CPSC 6300 I Applied Data Science I Spring 2023

Migraine: Classification Based on Gene Expressions

Suraj Khamkar	Sravani Pati	Pragathi Pendem	Rahul More
MS in Computer Science	MS in Computer Science	MS in Computer Science	MS in Industrial
Clemson University	Clemson University	Clemson University	Clemson University
skhamka@g.clemson.edu	spati@g.clemson.edu	ppendem@g.clemson.edu	rsmore@g.clemson.edu
Checkpoints, Modeling, Prediction	EDA, Feature Extraction	Data pre-processing, Report	Data Collection, Presentation

ABSTRACT

Migraine is a debilitating neurological disorder affecting a significant portion of the population worldwide. In this project, we aim to predict the type of migraine using machine learning classification models based on gene sequence data. We applied various machine learning classification algorithms, including K-Neighbor, Decision Tree, MLP Classifier, Support Vector Machines, Random Forest, and Grid Search CV, to the selected gene set. In conclusion, our study demonstrates the potential of machine learning classification models based on gene sequence data for predicting migraine type. These models may help improve diagnosis and treatment of migraine patients, leading to better outcomes and quality of life for those suffering from this debilitating disorder.

Keywords

M – Millions; ASCII, DNA, RNA, adenine (A), guanine (G), cytosine (C), and uracil (U), and UFOLD, Condon, etc.

1. INTRODUCTION

1.1 Fundamental Problem to Address

Migraine is a neurological disorder that affects millions of people worldwide, with a significant impact on quality of life. There are various types of migraines, including migraine with and without aura, Familial hemiplegic migraine, Typical aura without migraine, Basilar-type aura, and others. While the exact cause of migraine is not fully understood, genetics is believed to play a significant role. We evaluated the performance of each model and calculated the accuracy.

Advances in technology and the availability of large-scale genetic datasets have made it possible to study the genetic basis of migraine in more detail. In recent years, machine learning techniques have been applied to gene sequence data to predict various migraine type. In this study, we aimed to predict the type of migraine using machine learning classification models based on gene sequence data.

1.2 Motivation

Improved prediction of migraine type using machine learning classification models could have significant implications for the diagnosis and treatment of migraine patients. Accurate diagnosis of migraine type could help to tailor treatment to the specific needs of each patient, leading to better outcomes and an improved quality of life. This study highlights the importance of using machine learning techniques to identify the genetic basis of complex neurological disorders and to develop more effective treatments.

1.3 Dataset

We used publicly available gene expression datasets [1] for migraine patients and healthy controls to identify differentially expressed genes. Also, we have used processed data [2] to train models. In our initial observation, we saw that every sequence had a distinct sequence identifier. An organism's genetic code is determined by the precise arrangement or sequence of these nucleotides, which serve as the DNA molecule's building blocks. Before the data is transformed from ASCII values to integer values, the fastq file's ASCII values for each character are first extracted. Each fastq file consist of more than 10M of data.

2. SUMMARY OF EDA

2.1 Unit of Analysis

The genetic information in RNA is encoded by a sequence of nucleotides, which are the building blocks of RNA. RNA is composed of four types of nucleotides: adenine (A), guanine (G), cytosine (C), and uracil (U). The sequence of these nucleotides determines the sequence of amino acids in a protein, which in turn determines the structure and function of the protein. This sequence with its relevant quality score can be used to extract Condon information from sequence based on location of gene pattern with number of times it repeated. This can be help to determine the type of migraine and this type considered as unit of analysis to train models.

2.2 Observations

Basically, each fastq file consists of more than 10M of data. But, after data cleaning steps like deletion of null and duplicate data, and deletion of sequence having no quality score it eventually ends up having 1M of data. After preprocessing of the remaining data to get its type and symptoms based on gene expression (which required medical software's like UFOLD, InShot, etc.) we will get formatted data to process.

There are seven unique types of migraine based on processed data which are migraine with and without aura, Familial hemiplegic migraine, Typical aura without migraine, Basilar-type aura, and others. This data is taken from peoples starting from age 13 and sex matched volunteers. The time period in which this data has been recorded was two years of span.

2.3 Data cleaning steps

One common aspect of data cleaning is removing null data or missing values from the dataset to ensure that the data is consistent and reliable. Another important step is removed duplicate data to avoid over-representation of certain observations, which can lead to bias in the analysis. Additionally, removed sequences that do not have quality score information is essential to ensure that the data is

of high quality and reliable for downstream analyses. By performing these data cleaning steps, the resulting dataset became more accurate and reliable, which will improve the quality of any subsequent analyses or models.

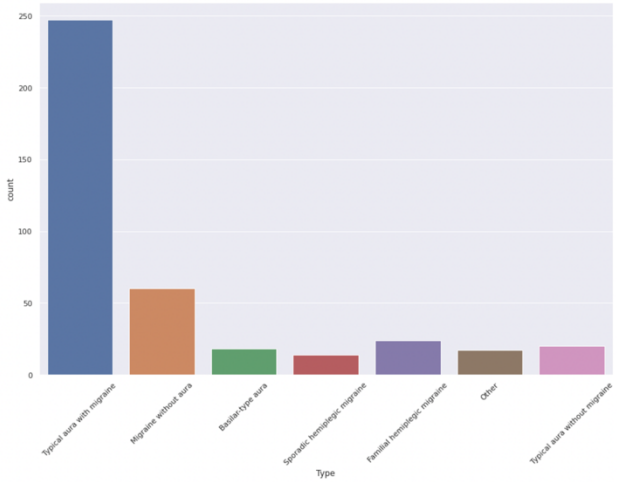

Fig 1. Histogram for Types of migraine

In the Fig. 1 above, histogram is plotted for types of migraine derived from dataset [2].

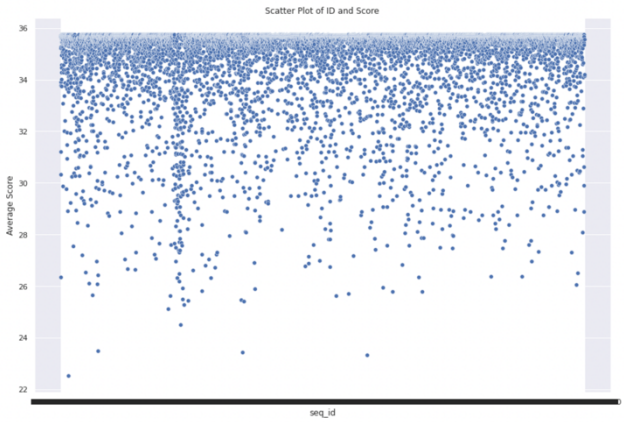

Fig 2. Scatterplot for Average score

In the Fig. 2 above, Scatterplot is shown for average score of each gene sequence id derived from dataset [1]. Using this average score and pattern of Condon we can determine the type of migraine, and symptoms for each gene expression.

We utilized `SeqIO.parse()`, as seen below Fig. 3, to get the same result. The data is then converted into dataframe. Here, we've used `letter_annotations['phred_quality']` to convert the ASCII values of quality scores to integers. The most important observation is that RNA sequence is unique for each sequenceID, causes data points to be unique. we have calculated average for all the quality scores points so that it would be easy to predict quality scores based on training and testing data. The DNA sequence is GATTTGGGGT…, and the sequence ID is ERR4796171.1. The quality ratings in this instance are [32, 32, 32, 32, 32, 36, 36,...]. Greater confidence in the base call is indicated by higher quality scores, and greater uncertainty is indicated by lower quality scores.

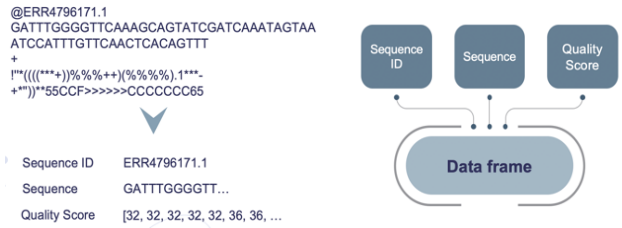

Fig 3. Data pre-processing

3. SUMMARY OF MODELS

We then applied various machine learning algorithms, including K-Neighbor, Decision Tree, MLP Classifier, support vector machines, Random Forest, and Grid Search CV, to predict migraine type based on the selected gene set. Here is summary about three best models out of all these six models.

3.1 Multi-Layer Perceptron (MLP) Classifier

The MLP Classifier is a neural network-based algorithm that can learn non-linear decision boundaries, handle various data types and large datasets, and offers several regularization techniques and tunable hyperparameters. It is a good choice for classification tasks that require a flexible and powerful algorithm, but its performance may depend on the data characteristics and hyperparameter tuning.

F1-Score: 0.73
Accuracy Score: 0.89

3.2 Support Vector Machine (SVM) Classifier

Support Vector Machine (SVM) classifier is a popular algorithm for classification tasks because it has the ability to handle complex data and learn non-linear decision boundaries. Additionally, SVMs offer different kernel functions that can be used to handle various data types and achieve better separation between classes. Overall, SVM is a powerful algorithm that can be a good choice for classification tasks, especially when dealing with complex and non-linear data.

F1-Score: 0.85
Accuracy Score: 0.92

3.3 Random Forest (RF) Classifier

Random Forest (RF) classifier is a popular algorithm for classification tasks because it has the ability to handle high-dimensional data and learn complex decision boundaries. RFs can also work well with both small and large datasets and provide good generalization performance. Additionally, RFs use an ensemble of decision trees to reduce the risk of overfitting and can provide feature importance measures that can help in feature selection.

F1-Score: 0.82
Accuracy Score: 0.93

3.4 Benchmark Model

Based on the provided accuracy values, the RF model outperforms the MLP and SVM models, achieving the highest accuracy of 93%. This indicates that the RF model is better at classifying the target variable and predicting the correct class labels for new, unseen data.

The RF model's superior performance may be due to its ability to handle high-dimensional data and learn complex decision boundaries. RF's use an ensemble of decision trees to reduce the risk of overfitting and provide robustness against noise and outliers.

Additionally, RF's can provide feature importance measures that can help in feature selection and interpretation of the model.

Overall, the RF model is a powerful and versatile algorithm that can be a good choice for classification tasks, especially when dealing with complex and high-dimensional data.

```
rfr = RandomForestClassifier(random_state= 32)
rfr.fit(X_train_scale, ytrain)
ypredrfr = rfr.predict(X_test_scale)
Acc_RF=accuracy_score(ypredrfr,ytest)
print('accuracy',Acc_RF)
```

```
accuracy 0.9333333333333333
```

```
# RF
from sklearn.metrics import classification_report
print(classification_report(ytest, ypredrfr, target_names=target_names))
```

```
                              precision    recall  f1-score   support

    Typical aura with migraine       0.60      1.00      0.75         3
         Migraine without aura       1.00      0.38      0.55         8
             Basilar-type aura       0.94      1.00      0.97        15
 Sporadic hemiplegic migraine        1.00      0.75      0.86         4
 Familial hemiplegic migraine        0.67      0.67      0.67         3
                         Other       0.95      0.99      0.97        80
Typical aura without migraine        1.00      1.00      1.00         7

                     accuracy                           0.93       120
                    macro avg       0.88      0.83      0.82       120
                 weighted avg       0.94      0.93      0.93       120
```

Fig 4. Random Forest Model Code and Results

3.5 Predictions on Benchmark Model

Prediction 1

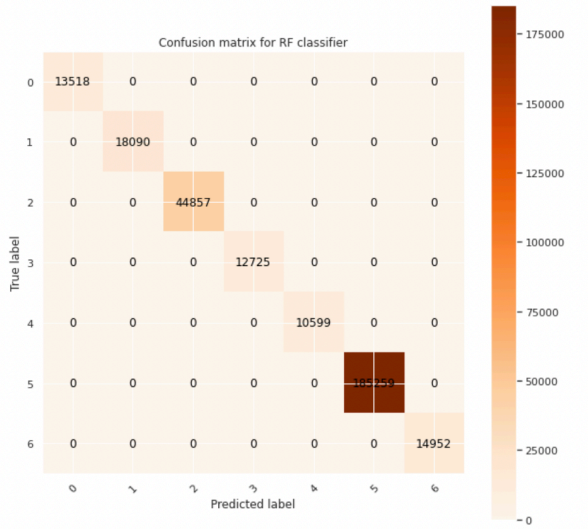

Fig 5. Confusion Matrix for Prediction of type of migraine

This Fig. 5, show prediction using confusion matrix to predict the type of migraine. This prediction has been made on 1M of data. But, it takes long to process that much data. This predicted data also shows accuracy of approximate 90% which we have mentioned earlier. With more advanced data and better accuracy this kind of prediction will also help to predict medications for that particular patient.

Prediction 2

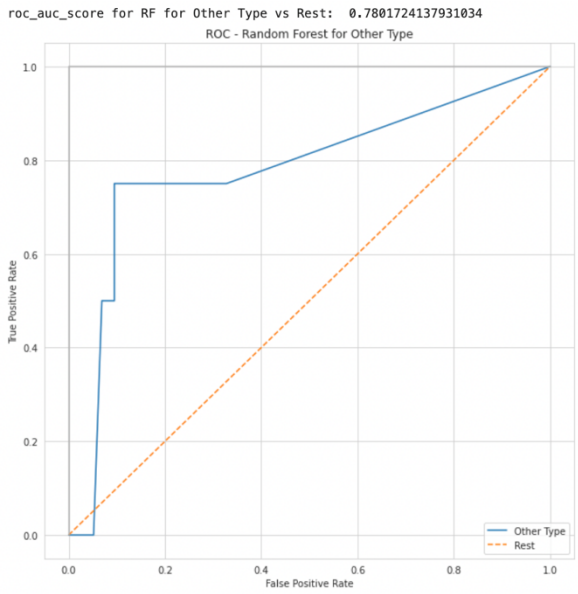

Fig 6. ROC Plot for Random Forest

This Fig. 6, show prediction using ROC curve to predict the Other type over rest of migraine types. This prediction has been made on just 400 records. But, we have got 0.78 accuracy. With more number of records accuracy will be better.

Prediction 3

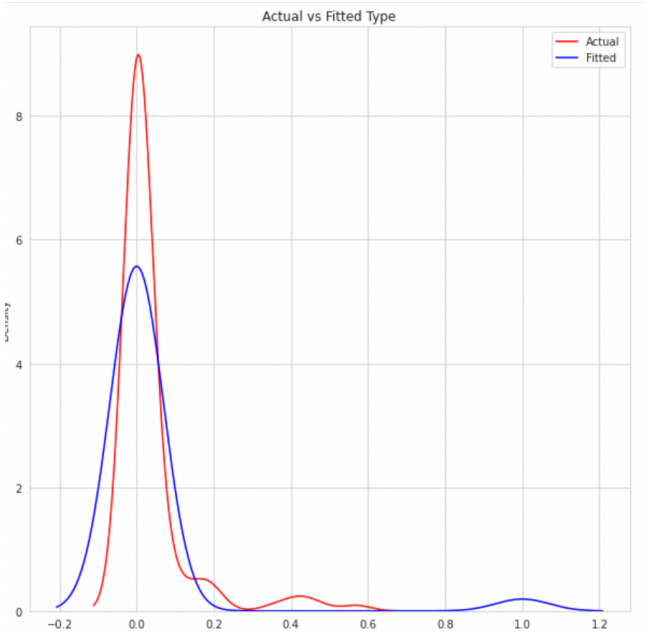

Fig 7. Actual vs Fitted Type for Basilar Type Aura Migraine

This Fig. 7 illustrates prediction using histogram to predict actual and predicted values of migraine type Basilar aura migraine. As we have used few records the accuracy of the data predicted is 75 percent.

4. CONCLUSION

4.1 Learnings from Project

The successful training of the Random Forest model with 93% accuracy on classifying the types of migraine and predicting the quality score for gene expression is a significant achievement. This accuracy indicates that the RF model has learned the underlying patterns and relationships between the features and the outcomes in the data. The high accuracy also implies that the model can make reliable predictions for future data. However, it is important to note that the model may have limitations in its generalization ability, meaning that it may not perform well on unseen data from different sources. Therefore, it is necessary to evaluate the model's performance on external validation datasets to ensure its reliability and robustness. Overall, the successful application of the Random Forest model demonstrates the potential for using machine learning algorithms to aid in the diagnosis and treatment of migraine, as well as in predicting the quality score for gene expression.

4.2 Future Scope for Health Department

Our project of classifying types of migraines and predicting quality scores for gene expression could be valuable for domain experts in medical field. For example, medical professionals could use the classification of different types of migraines to better understand and diagnose patients, leading to more targeted treatment plans. Additionally, understanding the relationship between gene expression and migraine quality scores could potentially provide insights into the underlying mechanisms of migraines and lead to the development of new treatments. The results of your analysis could inform the work of domain experts by highlighting potential biomarkers or genes that are associated with certain types of migraines or quality scores. These insights could then be further investigated and potentially used to develop more effective treatments or therapies. Overall, project has the potential to provide valuable insights for medical professionals and researchers working in the field of migraine research.

4.3 Possible Project Improvement

One way that the project could be improved is by gathering more data from a larger and more diverse population to increase the generalizability of the results. Additionally, incorporating more relevant features or using alternative feature selection methods could improve the performance of the models. Another possibility is to explore other machine learning algorithms or ensemble methods to potentially achieve even higher accuracy. Further, incorporating external data sources, such as genetic data or medical records, could provide additional insights into the underlying mechanisms of migraine and lead to more personalized treatment options for patients.

5. ACKNOWLEDGMENTS

Our thanks to Prof. Carlos Toxtli, Mr. Ravi Teja, and Ms. Nushrat Humaira for all the help related to understanding the dataset and clarifying the doubts we had throughout the project development.

I would also like to thanks Mr. Hussain M. S. and Mr. Shailesh Alluri for allowing us to use their processed dataset and to refer their code.

6. REFERENCES

[1] ENA Dataset
https://www.ebi.ac.uk/ena/browser/view/PRJEB40032

[2] Processed Dataset
https://github.com/hussainthedatasufi/migrainKNN-NB-SVM/blob/main/data.csv

[3] Gnome Sequence
https://www.ncbi.nlm.nih.gov/books/NBK21136/

[4] Classification with Neural Network
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8564744/

[5] Transcriptome Extraction
https://www.nature.com/scitable/topicpage/ribosomes-transcription-and-translation-14120660/

[6] MLP Classification
https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

[7] Random Forest Classification
https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[8] Identification of disease- and headache-specific mediators and pathways in migraine using blood transcriptomic and metabolomic analysis