January 2023

Developing Gradient Boosting Machine Learning Model For Predicting Headaches Among Adult Headache Patients

Hye Sun Kim
hyeskim10@gmail.com

Follow this and additional works at: https://elischolar.library.yale.edu/ysphtdl

Recommended Citation
https://elischolar.library.yale.edu/ysphtdl/2280

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
Developing Gradient Boosting Machine Learning Model for Predicting Headaches among Adult Headache Patients

A Thesis
Presented to the Department of Chronic Disease Epidemiology Of Yale University in Candidacy for the Degree of Master of Public Health in Chronic Disease Epidemiology Year Completed: 2023 Year Degree Awarded: 2023

By Hye Sun Kim

Advisor/Committee Chair: Judith Lichtman, PhD, MPH Committee Member: Christopher Metts, MD

May 2023
ABSTRACT

Background – Headache is one of the significant public health threats in the world. Patients often face difficulties preparing for medications before headache attacks due to their unexpected nature. A headache diary was previously employed in studies to develop a machine-learning prediction model. However, previous research used variables related to a narrow range of topics or used ensemble methods where the decision trees were not added sequentially. Therefore, this study aimed to use a gradient-boosting approach to find factors from a headache diary with a broader range of factors with strong predictive values for headache prediction.

Methods – The three-month headache diary data were collected thrice daily using the Status/Post Apple device application. A total of 23 adult patients’ self-reported data was used in a gradient-boosted classification machine-learning model. The primary outcome measure was the self-reported absence/presence of a headache, and the features used were self-reported symptoms and lifestyle variables asked in a headache diary.

Results – The gradient-boosting classifier model’s area under the curve (AUC) for 23 adult headache patients was 0.94, which shows a strong differentiating ability between the probability of having a headache and not having a headache. Also, the model’s ability to accurately predict headaches was 0.80, shown by the F1 score of the model. The study also discovered that premonitory symptom variables were more predictive than others.

Conclusion – This study shows that future headache attacks can be accurately predicted for adult headache patients using the GBM classification model. Additional research is needed to explore whether the model can be used in other populations and whether a strong predictive model with fewer and stronger predictive variables found in this study can be developed.

Keywords: Chronic Disease Epidemiology, Headache, Migraine, Prediction, Machine Learning
ACKNOWLEDGMENT

I would like to express my sincere gratitude to my advisor, Professor Judith Lichtman, for her valuable advice and guidance throughout my master's program. Her mentorship has been crucial in shaping my academic journey and professional skills. Additionally, I am also deeply grateful to Professor Christopher Metts for his guidance and expertise in developing the machine learning model, which has been a key component of my research and interest. I would like to extend my profound appreciation to Professor Elizabeth Seng for generously sharing her datasets and offering insightful advice on formulating research questions and methodologies throughout my thesis.

I would also like to express my heartfelt gratitude to my parents and grandparents for their unwavering love and support, which has been a source of strength and encouragement throughout the whole process. Lastly, I am grateful for the support of my friends, who have provided a sense of community and motivation during challenging times.

LIST OF FIGURES

Figure 1. Flow chart of the study population........................................................................................................7
Figure 2. Correlation heat map between each possible pair of features. Blue means negative correlation. Red means positive correlation. Darker color means a higher correlation..................11
Figure 3. Imbalance in the primary outcome variable (absence/presence of headaches) .............12
Figure 4. Confusion matrix of the model. .............................................................................................................16
Figure 5. AUROC curve of the model.....................................................................................................................17
Figure 6. Top 10 feature importance ranking ......................................................................................................18

LIST OF TABLES

Table 1. Classification report of the model...........................................................................................................17
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................................................. 2  
ACKNOWLEDGMENT ................................................................................................................................. 3  
LIST OF FIGURES ....................................................................................................................................... 3  
LIST OF TABLES ......................................................................................................................................... 3  
INTRODUCTION .......................................................................................................................................... 5  
METHODS.................................................................................................................................................. 6  
  1. STUDY POPULATION ............................................................................................................................. 6  
  2. INCLUSION / EXCLUSION CRITERIA .................................................................................................... 8  
  3. DATA COLLECTION .......................................................................................................................... 9  
  4. DATA PRE-PROCESSING ................................................................................................................... 10  
  5. FEATURE SELECTION ......................................................................................................................... 12  
  6. PRIMARY OUTCOME ......................................................................................................................... 12  
  7. GRADIENT BOOSTED CLASSIFICATION MODEL DEVELOPMENT .................................................. 13  
  8. OVERFITTING ................................................................................................................................... 14  
  9. EVALUATION METRICS ................................................................................................................... 14  
RESULTS .................................................................................................................................................... 15  
DISCUSSION .............................................................................................................................................. 18  
CONCLUSION ............................................................................................................................................ 20  
REFERENCE ............................................................................................................................................. 21
INTRODUCTION

Headache is one of the significant public health threats in the world (Stovner et al., 2018). Research has found that having a headache was the fourth negative contributor to changes in health-adjusted life expectancy between 1990 and 2013 (Burch et al., 2018). It was also found that in 2016, 4 million emergency department visits in the United States were due to headaches (Burch et al., 2021). Headaches are treated using preventative methods that reduce the headache frequency over time or treat attacks after they occur (Lawrence, 2004). However, as headaches can strike without warning, patients sometimes have trouble taking their medication until their pain intensifies. Therefore, predicting headache attacks using proactive strategies is needed for better treatment at an appropriate time.

Developing machine learning models is one of the proactive strategies used for predicting headaches. Machine learning is a growing field of computational algorithms that aim to mimic human intelligence by learning from their surroundings (El Naqa et al., 2015). It has been used to develop headache prediction models that can forecast an individual’s future headache attacks based on the patient’s input data. Data involving headache disorders are often obtained from a self-monitored headache diary, which is often used to enhance diagnostic accuracy and evaluate the effectiveness of headache therapies (Armstrong & Gossard, 2016). However, only less than one-third of the participants completed electronic headache diary recordings, which may be related to the burdensome number of daily questions that patients are required to record (Ramsey et al., 2014; Jamison et al., 2017; Seng et al., 2018). As missing data in the headache diary caused by lack of adherence can result in an inaccurate diagnosis, a misunderstanding of the clinical efficacy of a new treatment regimen, and erroneous patient beliefs regarding factors related to their headaches, it may be beneficial to see which questions in the diary are more
important than others when developing predictive models to reduce the number of questions asked in a daily headache diary (Seng et al., 2018).

There were previous studies that aimed to develop a prediction model for headache attacks using headache diary data. For example, Houle et al. (2017) used headache diaries to develop headache prediction models with stress-related variables and a generalized linear mixed-effects forecast model. Gago et al. (2018) used random forest and logistic regression to predict migraine crises with premonitory symptom features. Decision trees and Markov chains were also used to predict migraine attacks and build a decision support system to follow up and diagnose primary headache patients (Vandewiele et al., 2018; Barra et al., 2020; Alves et al., 2021). However, previous studies focused on features related to certain topics to predict headaches, which may fail to capture the importance of other lifestyle factors in headache attack predictions. Also, previous studies did not use a gradient-boosting model that is known to be more accurate than a random forest method. While there was one previous study that examined several machine-learning methods, including random forest, gradient boosting, decision tree, logistic regression, and support vector machine model, the aim of the study was to build a machine-learning model to build a clinical decision-making system for distinguishing migraine and tension-type headaches, not predicting whether there will be a headache or not (Liu et al., 2022).

Thus, this study aimed to use a gradient-boosting machine learning approach to identify factors with strong predictive values for headache prediction from a broader range of factors from a headache diary.

**METHODS**

1. Study population
This study used prospective observational data from phase 2 of the Clinical Decision Support Tool (CDST) study by Yeshiva University. Potential subjects were enrolled in the Montefiore Medical system after they completed their online informed consent document. Those who were enrolled in the Montefiore Medical System for at least one year and who have been prescribed a triptan within one year were identified by using Clinical Looking Glass. Potentially eligible patient providers were contacted, and those who agreed to be involved in the recruitment process for this study sent their identified patients letters informing their eligibility for this study. After two weeks, email and phone recruitment were done to select patients interested in study participation and meeting edibility criteria. Patients were then enrolled using REDCap (n = 57).

**Figure 1.** Flow chart of the study population
2. Inclusion / Exclusion criteria

Figure 1 demonstrates how the study population was selected. Before the initialization of the diary, participants needed to meet the following requirements to fulfill the inclusion criteria: 1) Between the ages of 18 and 65; 2) An International Classification of Headache Disorders-3 beta migraine diagnostic; 3) Self-report and diary confirmation of 6 to 14 headache days per month; 4) Triptan prescription for acute migraine treatment; 5) Stability on a current migraine preventative and acute treatment plan; 6) Read English; 7) Have the capacity to consent.

Also, to participate in the trial, participants must not have any of the exclusion criteria: 1) A possible or confirmed medication overuse headache; 2) A plan to change or be changing a preventative or acute migraine medication; 3) Pregnancy or expecting to become pregnant while participating in the study 4) Psychiatric or cognitive issues; 5) Participation in phase 1 of the CDST study.

Participants who self-reported that they met the study inclusion criteria were enrolled in the study and monitored with a headache study diary three times daily for 30 days. After 30 days, participants who did not record at least one migraine episode in 30 days, recorded fewer than six or greater than 14 headache days in 30 days, and did not fulfill the inclusion criteria were excluded. Participants who completed 80% of diary entries in the first 30 days and fulfilled the inclusion criteria above at this time were included and invited to continue with the diary for another two months. The research coordinator sent emails to each participant every two weeks with feedback on completed diary days, application problem advice, and a risk assessment of unfavorable outcomes. Participants received extra contact if they failed to record data for two days in a row. There were zero
dropout during the follow-up. During the analysis phase, 17 participants got excluded for failing to meet the inclusion criteria. A total of 23 participants’ three-month data was used for this study.

3. Data collection

Participants entered daily headache data through the daily electronic headache diary, collected through the Status/Post Apple device app, a secure ecological momentary assessment application that interacts with the REDCap data capture system (Baird et al., 2018; Tomoko et al., 2020). Participants completed a daily headache diary three times daily: in the morning, afternoon, and evening. The diary was designed using related factors that may be associated with migraine with multiple-choice questions and free-response questions.

Morning diary included measures assessing a) women-specific menstruation status, b) headache symptoms using ICHD-3b migraine criteria (Headache Classification Committee of the International Headache Society, 2013), c) acute medication usage with items used in other headache medication adherence studies (Seng et al., 2016), and d) sleep behavior, using selected items from the consensus sleep diary (Carney et al., 2012).

The afternoon diary included measures assessing a) headache symptoms using ICHD-3b migraine criteria and b) acute medication use with items used in other headache medication adherence studies. Evening diary included measures assessing a) preventative medication usage with an item used in other headache medication adherence studies, 2) headache symptoms using ICHD-3b migraine criteria, c) acute medication use with items used in other headache medication adherence studies, d) eating and hydration behavior using one item asking participants to check off each hour during which they consumed
food in the past 24 hours, and one item asking participants how many glasses of liquids they consumed in the past 24 hours, e) Perceived Stress Scale (PSS), which is a 10-item survey evaluating the frequency of perceptions of stress (feeling life is unpredictable, uncontrollable, or overloaded) in the past month, and d) migraine disability using an abbreviated daily version of the MIDAS (Stewart, 2001).

4. Data pre-processing

R program was used for data cleaning, and the missing data imputation, correlation analysis, and model development were done in Python.

During the data collecting processes, we deleted rows with more than nine consecutive N/A replies by participants from the question asking if the respondent completed or did not complete the questionnaire. This decision was made based on the reasoning that if people had three consecutive days of missingness, even though they received an extra contract if they failed to record data for two days in a row, it means the missingness is not because the participants did not fill out the form but because they did not have access to the data form. Then, we subsetted the data that did not have a missing outcome variable, whether they had a headache or not. Then, using the presumption that the missing data are not completely random, we did a multivariate imputation for missing data in a subset of the features.

After processing missing value data, the research conducted a correlation analysis for the characteristics to avoid a strong correlation between distinct features. Each feature's Pearson correlation coefficient was calculated for every other feature. Figure 2 displays the visualization outcome of the correlation analysis. No features were removed because of the high correlation value since none had a correlation value greater than 0.9.
Figure 2. Correlation heat map between each possible pair of features. Blue means negative correlation. Red means positive correlation. Darker color means a higher correlation.

We looked for data imbalances in the outcome variable. There were 5793 headache occurrences and 989 absences of headache (Figure 3). We discovered that there are almost 5.85 times as many observations in the majority class (presence of a headache) as in the minority class (absence of a headache). However, we decided not to balance the data as it can introduce bias and cause the model not to reflect the real world. Also, the outcome was not truly rare, with an 18% prevalence, which lowers the need to balance the data as the minority class count is already sufficient to inform our classifier.
Figure 3. Imbalance in the primary outcome variable (absence/presence of headaches)

5. Feature selection

The main goal of feature selection was to eliminate features gated by the outcome variable. We removed questions only posed when the primary outcome variable is present. Moreover, because the model seeks to provide a forecast within 24 hours, the response was expanded or merged to include all three-time points for questions only asked in the morning, midday, or evening. Afterward, 45 features in total were applied.

6. Primary outcome

The primary analysis is predicting a future headache attack in 24 hours based on the features in the headache diary. The existence or absence of any headache attacks over the 24 hours noted in the diary entry serves as the primary outcome of this investigation. The outcome variable question asked whether the participant was experiencing the headache now. Values 1 and 2, "not anticipating a headache" and, "I think I might develop a headache," were recoded as the absence of the headache. Values 3 and 4, "I am presently
having a headache" and "I had a headache after last entry but now I am free of pain," respectively, were recoded as the presence of the headache. A 24-hour headache absence/presence variable was created by combining the outcome variables gathered at three different time intervals.

7. Gradient-boosted classification model development

This study developed the model using a gradient boosting classifier using GBM Python scikit-learn library. The model was trained by passing 80% of the dataset to the gradient-boosted classifier, and 20% was used for testing. Gradient boosting classifier is a supervised machine-learning algorithm that combines many weak learning models to create a robust predictive model. Furthermore, the decision trees in gradient boosting are built sequentially, meaning each decision tree is built one after another. Previous clinical research that used gradient boosting has been successfully used to predict cardiovascular events, diabetes, and the development of sepsis, which shows that gradient boosting can capture complex relationships with non-linearity (Zhao et al., 2019; Delahanty et al., 2019; Raja, 2019). Thus, we chose to use gradient boosting classification as it results in more accurate models and the ability to deal with non-linear data.

We used Grid Search to fine-tune the hyperparameters to prevent overfitting. Grid Search is a tuning method that iterates all possible combinations of a model's hyperparameters to identify the best performance. As the testing data was not balanced, the study chose the f1 score as a single metric for Grid Search to get the optimal hyperparameters. The following was determined to be the ideal combination of hyperparameters in this study: learning rate = 0.3, max_depth = 5, max_features = sqrt, n_estimators = 150, subsample = 1.0. The outcome demonstrates that 150 decision trees
with a maximum depth of five make up the gradient-boosting classification model that is most optimal. In addition, the square root of the total number of features was used at each split with the subsample of 1.0, indicating that each tree will use the entire dataset.

8. Overfitting

While GBM tends to be more powerful in terms of accuracy, it tends to have a higher risk of overfitting when dealing with many variables. We choose to run 5-fold cross-validation to identify overfitting to ensure that our model is effective in being generalized and performs well in novel data. We divided the data points into five subgroups of equal size (folds). At each turn, one split act served as the testing set and the others as the training set. We averaged the results after each iteration to determine the performance of the entire model. Finally, we looked to check if there was a significant discrepancy between the cross-validation F1 score and the F1 score of our model than 0.05, which denotes an overfit.

9. Evaluation Metrics

The study used a confusion matrix to assess the performance of the gradient-boosting classification model through specificity and sensitivity. Further performance measurements, including precision, sensitivity (recall), f1 score, and support of the classification system, were reported by the classification report. The study also measured the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). ROC curve is an evaluation metric for binary classification. The AUC summarizes the ROC curve and measures how well a binary classifier can distinguish between classes. The greater the AUC, the better the model's performance differentiating between positive and negative classifications.
The feature importance provides a score indicating how useful or valuable each feature was in constructing the boosted decision trees within the model. For a single decision tree, importance is determined by the improvement in performance that each attribute split point makes, weighted by the number of observations that the node is in charge of. The purity used to choose split points or other, more specialized error functions may be considered performance. The average feature significance across all the model's decision trees is then calculated.

RESULTS

We used a confusion matrix to describe the classification model's performance and provide insight into what the model misclassifies (Figure 4). Using the confusion matrix, we calculated the sensitivity and specificity of the model.

\[
\text{Sensitivity (Recall)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = 0.72
\]

\[
\text{Specificity} = \frac{\text{True Negative}}{\text{False Positive} + \text{True Negative}} = 0.99
\]

A sensitivity of 0.72 indicates that the model can reliably identify those with headaches as positive with few false negative findings. With a specificity of 0.99, the model can well identify people without headaches as negative with few false positive outcomes. The model's goal was to forecast a headache so that the patient may take preventative medicine for a probable headache attack, which includes more harm when there are a lot of false positives than false negatives (Silberstein & Liu, 2002). Thus, having lower sensitivity than specificity was not a huge issue.
The study also generated a classification report demonstrating precision, recall, F1 score, and accuracy for model evaluation. Table 1 shows that the precision, the proportion of true positive predictions among all positive predictions, was 0.90. The precision of 0.90 means that out of all the participants that the model predicted to get a headache, 90% had an actual headache. Table 1 also shows a recall of 0.72, which means that out of all the participants that did get a headache, the model only predicted the outcome correctly for 72%. The report also shows an F1 score of 0.80, a weighted average of precision of 0.90, and a recall of 0.72. It provides a single measure of the model's accuracy, which considers both precision and recall. Since the value is close to 1, it tells us that the model does a good job of predicting whether participants will get a headache or not in 24 hours. Finally, the report shows that the model has an accuracy of 0.96, which is the sum of true positives and true negatives divided by the total number of samples. As the testing
data was not balanced, the study chose the f1 score as a single metric for Grid Search to get the optimal hyperparameters.

Table 1. Classification report of the model.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>No headache</td>
<td>0.96</td>
<td>0.99</td>
<td>0.98</td>
<td>1187</td>
</tr>
<tr>
<td>Had headache</td>
<td>0.90</td>
<td>0.72</td>
<td>0.80</td>
<td>170</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.96</td>
<td>1357</td>
</tr>
<tr>
<td>Macro average</td>
<td>0.93</td>
<td>0.86</td>
<td>0.89</td>
<td>1357</td>
</tr>
<tr>
<td>Weighted average</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
<td>1357</td>
</tr>
</tbody>
</table>

The study also looked at the AUROC curve to evaluate the model. The AUROC curve is a probability curve that plots the true positive rate against the false positive rate. Figure 5 showed an AUC of 0.94, indicating a high chance that the classifier will be able to distinguish the positive primary outcome values from the negative primary outcome values as the classifier can detect more numbers of true positives and true negatives than false negatives and false positives. The confusion matrix, classification report, and AUROC curve results demonstrate that the developed model can accurately predict headache risk based on the headache diary’s question on physiological symptoms and lifestyles.

Figure 5. AUROC curve of the model.
Figure 6 shows the relative importance of features in predicting headache attacks. The figure suggests that having no premonitory symptoms is the most important feature contributing to headache prediction, followed by feeling tired or weary, stiff neck, dizziness, predictive medication, menstruation, nausea, difficulty thinking or concentrating, quality of sleep, and total stress score variable.

![Top 10 Feature Importance](image)

**Figure 6.** Top 10 feature importance ranking

**DISCUSSION**

The study found that future headache attacks in 24 hours can be accurately predicted for 23 adult headache patients between the ages of 18 and 65 using the GBM classification model with an F1 score of 0.80. Among the 45 features used, the most influential variable in predicting the possibility was the state with no premonitory symptoms. The list was followed by feeling tired or weary, stiff neck, dizziness, predictive medication, menstruation, nausea, difficulty thinking or concentrating, quality of sleep, and total stress score variable.
The study used a gradient-boosted classification model with a broader range of features from a headache diary, including lifestyle factors and premonitory symptoms. Most previous studies used the decision tree or random forest model, in which decision trees are not added sequentially like a gradient-boosting algorithm (Vandewiele et al., 2018; Barra et al., 2020; Alves et al., 2021). While gradient-boosted classification is more prone to overfitting, our study found no overfitting in our model, which means the model can be applied to unknown data. Also, unlike some previous studies that only focused on premonitory symptoms or stress-related factors, this study created a model that can predict a given person's likelihood of experiencing a headache attack using various information from the headache diary, ranging from sleep quality to time of food consumption (Houle et al., 2017; Gago et al., 2018). The study also narrowed down features that are more important than others in predicting headaches. This finding may be advantageous since it allows us to limit the number of questions included in the dairy to those that are essential.

There are, however, a few limitations to this study. The first limitation of this research is that the feature importance report does not inform us of the feature's direction of the association with headache risks, whether they are positive or negative. Therefore, tools that can understand the prediction output of machine learning models, such as SHAP (Shapley Additive Explanations), should be used to identify the positive and negative impacts of the feature on the prediction to make the model more interpretable (Zoabi et al., 2021). Another limitation of this study is the need for more generalizability due to the lack of information on demographic data and the low sample size. As demographics, such as age, gender, race/ethnicity, and socioeconomic status, can influence how patients respond to diaries, it is difficult to determine if the result of the current study can be generalized to a broader population. Additionally, small sample size can weaken the study's statistical power, making it more difficult to draw firm conclusions regarding the
intervention's efficacy. Future research could concentrate on gathering comprehensive demographic information and enlisting a more extensive and varied sample of participants to improve the study outcome's generalizability. Also, based on the current study's result on feature importance, future studies should seek to create a headache diary with fewer items with high predictive values and observe whether adherence can be elevated. Lastly, given that headache diaries are self-reported data, there is a higher chance of recall bias. Although the recollection interval was shorter since the diary was taken three times a day, other factors, such as older age, may also contribute to recall bias. Therefore, future studies should examine whether the classification differs by stratified age, sex, and socioeconomic characteristics.

CONCLUSION

The study shows that future headache attacks can be accurately predicted for adult headache patients using the GBM classification model that used headache diary data, with an F1 score of 0.80. The study also found that premonitory symptoms, predictive medication, menstruation, quality of sleep, and total stress score had a higher predictive ability than other headache diary questions. Additional research is needed to explore whether the model can be used in other populations and whether a headache diary with fewer variables with higher feature importance can be developed.
REFERENCE

Armstrong, L., & Gossard, G. (2016). Taking an integrative approach to migraine headaches: with about half of migraine sufferers using CAM, it's important to know which alternative approaches are most likely to help and what to tell your patients. Journal of Family Practice, 65(3), 165-173.


