Classification comparative analysis of machine learning algorithms using the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) data

Tomide Victor Afolabi
Afolabi-t@ulster.ac.uk
School of Computing, Engineering and Intelligent Systems
Ulster University
Londonderry, United Kingdom

Abstract— Machine learning algorithms are increasingly being used effectively in the medical field, particularly for diagnosis and prognosis. In this study, the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) dataset was used to classify Alzheimer's disease and mild cognitive impairment using four ML algorithms: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT). The dataset included 862 observations with heterogeneous variables that were adjusted. The performance of each ML algorithm was evaluated using metrics such as Area Under the Receiver Operating Characteristic (AUROC), confusion matrix (CM), Accuracy (ACC), and F1-score.

The results indicate that SVM and Random Forest achieved the highest accuracy with AUROC and ACC of 98% and 95%, respectively, followed by DT with 94% and LR with 92%. Additionally, RF had the highest correct classification rate for the AIBL dataset among the four with 100% and 95% in sensitivity and specificity, respectively.

In summary, the study concludes that Support Vector Machine (SVM) and Random Forest outperformed the other algorithms. The findings suggest that machine learning can be a useful tool for the diagnosis and prognosis of Alzheimer's disease and mild cognitive impairment.

Keywords—Machine learning, SMOTE, Mild cognitive imperative, Health control, Non- Healthy control, Feature selection, Performances evaluation.

I. INTRODUCTION

Machine learning (ML) is a rapidly growing field that utilizes predictive algorithms to extract general concepts from large datasets (1). In the healthcare industry, ML is increasingly gaining recognition for its potential to efficiently and accurately identify the underlying causes of illnesses using data. As a subset of artificial intelligence, machine learning has demonstrated significant potential in organizing, analyzing, and normalizing healthcare data, thereby enabling fast and precise decision-making. For example, custom machine-learning models have been shown to aid in precision diagnosis using genomic sequencing, early detection of cancer, and advanced cardiac visualization (2). Alzheimer's disease (AD) is a neurodegenerative disorder that affects millions of people worldwide. Accurate and early diagnosis of AD is critical for effective treatment and management of the disease. Machine learning algorithms have gained increasing attention as a tool for improving the accuracy of AD diagnosis. However, the performance of these algorithms in AD classification remains a subject of debate. In this study, we compare the performance of four machine learning algorithms - Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT).

In recent years, the use of classification algorithms, such as Support Vector Machines, Logistic Regression, and Artificial Neural Networks, has become increasingly popular for medical diagnosis. Researchers have also proposed novel feature selection algorithms to improve classification accuracy. For instance, Jian ping Li et al. (2020) utilized a fast conditional mutual information feature selection algorithm with SVM to diagnose heart disease with high accuracy (3). However, to date, no comparative analysis of machine learning algorithms has been conducted for the classification of MCI using the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) dataset.

The objective of this study is to utilize machine learning algorithms to determine the most effective method for classifying cognitive impairment disease by analyzing the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) data. The clinical diagnostic results will be adjusted into Health Control (HC) and Non-Healthy Control (Non-HC), which combines both Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD). The AIBL dataset is characterized by heterogeneity, particularly in relation to AD, which is the most prevalent form of dementia. To assess the strengths and weaknesses of each ML algorithm, the confusion matrix, Sensitivity, Specificity, Accuracy, and F1 score will be evaluated.

II. MATERIAL AND METHOD

The main objective of the study is to compare and evaluate different ML algorithms to classify diagnosis into HC (0) and Non-HC (1). Using Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL), this is a classification problem.

A. Data Description

The Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL) dataset is a large and diverse dataset that includes data from participants aged 55 years and 96 years.

The AIBL dataset is accessible for research and has played a significant role in the advancement of machine learning models used in identifying and diagnosing cognitive impairment diseases.

It contains a wide range of data types which include Demographic, Genotypes, Biomarkers, Cognitive assessments, and Clinical classification.

Category	Description
Demographics	1) age: 55~96 years 2) gender: Female/Male
Medical history	3) psychiatric (MH_PSYCH) 4) neurologic (MH_NEURL) 5) cardiovascular (MH_CARD) 6) hepatic (MH_HEPAT) 7) musculoskeletal (MH_MUSCL) 8) endocrine-metabolic (MH_ENDO) 9) gastrointestinal (MH_GAST) 10) renal-genitourinary (MH_RENA) 11) smoking (MH_SMOK) 12) malignancy (MH_MALI). Each medical history is a binary feature (i.e., Y/N)
ApoE genotypes	13) 2 alleles genotype. Each allele holds one of three genotypes: ε2, ε3, ε4
Neuropsychology assessments	14) clinical dementia rating (CDR). Five categories: healthy (0), very mild dementia (0.5), mild (1), moderate (2), and severe (3) 15) mini-mental state exam (MMSE): 0-30 4 categories: may be normal (30-25), mild/early (24-21), moderate (20-10), and severe (9-0) 16) the total number of story units recalled - logical memory immediate recall (LMIR): 0~25 17) the total number of story units recalled - logical memory delayed recall (LMDR): 0~25
Blood analyses	18) thyroid stim. Hormone (AXT117) 19) vitamin B12 (BAT126) 20) red blood cell (HMT3) 21) white blood cell (HMT7) 22) platelets (HMT13) 23) haemoglobin (HMT40) 24) mean corpuscular haemoglobin (HMT100) 25) mean corpuscular haemoglobin concentration (HMT102) 26) urea nitrogen (RCT6) 27) serum glucose (RCT11) 28) cholesterol (high performance) (RCT120) 29) creatinine (rate blanked) (RCT329)
Diagnosis	30) diagnostic results: 2 categories, i.e., healthy control (HC) and Non-Healthy(NHC)

Table 1 The AIBL data description

Initially, the dataset contained 862 observations and 31 features, with the target variable divided into three categories: Healthy Control (HC), Mild Cognitive Impairment (MCI), and Alzheimer's disease (AD). However, for this study, the target variable was recategorized into two classes - Healthy Control (HC) represented by 0, and Non-Healthy Control (Non-HC) represented by 1, which included both MCI and AD. After preprocessing, the dataset was reduced to 570 observations and 26 features.

B Experimental Stages::

The Python environment was employed for the research study.

1. Stage 1 - Data Processing and Exploration

During the preprocessing stage, measures were taken to ensure that the data was fit for the research. for example, the correction of incorrectly spelt variables, such as gender. Min-Max normalization was applied to the independent variables to make their values contain a common range. Data visualization, such as a box plot, helped in identifying outliers. Outliers were properly managed by converting them to a NAN (not a number) data type and subsequently replacing the NAN with the mean value of the outliers. As the study focused on healthcare, negative values were eliminated since they could have an adverse effect on the outcome of the results. Furthermore, to address the imbalance in the data, we utilized SMOTE (Synthetic Minority Oversampling Technique).

a) Synthetic Minority Oversampling Technique (SMOTE): is a technique employed for augmenting the minority class of the data distribution (4). SMOTE redistributes data by randomly feeding the minority class with more instances. This method was used to balance leaf data.

b) Hold-out: is the process of splitting the data into training and testing datasets. The study made use of 70% of the data as a

training set and 30% as a testing set. The model trained on what is known as seen dataset and evaluated itself with the unseen dataset.

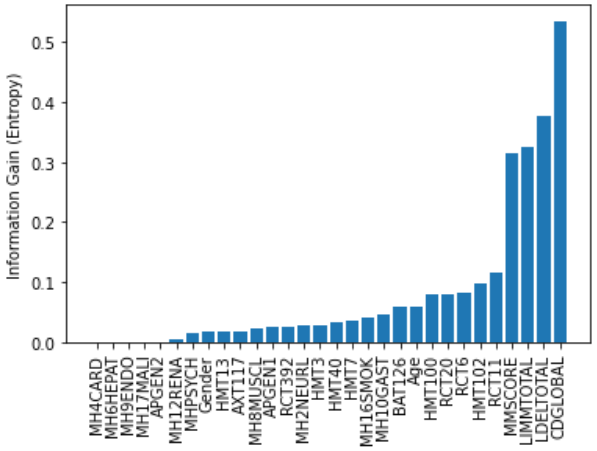

Figure 1 – Entropy-based method

2. Stage 2 – Features selection

The study utilized an entropy-based method for feature selection.The significance of all explanatory variables was plotted against the target variable in Figure 1. According to information gained from the entropy-based method, five variables (MH6HEPAT, MH9ENDO, MH4CARD, MH17MALI, APGEN2) that did not contribute to the target variable were removed. The remaining data was divided into a training set comprising 399 observations and a testing set comprising 171 observations.

C Methods

1. Machine learning (ML) is a field of artificial intelligence that enables computers to learn and improve their performance by analyzing data and identifying patterns(5). In essence, ML involves feeding data into an algorithm and allowing it to learn how to carry out a specific task. four machine learning algorithms - Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT) were used.

a) Logistic Regression (LG) is a classification algorithm that predicts a binary outcome using independent variables. It calculates target values as probabilities and is based on the cost function (logistic function), which has an S-shaped curve (6).

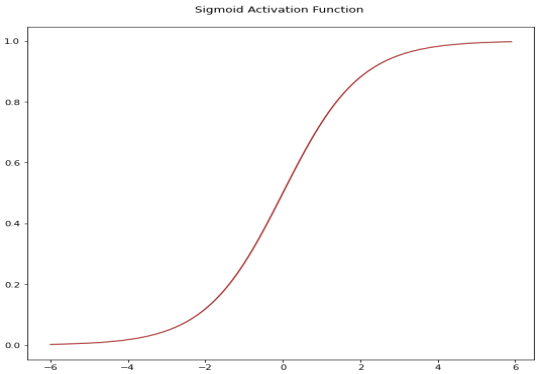

Figure 2 – Sigmoid function for binary logistic regression

In the case of multiple outcomes, it is called multi-logistic regression. This model is used to illustrate the relationship between the target variable and the independent variables. Equation 1 shows the hypothesis representation of the output.

$$\text{Want } 0 \le h_\theta(x) \le 1 \quad 0 \le h_\theta(x) \le 1, \text{ where}$$
$$h_\theta(x) = g(\theta^T x), \text{ and } g(z) = \frac{1}{1+e^{-z}} \qquad (1)$$
$$i.e \quad h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

The study focused on binary outcomes where Diagnosis will always be 1 (Non-HC) or 0 (HC) for every value of independent variables.

b) Decision Tree (DT) is like an ordered tree that is similar to a flowchart, consisting of three basic elements, namely: nodes that portray attributes, branches that represent attribute values and leaves that consist of objects that normally belong to the same class (7). It is a supervised learning model that uses discrete values of the target variable for classification.

c) Random forest: is a method introduced by Leo Breiman in the 2000s for building a predictive model using an ensemble of decision trees that increase in random samples of the data (8). It utilizes ensemble learning, which involves generating and combining multiple classifiers or models to solve a problem. The random forest comprises several decision trees and is also referred to as a random decision tree algorithm.

d) Support vector machine (SVM), also known as support vector networks, is a classification algorithm developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. It belongs to the generalized linear classification (9). SVM uses hyperplanes and support vectors to classify data into different categories. If the data points are not linearly separable, SVM transforms them to a higher dimension to separate them. Support vectors are data points that are near the decision boundary or hyperplanes. The location of the hyperplanes depends on the position of support vectors. Hyperplane is the decision boundary line that separates data points. The objective of SVM is to find the optimal hyperplane, and it does this through positive and negative hyperplanes. The line between negative and positive hyperplanes is known as the maximum margin hyperplane. In higher dimensions, such as three dimensions, a hyperplane is called a plane.

2. Performance Evaluation

a. Confusion Matrix (CM): also known as a matching matrix in unsupervised learning, is a table that displays the various outcomes of predictions and results in a classification problem, providing a visual representation of the outcomes (10). Each row of the matrix represents the number of instances in a target

variable class. Evaluation metrics that can be derived from the confusion matrix include

i. Condition positive which shows the number of real positives in the data.

ii. Condition negative which shows the number of real negatives in the data.

iii. True positive (TP): shows the result values of variables that really demonstrate the presence of a condition. In this case, TP is a percentage of people who are truly HC (0). And it is calculated through a true positive rate or recall.

TPR = TP/ (TP+FN). Where FN is False Negative.

iv. True negative (TN): shows the result values of variables that truly demonstrate the absence of a condition. It is a percentage of people who are truly Non-HC (1). To calculate it, TNR or specificity is used:

TNP = TN/FP+TN. where FP is False Positive

b. Accuracy: is the percentage of all observations that correctly predict the target variable Diagnosis. It is calculated by:

ACC = TR + TN/ (Total examples)

c. F1-score: is an evaluation metric used in machine learning to measure the overall accuracy of a model. It considers both the precision and recall scores of the model (11). It is calculated by

(2 * precision * Recall) / (Precision + Recall).

d. AUR0C(Area Under the Receiver Operating Characteristic): this is a metric that indicates how well a model can rank the observations in the data using the ROC curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR). A higher AUROC value indicates that the classifier is better at accurately ranking the observations, while a lower AUROC value indicates a poorer classifier.

III. EXPERIMENT RESULTS

The final data consisted of 570 observations and 26 features after data preprocessing. It was later split using a split ratio of 7:3 that is. 70% for the training dataset and 30% for the testing dataset. The data at the baseline was not used for the study because it contained multiple features and needed cleaning.

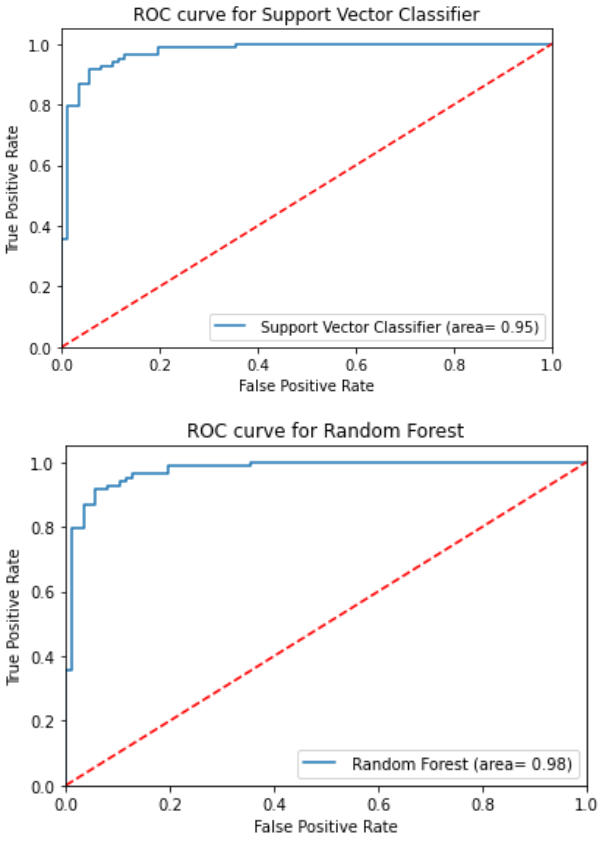

Figure 3 – A ROC curve for the most efficient algorithms

Algorithms	Accuracy	AUROC	Sensitivity	Specificity	F1 score
LG	0.9240	0.9240	0.9286	0.9195	0.92
RF	0.9766	0.9770	1.0	0.9540	0.98
SVM	0.9532	0.9532	0.9524	0.9540	0.95
DT	0.9415	0.9419	0.9643	0.9195	0.94

Table 2. Performance results of all ML algorithms

DISCUSSION AND CONCLUSION

According to the importance gain plot presented in Figure 1, it was observed that removing significant variables such as CDGLOBAL would significantly impact the performance of the models since these variables have a strong influence on the Diagnosis outcome.

The research used four ML algorithms to improve the diagnosis of cognitive impairment disease, and all of them achieved satisfactory results when the threshold for various performance evaluation strategies was set to 0.92. However, when ranked, Random Forest performed the best, followed by Support Vector Machine with a linear kernel and Decision Tree. The performance evaluation metric was above 0.94.

SVM's outstanding performance can be attributed to the dimension of the data, as the number of features is much larger than the number of observations.

In conclusion, the study recommends modeling and evaluating multiple ML algorithms to improve healthcare decision-making, such as the diagnosis of cognitive impairment disease.

V RECOMMENDATIONS

The performance of Logistic Regression was not up to expectations. In situations where the number of features is large compared to the number of observations, it is recommended to use either Logistic Regression or SVM with a linear kernel. Further investigation can be done to determine the reason for this suboptimal performance.
Using other ML algorithms and understanding their performance in binary classification can help gain a better understanding of data science problems.

VI REFERENCES

[1] Douple et al, (2019). "Machine Learning for Health Services Researchers". Value of health, volume 22, issue 7, pages 808-815.

[2] Machine Learning for Healthcare & Life Sciences, https://aws.amazon.com/health/machine-learning/?trk=2ea414ad-9f4f-4832-9ae8-e5e80a782073&sc_channel=ps&ef_id=Cj0KCQjwlumhBhClARIsABO6p-xN7bDAjRFZyqW9l7Waz03Z1rxbXKpELPQ1_rydZ3zifvUd7QVby3MaApujEALw_wcB:G:s&s_kwcid=AL!4422!3!581117978379!p!!g!!machine%20learning%20in%20medicine!16161826769!131827883423

[3] Jian Ping Li et al (2020). "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare". Pages 107562 – 107582. [online]. https://ieeexplore.ieee.org/abstract/document/9112202

[4]https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[5]https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/

[6]https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

[7] Ilyes Jenhani et al, (2008), "Decision trees as possibilistic classifiers" [online]. Decision trees as possibilistic classifiers - ScienceDirect

[8] Gerard Biau, (2012), "Analysis of a Random Forests Model". [online] biau12a.dvi (jmlr.org)

[9] Durgesh K. Srivastava, Lekha Bhambhu (2010), "Data Classification Using Support Vector Machine" [online] (PDF) Data classification using support vector machine (researchgate.net)

[10] https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning#:~:text=A%20confusion%20matrix%20presents%20a,actual%20values%20of%20a%20classifier.

[11]https://www.v7labs.com/blog/f1-score-guide#:~:text=F1%20score%20is%20a%20machine%20learning%20evaluation%20metric%20that%20measures,prediction%20across%20the%20entire%20dataset