

Comparative Analysis for the Detection of Alzheimer's using Multiple Machine Learning Models

Satyabrat Bhol¹

¹*Department of Electronics and Communication Engineering,
Birla Institute of Technology, Mesra, India*

Abhinav Bhushan, PhD²⁺

²*Department of Biomedical Engineering,
Illinois Institute of Technology, Chicago, Illinois*

Abstract

Alzheimer's disease is a global health issue and its early detection could slow the progression of the disease. The classification of Alzheimer's using multiple Machine Learning Models is presented in this paper. Methods like Imputation, Balancing, Feature Selection, etc. were used for getting the best attributes. It was observed that Random Forest is performing best among all others for the classification of Alzheimer's.

1. INTRODUCTION

Alzheimer's is a neurodegenerative disorder affecting adults over the age of 60 years and continues for a long period. One out of every three people dies with Alzheimer's [1]. It kills more than Breast Cancer and Prostate Cancer combined. Alzheimer's leads to nerve cell death and tissue loss throughout the brain. With time, the brain shrinks dramatically, affecting nearly all its functions. Right now, there is no cure for this disease. But its early diagnosis could slow the progression of symptoms. It also prevents possible harmful treatment resulting from misdiagnosis.

2. PROPOSED METHODOLOGY

In our proposed work, the tabular dataset is collected from a Disease Center. The dataset contains Longitudinal Clinical data of Alzheimer's and Non-Alzheimer's patients. We have applied more than ten classification techniques starting from Naïve Bayes to Neural Networks to the dataset using multiple Python frameworks like Scikit-Learn, TensorFlow, etc. The attributes in the dataset include age, follow-up years of individual patients, blood biomarkers, medication status, and cognitive assessment tests.

The following are the steps involved in the proposed work:

2.1 Data Cleaning

The dataset contains 3309 unique subjects with each having multiple follow-up years. To get a unique patient's lists, we narrowed each patient to their last follow-up year. Several attributes including cognitive assessment tests and many medications, which were not a part of our research were removed. Attributes causing the problem of Data Leakage were also removed.

2.2 Data Imputation

Imputation is one of the most fundamental parts of the data pre-processing due to the incompleteness and inconsistency of real-world data. The missing entries were filled up using the K-Nearest Neighbor algorithm. It uses feature similarity to predict values of new data points. The new point is assigned a value on how closely it resembles the points in the dataset.

2.3 Data Balancing

Most machine learning classification algorithm is sensitive to unbalance in predictor class. They tend to bias towards the majority class while predicting using an unbalanced dataset. In our case, the ratio of Alzheimer's to Non-Alzheimer's patients was around 3 to 1. We used Sampling Techniques like Under-Sampling (reduces the majority class with respect to minority class) and Over-Sampling (create dummy sets for minority class to make it equal to the majority class) to balance the dataset.

2.4 Scaling

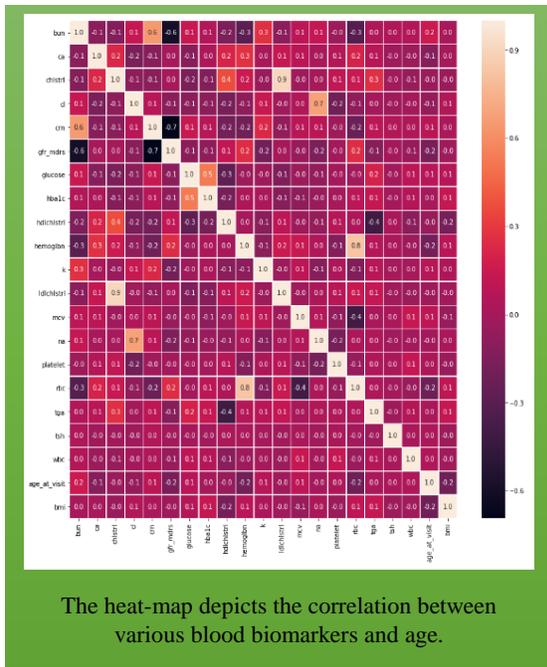
Scaling of data ensures that features whose numerical values are much greater than the others, the machine learning models won't use them as the main predictor.

The numerical features in our dataset were scaled to the range of [0,1] using MinMax Scaler. The MinMax Scaler uses the following formula:

$$Xi - \min(i)/\max(i) - \min(i)$$

2.5 Feature Selection

It is the process of selecting the attributes in our dataset that are relevant to the predictive modeling we are working on. We used the Recursive Feature Elimination (RFE) method to list out the optimal no. of features. RFE fits a model and removes the weakest feature(s) until the specified no. of features is reached.



The heat-map depicts the correlation between various blood biomarkers and age.

2.6 Classification

We used multiple classifiers in our model. Some of them are as follows:

- Logistic Regression

It is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable [2].

$$p = b_0 + b_1X_1 + b_2X_2 + \dots$$

$$odds = p/1 - p$$

$$logit(p) = \ln(p/1 - p)$$

- Random Forest

A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k=1, \dots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [3].

- Gradient Boosting

Boosting is a method of converting weak learners to strong learners, typically decision trees. It begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify.

- LightGBM

It is a fast, distributed, high-performance gradient boosting framework based on the decision tree algorithm [4]. Unlike other boosting algorithms, it splits the tree leaf wise with the best fit.

- XGBoost

It is a decision tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework [4]. It provides a parallel tree boosting method.

- Neural Networks

Neural Networks consist of input and output layers, as well as a hidden layer(s) consisting of units that transform the input into something that the output layer can use. A layer is called a node which has neuron-like switches that turn on or off as input is fed to the network. Each unit has its own set of parameters, called the weight(w) and bias(b). In each iteration, the neuron calculates a weighted average of the values of the vector x , based on its current weight vector w and adds bias [5].

$$\check{Z} = w_1x_1 + w_2x_2 + w_3x_3 + \dots$$

$$Z = \check{Z} + b$$

$$\hat{Y} = g(z)$$

2.7 Calculations

Finally, the classification accuracy, specificity, sensitivity and confusion matrix were calculated.

$$Tp - \text{True Positive} \quad Fp - \text{False Positive}$$

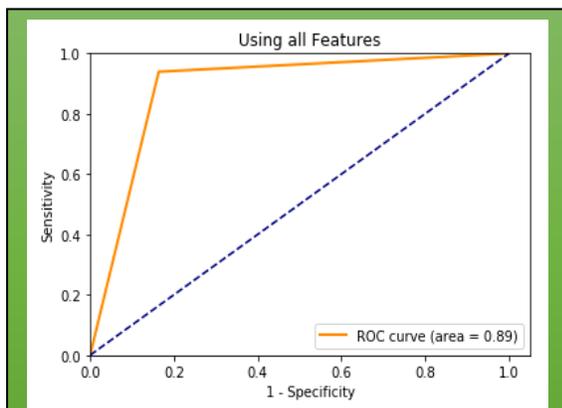
$$Tn - \text{True Negative} \quad Ffn - \text{False Negative}$$

$$\text{Accuracy} = \frac{Tp + Tn}{(Tn + Tp + Fp + Fn)}$$

$$\text{Specificity} = \frac{Tn}{(Tn + Fp)}$$

$$\text{Sensitivity} = \frac{Tp}{(Tp + Fn)}$$

3. RESULTS



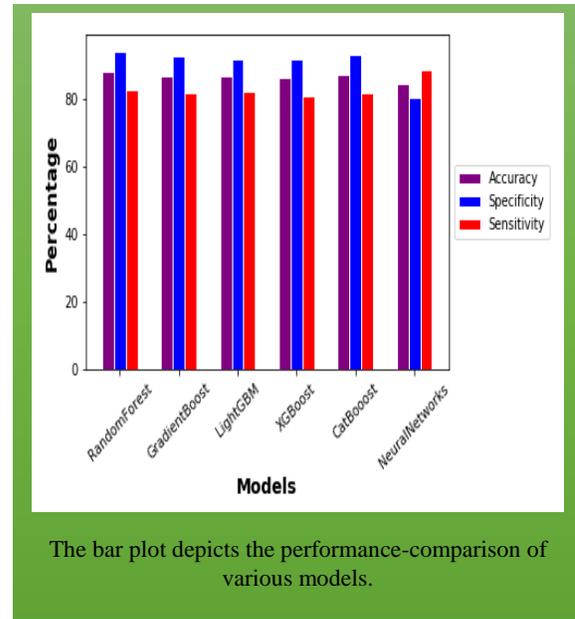
The graph shows the Region of Convergence (ROC) area of the final prediction (model – Random Forest).

We split our data into a training set, valid set and test set. Using the valid set, we tuned the hyper-parameters of the models. Against the test data, Random Forests acquired the highest accuracy score of 88.3% while Neural Networks obtained the highest sensitivity score of 87.2%.

5. CONCLUSION

Instead of opting for more complex examinations, a few blood tests would reveal the possibility of Alzheimer's with quite a good accuracy. However, there is still an utmost need for identification of relevant attributes for its early detection.

Our future work entails the use of Magnetic Resonance Imaging (MRI) scans and Positron Emission Tomography (PET) scans along with the Blood Biomarkers for classification of Alzheimer's.



The bar plot depicts the performance-comparison of various models.

6. ACKNOWLEDGEMENTS

The authors gratefully thank ACE Undergraduate Research Program for their support. We acknowledge the Birla Institute of Technology and BITMAA-NA (Alumni Association) for funding this program.

REFERENCES

1. Wu, Y. T., Beiser, A. S., Breteler, M. M., Fratiglioni, L., Helmer, C., Hendrie, H. C., & Matthews, F. E. (2017), "The changing prevalence and incidence of dementia over time [mdash] current evidence", *Nature Reviews Neurology*, 13(6): 327.
2. Jaeger, T.F. (2008), "Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and towards Logit Mixed Models". *Journal of Memory and Language*, 59, 434-446.
3. Breiman, L. (2001), "Random forests", *Machine learning*, 45(1): pp. 5-32.
4. Shenghui Yang, Haomin Zhang (2018), "Comparison of Several Data Mining Methods in Credit Card Default Prediction", *Intelligent Information Management*, pp. 115-122.
5. Anonymous, [A Beginner's guide to Neural Networks and Deep Learning](#).