Applied Machine Learning Project cover sheet

Please complete all sections below and attach the completed coversheet to the front of your electronic assignment before submission:

Student Number: 199110799 (as shown on student ID card)
Programme: Data Science Degree Apprenticeship
Module Tutor: Uche Onyekpe
Module Code: DSC5007M
Module Title: Applied Machine Learning
Assignment Title: Assessing the Performance of Classification Techniques in Predicting the Onset of Dementia
Word Count: 2994 (3000 Limit)

Declaration of Academic Integrity

Please complete before submitting your assignment:

x

By entering an 'x' in the box above, I confirm that I have read and understood the University regulations on cheating and plagiarism and that the work submitted is my own within the meaning of the regulations.

Est. 1841 | YORK ST JOHN UNIVERSITY

Assessing the Performance of Classification Techniques in Predicting the Onset of Dementia

Applied Machine Learning Project

Monday 30th May 2022

Table of Contents

Abstract

This report examines the efficacy of using machine learning techniques in the early detection of dementia. Dementia is a debilitating disease that results in a significant decline in cognitive function. For this study we used a freely available data set provided by Open Access Series of Imaging Studies (OASIS-2) and it is a longitudinal collection of 373 MRI scans across 150 right-handed subjects. After some pre-processing of the data including the correction of class imbalance using SMOTE, we found the Random Forest model to produce the best results with AUC scores between 0.99 and 1, and precision and recall scores all above 93%.

1 Introduction

Dementia is a debilitating disease that presents as a decline in the cognitive function of a patient above and beyond the normal decline experienced with aging. Currently diagnosing dementia requires extensive testing and evaluation of how a patient performs on various cognitive and functional assessments which are carried out across a series of consultations with a professional (Moreira and Namen, 2018). The difficulty in early detection is compounded by the lack of a suitable assessment that could establish the type of dementia (Duboisa *et al.*, 2016). The present study looks to appraise some of the previous research done in this area. Specifically, we shall be examining some previous research on a particular data set to see if we can further improve upon the results obtained using a wider range of features and machine learning techniques.

2 Literature Review

In a 2019 paper, Battineni et al used a support vector machine classifier on a longitudinal collection of MRI scans collected from the Open Access Series of Imaging Studies (Battineni, Chintalapudi and Amenta, 2019). The sample included 150 right-handed men and women between the ages of 60 and 96 years old. Subjects attended at least two MR sessions separated by at least one year for a total of 373 MR sessions. Overall, the subjects were divided into three categories: demented, non-demented, and converted. The key attributes they chose were Mini-Mental State Examination, Clinical Dementia Ratio, MR delay – the time between MRI scans measured in days - and normalized Whole Brain Volume. They purposely didn't use other demographic values like Gender, Social Economic Status, education level, and Atlas Scaling Factor as it was felt these wouldn't be useful predictors; this is in addition to the fact that the use of too many predictors might negatively affect the performance of the model. They filled in missing values with the average of that feature. In terms of their results, they found that the SVM model had an accuracy of about 70% with the recall and sensitivity values ranging between 65-82% depending on the which category the subject was in. This is the same data set that will be used in my own study and so it will be interesting to later compare their approach with mine and see where the results differed.

In a meta-analysis of ML techniques used to detect dementia, Pellegrini et al found that all models used showed severe limitations (Pellegrini *et al.*, 2018). For example, the over reliance on a single data set meant that the results weren't generalisable; data was generally collected from populations with a higher number of cases compared to control individuals; finally, only a single ML algorithm is employed. They also found that whilst ML models were quite successful at differentiating between subjects with and without dementia, they had more trouble determining if an individual was at a high risk of developing the disease.

In a similar literature review, Goerdten et al reviewed 137 studies looking at models predicting dementia risk (Goerdten *et al.*, 2019). Machine Learning was used for 55 of those models. The most commonly used algorithms were SVM (n=17), Disease State Index (n=6), and Random Forest (n=5). Logistic Regression was also used. Across the entire sample of studies, the most commonly used prediction model approaches were machine learning, cox regression, and linear regression. They had similar findings to Pellegrini et al and found that over 60% of the studies reviewed used the same data source and only 12.7% investigated the predictive power of the models in non-affected populations. Overall, Goerdten et al found that the major drawbacks of the previous studies were the over reliance on just a single data source, a general lack of evaluating the assumptions underlying the models employed, and little to no internal or external validation of their prediction models (Goerdten *et al.*, 2019).

To address the use of very few data sources, they suggested the use of a broader number of data sources and more diverse samples when building predictive models for dementia. They further made the point that by discriminating the subjects according to their type of dementia that one would be better able to identify the risk factors specific to each type of dementia. The inclusion of various follow-up periods will also allow the tracking of the disease's progression over multiple time frames. They also emphasised the need for internal and external validation methods to be employed to check for overfitting - and therefore its predictive ability is too optimistic - as well as to check its performance in a comparable population. They specifically recommended bootstrapping the data when a large number of variables are used for the prediction. For external validation, a similar but different population of subjects would be required.

As described above, there are numerous difficulties in trying to use ML techniques in determining whether a patient has dementia. These are a lack of diversity in the data sources employed, little to no evaluation of the limitations of the models used so far, as well as a lack of proper validation methods being employed. The first of these difficulties will no doubt be exhibited in the present study given the small size of the data set which was also used by one of the other papers. An attempt will be made to see if it’s possible to find a larger overall data set to work with – or potentially to find other smaller datasets and see if there’s a possibility of combining the data into a larger overall set. Keen attention will also be paid to ensuring that the assumptions of the models used as well as proper validation of results is employed.

3 Problem and Dataset Description

3.1 Dataset Description

As stated in the literature review above, this study will be using the same data set as that used in the Battineni study which is freely available and can be found supplemented to their article (Battineni, Chintalapudi and Amenta, 2019). To quickly recap the features of the dataset here, the sample included 150 right-handed men and women between the ages of 60 and 96 years old. Subjects attended at least two MR sessions separated by at least one year; in total there are 373 MR sessions in the dataset. Overall, the subjects were divided into three categories: demented, non-demented, and converted. Table 1 below outlines each of the 15 features included in the data set.

Table 1 Column name and description

Column name	Description	Total Values	Value type
Subject ID	Subjects unique ID	373	Text
MRI ID	Subjects MRI ID	373	Text
Group	Whether the subject is nondemented, demented, or Converted (meaning their status changed from non-demented to demented during the course of the study)	373	Categorical
Visit	Subjects visit number	373	Numerical
MR Delay	The time between the initial appointment and the current scan. Measured in days	373	Numerical
M/F	Subjects gender	373	Categorical
Hand	Whether subject is right- or left-handed	373	Categorical
Age	Subjects age at the time of the MRI scan	373	Numerical
EDUC	Years of education	373	Numerical
SES	Social Economic Status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)	354	Ordinal
MMSE	Mini-Mental State Examination score (range is from 0 = worst to 30 = best)	371	Ordinal
CDR	Clinical Dementia Ratio (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD)	373	Ordinal
eTIV	estimated Total Intracranial Volume	373	Numerical
nWBV	normalized Whole Brain Volume	373	Numerical
ASF	Atlas Scaling Factor	373	Numerical

As we can see from Table 1, all the columns except for MMSE and SES had the complete set of values. 10 of the 15 columns were numerical and produced the basic statistics as shown in Table 2 below. In a further analysis of the data, it was found that of the 373 sessions, 213 were for females and 160 males, approximately 57% and 43% of the entire data respectively. As can be seen in Figure 1 below, Age, EDUC, eTIV, nWBV, and ASF all followed a normal distribution pattern. The time between visits, MR delay, shows a large negative skew towards shorter periods between the subject's scans.

Table 2 Basic Statistics of the Numerical Columns

	Visit	MR Delay	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
count	373.00	373.00	373.00	373.00	354.00	371.00	373.00	373.00	373.00	373.00
mean	1.88	595.10	77.01	14.60	2.46	27.34	0.29	1,488.13	0.73	1.20
std	0.92	635.49	7.64	2.88	1.13	3.68	0.37	176.14	0.04	0.14
min	1.00	0.00	60.00	6.00	1.00	4.00	0.00	1,106.00	0.64	0.88
25%	1.00	0.00	71.00	12.00	2.00	27.00	0.00	1,357.00	0.70	1.10
50%	2.00	552.00	77.00	15.00	2.00	29.00	0.00	1,470.00	0.73	1.19
75%	2.00	873.00	82.00	16.00	3.00	30.00	0.50	1,597.00	0.76	1.29
max	5.00	2,639.00	98.00	23.00	5.00	30.00	2.00	2,004.00	0.84	1.59

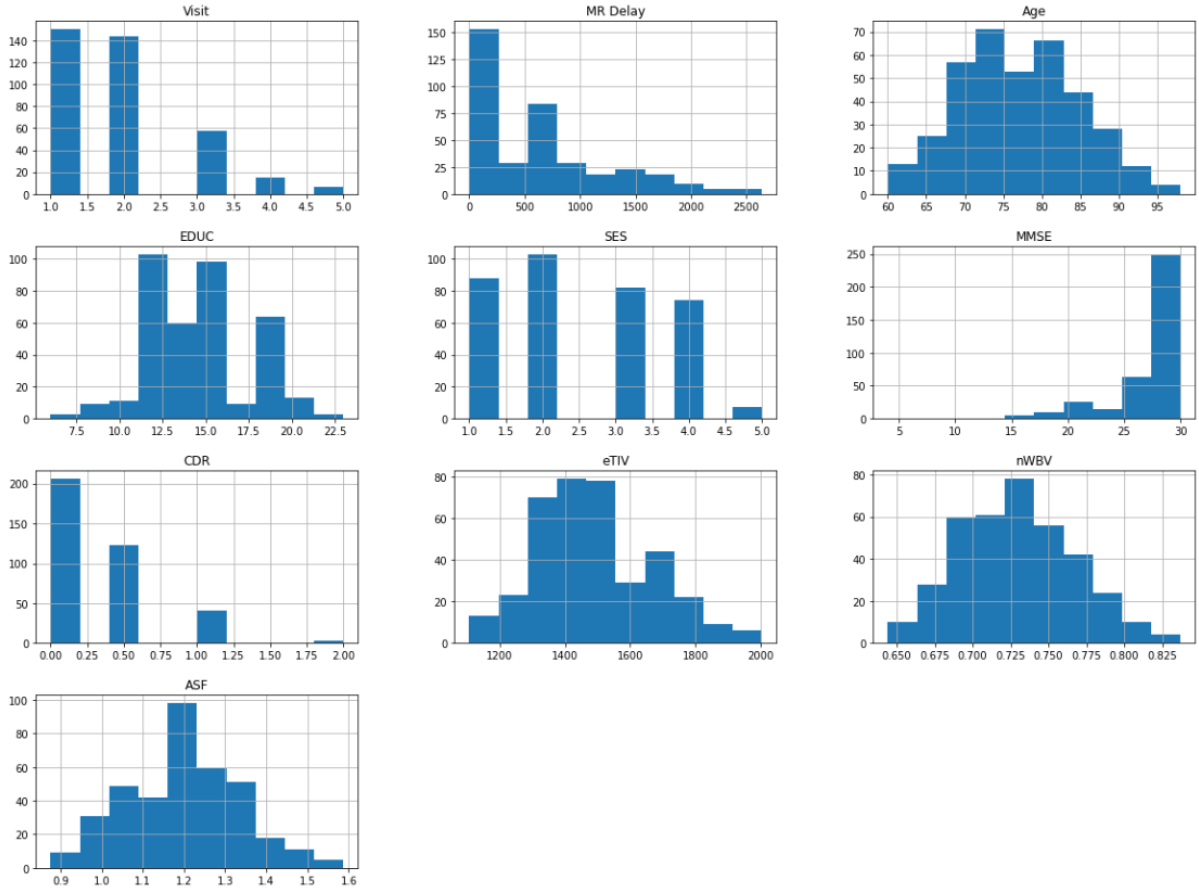

Figure 1 Plots of the Numerical Features to Ascertain Distribution

3.2 Multiclass Classification

There are three different classes in this data set: non-demented, demented, and converted. These shall be represented in the data set as 0, 1, and 2 respectively. Supervised machine learning classification models will be used to label the data.

The aim will be to ensure that False Positives and False Negatives are reduced to ensure model predictions align as much as possible with the reality.

3.3 Class Imbalance

190 of the sessions were labelled as Nondemented (~50%), 146 as demented (~39%), and 37 (~11%) as converted. There is a class imbalance against the 'converted' class. Given that the main purpose of this study is the early detection of patients developing dementia, the 'converted' class is therefore the focus of the study. Considering this, the ROC AUC score shall be used to measure the efficacy of the models. Being able to detect the onset of dementia at an as early stage as possible will result in better support and with the right treatment a delay in the onset of symptoms (Rasmussen and Langerman, 2019).

4 Methods

For this study we are going to be using the Support Machine Vectors (SVM) and Random Forest models from the Scikit package in python. Both models work well with high-dimensional data (Aria, Cuccurullo and Gnasso, 2021; Gaye, Zhang and Wulamu, 2021). The Battineni et al study also employed SVM so this will be a useful point of comparison for the result we achieve. For both models, we used a Randomised Search and a 5-fold cross validation to fine tune our models.

4.1 Support Vector Machines

SVM's are a powerful supervised learning models that work very well for classification and regression problems on small- to medium-sized data sets (Géron, 2017). The randomised search was carried out with the following parameters: the kernels were linear, polynomial, and sigmoid; C values were chosen at random from a continuous reciprocal distribution from 20 up to 200,000 – this was achieved using NumPy's reciprocal function; degrees tested for the polynomial kernel ranged from 0 to 6.

4.2 Random Forest

Random Forest is an ensemble learning method that creates many Decision Trees to train its model. To measure the quality of the split, our model used the Gini index impurity measure (Eq. 1)

$$Gini\ Index = 1 - \sum_{i=1}^{n}(P_i)^2 \qquad (1)$$

For the randomised search, the number of estimators was chosen from a list of values ranging from 200 to 2000 increasing by 200 at a time (ie 200, 400, 600 etc). The maximum number of features ranged from 2 to 12 increasing by 2 (ie 2,4,6 etc). Bootstrapping was set to generate as either True or False.

4.3 Metrics

The standard assessment metrics to ascertain how well a model is performing are accuracy, precision, recall, AUC score, and an F1 score. However, due to this being a multi-class classification problem with a class imbalance away from the group of interest (the converted patients), we won't be relying so much on the F1 score which also performs worse with multi-class classification. The main metric we shall employ is the ROC AUC score, as mentioned above in the class imbalance section; however, we shall still analyse the other metrics so as the allow for proper comparison to the results found in the Battineni et al study. The equations for these metrics can be found below in equations 2 - 5:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (5)$$

5 Data Pre-processing

5.1 Missing Values, Dropped Parameters, and Categorical Variables

As noted in the Data Description section, the SES and MMSE columns are missing 19 and 2 values respectively. These were filled in using the median value of each column.

The columns 'Subject ID' and 'MRI ID' were both dropped as they are just identification columns, and the 'Hand' column was dropped since the data set only contained right-handed individuals.

The values of the 'Group' column which detail the patient's dementia class were replaced with the values 0 – 2 as outlined above.

The M/F column that outlines the subject's gender was converted into two dummy columns 'Male' and 'Female' which only Boolean values.

5.2 Splitting the Data set and SMOTE

The data set was split into a training and test set using the StratifiedShuffleSplit function from the sklearn package. This allowed us to ensure that the split data sets accurately represented the original data. The test set size was set to 20%.

To address the class imbalance, we also ran both models with a synthetic minority oversampling technique (SMOTE) applied to the training data set to balance the data set. With this version of the data, the data was split using Scikit's

train_test_split function with the test set size set to 30% as there was now a larger amount of available data.

5.3 Outliers

Box plots of the data set were created to assess whether the data contained any outliers, the plot for which can be found in figure 2 below. Only Age, nWBV, and ASF were found to not contain any outliers. MR delay and MMSE both seemed to contain a high preponderance of outliers.

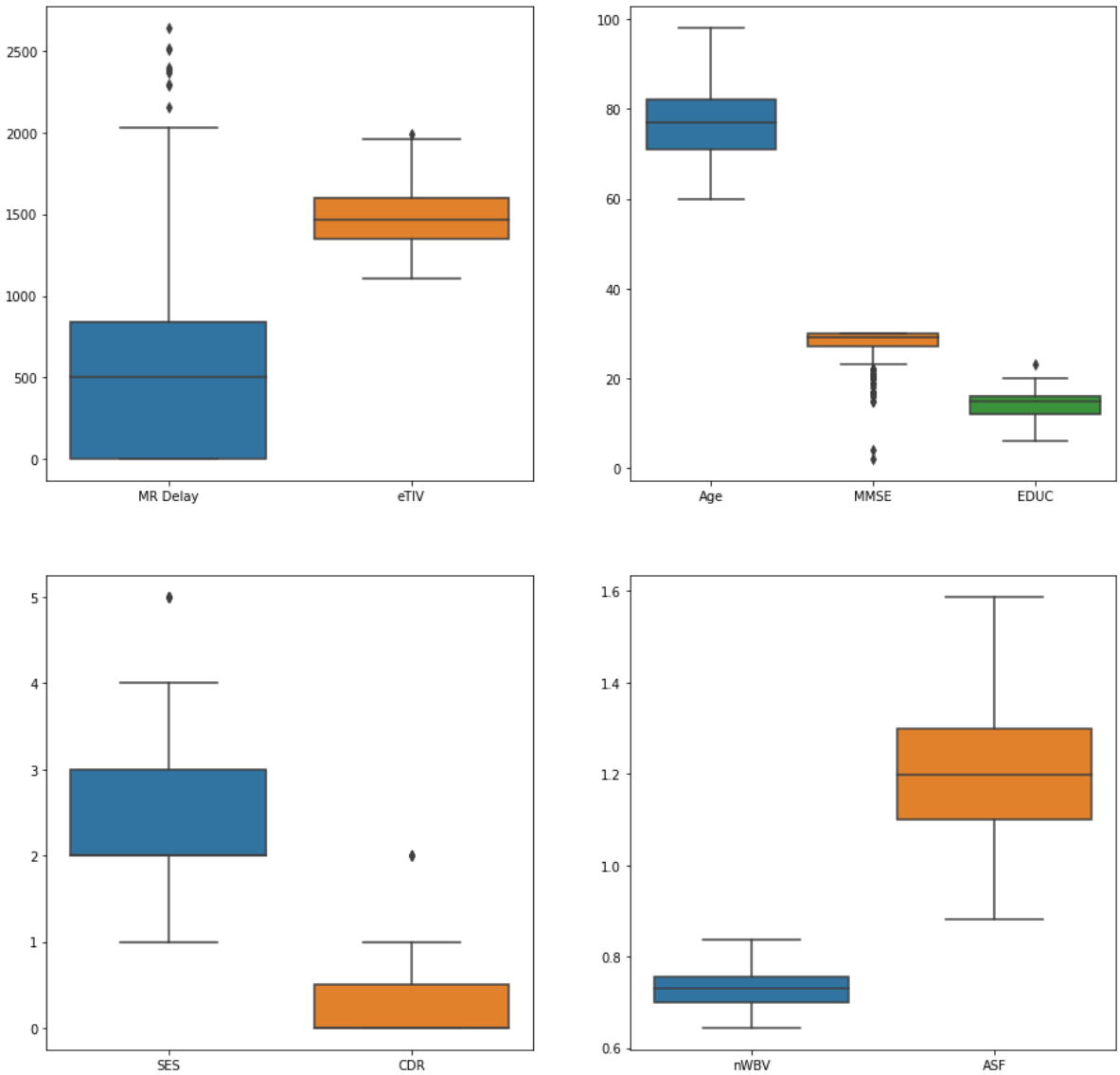

Figure 2 Box plot of the 9 numerical variables.

5.4 Checking for Correlation and Collinearity

A correlation matrix was plotted, Figure 3, to ascertain the collinearity between the difference features as well as the correlation between the individual features and the outcome. Most of the variables showed very little significant correlation between

the features and the outcome. The strongest correlations with the outcome were found for the CDR (0.55), MMSE (-0.34), and nWBV (-0.27) features. Collinearity was found between several of the features. MR Delay and Visit were highly correlated which makes sense as the higher the visit number the longer the number of days since the beginning of the process. ASF and eTIV showed a very strong negative correlation (-0.99); SES was negatively correlated to EDUC (-0.69); CDR and MMSE also showed a strong negative correlation (-0.76)

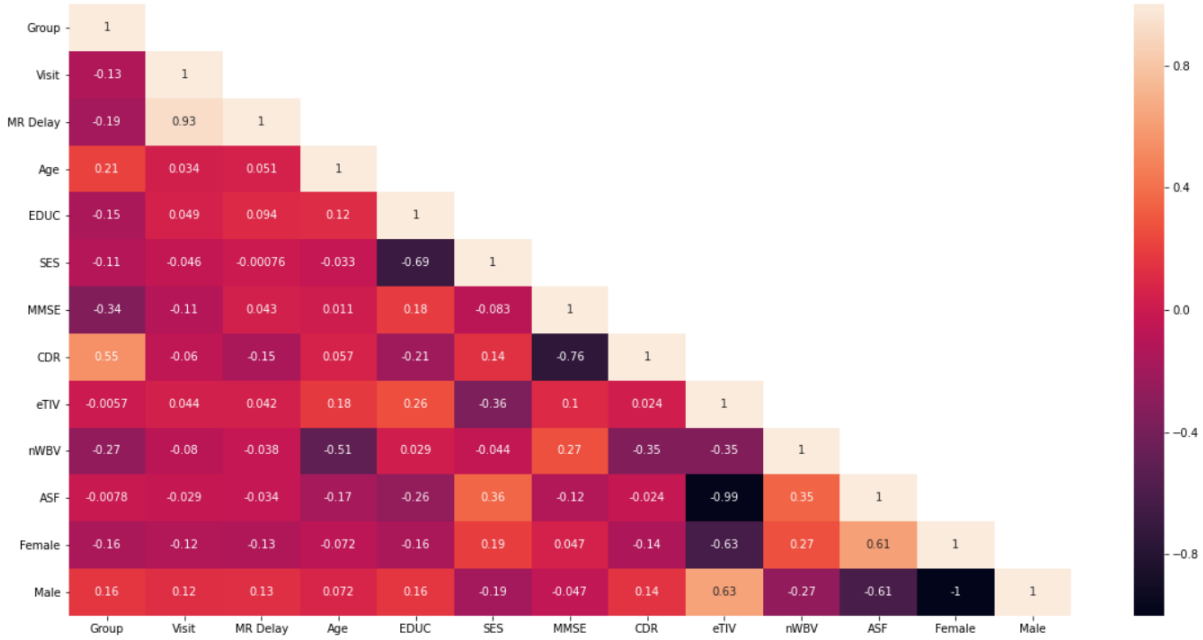

Figure 3 Correlation Matric for All Features

5.5 Standardising Figures / Scaling data

There was a quite a large disparity in scale sizes between the different features, a standard scaler was therefore applied. It was decided to not use a MinMax Scaler since there were many outliers in some of the features that would affect the scaling. A standardisation approach was therefore used.

6 Results

Both the SVMs and the Random Forest models were initially run on the non-SMOTE data with a randomised search function. The confusion matrices and ROC curves for both can be found below in Figure 4 and 5 along with their classification reports in tables 3 and 4. The randomised search for the SVMs performed 250 fits of the data set for both models. The same graphs and tables can also be found for the same models run on the SMOTE version of the data in figure 6 & 7 and tables 5 & 6.

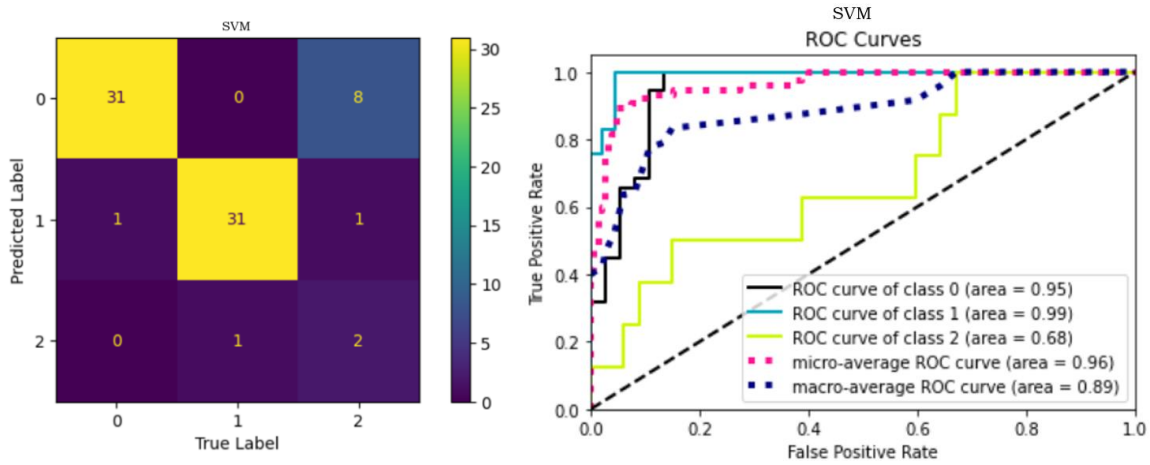

Figure 4 Confusion Matric (left) and ROC curve (right) for the Support Vector Machine

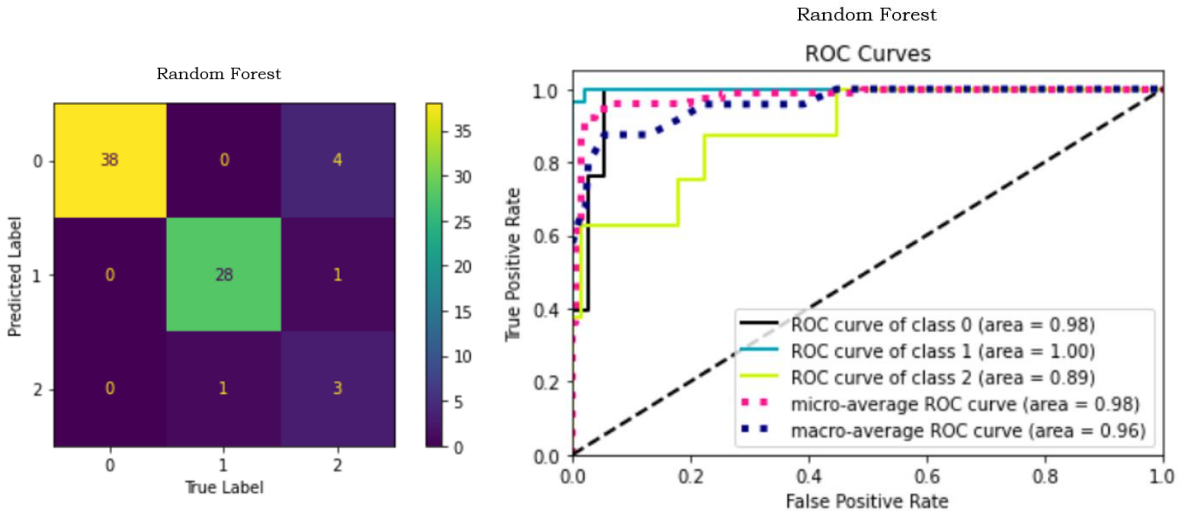

Figure 5 Confusion Matric (left) and ROC curve (right) for the Random Forest

Table 3 Classification Report for SVM Randomised Search

Classification Report:

	precision	recall	f1-score	support
0	0.795	0.969	0.873	32
1	0.939	0.969	0.954	32
2	0.667	0.182	0.286	11
accuracy			0.853	75
macro avg	0.800	0.706	0.704	75
weighted avg	0.838	0.853	0.821	75

Table 4 Classification Report for Random Forest Randomised Search

Classification Report:

	precision	recall	f1-score	support
0	0.905	1.000	0.950	38
1	0.966	0.966	0.966	29
2	0.750	0.375	0.500	8
accuracy			0.920	75
macro avg	0.873	0.780	0.805	75
weighted avg	0.912	0.920	0.908	75

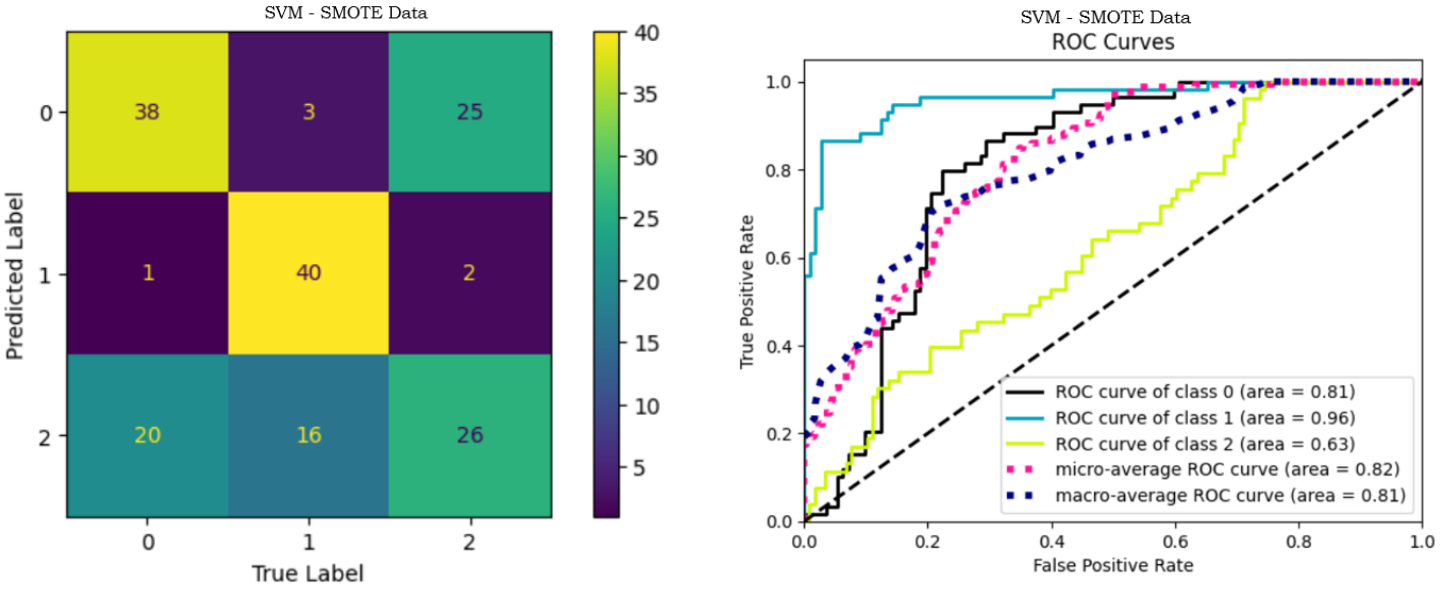

Figure 6 Confusion Matric (left) and ROC curve (right) for the Support Vector Machine (SMOTE)

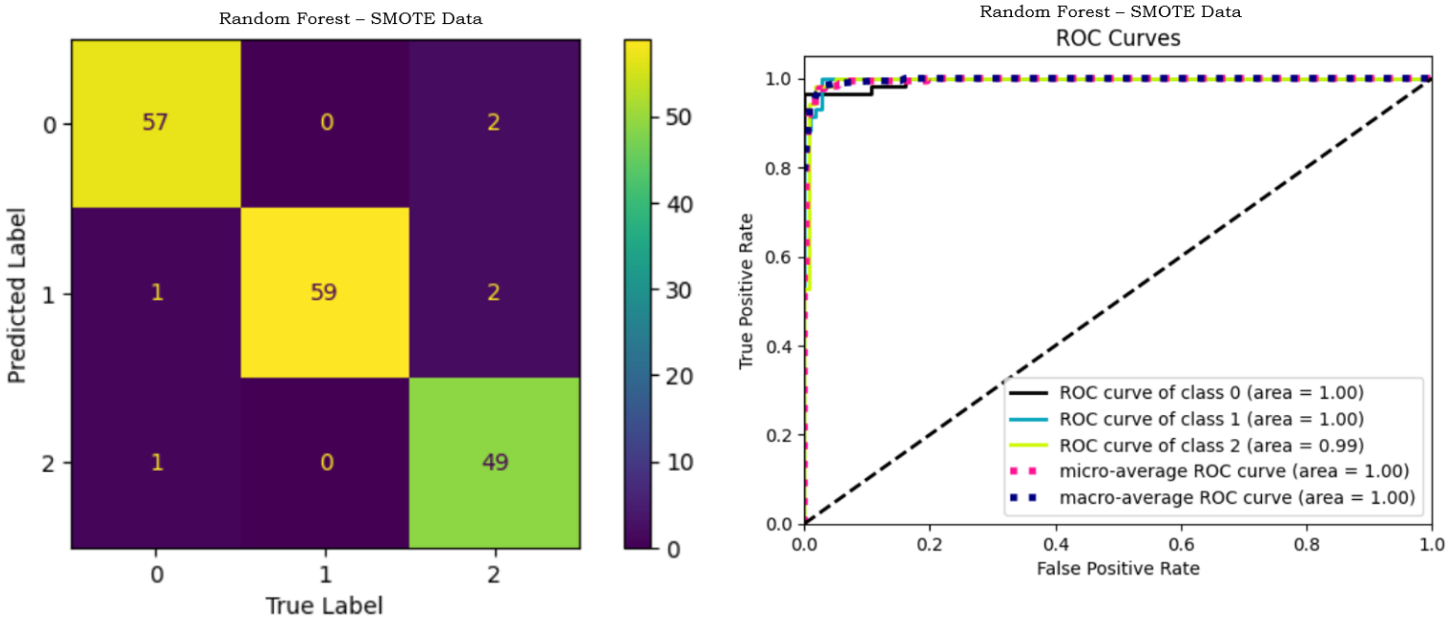

Figure 7 Confusion Matric (left) and ROC curve (right) for Random Forest (SMOTE)

Table 5 Classification Report for SVM Randomised Search (SMOTE)

Classification Report:

	precision	recall	f1-score	support
0	0.576	0.644	0.608	59
1	0.930	0.678	0.784	59
2	0.419	0.491	0.452	53
accuracy			0.608	171
macro avg	0.642	0.604	0.615	171
weighted avg	0.650	0.608	0.621	171

Table 6 Classification Report for Random Forest Randomised Search (SMOTE)

Classification Report:

	precision	recall	f1-score	support
0	0.966	0.966	0.966	59
1	0.952	1.000	0.975	59
2	0.980	0.925	0.951	53
accuracy			0.965	171
macro avg	0.966	0.964	0.964	171
weighted avg	0.965	0.965	0.965	171

We also include in table 7 the results found from Battineni et al study with which we shall compare our results.

Table 7 Confusion Matrix of given subjects TND: True Non-Demented; TD*: True Demented; TC*: True Converted; PND*: Predict Non-Demented; PD*: Predict Demented, and PC*: Predict Converted.*

	TND	TD	TC	precision
PND	43	14	10	64.18%
PD	8	27	1	75.00%
PC	2	0	0	0.00%
Recall	81.13%	65.85%	0.00%	0.00%

7 Discussion

Overall, the random forest models performed a lot better than the SVM ones. For the non-SMOTE-SVM model, it was able to categorise the demented and non demented cases very well with the precision, recall, and AUC scores all being well above 80%/0.8; however, it couldn't categorise the converted cases very well with precision, recall, and AUC scores of 67%, 18%, and 0.68 respectively. With such a low recall score, the model is very unreliable for our needs as it is unable to detect a high percentage of the true converted dementia cases. The SMOTE-SVM model, whilst performing better in terms of recall for the converted case (49%) performed worse across all other metrics, especially in categorising non-demented cases.

Both random forest models performed very well with AUC scores ranging between 0.89 to 1, the highest score possible. The SMOTE random forest model (SMOTE-RF) performed the best of the two with AUC scores between 0.99 and 1. The precision scores for SMOTE-RF ranged between 95% and 98%; the recall scores were between 93% and 97%; the accuracy scores was 97%.

If we compare our results with those of the Battineni et all study in Table 7, we can see that our non-SMOTE-SVM and both RF models performed much better than the Battineni results with their precision score for converted cases at 0, though recall was 81.13%. Unfortunately, their study did not include ROC-AUC scores which means we cannot make a full comparison between the studies. It would ostensible appear though that by including more features in our model than the Battineni study we have produced a better model. It would also appear that the RF models provide a more promising avenue of future research than the SVM models which did not appear to work as well as hoped.

There are some drawbacks to the current study. First and foremost is the limited size of and the class imbalance within the dataset. With only 373 data points

across 150 subjects and classifications heavily skewed away from converted patients, there simply is not enough data to run our models on and be sure of producing a reliable model that will work well in other instances. Generally, having larger data sets allows for a better predictive model (Raudys and Jain, 1991; Zhang and Ling, 2018; Vabalas *et al.*, 2019), though larger data sets are not without their own issues (Fan, Han and Liu, 2014). The author did try to find larger publicly available data sets as well as contacted various institutions to see if access could be obtained for some larger data sets but unfortunately nothing suitable could be found. Even without a larger data set, it would also still be useful to further investigate feature importance and selection to a much greater degree though this was not possible under current time constraints.

The ability to detect the onset of dementia in its early stages would be of great benefit to many patients and their families. It would allow for support and intervention practices to be provided and employed at the earliest stage possible and hopefully produce a much better outcome in terms of morbidity.

8 Conclusion

The focus of this study has been on the early detection and diagnosis of dementia. In addition to carrying out our own modelling research, the purpose of this paper was to compare those results we obtained to previous research in this field and on this particular data set. We ran Support Vector Machine and Random Forest models both with and without SMOTE implementation. The SMOTE Random Forest model produced the best results with AUC scores between 0.99 and 1, and precision and recall scores all above 93%. The models produced in this study also performed much better than those from previous studies. The early detection of dementia in patients is vital in helping us provide treatment to slow the onset of further symptoms and providing the necessary support to both patients and their families.

9 Appendix

Please find the code for this project in the following GitHub Repository: https://github.com/199110799-DSDA-yorksj/AppMlProject.git

10 References

Aria, M., Cuccurullo, C. and Gnasso, A. (2021) 'A comparison among interpretative proposals for Random Forests', *Machine Learning with Applications*, 6, p. 100094. doi: 10.1016/J.MLWA.2021.100094.

Battineni, G., Chintalapudi, N. and Amenta, F. (2019) 'Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)', *Informatics in Medicine Unlocked*, 16, p. 100200. doi: 10.1016/J.IMU.2019.100200.

Duboisa, B. *et al.* (2016) 'Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges', *Journal of Alzheimer's Disease*, 49(3), pp. 617–631. doi: 10.3233/JAD-150692.

Fan, J., Han, F. and Liu, H. (2014) 'Challenges of Big Data Analysis', *National science review*, 1(2), p. 293. doi: 10.1093/NSR/NWT032.

Gaye, B., Zhang, D. and Wulamu, A. (2021) 'Improvement of Support Vector Machine Algorithm in Big Data Background', *Mathematical Problems in Engineering*, 2021. doi: 10.1155/2021/5594899.

Géron, A. (2017) 'Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019, O'reilly)', in *Hands-On Machine Learning with R*. O'Reilly Media, pp. 153–175.

Goerdten, J. *et al.* (2019) 'Statistical methods for dementia risk prediction and recommendations for future work: A systematic review', *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5, pp. 563–569. doi: 10.1016/J.TRCI.2019.08.001.

Moreira, L. B. and Namen, A. A. (2018) 'A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia', *Computer Methods and Programs in Biomedicine*, 165, pp. 139–149. doi: 10.1016/J.CMPB.2018.08.016.

Pellegrini, E. *et al.* (2018) 'Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review', *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10, pp. 519–535. doi: 10.1016/J.DADM.2018.07.004.

Rasmussen, J. and Langerman, H. (2019) 'Alzheimer's Disease – Why We Need Early Diagnosis', *Degenerative Neurological and Neuromuscular Disease*, 9, p. 123. doi: 10.2147/DNND.S228939.

Raudys, S. J. and Jain, A. K. (1991) 'Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3), pp. 252–264. doi: 10.1109/34.75512.

Vabalas, A. *et al.* (2019) 'Machine learning algorithm validation with a limited sample size', *PLOS ONE*, 14(11), p. e0224365. doi: 10.1371/JOURNAL.PONE.0224365.

Zhang, Y. and Ling, C. (2018) 'A strategy to apply machine learning to small datasets in materials science', *npj Computational Materials 2018 4:1*, 4(1), pp. 1–8. doi: 10.1038/s41524-018-0081-z.