HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 3 (2021)

# Predicting Aphasia from Strokes

Joseph Jia[1] and Joanna Gilberti[#]

[1]Lexington High School, Lexington, MA, USA
[#]Advisor

## ABSTRACT

Strokes can occur when someone's blood vessels get blocked and the nutrients and oxygen being transported will not reach the brain. When a stroke happens, the brain cells don't get the nutrients they need and start to die [3]. This could cause different side effects after a stroke. In this study, we try to predict the possibility of one type of after-stroke side effect, aphasia, using Machine Learning (ML) techniques. Using the data of a study about brain lesion damage after a stroke and what affects the patients were experiencing afterwards, we trained a model to predict whether a person may have aphasia based on where their lesion was, how big the lesion was, how long ago their stroke was, and some other factors. We evaluated several classification methods and found that using linear discriminant analysis was the most accurately predicting when we used age, sex, lesion location, lesion volume, and many more. By linear discriminant analysis, we were able to have a 91% overall predictive rate of patients having aphasia or not after experiencing a stroke.

## Introduction

Even though strokes happen most commonly in elderly people, there is still a big group of younger patients who have experienced a stroke. Young people are also at risk and, in our dataset, there were almost 15% of patients who were 50 or younger. In 2017, the CDC released data, showing that approximately 795,000 people in the US experience strokes every year [2]. Based on data in the US, every 40 seconds a person has a stroke and every 4 minutes a person in the US dies from a stroke. Stroke is one of the lead-causes of death in the United States [2,3]. It is important to understand what it can cause to your body, especially your brain. Having a stroke can lead to many side effects.

In this study, we focused on the side effect of stroke, aphasia. People may experience aphasia when the portions of their brain responsible for language is damaged, possibly by a stroke or head injury. When this happens, the person will struggle with the understanding and expression of languages, affecting their reading and writing abilities [4]. We built a ML model to predict whether or not someone, who had experienced a stroke, would have aphasia or not. In order to accomplish this, the ML model learnt the differences in having aphasia versus not having aphasia. The learning requires sufficient data to formulate a pattern so that it is able to make good predictions about what defects, especially aphasia, a person may have from a stroke. It is also important to emphasize that aphasia is not the only after effect of strokes, and that we chose it because it changes our lives and forces people around us to adapt.

## Review of Literature

There is currently a great understanding of the effects of aphasia and methods to cure it. There are also many studies about predicting the recovery of aphasia of patients as well as predicting aphasia type based on MRI scans. The main focus of aphasia is the recovery of a patient from aphasia based on pictures and scans to determine the severity and recovery for the patient. However, I am not able to find any articles or studies about being able to predict if a patient has aphasia.

# Methods

We built an ML model by using the individual patient data from doi.org[1] that did an experiment on people who had experienced a stroke before. The study found out what types of defects people had and one of the defects they had tested for was aphasia [1]. We split the data into two groups: 75% being training data and 25% being test data. In order for a model to learn well, we needed to determine the important features which would allow us to predict aphasia in patients as accurately as possible. Based on the data table given by this source, we used 10 features: age, sex, phase, lesion type, lesion location, lesion area, time since stroke, FM A, FM B+C, and FM Sensation.

For features with limited numbers of values, we need to have enough data for each individual value so that the model can learn effectively. For some features, we removed values, which have very limited data. An example of data removed was the left hemisphere basilar artery for the lesion location of L-BA since there was only one person with this lesion location. If such data were used to train a model, it may be thrown off since this is a unique case, possibly making it more inaccurate. This allowed the model to focus on cases with rich enough data.

We tested multiple different classification methods to determine the best method for the predictive model to be successful by using certain features of each patient in the data and modifying which features we would use and then running a program to see how accurate each classification method was. The classification methods we tested were logistic regression classifier, decision tree classifier, k-nearest neighbor classifiers, linear discriminant analysis, gaussian naive bayes, and support vector classifiers [5]. Logistic regression classifier uses a logistic model to help predict outcomes. Decision tree classifiers use parameters to continuously categorize data and eventually predict the result based on how previous data with similar traits resulted in, creating a tree-like graph. K-nearest neighbor classifiers determine what group data points are part of by looking at data points around them and find the best fit. Linear discriminant analysis is a classification method that uses dimensionality reduction, which removes redundant and dependent features by transforming the features into a lower dimensional space. Gaussian naive bayes is a classification method that uses the Gaussian distribution on continuous data to be able to predict the outcome. Finally, support vector classifiers use a support vector machine to be able to build a learning model by assigning data to one group or the other, therefore, forming a binary linear classifier. The model generated can then be used to predict unknown outcomes of data. We noticed that some classifiers were more effective than others based on the features being used and that the accuracy fluctuated when we changed the features being used. For a model, we want it to have high precision, or correct predictions, for both positive and negative cases. However, we can hardly get both to be high due to false predictions. Interestingly enough, we noticed that the gaussian classifier has 100% precision for negative cases but very low precision for positive cases. This means that it would only choose that a patient has aphasia if it was very confident, but otherwise it would always say the patient didn't have aphasia. It would be able to guarantee 100% recall, accuracy of the actual results when compared to the results of the learning model, on people having aphasia, with the cost of only 65% recall on people not having aphasia. This showed that the Gaussian was not the best fit for the classification method because there was not much for the model to learn on different cases if it was always choosing one result, unless it was extremely confident about the other. When it came down to choosing which classifier we were going to use, we considered what type of accuracy we wanted the model to have. The three we considered were: having higher accuracy of positive recall and negative precision, having equal rates of precision and recall, as well as having high negative recall accuracy and positive precision accuracy. Each of these types had their own benefit but we decided having equal, high rates of precision, the accuracy of the learning model, and recall, accuracy of the model compared to the model predictions, was the best option because it allowed for higher overall accuracy. The one downside to this option is that the prediction could be wrong no matter what option they choose more often compared to the other more conservative approaches. We also noticed that a model could overfit based on the training data but predicts not well using test data. Decision tree model reached 100% accuracy on the training data, but only 82% on the test data. With a significant drop on the set of test data, we think the model overfits. Based on our evaluation by

looking at the performance on both training data and test data, the best classification method is linear discriminant analysis.

For the model, we want to use the most effective features, or the features that were most important to making accurate predictions. We tested which feature we could remove without significantly changing the overall accuracy by limiting what features about the patients that the model could learn and testing how accurate linear discriminant analysis classification would be. Testing for a classification method would also allow us to check features, as we would be able to see the changes in accuracy of the different methods when we changed which features were incorporated. After we chose to use linear discriminant analysis because it was the most consistently accurate, we tested the model by removing some features. Sex and age are good features to include because they can help categorize the patients since they are limited answers to sex and age, which may be useful for understanding which age group may have the worst lesions. The area of lesion and size of lesion are important because it can change the results of what effects people will experience. Time after the stroke is also important because the longer the time after, the more likely the patient could have time to either have worse aphasia or possibly have time to recover from it. The rest of the features are tests that the patients perform which can help determine what kind of side effects they are. At the end, we concluded that it was best to include almost all the features because all of the features included in the dataset were important, in some way or another, to help determine if a patient had aphasia or not.

Linear Discriminant Analysis is a method of classification that uses the data given to create a formula. We would use that formula and test it with specific data in our dataset to see if the program was as reliable.

# Results

We did not choose Linear Discriminant Analysis randomly, but instead we tested each of the types of classification using the train dataset. The predictive accuracy rates for the 75% portion of the data for training is shown below:

| Classification Type | Positive Case Precision | Negative Case Precision | Positive Case Recall | Negative Case Recall | Overall Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 73% | 83% | 44% | 94% | 81% |
| Decision Trees | 100% | 100% | 100% | 100% | 100% |
| K-Nearest Neighbors | 64% | 85% | 56% | 89% | 80% |
| Linear Discriminant Analysis | 69% | 90% | 72% | 89% | 85% |
| Guassian Naive Bayes | 50% | 100% | 100% | 65% | 74% |
| Support Vector | 79% | 83% | 44% | 96% | 82% |

After using 75% of the data to train the machine to create a formula to use when data is entered from a user. Once this was completed, we used the remaining data to test the program out, which was 25% of the data. Below are results from the test data:

| Classification Type | Positive Case Precision | Negative Case Precision | Positive Case Recall | Negative Case Recall | Overall Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | 100% | 79% | 45% | 100% | 82% |

| | | | | | |
|---|---|---|---|---|---|
| Decision Trees | 69% | 90% | 82% | 82% | 82% |
| K-Nearest Neighbors | 100% | 81% | 55% | 100% | 85% |
| Linear Discriminant Analysis | 90% | 91% | 82% | 95% | 91% |
| Guassian Naive Bayes | 69% | 100% | 100% | 77% | 85% |
| Support Vector | 100% | 73% | 27% | 100% | 76% |

These tables show the precision, recall, and overall accuracy by splitting the dataset into a training set and test set randomly. Precision is determined by answers from the model and how accurate they are compared to actual results, while recall is determined by the actual results and how accurate those answers are based on the answers the model predicted. Because our goal was to predict whether someone had aphasia after experiencing a stroke, it would be best if both recall and precision were high for both positive and negative cases. In the training dataset, Linear Discriminant Analysis had an 85% accuracy rate, second to decision trees with 100%. Based on those conditions, the results of the testing show that Linear Discriminant Analysis is the best. This is because the precisions and recalls in the test set are 90%, 82%, 91%, and 95%, which are all high or relatively high accuracy rates. This method of classification also has the highest overall accuracy in the test data of 91% and there is a 6% difference between Linear Discriminant Analysis and the second-most accurate methods, which are K-Nearest Neighbors and Guassian Naive Bayes. We used Linear Discriminant Analysis because it had high overall accuracies compared to the other methods of classification and it proved decently consistent.

Here are a few of examples from the test data in the table below by Linear Discriminant Analysis:

| Phase | Lesion side and Territory | Lesion type | TAO (months) | Lesion volume (cc) | FM A (-/30) | FM B+C (-/24) | FM T (-/66) | FM Sensation (-/12) | Aphasia | Predicted Result |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 1.38 | 0.67 | 30 | 24 | 66 | -1 | -1 | -1 |
| 3 | 1 | 1 | 1.97 | 11.61 | 10 | 6 | 22 | -1 | 1 | 1 |
| 2 | 2 | 2 | 45.93 | 1.45 | 27 | 20 | 53 | 4 | -1 | -1 |
| 3 | 2 | 2 | 1.21 | 11.87 | 28 | 18 | 54 | -1 | -1 | -1 |
| 3 | 1 | 2 | 1.64 | 7.27 | 20 | 16 | 46 | -1 | 1 | -1 |
| 2 | 2 | 2 | 31.02 | 92.41 | 30 | 24 | 64 | 12 | -1 | -1 |
| 1 | 1 | 1 | 0.82 | 22.23 | 2 | 0 | 6 | 1 | -1 | 1 |
| 1 | 2 | 2 | 0.56 | 72.48 | 0 | 0 | 4 | 9 | -1 | -1 |
| 1 | 2 | 2 | 0.85 | 2.23 | 19 | 14 | 41 | 12 | -1 | -1 |

As you can see, most of the outcomes are accurate. However, there are still some errors in the prediction formula, due most likely to a smaller size of data. When there is a smaller size of data, the formula will have less data to create the formula, therefore creating an equation that is not as accurate as a formula created with thousands of data. Overall,

the linear discriminant analysis model was accurate and provided good predictions to patients if they had aphasia or not.

# Discussion

In this study, there was one major unexpected finding and a few other smaller unexpected findings. The major finding was that the decision trees had 100% accuracy during the train data but only 82% accuracy in the test data. I found this relatively surprising that a type of classification method even had 100% accuracy but then I had much higher expectation for decision tree classification in the test data but it only got 82% accuracy. This may have happened because the train data did not cover every case presented in the test data and the decision tree may have had to face those cases more often and wasn't able to be successful, which demonstrates that it wasn't able to create a successful method to classify patients. Another finding I found interesting was the accuracy of linear regression classification. The overall accuracies of the train and test data of the linear regression model was almost exactly the same but when looking at the accuracies of recall and precision for positive and negative cases, the accuracies are much different in the test data than they are in the train data. This may have happened because the linear regression model formed from the train data was able to form a model that was much better suited for the positive case predictions in the test data. A similar change is also noticeable in the K-nearest neighbor model and the support vector model. The other notable finding was the consistency of Guassian Naive Bayes classification with negative case precision with 100% accuracy in both the train and test data, however it failed to produce a model to successfully predict positive cases at a high accuracy.

# Conclusion and Implications

In this study, we used the dataset of an experiment that tested for aphasia in patients who previously had strokes. We used the data of the patients, such as the lesion area or lesion size, to allow the program to identify what patients had aphasia and which ones didn't. We used Linear Discriminant Analysis because it was the most accurate in our testing among 6 different classifier methods. By using the most accurate method overall, we were able to predict whether or not certain patients have aphasia with 91% accuracy. However, one challenge to this model is that it asks for lots of input that not everyone may have, so this model may be useful to only a limited group of people. Another challenge was the limited amounts of data. Overall, we were able to create a ML model with a good accuracy rate and tested it with specific individual data on individual patients to truly see how well our predictor does. This model can hopefully be the start to a research project that can help predict aphasia in patients and alert them early on so that they are aware of having aphasia. Because aphasia is very impactful to people, I believe that it is extremely important that we can alert people if they do have aphasia as early as possible.

# Future Research

In future research, studying the relationship between having aphasia and demographics would be very helpful to determine how aphasia may be able to be identified and what type of people are at greater risks than others. Future research should certainly further determine what characteristics of a patient would lead to aphasia and if there is anything to help prevent a patient from obtaining aphasia. Additionally, further research should also take a deeper dive into the effects of drugs or alcohol on aphasia because the use of drugs and alcohol does impact the brain and it would be interesting to see what impacts that has on a patient with aphasia.

# Limitations

This work used a data set of 130 samples, which is limited. We also covered a subset of cases as some cases have less than 5 samples. Finding more data and possibly a more diverse and balanced data set would help improve the result of the machine learning model and cover all possible cases. We think the reason why we cannot remove any features could also be because of the limited data we have in this study. After getting a larger balanced data set, we could test more combinations of features to see if we could reduce the features without having a significant change in the accuracy of the model prediction. Finally, we believe this method could be expanded for other effects from the stroke once there is a good set of data to train a model.

# Acknowledgements

Thank you to Professor Ramin Ramezani, who teaches at UCLA, for giving me engaging and wonderful meetings about AI and Machine Learning in the world of healthcare to help me gain a greater understanding of the AI world! Huge thank you to TA and advisor, Joanna Gilberti who works at NYU, for helping me edit and maintain progress with this project! Finally, I would like to give a big thank you to my parents who also helped me through this process and shared some of their own ideas as well as also helping me with some of the editing process!

# References

[1]Frenkel-Toledo S, Fridberg G, Ofir S, Bartur G, Lowenthal-Raz J, et al. (2019) Lesion location impact on functional recovery of the hemiparetic upper limb. PLOS ONE 14(7): e0219738. https://doi.org/10.1371/journal.pone.0219738

[2]Benjamin, Emelia J., and Michael J. Blaha, et al. “Circulation.” *CDC Stacks*, 25 Jan. 2017, stacks.cdc.gov/view/cdc/45425. Accessed 26 Dec. 2020.

[3] ”About Stroke.” *Stroke.org*, American Heart Association, https://www.stroke.org/en/about-stroke

[4]”Aphasia.” *National Institution on Deafness and Other Communications Disorders*, U.S. Department of Health and Human Services, Dec. 2015, www.nidcd.nih.gov/ health/aphasia. Accessed 26 Dec. 2020.

[5]Li, Susan. “Solving a Simple Classification Problem with Python—Fruits Lovers’ Edition.” *Towards Data Science*, Medium, 4 Dec. 2017, towardsdatascience.com/solving-a-simple-classification-problem-with-python-fruits-lovers-edition-d20ab6b071d2. Accessed 26 Dec. 2020.

[6]http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html, Pedregosa *et al.*, JMLR 12, pp. 2825–2830, 2011.