can affect whether a correlation is found. Furthermore, there are more triggers for migraine than the weather. What these studies have in common is that each examines an entire dataset of migraine patients [ZRF+11] [GS73] [LBM+19] [YFH+15]. However, not all patients might experience weather dependent migraine. The hypothesis suggested in several studies is that only a subgroup of migraineurs have weather dependent migraine [HSHL+15] [PRS+04].
However, there is little research done into identifying the subgroup of migraineurs with weather

dependent migraine. The article by Hoffmann et al. describes how a part of their patients show a correlation with weather and identifies these migraineurs as the subgroup. Nonetheless, there is no additional description of the patients in the subgroup [HSHL+15].
This research will expand the current research area of the relation between migraine and the weather by using machine learning to identify migraineurs with weather dependent migraine. After identifying the subgroup, we will analyse and describe the patients in it. As a result, the gap in research to predict migraine using the weather and identify patterns in the subgroup of patients with weather dependent migraine will be filled.
Additionally, other related research done to find a correlation between migraine and the weather has not used machine learning to predict migraine using weather variables. In this research we will use machine learning models to predict migraine based on weather conditions and thereby add to what is known about the correlation between weather and migraine.

1.2 Research question

The goal of this thesis project is to gain a better understanding of the relation between episodic migraine and the weather.
To that end, we will use weather data to train machine learning models on migraine patients in order to identify what patterns can be found that are a good predictor for a subgroup of patients with weather dependent migraine.
This leads to the following **research question**:

> To what extent is it possible to predict migraine on the basis of weather factors for migraineurs and thereby identify patterns that are a valid predictor for a subgroup of migraineurs with weather dependent migraine?

The following **sub-questions** will be addressed in order to answer the research question:

1. How can classification models be used to develop a predictive model for migraine based on weather data?

2. What patterns can be identified within the subgroup of patients for which the model performs better?

To answer the research question we will use machine learning to predict migraine on episodic migraine patients using weather data. The performance of the machine learning models are then evaluated per patient in order to identify patients with weather dependent migraine. The subgroup of migraineurs will then be described in order to identify patterns that are an adequate predictor for this subgroup. This can also be useful for future research in order to identify the subgroup of patients with previously established weather dependent migraine and train models on the subgroup. This way, there would be no need to train and test models on the entire dataset of migraine patients.

1.3 Thesis overview

This chapter contains the introduction. In section 2, related work into the use of machine learning methods in the medical field and the correlation between migraine and the weather will be discussed. Section 3 includes a description of the sources of the weather and migraine data used in this research

and of the data itself. Section 4 describes the method used and in section 5, the results obtained by using this method are shown. In section 6, the meaning of the results and the limitations of this research are discussed. Section 7 gives a conclusion of the results and answers the research question. This section also contains suggestions for further research.
This is the outline of my bachelor thesis, done within LIACS and LUMC and supervised by Matthijs van Leeuwen, Hermes Spaink and Marieke Vinkenoog.

2 Related Work

In this section related research is discussed. Firstly, the application of machine learning in the medical field related to migraine is discussed. After that, a description of the research into the correlation between weather and migraine is given.

2.1 Machine learning in the medical field

The usage of machine learning techniques for disease prediction has shown a potential application area for these methods [SKHM19]. "Data mining techniques have been widely used in developing decision support systems for disease prediction through a set of medical datasets" [NIAS17]. This is also the case for migraine. Machine learning is successfully used to identify automatic predictors in migraine classification [PZS+20]. A research performed by Garcia-Chimeno et al. uses random forest for feature reduction, followed by support vector machine for the classification of migraine patients. They concluded that this method can be used to "support specialists in the classification of migraines in patients undergoing magnetic resonance imaging" [GCGZGB17]. However, as there are many triggers for migraine, there is fewer research into predicting migraine attacks compared to classifying migraine patients.

2.2 Correlation between weather and migraine

There have been multiple studies investigating the correlation between migraine and the weather although with contradicting results. Table 1 gives an overview of several related studies where we focus on the study location, sample size, included meteorological factors, and a summary of the results.

Table 1: Information and findings of research into the correlation between migraine and the weather

Study	**Location**	**Sample size**	**Used weather variables**	**Findings**
Migraine and weather: a prospective diary-based analysis. [ZRF+11]	Vienna, Austria	238 patients	Minimum, mean and maximum air temperature. Minimum, mean and maximum atmospheric pressure. Minimum, mean and maximum wind speed. Sum of sunshine duration. Mean relative humidity. Sum of precipitation.	"Influence of weather factors on migraine and headache is small and questionable". [ZRF+11]
Variations in migraine attacks with changes in weather conditions. [GS73]	Aberdeen, United Kingdom	56 patients	Bright light. Cooling. Thunderstorms. Blizzards. Specific winds.	No association between the prevalence of migraine and the changes in weather conditions.
Migraine and weather. [WW79]	London, United Kingdom	310 patients	Wind direction. Velocity. Barometric pressure. Temperature. Humidity.	The frequency of migraine attacks is not influenced by the weather conditions.
Weather, ambient air pollution, and risk of migraine headache onset among patients with migraine. [LBM+19]	Boston, United States of America	98 patients	Temperature. Relative humidity. Barometric pressure.	In the warm season, a higher relative humidity is correlated with higher odds of migraine attacks.
Patients with migraine are right about their perception of temperature as a trigger: time series analysis of headache diary data. [YFH+15]	Taipei, Taiwan	66 patients	Temperature.	There is a relation between the temperature and migraine, but whether it is positive or negative is not stated.
Barometric Pressure and Other Factors in Migraine. [Cul81]	Edinburgh, United Kingdom	44 patients	Barometric pressure.	When there is a lower barometric pressure, migraineurs get fewer migraine attacks. However, no evidence was found that for a higher barometric pressure there are more migraine attacks.

The studies summarized in Table 1 all investigated the correlation between weather and migraine. However, the first three studies in Table 1 conclude that there is no correlation between migraine and the weather, which is the opposite of what the last three studies in Table 1 conclude.

Becker provides reasons for explaining these contradictory results and why multiple studies have not been able to find a significant association between weather factor and migraine. In summary, the five reasons for this are:

1. "Most migraine patients report multiple triggers" [Bec11].
2. "A specific trigger may not precipitate an attack with each exposure" [Bec11].
3. "The mechanisms by which weather-related factors or indeed any trigger factor precipitates migraine attacks are not understood" [Bec11].
4. "Many weather changes do not happen abruptly, may occur at different times in neighbouring locations as a system moves across the landscape, and the lag time between a trigger and migraine onset is not well established and may be variable" [Bec11].
5. "Migraine populations are heterogeneous, and what may be a trigger for one individual may not be a trigger for another" [Bec11].

Taking patient perceptions into account, it is likely that weather is a trigger for migraine, only not for every migraine patient [Bec11].
In addition, J. Hoffmann et al. concluded that there is a subgroup of migraineurs sensitive to weather. During their study, they assessed the response of 100 migraine patients to differences in atmospheric pressure, relative air humidity, and ambient temperature in 4-hour intervals over 12 consecutive months. They analysed the weather factors in the 24 hours preceding a migraine attack. If a patient showed a positive correlation, they determined the predictability by using logistic regression analysis [HSHL+15].
In this thesis, we will presume the existence of a subgroup within migraine patients being sensitive to weather.

3 Data

For this research,electronic headache diaries from patients of the Leiden University Medical Centre Migraine Neuro Analysis Programme (LUMINA) are used. This is a dataset obtained by the Leiden University Medical Centre (LUMC) containing data from clinically diagnosed migraine patients. Additionally, meteorological data from the Royal Dutch Meteorological Institute (KNMI) was used. This is a publicly available weather dataset collected from multiple weather stations throughout The Netherlands. In this section, an explanation of the resources of the data and a description of the content of the data is provided.

3.1 Data resources

LUMC: LUMINA migraine e-diaries. LUMC migraine patients have kept e-diaries in which they kept track of when migraine attacks occur, the headache characteristics (if applicable), the intensity, the use of medication and whether they experienced aura effects. The e-diaries are filled in by the patients daily by a link they received every morning, 9.00 am, about the previous 24 hours. The patients had to answer 6-31 questions, depending on the presence of headache and associated symptoms [vCVdB+21]. If the daily diary was not completed by 6.00 pm, the patients received a reminder, once again including the link to the e-diary. Also, monthly diaries are filled in every 28 days, keeping track of medication use, menstrual cycle regularity and (post)menopausal status [vCVdB+21]. In addition, patient meta-data including their home address, age and sex are noted. The e-diaries are proven to be useful in diagnosing migraine and obtaining reliable information [vCVdB+21].
For this research the following data was used:

- the dates of the start of migraine attacks,
- the number of months in the diaries,
- the age of the patients,
- the patients' response to whether they think weather triggers their migraine,
- the zip codes to their home addresses.

Migraine headache data is collected from patients included at the LUMC headache clinic or patients recruited for research purposes. In this research we will look at the weather prior to the start dates of the migraine attacks.
KNMI: Weather data. Weather data is publicly available via the KNMI site, containing the daily weather conditions, collected by multiple weather stations throughout the Netherlands [KNMa]. There are a total of 48 weather stations in the Netherlands, of which 34 are on land and 14 in the sea [KNMb]. The locations of the weather stations are based on requirements of the World Meteorological Organisation about the spatial distribution [KNMb]. These weather stations actualise the weather conditions every 10 minutes [KNMc]. We will consider combinations of the following weather factors: minimum, maximum and average temperature, maximal sun percentage, total precipitation, average air pressure, average cloudiness and average humidity. The average of

Table 2: Baseline characteristics of the LUMINA patient population

	Total migraine population
Patients, n	2004
Age, mean ± sd	43.0 ± 10.4
Migraine attack days per month*, mean ± sd	4.3 ± 3.6
Females, n (%)	1726 (86.1)
Patients with aura, n (%)	1145 (57.1)
Think to be sensitive to weather, n (%)	876 (43.7)

*A month was defined as a regular period of 28 days

24 hours is calculated over the 144 daily measurements of the weather condition. The minimum and maximum of 24 hours are respectively the minimum and maximum of the 144 daily measurements.

3.2 Data description

Patient data: Only a part of the total population of migraine patients is used for training and testing the models. This will be explained in the section 4. The baseline characteristics of the LUMINA patients are given in (Table 2).

Table 3: Measurement method used by the KNMI per weather condition

Weather condition	**Measurement method**
Average temperature	Average of 24 hours, measured in 0.1 degrees Celsius
Maximum temperature	Maximum of 24 hours, measured in 0.1 degrees Celsius
Minimum temperature	Minimum of 24 hours, measured in 0.1 degrees Celsius
Sunshine	Percentage of the longest possible sunshine duration
Precipitation	Total of 24 hours, measured in 0.1 millimeters
Air pressure	Average of 24 hours, reduced to sea level in 0.1 hectopascal
Cloudiness	Average of 24 hours of upper air coverage
Relative humidity	Average percentage of 24 hours

Weather data: There are 8 weather conditions measured by the KNMI and used for this study (Table 3).

4 Methods

The data preparation and analysis were done in Python, using pandas for data handling and -manipulation and NumPy for performing fast calculations. The used method consists of three main phases. In the first part, the data preprocessing phase, the LUMC migraine diaries and patient info are combined and prepared for the analysis. Also, the weather data is linked to the correct dates and zip codes in the patient diaries. In the second phase, two classification models, random forest and support vector machine, were applied to make predictions about the start of migraine attacks. In the third and final phase, the predictions made by both models were analysed and compared to each other and a baseline method.

4.1 Data preprocessing

In the data preprocessing phase we began with 2004 migraine patients. In this study, we leave out patients with less compliance than 80% for 3 consecutive months so that there are sufficient data points to train / test the model on. Also, solely EM patients are selected, as the main cause for migraine in MOH patients is medication overuse and CM has by definition too many attacks to have one explicit trigger such as weather. Lastly, patients without a zip code in the meta data cannot be linked to weather stations. In Figure 1, the removal method of the patient selection is illustrated.

In the data we distinguish between two types of days, migraine start days and non-migraine start

Figure 1: Block diagram showing the patients selection process

days. Migraine start days are labeled as 1 in the data. Within the non-migraine start days, we have two subcategories, namely days that could have been a migraine start day and days that could not have been a migraine start day. The days on which it is by definition impossible for a migraine attack to start are dropped from the analysis, as those days are not a positive nor a negative label. This is because it is not due to the weather that a migraine attack is not starting, but it is because a migraine attack already has started. Therefore, those are migraine days that are not a migraine start but the 48 hours succeeding the last day of a migraine attack. The succeeding 48 hours are disregarded because if there is one non-migraine day between two migraine days, this is seen as the continuation of the same attack.

From the patients, we recorded the migraine start date information, their respective age, whether they reported having weather dependent migraine and their zip code.

4.2 Linking migraine diaries to weather data

The zip codes of the patients and weather stations are used to link the patients to the two closest weather stations. The distance is measured based on the zip code areas and the longitude and latitude of the weather stations. There are different zip code areas in the Netherlands, shown by the two first numbers of a zip code. The longitude and latitude of the weather station locations is used to find the Euclidean distance of a weather station to the centre of a zip code area. Thereby, the two closest weather stations are linked to each zip code area.
The weather stations do not all keep track of every weather variable. To enlarge the weather variables used per patient, the patient's diaries are linked to the two weather stations closest to their zip code area.
The correct weather conditions measured by those stations are linked with the matching dates per patient. Also, the absolute differences of those weather conditions between the date in the diary and the day before are calculated to see whether the migraine attacks are triggered by a change in the weather.
To properly make a prediction and evaluate the performance of the model, a patient needs to have at least 20 days in their diaries left after the merge with the weather data (Figure 2). After the

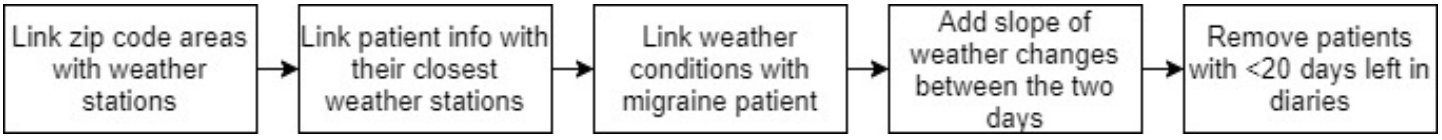


Figure 2: Block diagram showing the process of linking with weather data with the patients

patient selection, there were on average 147 observations per patient of 860 patients left in the dataset. Table 4 shows the deviation of the variables per observed day. As can be seen, there are sixteen numeric input variables of the weather conditions and there is one binary output variable, namely the migraine start. All these variables are measured per day per patient, so all the days are single data points.

4.3 Modelling

In the modelling phase, the start of a migraine attack will be predicted. Machine learning models will be trained on the weather data and migraine start data. In order to make this a solid prediction, several steps are taken. Each main step will be explained separately to clarify the process, starting with the general steps of the stratified train-test split and cross validation, and after that explaining the baseline and the two machine learning models used for the predictions.
Stratified train-test split & 5-fold cross validation: To make a model the data needs to be split into train data and test data. The train data is what the model will learn from, for which the feature of whether it is a migraine start day is given. The test data is what the model will make the prediction on, without knowing whether it is a migraine start day. To reiterate, the test data is used to test how well the model has learned to make correct predictions from the train data.
To train the models, five-fold cross-validation was applied using a stratified train-test split. A stratified train-test split means that based on a specific set of features, all splits contain equal ratios of those features. Patients are divided into five folds in such a way that the distributions of age,

Table 4: Overview of weather and migraine data used for training and testing the models. The parameter value are the weather conditions and the two-day gradient value are the changes in the weather conditions compared to the previous day

	Parameter value in dataset	**Two-day gradient value in dataset**
Average temperature (°C), mean ± sd	11.3 ± 5.5	0.0012 ± 2.0
Minimum temperature (°C), mean ± sd	7.0 ± 5.1	-0.0028 ± 2.7
Maximum temperature (°C), mean ± sd	15.2 ± 6.6	0.0054 ± 2.8
Sun percentage (%), mean ± sd	41.3 ± 31.1	41.26 ± 33.2
Total precipitation (mm), mean ± sd	2.4 ± 4.6	0.019 ± 6.0
Average pressure (hPa), mean ± sd	1014.7 ± 10.7	-0.03 ± 6.5
Average cloudiness, mean ± sd	6.0 ± 2.2	0.00068 ± 2.2
Average humidity (%), mean ± sd	78.9 ± 11.5	-0.0076 ± 8.05
Start of migraine attack (%)	True (10.2) / false (89.8)	

the number of months in the diary and whether the patient indicates to have weather dependent migraine are equal in all folds. These metrics are used because they might be of influence to the sensitivity to weather and make sure that a similar amount of information (number of months) is known of the patients in different folds. For the distribution of age, we distributed the patients in bins with 10 years per bin. Thereby, there were no singular values and the patients could be distributed in the five folds equally. However, there were 3 patients with a unique combination of age category, number of months in diary and whether the patient indicates to have weather dependent migraine. These patients are distributed over the different folds randomly. Each fold is used as the test set once, using the other four folds as a train set.

Baseline: A baseline was made to compare the accuracies, which will be discussed more broadly in the evaluation, of the predictions of random forest and support vector machine to. The baseline is always predicting that it is not a migraine start day, so predicting 0 for every patient every day, which leads to an accuracy equal to the percentage of non-migraine start days (87.13%).

Random forest classifier: Random forest is a supervised learning algorithm that works by way of an ensemble of decision trees. An ensemble is a collection of models that works by merging the individual results. In other words, random forest is a model that builds multiple decision trees and puts them together to make an accurate decision [LW02] [Don21]. For classification, random forest takes the majority vote of all models and gives this back as the prediction. This model is implemented in python using the sklearn library [skl]. To optimize the result, hyperparameter optimization was applied on the train data using sklearn's built in GridSearch to optimize the number of trees built and the depth of the trees [skl]. The depth of a tree means the number of layers of decisions the decision tree has to make an accurate decision, that is still generalizable. If the tree has a very high depth and number of trees, a very good decision can be made for a specific case. However, when there is a different case, the model no longer works and it is not generalisable. This is called overfitting.

5 random forests are learned, as explained in the 5-fold cross validation section. Thereby, the model is tested on a different group of migraineurs every time and makes predictions for all patients.

Support vector machine classifier: Support vector machine is a supervised learning algorithm

Table 5: Confusion matrix

True value	**Prediction**	
	Non-migraine start	Migraine start
Non-migraine start	True negative (TN)	False positive (FP)
Migraine start	False negative (FN)	True positive (TP)

that finds the optimal hyperplane to distinctively classify the data points. This is done in an N-dimensional space. The dimension of the hyperplane depends on the number of features [Gan18]. The hyperplane divides the different data points, for this classification task, in two categories. The optimal dimensionality differs per task. Using the sklearn library, which includes a SVM model for classification, the dimensionality can be optimized, using hyperparameter optimization GridSearch, like done with the random forest.
5 support vector machines are learned, as explained in the 5-fold cross validation section. The model is tested on a different group of migraineurs every time and makes predictions for all patients.

4.4 Evaluation

To evaluate the model, several metrics are applied to measure the performance. The sensitivity, specificity and accuracy for the baseline prediction result and the random forest and support vector machine prediction results are calculated using the true positives (TP), false positives (FP), true negatives (TN) and the false negatives (FN) (Table 5).

Sensitivity: The sensitivity is the ability of the models to identify true positives. The formula used to calculate is as follows sensitivity = TP / (TP + FN). Thus, this shows the ratio of correctly predicted migraine start days to all actual migraine start days.
Specificity: The specificity is the ability of the models to identify TN's. The formula used to calculate is as follows sensitivity = TN / (TN + FP). Thus, this shows the ratio of correctly predicted not migraine start days to all actual not migraine start days.
Accuracy: The accuracy here is the ability of the models to correctly predict migraine start days. This shows the ratio of the total correctly predicted labels (TP & TN) of the total number of labels predicted, and as such of the total dataset. The formula to calculate is as follows: accuracy = (TP +TN) / (TP + TN + FP + FN). Thus, this shows the ratio of correctly predicted migraine days over the entire set of predictions.
The sensitivity, specificity and accuracy are calculated as an average of the entire dataset and per patient, to see for what patients the models work best.
Compare random forest & support vector machine: To find out whether the models both perform better for the same patients or not, the accuracy of patients obtained by using random forest will be compared with the accuracy of patients obtained by using support vector machine. If both models perform equally well for the same patients, a positive linear correlation can be expected.
Compare to the Baseline: The relative accuracy of each patient will be calculated by subtracting the accuracy of SVM and random forest of each patient with the baseline (naive) accuracy for that patient. If the accuracy of the patient is higher when using the baseline for prediction instead of

using the model, the relative accuracy will be negative. This way, patients for which the model works better than always predicting non-migraine start will be identified.
Sensitivity & specificity combination: The sensitivity of each patient will be compared with the specificity of each patient. For migraine patients, clinically it is most important to identify the migraine days and it is not too bad if the model sometimes says it is a migraine day while it actually is not. This is better than the model saying it is not a migraine day while it actually is. So therefore, the sensitivity, which shows the ability to identify the true positives, is important. However, it is useless if the model always predicts true, so therefore the specificity is also important to take into account. The patients for which the model works properly are patients for which both the sensitivity and the specificity is higher than 0.5, so the model works good more than half of the time. This is an arbitrary cutoff, made at 0.5 because here the model predicts more than half of the time true positives and more than half of the time true negatives. Therefore, the accuracy is automatically higher than 0.5, as the total number of correctly predicted values is more than half.
Sensitivity & specificity combination: The sensitivity of each patient will be compared with the specificity of each patient. For migraine patients, it is clinically the most important to identify the migraine days and it is not a terrible thing if the model sometimes indicates it to be a migraine day while it actually is not. This is better than the model saying it is not a migraine day while it actually is. Therefore, the sensitivity, which shows the ability to identify the true positives, is important. However, it is useless if the model always predicts true, so therefore the specificity is also important to take into account. The patients for which the model works properly are patients for which both the sensitivity and the specificity are higher than 0.5, so the model works correctly more than half of the time. This is an arbitrary cutoff, made at 0.5 because here the model predicts true positives more than half of the time and true negatives more than half of the time. Therefore, the accuracy is automatically higher than 0.5, as the total number of correctly predicted values is more than half.

5 Results

The results show the performance of random forest and support vector machine described in Section 4.4. First, the results of the data preprocessing are discussed, after which a discussion on the accuracy, sensitivity and specificity of patients using random forest and support vector machine follows. This will be followed by a comparison of the accuracy of the two models per patient. After that, the relative accuracy, defined as the absolute difference with the baseline accuracy, of the two models will be looked at. Lastly, an analysis of the patients with a sensitivity and specificity higher than 0.5 will be discussed.

Data preprocessing: The data preprocessing phase started with 2004 migraine patients. Based on the requirements discussed in Section 4 patients are discarded from the analysis as can be seen in Figure 3. Some of the weather stations considered in this research only periodically measure the weather. Therefore, if the stations are non-functional on particular patient diary dates, some patients have no recorded weather data to train on and are therefore removed from the dataset. After linking with the weather variables as shown in Figure 2, 932 patients are left. This phase ends with 860 patients whose data will be used to train and test the models (Figure 3).

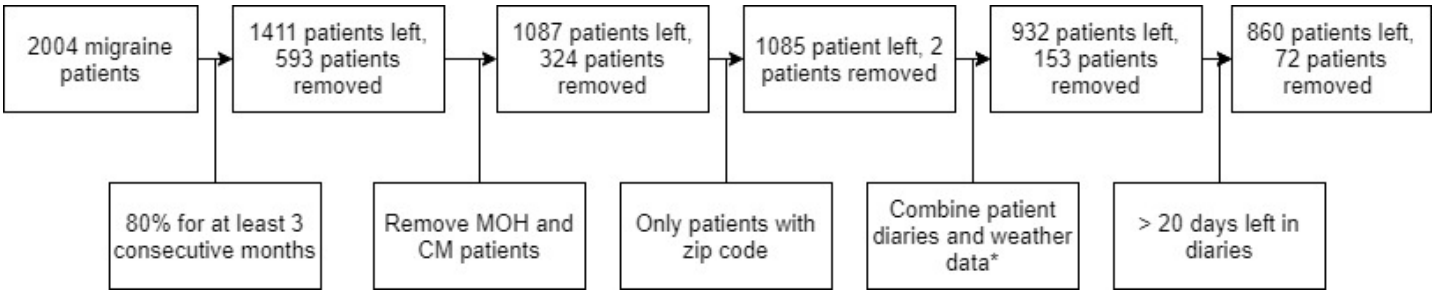
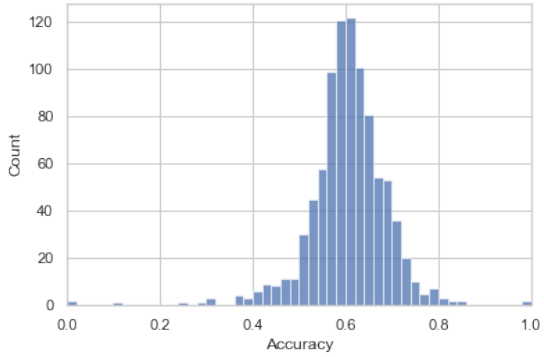
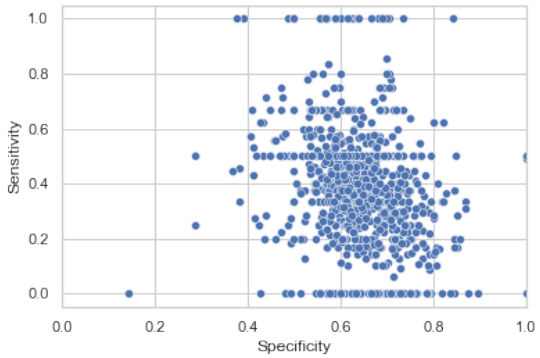


Figure 3: Block diagram showing the data preprocessing phase in which solely the patients that can be used for training and testing the models are kept

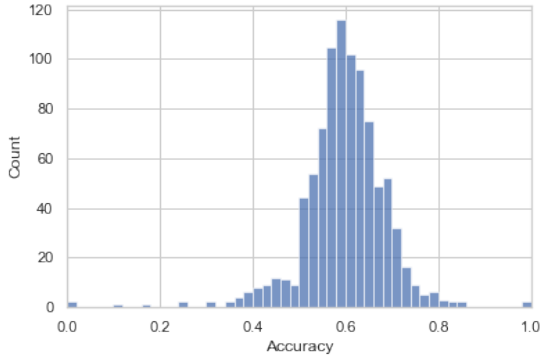
Performance random forest & support vector machine: For the hyperparameter optimization of the number of estimators for random forest, the search space was [5, 10, 15, 20] where the optimal value found was 10 estimators. This was identical for all folds. For the hyperparameter optimization of random forest's max depth of a tree the search space was [7, 9, 12, 15]. The optimal value found was 9. For the hyperparameter optimization of the support vector machine's degrees, the search space was [1, 2, 3]. The optimal value found was 3. This was identical for all folds. It is possible that a higher degree would have worked better. However, with the available computation power this was not possible. The prediction on the dataset using random forest and support vector machine have a mean accuracy over all the folds of 60.5% for random forest and 59.7% for support vector machine. Both models show an approximate normal distribution (Figure 4a, Figure 4c) of the accuracy, not necessarily showing a group of patients performing significantly better. The sensitivity and specificity differ substantially for each patient. Figure 4b and 4d show that the majority of the patients lie somewhat around the middle. For 118 patients for random forest and 47 patients for support vector machine, the models have a sensitivity and a specificity higher than 0.5. These



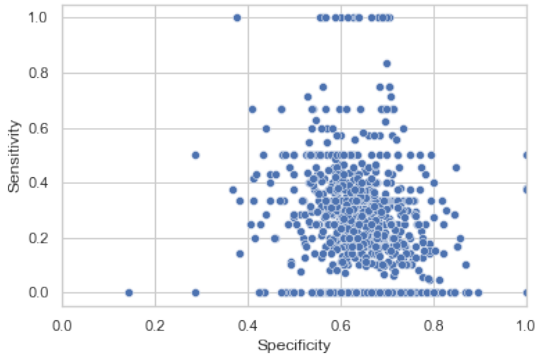
(a) accuracy per patients using random forest



(b) sensitivity & specificity per patient using random forest



(c) accuracy per patient using support vector machine



(d) sensitivity & specificity per patient using support vector machine

Figure 4: Results of support vector machine and random forest

patients will be categorized as a subgroup as the models provide an accurate prediction in more than half of the migraine start cases and still regularly predict that it is not a migraine start, as it also predicts more than half of the non-migraine start days correctly.

Comparison in performance random forest & support vector machine: The models have similar accuracies for all patients and make almost equal predictions, except that random forest often performs slightly better. However, the accuracy of every patient shows a linear correlation in both models (Figure 5). This shows a pattern identifying patients for which it is easier to predict migraine based on weather conditions.

Relative accuracy random forest & support vector machine: The baseline has a mean accuracy of 87.1%. When the relative accuracy is higher than 0, the model performs better than the baseline. In Figure 6a and Figure 6b you see the relative accuracies of patients using respectively random forest and support vector machine. This is of all patients before removing patients with <20 days left in the diary after the linking with the weather data. After taking into account that patients must have at least a compliance of 20 days after linking with the weather data, only one

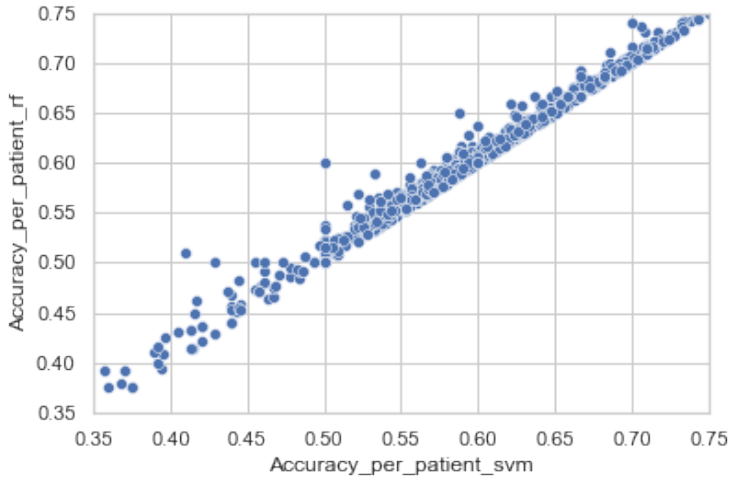
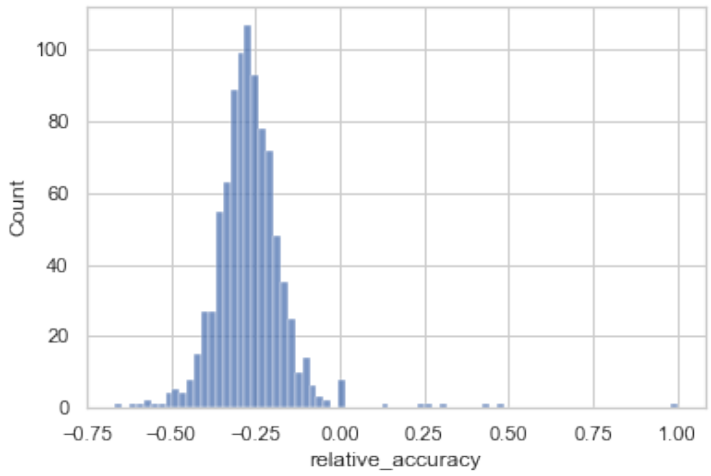
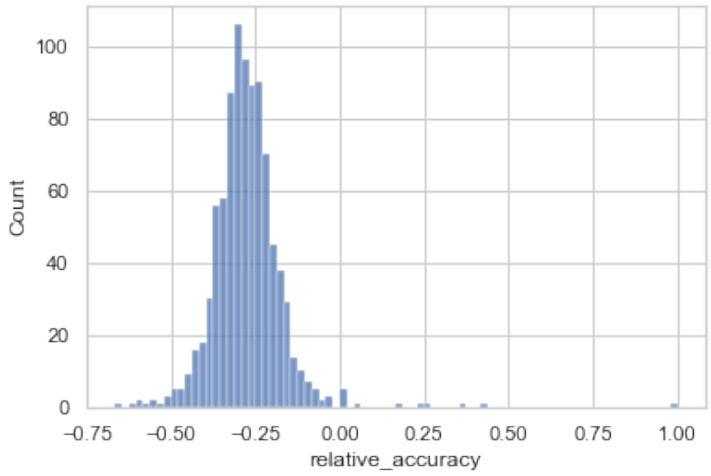


Figure 5: Accuracy of prediction per patient in random forest and support vector machine

patient with a positive relative accuracy remains. This is the same patient for random forest and support vector machine. The patients identified as the subgroup with a specificity > 0.5 and a sensitivity > 0.5 are distributed on the relative accuracy scale over the area between -0.3 and 0.
Patients subgroup: The information found in Table 6 is based off patients in the subgroups identified by having a specificity and a sensitivity higher than 0.5 for random forest and support vector machine. For random forest there were 118 patients, which is more than twice as many as the 47 patients found with support vector machine. All patients found with support vector machine are also included in the set found with the random forest. However, random forest identifies another 71 patients with a good performance on top of those 47 identified by the support vector machine.



(a) Relative accuracy of random forest compared to the baseline, including patients with less than 20 days left in the diary after linking the diaries with the weather data



(b) Relative accuracy of support vector machine compared to the baseline, including patients with less than 20 days left in the diary after linking the diaries with the weather data

Figure 6: Relative accuracy

Table 6: Baseline characteristics of patients with a sensitivity & specificity ¿ 0.5 and the selection used for model development

	Subgroup random forest	**Subgroup support vector machine**	**Patient selection used in model development**
Patients, n	118	47	860
Age, mean ± sd	43.7 ± 10.0	42.0 ± 8.9	43.7 ± 11.6
Migraine attack days per month*, mean ± sd	2.5 ± 2.2	2.8 ± 1.8	3.0 ± 3.5
Females, n (%)	111 (94.1)	46 (97.9)	782 (90.9)
Patients with aura, n (%)	49 (41.5)	14 (29.8)	457 (53.1)
Think to be sensitive to weather, n (%)	48 (40.7)	18 (38.3)	368 (42.8)

*A month was defined as a regular period of 28 days

6 Discussion & limitations

Performance random forest & support sector machine: As the hypothesis followed in this research is the existence of a subgroup of patients with weather dependent migraine, it would be expected that for the majority of the patients the models would not give a high accuracy but that the subgroup would show a second peak at a higher accuracy in Figure 4a and 4c. However, the results found do not give a distinctive subgroup of patients with weather dependent migraine, based on the accuracy of the patients. This is the case using the weather on the day of the attack and the day before as predictive variables with random forest and support vector machine. This result can have multiple causes and therefore does not imply that there is no distinctive subgroup. The weather or weather changes may have a more of a long term influence on patients, meaning that the day of the attack and the day before is too short of a period to look at. It can also be the other way around. The day before and the day of the attack may be too broad and a shorter, more specific time frame, of for example two hours before the attack, gives a better result. However, for migraine patients, clinically it is most important to identify the migraine days and it is not too bad if the model sometimes says it is a migraine day while it actually is not. Therefore, the patients for whom the models have a sensitivity > 0.5 and specificity > 0.5, the models give a useful prediction.
Comparison in performance random forest & support vector machine: Comparing the accuracy of patients using random forest and support vector machine, a linear correlation is found. This shows that the predictions are not done randomly, as both perform better for the same patients, so apparently for some patients it is easier to use weather to make a prediction for the start of migraine attacks. However, random forest gives a slightly higher accuracy for most patients. This is also seen when looking at the number of patients with a sensitivity and specificity higher than 0.5 by random forest compared to support vector machine. From this we conclude that random forest finds a better pattern in the subgroup leading to 71 more patients with a sensitivity > 0.5 and specificity > 0.5 than support vector machine.
Relative accuracy random forest & support vector machine: There is only one patient with a positive relative accuracy. Because this is solely one patient, this gives no relevant information. However, solely looking at the relative accuracy is not useful, because it is clinically more important

to detect true positives than it is to have the highest accuracy. As such, the results with a high sensitivity provided by the model are more interesting than the high accuracy provided by the baseline by always predicting 0. Also, as there are more triggers for migraine than the weather, an accuracy of 1 is unrealistic. Even for patients with weather dependent migraine it is likely that not all attacks are caused by the weather, so a model could never predict every migraine attack solely looking at weather conditions.
Patients subgroup: The subgroup identified by taking patients with a sensitivity and specificity higher than 0.5 does not differ much compared to the total patient dataset much for age, number of attacks per month, percentage of females, percentage of patients with aura or the percentage of patients who report to have weather dependent migraine. Thereby, no distinctive pattern could be found to distinguish the subgroup from the entire dataset of migraine patients based on baseline characteristics. However, there seems to be a pattern identifiable for the subgroup of patients when looking at the performance of random forest and support vector machine using the weather, as both models return the same patients for the subgroup.

Our results can be compared to studies under different conditions, such as, for example, a study conducted by Zebenholzer that took place in the area near Vienna [ZRF+11]. Vienna has an oceanic climate (Cfb), according to the Köppen climate classification [Vie]. The Netherlands has the same climate classification, which means this research took place under similar circumstances [NL]. However, the method applied in this research deviates from the method that will be used in this research. Also, the quantity of patient data for this research is significantly more than the data used by K. Zebenholzer et al. The study by K. Zebenholzer et al. did not find any correlation between migraine patients and weather factors [ZRF+11].
Another study, by J. Gomersall and A. Stuart, looked at "variations in migraine attacks with changes in weather conditions". They mainly focused on glare and bright light, cooling, thunderstorms, blizzards and specific winds [GS73]. This study differs in the fact that they only take weather changes into account and the weather factors they focus on are quite specific and will not be looked at in this study. They did not find a correlation between migraine and these factors.
Another study by Li et al. took place in the area of Boston, Massachusetts. Boston had a hot humid continental climate (Dfa) according to the Köppen climate classification [mas]. This means they have no dry season, which is similar to The Netherlands. However, the summers in Boston are classified as hot, while the summers in The Netherlands are classified as warm. These climates are still relatively similar and comparable. This study concluded that there is a correlation between weather conditions and the likelihood of migraine [LBM+19].
Lastly, the study performed by Hoffmann et al concluded that there is a subgroup of migraineurs with weather dependent migraine. This research is performed in Berlin [HSHL+15]. Berlin has a similar climate to the Netherlands. However, this study looked at many more time points in the 24 hours previous to the exact time of the start of a migraine attack. This implies that the weather in the shorter time period before an attack is a trigger for migraine, compared to the results of this study.
This shows that conditions under which a research takes place are of influence on the results on whether a correlation between migraine and the weather is found.

6.1 Limitations

There are several limitations to the data and method used for this research. Firstly, there was a lack of weather data available. Because not all stations measure weather variables every day, there are patients with no available weather data to link them with. This leads to the removal of patients that otherwise could have been used in the models.
Secondly, the data is linked to the home address zip codes of patients. However, it is not known where the patients are when they fill in the diaries and where they are when they have a migraine attack. This gives the predictions more uncertainty.
Thirdly, solely the date of the start of a migraine attack is known. If a migraine attack starts early in the morning, the weather conditions of the day of the migraine attack might not be important. Also, this makes it impossible to focus on a more specific time frame shortly before an attack.
Lastly, due to lack in computational power, the degrees of support vector machine could not be higher than 3. If the optimal degree is higher than 3, this could explain the difference in performance of random forest and support vector machine.

7 Conclusions and Further Research

The goal of this research was to use weather and migraine data to train a model on to identify patterns that are a good predictor for a subgroup of patients that has weather dependent migraine. For this purpose, two machine learning classification models, namely random forest and support vector machine, were trained on a dataset of episodic migraine (EM) patients.
Random forest and support vector machine were used to predict a migraine start day for the entire dataset of migraineurs and then evaluated to find a subgroup of patients for which the models work better. Based on the results obtained from this research, both random forest and support vector machine work equally well for this problem, although random forest gives slightly better predictions. But since both models result in similar overall accuracies per patient, both are able to identify patients with weather dependent migraine. Because both perform better for the same subgroup of patients, and the predictions are therefore not random, it can be concluded that for some patients it is indeed easier to use changes in the weather to accurately predict the start of a migraine attack.
The evaluation to identify the subgroup of migraine patients with weather dependent migraine is done using the specificity and sensitivity. For this problem it is more important to identify true positives, which therefore means that having a high sensitivity and predicting some false positives as well is preferred over not predicting any starts and having a high accuracy. This is because it is clinically more relevant to identify migraine start days in order to give migraine patients better insight into their condition.
Globally, a subgroup of migraine patients was found by using a cut-off of 0.5 for the sensitivity and specificity. However, many patients were close to being in the subgroup, meaning the subgroup was not very distinctive. The model does not perform much better on migraineurs in the subgroup than on other patients close to the subgroup. Additionally, the baseline characteristics of this subgroup were similar to those of the total population. This implies that no distinctive patterns are found in the subgroup. Based on the screened parameters, there is no characteristic found based on which patients with weather dependent migraine can be identified.
Using classification models, we found that for approximately 10% of the patients we were able to make predictions with a sensitivity > 0.5 and a specificity > 0.5 solely based on weather and weather changes. This shows that there are patterns in weather conditions which could be a trigger for patients with weather dependent migraine.

7.1 Further research

It would be useful to do further research that takes a shorter or longer period into account before the migraine attack. A useful avenue of further research would be to find out how the weather variables in the hours before an attack influence the performance of the models. It is however also possible that the weather has more of a long term influence on the start of a migraine attack, so the weather conditions a week before the attack could be taken into account, as well as how they changed in the week before the attack.
Furthermore, it would be useful to find out which weather variables have the most influence on the performance of the models, or in other words feature importance. Thereby, it can be found whether the changes in the weather or just the current weather conditions are the most influential on the patients. Furthermore, it would be interesting to train the model on the identified subgroup and

see how this influences the performance.
Lastly, it is also possible that migraine attacks are not triggered by weather conditions, but that the weather does influence the migraine attack. It would be interesting to look at the duration and intensity of attacks to see whether they are influenced by specific weather conditions or weather changes.

References

[Bec11] W. J. Becker. Weather and migraine: Can so many patients be wrong? pages 387–390, 2011.

[BSLS15] R. C. Burch, E. Loder S. Loder, and T. A. Smitherman. The prevalence and burden of migraine and severe headache in the united states: updated statistics from government health surveillance studies. *Headache: The Journal of Head and Face Pain*, 55:21–34, 2015.

[Cul81] R. E. Cull. Barometric pressure and other factors in migraine. *Headache: The Journal of Head and Face Pain*, 21:102–104, 1981.

[Don21] N. Donges. A complete guide to the random forest algorithm. *Built In. https://builtin.com/data-science/random-forest-algorithmhow*, 2021.

[Gan18] R. Gandhi. Support vector machine — introduction to machine learning algorithms. *Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47*, 2018.

[GCGZGB17] Y. Garcia-Chimeno, B. Garcia-Zapirain, and M. Gomez-Beldarrain. Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data. *BMC Med Inform Decis Mak*, 17:38, 2017.

[GS73] J. D. Gomersall and A. Stuart. Variations in migraine attacks with changes in weather conditions. *International journal of biometeorology*, 17:285–299, 1973.

[HHLMR11] J. Hoffmann, L. Neeb H. Lo, P. Martus, and U. Reuter. Weather sensitivity in migraineurs. *Journal of neurology*, 258:596–602, 2011.

[HSHL+15] J. Hoffmann, T. Schirra, L. Neeb H. Lo, U. Reuter, and P. Martus. The influence of weather on migraine–are migraine attacks predictable? *Annals of clinical and translational neurology*, 2:22–28, 2015.

[IHS21] IHS. The international classification of headache disorders 3rd edition. *https://ichd-3.org/*, 2021.

[Kel07] L. Kelman. The triggers or precipitants of the acute migraine attack. *Cephalalgia*, 27:394–402, 2007.

[KNMa] https://www.daggegevens.knmi.nl/klimatologie/daggegevens.

[KNMb] https://www.knmi.nl/kennis-en-datacentrum/uitleg/automatische-weerstations.

[KNMc] https://www.knmi.nl/nederland-nu/weer/waarnemingen.

[LBM+19] W. Li, S. M. Bertisch, E. Mostofsky, C. Buettner, and M. A. Mittleman. Weather, ambient air pollution, and risk of migraine headache onset among patients with migraine. *Environment international*, 132:105100, 2019.

[LR13] M. Leonardi and A. Raggi. Burden of migraine: international perspectives. *Neurological Sciences*, 34:117–118, 2013.

[LW02] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

[mas] Massachusetts. *Climate-Data. https://en.climate-data.org/north-america/united-states-of-america/massachusetts-1062/.*

[Mig] Migraine. volksgezondheidenzorg.info. *https://www.volksgezondheidenzorg.info/onderwerp/migraine.*

[NIAS17] M. Nilashiab, O. Ibrahim, H. Ahmadi, and L. Shahmoradi. An analytical method for diseases prediction using machine learning techniques. *Computers and Chemical Engineering*, 106:212–223, 2017.

[NL] Netherlands climate: Average temperature, weather by month weather for the netherlands. *Climate-Data. https://en.climate-data.org/europe/the-netherlands-40/.*

[PM13] D. Pietrobon and M. A. Moskowitz. Pathophysiology of migraine. *Annual review of physiology*, 75:365–391, 2013.

[PRS+04] P. B. Prince, A. M. Rapoport, F. D. Sheftell, S. J. Tepper, and M. E. Bigal. The effect of weather on headache. *Headache: The Journal of Head and Face Pain*, 44:596–602, 2004.

[PZS+20] P.Ferroni, F. M. Zanzotto, N. Scarpato, A. Spila, L. Fofi, G. Egeo, A. Rullo, R. Palmirotta, P. Barbanti, and F. Guadagni. Machine learning approach to predict medication overuse in migraine patients. *Computational and Structural Biotechnology Journal*, 18:1487–1496, 2020.

[Rop20] A. H. Ropper. Migraine. *The New England Journal of Medicine*, 383:1866–76, 2020.

[SKHM19] S.Uddin, A. Khan, M. E. Hossain, and M. A. Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Maky*, 19:281, 2019.

[skl] scikit-learn: machine learning in python. *Scikit-Learn. https://scikit-learn.org/stable/.*

[VAA+17] T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, and M. H. Criqui. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390:1211–1259, 2017.

[vCVdB+21] D. S. van Casteren, I. E. Verhagen, I. de Boer, S. de Vries Lentsch, R. Fronczek, E. E. van Zwet, and G.M. Terwindt. E-diary use in clinical headache practice: A prospective observational study. *Cephalalgia*, 2021.

[Vie] Vienna. *Climate-Data. https://en.climate-data.org/europe/austria/vienna-1461/.*

[WW79] M. Wilkinson and J. Woodrow. Migraine and weather. headache: The journal of head and face pain. *The Lancet*, 19:375–378, 1979.

[YFH+15] A. C. Yang, J. L. Fuh, N. E. Huang, B. C. Shia, and S. J. Wang. Patients with migraine are right about their perception of temperature as a trigger: time series analysis of headache diary data. *The journal of headache and pain*, 16:1–7, 2015.

[ZRF+11] K. Zebenholzer, E. Rudel, S. Frantal, W. Brannath, K. Schmidt, C. Wober-Bingol, and C. Wober. Migraine and weather: a prospective diary-based analysis. *Cephalalgia*, 31:391–400, 2011.