Prediction of Alzheimer’s disease (AD) Using Machine Learning Techniques with Boruta Algorithm as Feature Selection Method

Lee Kuok Leong¹ and Azian Azamimi Abdullah¹*

¹ Biomedical Electronic Engineering Programme, School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Pauh Putra Campus, 02600 Arau, Perlis Malaysia.

*azamimi@unimap.edu.my

Abstract. Alzheimer’s disease (AD) is the most frequent incurable neurodegenerative disease, a general term for memory loss and other cognitive abilities. Early detection of AD can help with proper treatment and prevent brain tissue damage. Traditional medical tests are time consuming, fail to recognize early signs and lack of diagnosis sensitivity and specificity. To achieve promising prediction accuracy, the best predictive machine learning model is selected based on initial pre-processing step followed by vital attributes selection and performance evaluation for five proposed supervised machine learning algorithms. In the pre-processing, all the missing values have been removed since the overall percentage only covered 5.63%. Boruta algorithm as feature selection method resulted Atlas Scaling Factor, Estimated Total Intracranial Volume, Normalized Whole-brain Volume, Mini-Mental State Examination and Clinical Dementia Rating must be included as primary features. With Boruta algorithm, it has been shown that Random Forest Grid Search Cross Validation (RF GSCV) outperformed with 94.39% of accuracy, 88.24% sensitivity, 100.00% specificity and 94.44% AUC among other 12 models that includes conventional and fine-tuned models even for the small OASIS-2 longitudinal MRI dataset. Finally, our developed Graphical User Interface (GUI) prediction tool was evaluated through prediction over OASIS-1 cross-sectional MRI dataset containing 216 samples of imaging sessions that have been pre-processed. Prediction results were closed with the dementia status provided in OASIS cross-sectional data fact sheet.

1. Introduction
Alzheimer’s disease (AD), is a type of dementia which is incurable and affecting millions of people worldwide [1], characterized as progressive neuropsychological disorder that have a dramatic consequences on thinking ability and behaviour functions which greatly impact the quality of patient’s life with AD and undergo dramatic personality changes [2]. The current issue mainly highlighted that the detection of the AD is still not accurate until the patient reaches the sensible stage of AD. Thus, the accuracy of the early stage diagnosis will directly relate the validation of neuropathological, which should detect the fundamental characteristics neuropathology changes before to the onset of clinical symptoms where with the sensitivity and specificity of no less than 80% [3].

The relationship between combinations of features with the accuracy of prediction model have been concerned in this recent years specifically by Schouten et al. [4], presented a combination of different neuropathological features to prove be a method for improving the model performance. Machine learning techniques have been extensively developed with the aim of automatically
classifying AD, producing encouraging results (e.g., Sarraf et al. [5]; Lebedev et al. [6]; Deepika Bansal et al. [7]; Amoroso et al. [8]). This paper mainly focused on detailed study of binary classification (demented or non-demented) and diagnosis of AD which includes initial pre-processing step and conduct of the imperative attributes selection by Boruta algorithm, evaluation of different classifiers and decide the best classifier for prediction on Open Access Series of Imaging Studies (OASIS) imaging datasets. Finally, a high degree of accuracy prediction tool that can help clinicians to evaluate and classify AD with normal individuals in early stage using MRI features based data.

2. Methodology
An overview of the research workflow for this project is shown in Figure 1 below.

![Image of research workflow](image_url)

**Figure 1.** Research workflow

2.1. Data Acquisition
The MRI related data was generated by the Open Access Series of Imaging Studies (OASIS) project. The released data for OASIS-1 cross sectional and OASIS-2 longitudinal MRI datasets were utilized for training and testing with various machine learning models. The subjects from longitudinal dataset include both men (n=62) and women (n=88) with all right-handed dominance and for each subject, at least three or four acquired T1-weighted MRI scan brain imaging images took in single MRI scan session. For a total of 373 imaging sessions in the longitudinal dataset, each subject was scanned on two or more separate instances with separated by at least one year, with an average delay of 719 days (standard delay period within 183 to 1707 days) between visits. Table 1 shows all the features that included in the longitudinal data set.

| Feature | Description |
|---------|-------------|
| Age | Age at time of image acquisition (years). |
| MR Delay | Delayed T1 post contrast |
| Gender | Gender (M or F) |
| Education | Years of education |
| SES | Socioeconomic status classified into categories from 1 (highest status) to 5 (lowest status) |
| MMSE | Mini-Mental State Examination score (range is from 0 [worst] to 30 [best]) |
| CDR | Clinical Dementia Rating. (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD) |
| ASF | Atlas scaling factor (unit less). Computed scaling factor that transforms native-space brain and skull to the atlas target. |
| eTIV | Estimated total intracranial volume (cm³) |
| nWBV | Normalized whole brain volume, using automated tissue segmentation process to expressed as a percent of all voxels in the atlas-masked image that are labelled as grey or white matter |

Table 1. Feature measures counted in the dataset
2.2. Data Pre-processing

It can be summarized that there are few preferable alternatives that can easily handle the missing values such as (i) remove the data row or column, (ii) use a global constant to fill in for missing values and (iii) use feature mean (or median/ mode/ mean) for all samples fitting to the same class depending on the percentages of missing values in the datasets.

2.3. Boruta Algorithm as Feature Selection Method

2.4. In this paper, Boruta algorithm [9] is used as feature selection method, which is able to select optimal features carrying usable information for model’s prediction. Boruta algorithm basically works with Scikit-learn dependency which act as interface medium on Random Forest (RF) classifier which utilizes the importance measure generated by the original algorithm. It is based on the implementation of Python’s random forest which runs on a single code to identify all the importance features in the dataset with respect to an outcome variable (categorization of demented or non-demented). In this paper, number of estimators is set to be ‘auto’ since Boruta algorithm in Python offers automatic number of estimator’s selection instead of finding a possibly compact subset of features in dataset since it’s applicable to all type of features. To achieve the minimal optimal set of features, Boruta algorithm is used with the following non-default parameters: depth of each tree in the forest is set be 5 since recommended depth in Boruta algorithm is using pruned trees with a depth between 3 to 7 and verbose is set be 2 to controls the verbosity of output. Other than that, parameters such as percentile (perc), alpha, two step correction for multiple testing (two_step) and maximum iterations (max_iter) are set be default where perc=100, alpha=0.05, two_step=True and max_iter=100. However, for the attribute’s decision, support and ranking parameters are executing to mask the confirmed selected features as True and assign the features ranking correspond to the ranking position of ith of features (e.g. the best features assign as rank 1 and tentative features as rank 2 etc) respectively. The coded Boruta programme with pre-processed OASIS-2 longitudinal dataset is running over Intel Core i5-7200U 2.5GHz with Turbo Boost up to 3.1GHz and NVIDIA GeForce 940MX with 2GB dedicated VRAM installed with 120GB 530MB/s Solid State Drive.

In conclusion, Boruta algorithm approach is highly recommended for the exploration of high-dimensional data. The formula for model training before the feature’s selection is: Group ~ Gender + Age + Visit + Hand + Educ + MR Delay + SES + CDR + MMSE + eTIV + nWBV.

2.5. Supervised Machine Learning Method

2.5.1. Deep Neural Network (DNN)

DNN model is created using Tensorflow. The model will have 11 and 5 fully connected hidden layer with the same number of neurons as input variables for two models, without feature selection (11 features) and with Boruta feature selection (5 features). Thus, an input data is being trained by DNN with difference hidden layer and activation function. The proposed of two DNN models are trained over 100 training epochs with batch size of 10. All the parameters to sett up DNN models are summarized as shown in Table 2. Finally, binary cross-entropy (logarithmic loss function) is used during training, the ideal loss function for binary AD classification.
Table 2. Proposed default estimator for each Deep Neural Network

| DNN | Feature Selection | Activation Function | With Boruta algorithm | Without Boruta algorithm |
|-----|-------------------|---------------------|-----------------------|-------------------------|
|     |                   |                     | Input layer: ReLU     | Input layer: ReLU        |
|     |                   |                     | Hidden layer: ReLU    | Hidden layer: ReLU       |
|     |                   |                     | Output layer: Sigmoid | Output layer: Sigmoid    |
|     | Hidden Layer      | 11                  | 5                     |
|     | Hidden Neuron     | None, 11            | None, 5               |
|     | Epochs            | 100                 | 100                   |
|     | Input Dropout Ratio | 0.1                | 0.1                   |
|     | Hidden Dropout Ratio | 0.1                | 0.1                   |

2.5.2. Random Forest (RF)

RF can work well with the Boruta algorithm, as selection is based on the feature’s importance. Two types of cross validation techniques are used such as grid search cross validation (GSCV) and randomized search cross validation (RSCV) to search and obtain the best combination of hyperparameters to train RF models. The pre-defined arrays of ideal values for the grid parameters are set to search, specifically declared as Table 3.

Table 3. Set of hyperparameters for RF, RF GSCV and RF RSCV

| Grid Parameters | RF (Default) | RF GSCV | RF RSCV |
|-----------------|--------------|---------|---------|
| max_depth       | None         | [2, 4, 6, 8, 10, 12, 14] | [2, 4, 6, 9, 11, 14, None] |
| max_features    | "Auto"       | Without Boruta algorithm: [11] | Without Boruta algorithm: [11] |
| min_samples_split | 2          | [6, 8, 10, 12] | [6, 8, 10, 12] |
| min_samples_leaf | 1           | [2, 4, 6, 8, 10] | [2, 4, 6, 8, 10] |
| bootstrap       | True         | [True, False] | [True, False] |
| criterion       | "gini"       | ["gini", "entropy"] | ["gini", "entropy"] |
| n_estimators    | 10           | [1, 2, 4, 8, 16, 32, 64] | [1, 2, 4, 8, 16, 32, 64, 100, 200] |

2.5.3. Gradient Boosting Machines (GBM)

The most crucial parameters used to define a tree such as max_depth and min_samples_split is tuned and maintained the learning rate as default (learning rate= 0.1). Max_depth boundaries the maximum depth of the tree; deeper the tree, more data information is captured and min_samples_split sets the minimum number of samples to split. Table 4 below shows the initial estimates for boosting parameters and range of tree-based tuned parameters which will do a grid search.

Table 4. Initial estimates and range for tuning tree-based parameters

| Tune Parameters | GBM (Default) | GBM PT |
|-----------------|---------------|--------|
| max_depth       | 3             | [4, 6, 8, 10, 12, 14, 16] | [200, 400, 600, 800, 1000] |
| min_samples_split | 2           | [4, 6, 8, 10, 12, 14, 16] | [200, 400, 600, 800, 1000] |
| Default estimators |             | • learning_rate= 0.1, | • learning_rate= 0.1, |
|                  |              | • n_estimators= 100, | • n_estimators= 100, |
|                  |              | • max_features= None, | • max_features= “auto”, |
|                  |              | • subsample=1.0     | • subsample=1.0     |
2.5.4. Support Vector Machine (SVM)

As shown in Table 5, three types of SVM models have been proposed with different parameters for linear kernel (LK), radial basis function (rbf) kernel (RBFK), GSCV method and set the C parameter at different values which suite for the function optimization. The GSCV model is proposed to explore combinations of parameters of kernel, C, gamma, decision_function_shape, degree, coef0 and shrinking to minimize the misclassifications.

Table 5. Default parameters and tune parameter values of proposed SVM models

| Grid Parameters                | SVM LK  | SVM RBFK | SVM GSCV                      |
|-------------------------------|---------|----------|-------------------------------|
| kernel                        | “linear”| “rbf”    | ['linear', 'poly', 'sigmoid']|
| C                             | 1       | 1        | [0.0001, 0.001, 0.01, 0.1, 1, 10]|
| gamma                         | ‘auto’  | 0.01     | [0.0001, 0.001, 0.01, 0.1, 1]|
| decision_function_shape       | ‘ovr’   | ‘ovr’    | [‘ovo’, ‘ovr’, None]         |
| degree                        | 3       | 3        | [2, 4]                        |
| coef0                         | 0       | 0        | [0, 1, 2, 3]                  |
| shrinking                     | True    | True     | [True, False]                 |

2.5.5. Logistics Regression (LR)

Two LR models were proposed where the first model is designed with default parameters as shown in Table 6. The main purpose to find the optimal C in the hyperparameter space is to control underfitting and overfitting of the model since large value of C can lead to overfitting. And, the second model is mainly target on tuning of parameter C as the regularization parameter or regularization strength.

Table 6. Default parameters and tune parameter values of proposed LR models

| Grid Parameters                | LR (Default)                  | LR PT                             |
|-------------------------------|-------------------------------|-----------------------------------|
| C                             | 1.0                           | [0.00001, 0.0001, 0.001, 0.1, 1, 10, 100] |
| Default estimators            | penalty=’l2’,                 | n_splits=10,                      |
|                               | max_iter=100,                 | shuffle=True,                     |
|                               | multi_class=’ovr’,            | random_state=42                   |
|                               | max_iter=100,                 |                                   |
|                               | random_state=None             |                                   |

2.6. Performance Evaluation and Comparison

The classification results were calculated by means of three metric measurements, which are used for quantitative valuation and evaluation, including accuracy, sensitivity (recall) and specificity. Additionally, numerous optimal approaches such as receiver operating curve (ROC) and Area Under the Curve (AUC) are calculated as well.

Fundamentally, the binary classification will produce four outcomes such as True positive (TP), true negative (TN), false positive (FP) and false negative (FN). Terms associated with confusion matrix:

1. True positive (TP): Correct positive prediction
The case where the patient is actually having dementia (1) and the prediction model classified the patient as demented (1) comes under True Positives.

2. True negative (TN): Correct negative prediction

The case where the patient is actually NOT having dementia (0) and the prediction model classified the patient as non-demented (0) comes under True Negatives.

3. False positive (FP): Incorrect positive prediction

The case where the patient is actually NOT having dementia (0) and the prediction model classified the patient as demented (1) comes under False Positives. False (model has predicted incorrectly) and Positive (class predicted was a positive (1))

4. False negative (FN): Incorrect negative prediction

The case where the patient is actually having dementia (1) and the prediction model classified the patient as non-demented (0) comes under False Negatives.

3. Results and Discussion

In this study, the missing values were removed by dropping the 19 rows and 2 rows with missing values in the column of SES and MMSE respectively since in overall, only 5.63% of the values in the dataset were missing. Since the prediction of dementia status are totally relies on complex neuroimaging data, thus replace missing values with sensible values (constant values or column’ mean or median value) might affect the accuracy of the dementia classification.

With Boruta algorithm, first five with the highest importance features were selected such as MMSE, CDR, eTIV, nWBV and ASF for the training and testing model later. The formula for model training with the feature’s selection is: Group ~ SES + CDR + MMSE + nWBV. With the implementation of Boruta algorithm in the proposed work, 12 proposed machine learning algorithms performed better which shown with higher accuracies, sensitivities, specificities and AUCs as compare that without the feature selection. This is a positive sign for Boruta algorithm, which is suitable for OASIS-2 longitudinal MRI dataset to provide a combination set of features as optimal features for classification. RF GSCV with features selection is showing the best classification performance with accuracy of 94.39%, 88.24% sensitivity, 100.00% specificity and 94.44% AUC but without applying attribute selection, RF GSCV algorithm only perform with accuracy of 91.59%, 85.42% sensitivity, 96.61% specificity and 92.21% AUC as shown in Table 7. High sensitivity and specificity can promise high true positive and true negative rates respectively. And, in case of neurodegenerative diseases it is important to have a high true positive rate (sensitivity) so that all patients with Alzheimer's can identified as early as possible. When compared the experimental results with other recent classification experiments (binary or multi-class classifications) done with OASIS datasets, the current experimental results have shown a good sign in terms of classification accuracy, especially proposed RF GSCV algorithm with 94.39% and 94.44% of accuracy and AUC respectively for binary AD/HC (demented/non-demented) classification, based on MRI data [10]–[12].

| Model   | Feature Selection | ACC (%) | SEN (%) | SPE (%) | AUC (%) |
|---------|------------------|---------|---------|---------|---------|
| RF GSCV | No               | 91.59   | 85.42   | 96.61   | 92.21   |
|         | Proposed method  | 94.39   | 88.24   | 100.00  | 94.44   |
RF GSCV model works best with OASIS-2 longitudinal dataset and the fine-tuned parameters are preserved for the predictive model in the GUI prediction tool. Knowing that the performance of best prediction algorithm is based on 94.39% of accuracy rather than percentage lower than 90%, thus it is persuasive that can give more evidence and confidence for making the AD classification diagnosis. Whereas, with the developed prediction tool, the classification results were presented with 83 (38.43%) predicted as demented samples, while 133 (61.57%) as non-demented samples. The predicted subjects with demented were closely to the statement in OASIS fact sheet (rev. 2007-8-20) provided by Marcus et al. [13]. Lastly, three of the subjects with different classes were randomly selected from the OASIS-2 longitudinal MRI dataset for the prediction validation purpose. Five input values (MMSE, nWBV, CDR, eTIV and ASF) were manually key-in to the entries of proposed GUI prediction tool and the predicted classes of the patient were correctly matched with the expected status of patient as shown in Table 8.

| Subject ID | MRI ID   | Group         | MMSE | nWBV | CDR | eTIV | ASF     | Prediction  |
|------------|----------|---------------|------|------|-----|------|---------|-------------|
| OAS2_0001  | OAS2_0001_MR1 | Non-demented  | 27   | 0.696| 0   | 1987 | 0.883   | Non-demented|
| OAS2_0002  | OAS2_0002_MR1 | Demented      | 23   | 0.736| 0.5 | 1678 | 1.046   | Demented    |
| OAS2_0018  | OAS2_0018_MR1 | Converted     | 30   | 0.715| 0   | 1406 | 1.248   | Demented    |

4. Conclusion
This study is based on the comparison and evaluation of different machine learning algorithms with pre-processing and Boruta algorithm features selection in the prognosis and prediction of AD. In the proposed method, the missing values were cleansed by columns deletion in the column of SES and MMSE respectively. The experimental results validate that a combination of pre-processing and experimental results with Boruta features selection is a trustworthy technique for early prediction of AD. RF GSCV with Boruta algorithm is the best classifier outperformed among all proposed algorithms, with the accuracy of 94.39%. RF model has better feature propagation and provides better classification result even for the small dataset. The prediction results from the developed GUI prediction tool provided the prediction on the status of dementia based on the available patient information, including demographics, clinical and MRI derived anatomic volume features able to cover the limitations pointed out from the previous researches.

References
[1] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, ‘The global prevalence of dementia: A systematic review and metaanalysis’, Alzheimer’s and Dementia. 2013.
[2] C. Dillon, C. M. Serrano, D. Castro, P. P. Leguizamón, S. L. Heisecke, and F. E. Taragano, ‘Behavioral symptoms related to cognitive impairment.’, Neuropsychiatr. Dis. Treat., vol. 9, pp. 1443–55, 2013.
[3] S. Teipel et al., ‘Multimodal imaging in Alzheimer’s disease: Validity and usefulness for early detection’, The Lancet Neurology. 2015.
[4] T. M. Schouten et al., ‘Combining anatomical, diffusion, and resting state functional magnetic
resonance imaging for individual classification of mild and moderate Alzheimer’s disease’, *NeuroImage Clin.*, 2016.

[5] S. Sarraf, G. Tofighi, and for the A. D. N. Initiativ, ‘DeepAD: Alzheimer’s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI’, *bioRxiv*, 2016.

[6] A. V. Lebedev *et al.*, ‘Random Forest ensembles for detection and prediction of Alzheimer’s disease with a good between-cohort robustness’, *NeuroImage Clin.*, vol. 6, pp. 115–125, 2014.

[7] A. Khan and M. Usman, ‘Early Diagnosis of Alzheimer’s Disease using Machine Learning Techniques’, *2015 7th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag. (IC3K)*, 2015.

[8] N. Amoroso *et al.*, ‘Deep learning reveals Alzheimer’s disease onset in MCI subjects: Results from an international challenge’, *J. Neurosci. Methods*, 2018.

[9] M. B. Kursa and W. R. Rudnicki, ‘Feature Selection with the Boruta Package’, *J. Stat. Softw.*, 2015.

[10] G. Awate, S. Bangare, G. Pradeepini, and S. Patil, ‘Detection of Alzheimer’s Disease from MRI using Convolutional Neural Network with Tensorflow’, Jun. 2018.

[11] J. Islam and Y. Zhang, ‘A Novel Deep Learning Based Multi-class Classification Method for Alzheimer’s Disease Detection Using Brain MRI Data’, 2017, pp. 213–222.

[12] J. Islam and Y. Zhang, ‘Brain MRI analysis for Alzheimer’s disease diagnosis using an ensemble system of deep convolutional neural networks’, *Brain Informatics*, vol. 5, no. 2, p. 2, May 2018.

[13] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, ‘Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults’, *J. Cogn. Neurosci.*, 2007.