Dementia Classification with Neural Networks

11/01/2023

1 Abstract

Dementia is a descriptive word for a group of diseases that impact the brain causing a decline in cognitive abilities. The most common form of dementia in the UK is Alzheimer's. Its causes are still unknown. [1] It is important to identify dementia quickly to help patients and study them to uncover more. This paper looks at 2 machine learning models and a deep learning model to see how well they can predict a person's dementia status given a set of features.

2 Introduction

In the world, there are more than 55 million people with dementia. This number increases by 10 million every year. It has the 7th highest death toll among deadly diseases and is one of the biggest causes of disabilities in aging people. [2] There is currently no cure for all common dementia-causing diseases indicating urgency for more research, this is vital as dementia has everlasting impacts on the individual, their family, and carers. The UK currently has an aging population and with life expectancy increasing more of the UK will have dementia which could have detrimental costs to our society and economy. [3]

Key signs of dementia include memory loss, difficulties in timekeeping, speech impediments, and confusion. Symptoms occur in progressive stages early, middle, and late. The late stages of symptoms are more obvious, whereas the early stages are harder to identify. From the middle to late stages, the individual will become increasingly dependent on their carer, up to complete dependence, and have periods of total inactivity. [2]

The purpose of this model is to help medics identify dementia earlier in the hopes of reducing negative impacts. Much like other illnesses, there are advantages to early diagnosis though there is no available treatment to reduce the progression or a cure. Those who get diagnosed will have access to mental, social, and even financial support plus the opportunity to participate in medical research to shape the future. One of the suspected key reasons for failing trials is that it is being tested on those who have let the condition progress too far. [4]

3 Related Work

Health and deep learning have plenty of work together to identify relations that a human may not be able to spot. With healthcare, being one of society's most focused areas of development, it makes sense that the powerful technology of deep learning will be experimented with in that department.

The authors in [5] use a dataset specific to South Korea, to predict dementia with the use of a deep neural networks (DNN) structure containing 4 hidden layers. This model achieved a 0.805 accuracy score on the confusion matrix. At first glance, this was impressive, however, when looking at the recall value it is only 0.657. Meaning that, given you have dementia, the model will only predict correctly 65.7% of the time. The problem with comparing my report with theirs is that my dataset is significantly smaller.

Huaping Zhou in [6] showcases a deep-learning diabetes prediction model with greater success. The model they have proposed is used on two different datasets, one being orientated around Pima Indians, and the other is data on a more global scale. They conclude that they achieved 99% accuracy on the

Pima Indians and 94% on the other. These results are great, however, there is no showing of a confusion matrix to see how well it predicts test data.

4 Ethical Considerations

Generally, models created using deep learning comes with several ethical challenges that must be considered. In medical cases, there has always been the question of whether or not an AI will be able to replace a doctor. This can be problematic for employment and therefore raises ethical concerns. Studies have found that AIs are better than doctors at diagnosing a patient [7] and therefore could (and arguably should) replace humans, however, this is not to be the case. In America, there is currently a shortage of psychiatrists, and this shortage is projected to continue up to and beyond 2030. As a result, those in the workforce feel overworked. [8] This is the case in the UK and many other countries [9]. For this reason, AI must have a bigger role in the medical field to reduce workload and improve accuracy in diagnosis. Instead of AIs replacing people, people must learn to work alongside AI to produce better results.

5 Deep Learning and Neural Networks

Deep learning is a subset of machine learning, hence, a subset of AI. It's inspired by the human brain's way of thinking and aims to replicate that. Deep learning models remove the human element of algorithms. [10] It uses multiple layers of nodes and connections between the nodes, equivalent to the brain's neurons and synapses. These nodes are mathematical functions that execute nonlinear operations to weighted sum-of-values, known as perceptrons, to formulate an output value. These operations are needed to manipulate a function to best fit the data. [11] All neural networks have at least 3 layers, input, output, and hidden layers.

The use of deep learning networks has vastly increased since people began relying on digital methods of storing and collecting data. Deep learning can utilize large data better than any other method and it is now used commonly in medical diagnosis. [12]

6 Dataset

The dementia data, obtained from Kaggle by Shashwat Tiwari [13], contains 373 imaging session visit information from 150 different patients. There are 12 different features given for each case. The patients are classified into one of three groups, demented, non-demented, and converted. Converted means that in their first imaging session visit, they were classified as non-demented, most likely with mild cognitive impairment, but changed in a later visit. The first 5 rows of the raw data are shown in figure 1. In figure 2 you can see the list of attributes, their possible values, and a short descriptor. You will find that there's a combination of both basic information and clinical data most of these are supposedly related to dementia.

	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034

Figure 1: First 5 rows of the raw data

Dementia Dataset Description		
Attribute	Values	Description
Age	Integer, 60 - 96	Age of person
Sex	M/F	Gender
Hand	L/R	Preferred hand to use for tasks
EDUC	Integer, 6 - 23	Education
SES	Integer, 1 - 5	Socioeconomic Status
Visit	Integer, 1 - 5	Number of visits this patient has had
MR Delay	Integer >= 0	Time of delayed enhancement in Miliseconds
MMSE	Integer, 0 - 30	Mini Mental Status Exam score
CDR	Float , 0 - 2	Clinical Dementia Rating
eTIV	Integer > 0	Estimated Total Intracranial Volume
nWBV	Float > 0	Normalized Whole Brain Volume
ASF	Float > 0	Atlas Scaling Factor
Group	Demented, Non-demented, converted	Dementia Status

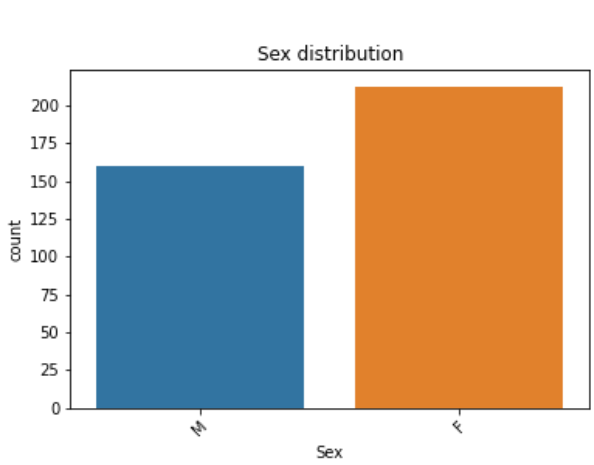
Figure 2: Description and values of attributes in dataset

6.1 Exploratory Data Analysis

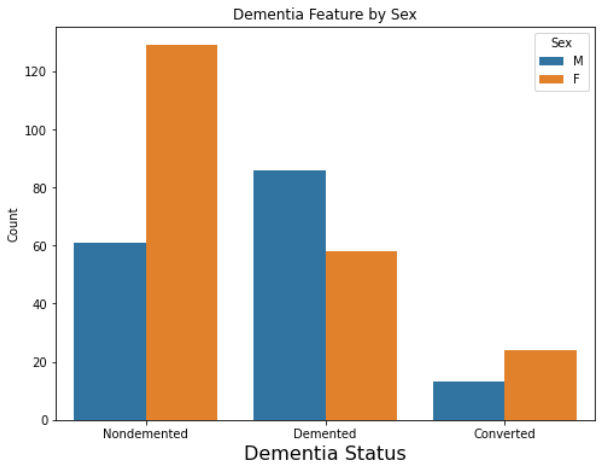
The easiest start would be to look at would be the binary attributes, in this case, it's hand and sex. For the particular cases of hand, sex socioeconomic status (SES), and education (EDUC), I have grouped them by their subject ID as I have assumed these attributes to be constant throughout all their visits. This would also be the case for their dementia status as the group of converted already represents any change throughout their visits.

When looking at the hand attribute, it is seen that all patients were right-handed. For this reason, we will drop and ignore this attribute.

Looking at figure 3, there is a higher count of females than males but yet figure 3b shows that more males are demented and there are more than double non-demented female patients. This appears to go against research that says that women are more likely to have dementia than men. [14] This could be a result of a small dataset that does not cover a broad enough sample to accurately represent this.



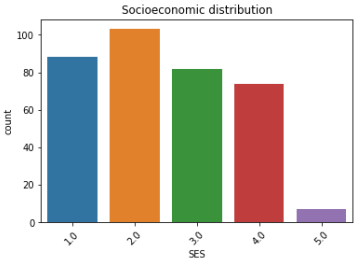
(a) Distribution of sex



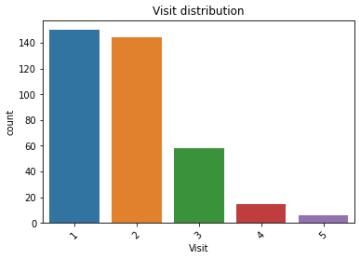
(b) Distribution of dementia and sex

Figure 3: Gender and dementia

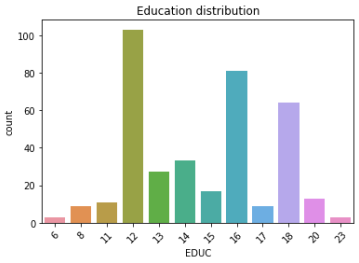
SES, the number of visits, and EDUC are shown in figure 4. For SES, 1.0 represents a low socioeconomic status and 5.0 is a very high socioeconomic status. Most people have an SES of 2 and very few have a rating of 5. As for the number of visits, almost everyone has had at least 2 visits and not many have had 3 or more. In EDUC we see large numbers in 12, 16, and 18.



(a) Distribution of Socioeconomic Status



(b) Distribution of visits



(c) Distribution of Education Levels

Figure 4: Categorical attributes

Looking at how our predictor variable is distributed in figure 5, there are slightly more non-demented patients than demented, and both of those are significantly more than those who are converted. This signals a bias in our data as the classifications are not equally distributed. Action must be taken on this, SMOTE (synthetic minority oversampling technique) will be applied later.

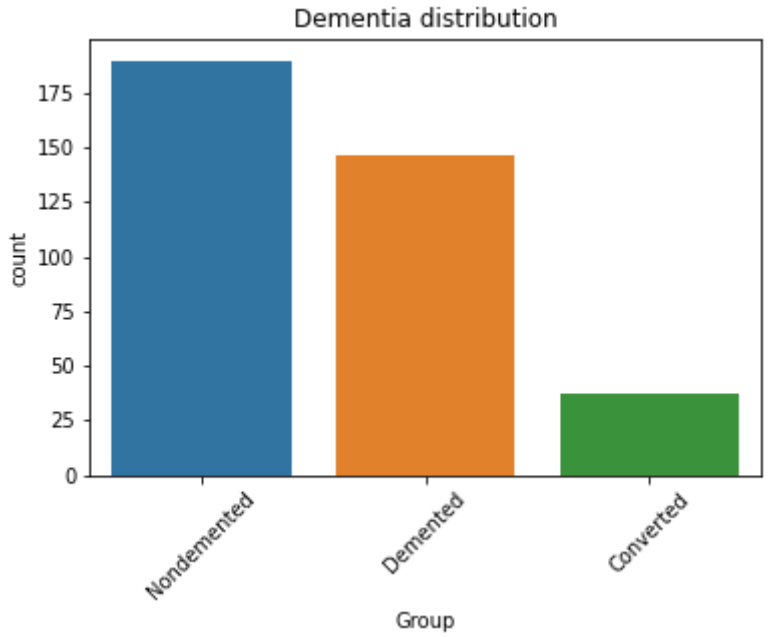


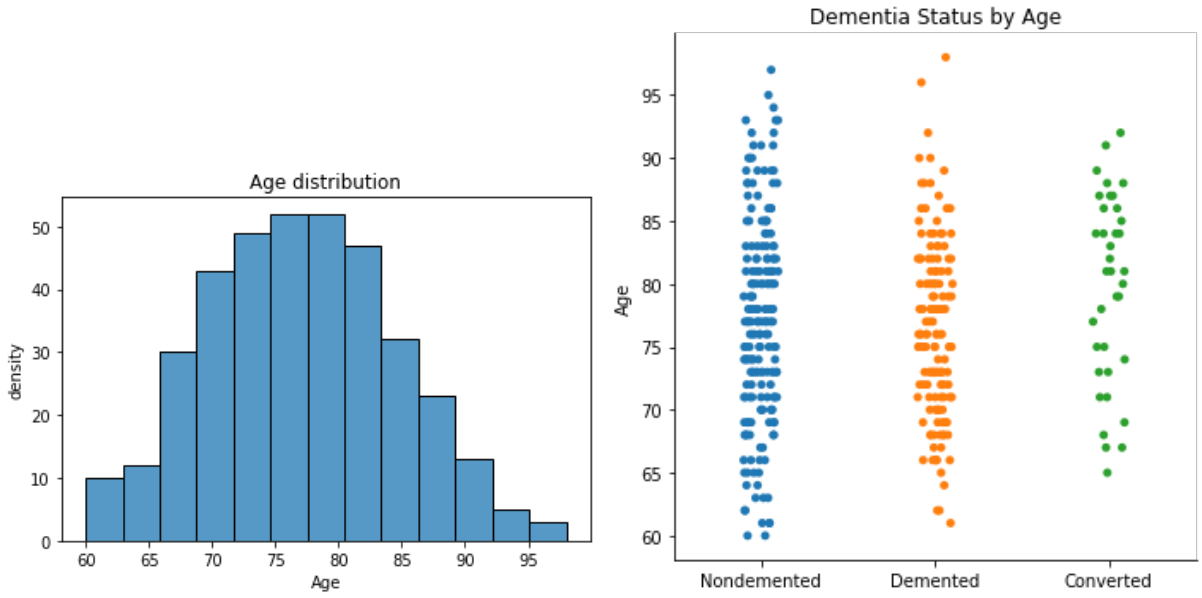
Figure 5: Distribution of dementia

The distribution of age is shown in the histogram below and looks quite normally distributed. The range for age is 60 to 96 years old with the mean being approximately 77. It is often thought that age and dementia have a correlation with each other, typically the older someone is, the more likely they are to have dementia. According to the NHS [15] the risk of Alzheimer's increases with age, from 1 in 14 for those above 65 to 1 in 6 beyond 80 years. To investigate this further we can look at the catplot in figure 6b. Surprisingly, there doesn't appear to be a clear relationship between these two. When calculating the Pearson's correlation coefficient (r) between the two, it results in -0.05 which again says that there is no correlation. Pearson's r is calculated using the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

[16]

The final value is between -1 and 1. If r is close to or equal to 1, this represents a strong positive correlation, -1 is a strong inverse correlation and the closer to 0 it is, the weaker the correlation is.



(a) Distribution of age

(b) Relation between age and dementia status

Figure 6: Exploration of age

6.2 Correlation Matrix

A good way to see which features correlate best with our predicting variable is to create a heatmap showing the r-value for each combination of two variables. This shows which features most influence the group a patient is labeled as and if there are any multicollinearities. MR Delay and visit have a high value of 0.92, we should consider dropping at least one of these.

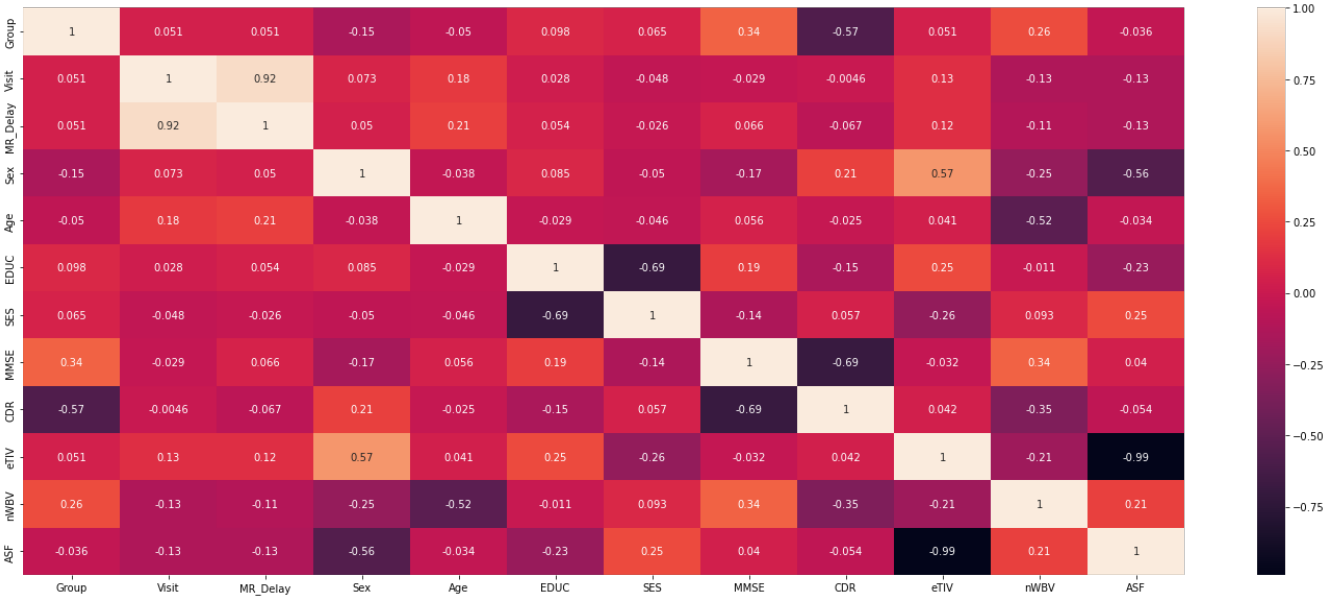


Figure 7: Distribution of dementia

The matrix shows that CDR is the most correlated feature to dementia with r being -0.57. Figure 8 exhibits this clearly as everyone those with a CDR of 1 or greater are demented. Patients with a CDR of 0 are either non-demented or converted. There are likely other factors that will differentiate the rest.

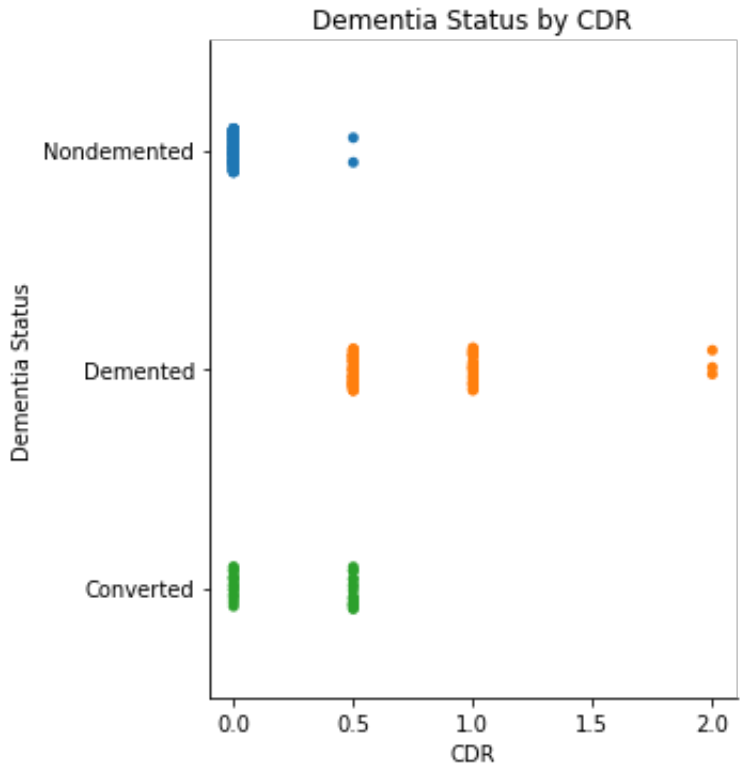


Figure 8: Distribution of dementia against CDR

With MMSE being the next most correlated feature, a graph plotting patients with a CDR of 0.5 and their MMSE should help differentiate their status. Figure 9 shows that a CDR of 0.5 and MMSE of less than 24 are demented.

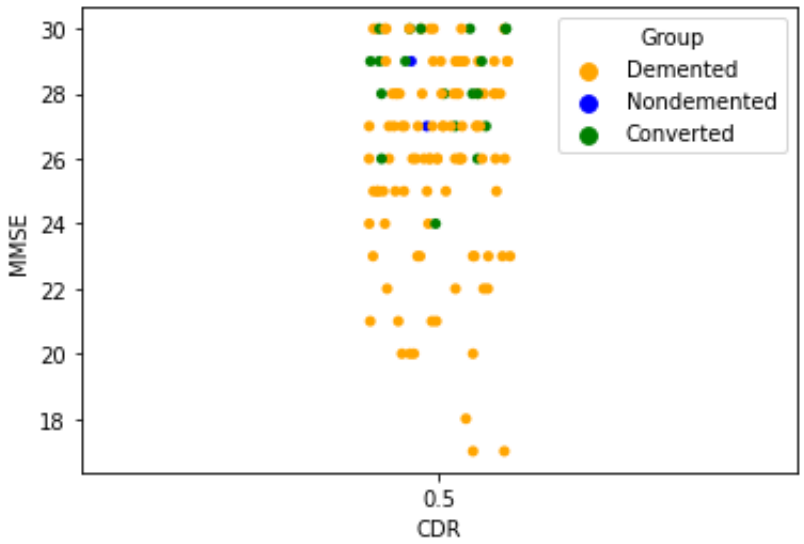


Figure 9: Distribution of dementia against MMSE (Only patients with CDR=0)

With everything discussed above, CDR and MMSE are the key features to use. Though the other features may seem irrelevant, they will still be kept for modeling since the dataset is very small, therefore memory and computational cost is not a worry. The feature 'visit' will be dropped since it has a high correlation with MR delay.

7 Pre-processing

Before modeling, steps need to be taken for the model to work efficiently and effectively. This mainly consists of manipulating the data with various transformations.

7.1 Label Encoder

The first transformation is the label encoder which must be applied to all the categorical features, especially the ones that are of a string data type. The two categorical features we will be encoding are sex and group. Encoding entails mapping the possible categories to integers. In the case of sex, 'F' will be replaced with 0 and 'M' with 1. The mapping can be seen in figure 10. [17]

```
Sex:
0 : F
1 : M


Group:
0 : Converted
1 : Demented
2 : Nondemented
```

Figure 10: Distribution of dementia against CDR

7.2 Standard Scaler

This function is used to standardize the data to a form that has a mean of 0 and a variance of 1. This brings all the features down to a common scale without distorting the distribution. This makes it easier for the model to compute. [17]

7.3 SMOTE

As mentioned earlier in the exploratory data analysis section, we need to apply SMOTE to synthetically create more demented and converted cases of data, enough to have an even distribution of each classification. Before applying SMOTE there were 142 non-demented, 108 demented, and 28 converted imaging sessions, now there are 142 cases in all 3 groups. [17]

8 Modelling Methods

In this investigation, there are two machine learning (ML) algorithms, an Extremely Randomized Trees Classifier (ET) and Gradient Boosting Classifier (GB), and one deep learning method, in the form of a 1-dimensional convolutional neural network (Conv1D). The ML models are created using scikit learn [17] and the Conv1D is built using TensorFlow. [18]

8.1 Hyperparameter Tuning

Hyperparameter tuning will be implemented, more specifically, 'GridSearchCV. This is a robust method of testing every possible combination of parameters, given a dictionary of parameters. Alternatively, 'RandomizedSearchCV' could have been used, but this is not necessary due to the small size of the dataset being worked on. [17]

8.2 Confusion Matrix

A confusion matrix is a way in which each model can be summarised. This will allow us to easily compare the performance of each model using metrics derived from it. The matrix consists of rows representing the predictions made by the model and columns corresponding to the known truths. In figure 11, the green boxes show what the algorithm has predicted correctly and the red boxes show the incorrect predictions. In our case, since we have 3 classifications, the matrix is 3x3 and the metrics are calculated slightly differently. The metrics used are accuracy, recall, and precision.

PREDICTED	Converted	True Converted (TC)	False Converted 1 (FC1)	False Converted 2 (FC2)
	Demented	False Demented 1 (FD1)	True Demented (TD)	False Demented 2 (FD2)
	Non-demented	False Non-demented 1 (FN1)	False Non-Demented 2 (FN2)	True Non-demented (TN)
		Converted	Demented	Non-demented
		TRUE		

Figure 11: Confusion Matrix

8.2.1 Accuracy

The equation to calculate accuracy is as follows.

$$Accuracy = \frac{TC + TD + TN}{TC + FC1 + FC2 + FD1 + TD + FD2 + FN1 + FN2 + TN}$$

or simply

$$Accuracy = \frac{correct}{all}$$

This gives a percentage of all correctly classified results.

8.2.2 Recall

There will be 3 recall values, one for each true classification. Assuming we wanted to calculate recall for the demented group, the equation would be the correctly predicted demented cases divided by all the actual demented cases.

$$Recall(Demented) = \frac{TD}{TD + FC1 + FN2}$$

The result indicates, given that a case is truly demented, how likely is it that the model predicted it correctly?

8.2.3 Precision

There are precision values for each predicted group. To calculate precision for demented cases, it would be the correctly predicted demented cases divided by the total number of predicted demented cases.

$$Precision(Demented) = \frac{TD}{TD + FD1 + FD2}$$

This result would answer the question of, given a case is predicted to be demented, how likely is it that the case is truly demented?

8.3 Extremely Randomized Trees

ET is similar to the random forest algorithm (RF) except it adds one more step of randomness. RF involves creating numerous bootstrapped datasets with a subset of random feature selection. Bootstrapped datasets are created from random sampling with replacement of the whole original dementia dataset. Having replacements allows for duplicate entries in the bootstraps. Decision trees are created for each of these bootstraps. The final prediction will be an average of all the results from each tree. With ET, each decision tree is pruned at random nodes.[19]

8.3.1 Results

Figure 12 demonstrates the performance of our ET model. Overall, it performs quite well with an accuracy of 88.2% after tuning, however, there are still places where it can improve. Particularly with the converted group as the recall and precision scores lack severely. The recall score on the demented group could also have been improved, though 83% is still okay, this is arguably the most important group to get correct.

```
Best hyperparameters:
 {'max_features': 1, 'min_samples_split': 6, 'n_estimators': 50}
Test accuracy of the best model (tuned):
 0.882
Recall:  [0.44444444 0.83333333 1.        ]
 Accuracy:  0.8817204301075269
 Precision:  [0.44444444 0.96774194 0.90566038]
classification report :
              precision    recall  f1-score   support

           0       0.44      0.44      0.44         9
           1       0.97      0.83      0.90        36
           2       0.91      1.00      0.95        48

    accuracy                           0.88        93
   macro avg       0.77      0.76      0.76        93
weighted avg       0.89      0.88      0.88        93
```

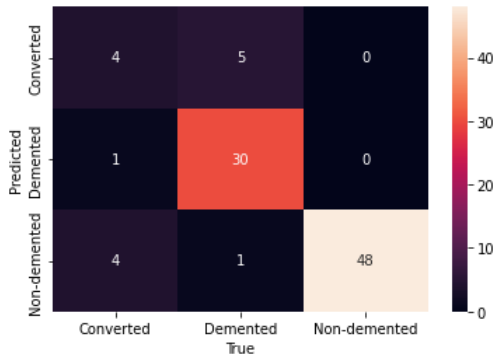


Figure 12: Metrics and Confusion Matrix of ET Model

8.4 Gradient Boosting

GB classification works by initializing the model with a first prediction. This prediction would be calculated using 'log of the odds'. Since this is not a binary classification, the log odds need to be calculated for each class label to find the 3 probabilities. This is known as one-vs-all. Whichever probability is highest will be the first prediction used for all cases. From here the pseudo-residuals for each case are calculated, this is the difference between the (mapped) observed values and predicted values. A decision tree is created using the features to predict these residuals. Since the predictions are in terms of the log of the odds and the residuals are derived from a probability, a transformation on the residuals is required. The transformed residuals are multiplied by a learning rate and then added to the original prediction to form new predictions for each case and therefore new residuals which a decision tree can be formed to predict again. The more iterations of this, the lower the error should be. [20]

8.4.1 Results

The GB model performed extremely well with a very high accuracy of 92.5% on the confusion matrix. Once again, it is the recall and precision scores on the converted group that lets the model down, however, it still performs much better than the ET model in every single aspect.

```
Best hyperparameters:
 {'criterion': 'squared_error', 'learning_rate': 0.78, 'loss': 'log_loss', 'min_samples_split': 8, 'n_estimators': 50}
Test accuracy of the best model (tuned):
 0.925
Recall:  [0.55555556 0.91666667 1.        ]
 Accuracy:  0.9247311827956989
 Precision:  [0.71428571 0.97058824 0.92307692]
classification report :
               precision    recall  f1-score   support

           0       0.71      0.56      0.63         9
           1       0.97      0.92      0.94        36
           2       0.92      1.00      0.96        48

    accuracy                           0.92        93
   macro avg       0.87      0.82      0.84        93
weighted avg       0.92      0.92      0.92        93
```

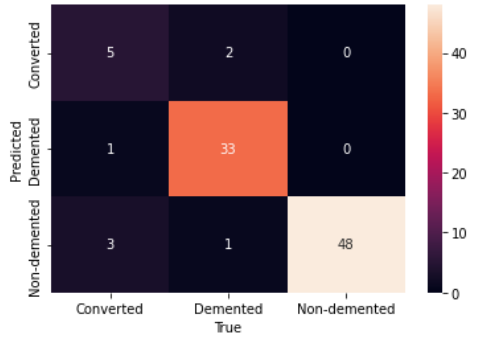


Figure 13: Metrics and Confusion Matrix of GB Model

8.5 Convolutional Neural Network (CNN)

CNN is a section of Deep Neural networks, typically used for image recognition in the form of a two-dimensional CNN (Conv2D). Here, a Conv1D is more appropriate for the data we are using as the kernel is only able to move in 1 dimension as opposed to images that are 2-dimensional.

Considering the small size of the dataset, the architecture of this model is very simple and contains only 2 hidden layers both with Rectified Linear Unit (ReLu) functions. Opting for this function over the sigmoid function means that the gradient descent process (the algorithm that optimizes the residuals of the model) can be done more quickly.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_2 (Conv1D)           (None, 9, 64)             192

 conv1d_3 (Conv1D)           (None, 8, 64)             8256

 dropout_1 (Dropout)         (None, 8, 64)             0

 max_pooling1d_1 (MaxPooling  (None, 4, 64)            0
 1D)

 flatten_1 (Flatten)         (None, 256)               0

 dense_2 (Dense)             (None, 100)               25700

 dense_3 (Dense)             (None, 3)                 303

=================================================================
Total params: 34,451
Trainable params: 34,451
Non-trainable params: 0
```

Figure 14: Conv1D Model Summary

The dropout layer rate has been set to 0.4 to prevent overfitting. This is a relatively high rate, but since overfitting is a major problem in neural networks, particularly on small datasets, it shouldn't be an issue.

The model was trained for 30 epochs. Model accuracy and loss for each iteration are plotted below. The model drastically improves after only 5 epochs and begins to plateau beyond that. Both the train and test improve after each epoch showing that overfitting is not a concern.

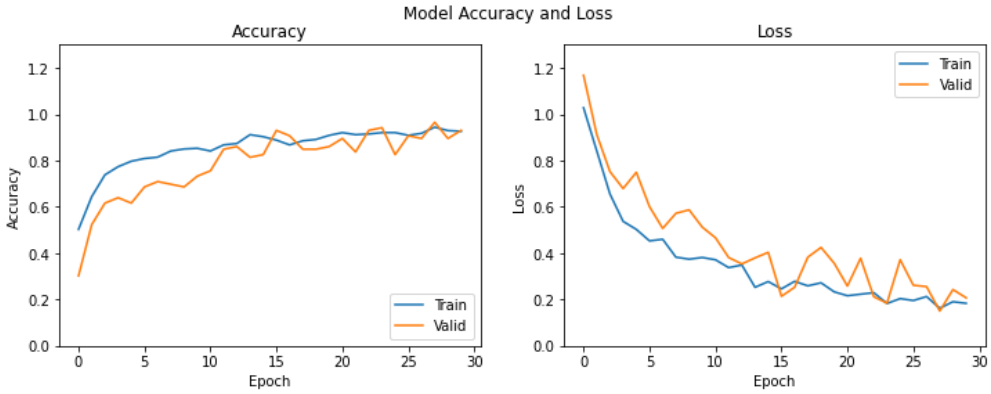


Figure 15: Conv1d Model accuracy(left) and loss (right) for train and test datasets

9 Results

The Conv1d model has performed well showing some good metrics, notable 90% accuracy, however, it falls short on the recall and precision of the converted group. Though performing better than the ET model, it does not beat the GB model in the majority of the metrics.

```
Recall:  [0.55555556 0.88888889 0.97916667]
 Accuracy:  0.9032258064516129
 Precision:  [0.5        0.94117647 0.95918367]
classification report :
              precision    recall  f1-score   support

           0       0.50      0.56      0.53         9
           1       0.94      0.89      0.91        36
           2       0.96      0.98      0.97        48

    accuracy                           0.90        93
   macro avg       0.80      0.81      0.80        93
weighted avg       0.91      0.90      0.91        93
```

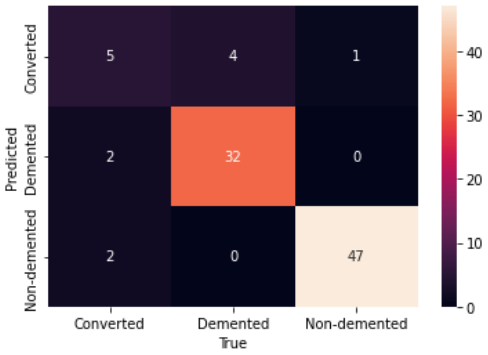


Figure 16: Conv1d Model Summary

10 Concluding Results & Final Thoughts

It appears that the GB model has performed the best, even against the Conv1D model. This does not discredit the usability of deep learning models in any way as there are possible explanations for this. For one, this is a small dataset and deep learning models are most effective on significantly larger datasets. I also found that the inaccuracies lay in the converted classification, this is perhaps because there is a blurred line where actually the patient was non-demented upon their first visit and then found themselves to be demented on a later visit. If maybe, there was more clarity in the dataset and simply kept as a binary classification where the imaging sessions were independent of each other rather than dependent on the patient, the models would perform better. Despite this, the models still performed very well and I can only imagine what the results would be if the data was clearer and larger. The results still show that it can accurately predict dementia in a patient which is significant in this field and there is hope that dementia can be spotted at earlier stages more often.

References

[1] Lewis F. The UK's leading alzheimer's research charity; 2022. Available from: https://www.alzheimersresearchuk.org/.

[2] Dementia. World Health Organization; 2022. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia.

[3] Xu J, Zhang Y, Qiu C, Cheng F. Global and regional economic costs of dementia: a systematic review. The Lancet. 2017;390:S47.

[4] Why Spotting Early Signs Dementia Important. Alzheimer's Society; 2020. Available from: https://www.alzheimers.org.uk/blog/why-spotting-early-signs-dementia-important.

[5] Kim J LJ. A Deep Neural Network-Based Method for Prediction of Dementia Using Big Data; 2021. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8158341/#B1-ijerph-18-05386.

[6] Zhou H, Myrzashova R, Zheng R. Diabetes prediction model based on an Enhanced Deep Neural Network - EURASIP Journal on Wireless Communications and networking. Springer International Publishing; 2020. Available from: https://jwcn-eurasipjournals.springeropen.com/articles/10.1186/s13638-020-01765-7.

[7] Leibowitz D. AI Now Diagnoses Disease Better Than Your Doctor, Study Finds. Towards Data Science. 2020 sep. Available from: https://towardsdatascience.com/ai-diagnoses-disease-better-than-your-doctor-study-finds-a5cc0ffbf32#:~:text=In\%20the\%20experiments\%2C\%20doctors\%20achieved,of\%20doctors\%20in\%20the\%20study.

[8] Dall TM, Reynolds RL, Chakrabarti R, Forte GJ, Langelier M, Wang S, et al. The physiatry workforce in 2019 and beyond, part 2: modeling results. American Journal of Physical Medicine & Rehabilitation. 2021;100(9):877-84.

[9] NHS medical staffing data analysis. British Medical Association 2022. 2022. Available from: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/workforce/nhs-medical-staffing-data-analysis#:~:text=High\%20vacancies&text=As\%20of\%20September\%202022\%2C\%20133\%2C446,vacancies\%20in\%20the\%20previous\%20quarter.

[10] Learning D. Deep learning. High-Dimensional Fuzzy Clustering. 2020.

[11] Jiang J, Chen M, Fan JA. Deep neural networks for the evaluation and design of photonic devices. Nature Reviews Materials. 2021;6(8):679-700.

[12] Latha R, Sreekanth G, Suganthe R, Selvaraj RE. A survey on the applications of Deep Neural Networks. In: 2021 International Conference on Computer Communication and Informatics (ICCCI). IEEE; 2021. p. 1-3.

[13] Tiwari S. Dementia Prediction Dataset; 2021. Available from: https://www.kaggle.com/datasets/shashwatwork/dementia-prediction-dataset.

[14] Prevalence by gender in the UK; 2018. Available from: https://www.dementiastatistics.org/statistics/prevalence-by-gender-in-the-uk/.

[15] Alzheimer's disease; 2021. Available from: https://www.nhs.uk/conditions/alzheimers-disease/#:~:text=The\%20risk\%20of\%20Alzheimer's\%20disease,under\%20the\%20age\%20of\%2065.

[16] Akoglu H. User's guide to correlation coefficients. Turkish journal of emergency medicine. 2018;18(3):91-3.

[17] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

[18] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Software available from tensorflow.org. Available from: https://www.tensorflow.org/.

[19] Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Academic pathology. 2019;6:2374289519873088.

[20] Son J, Jung I, Park K, Han B. Tracking-by-segmentation with online gradient boosting decision tree. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 3056-64.