Critical Analysis of Synthetic Alzheimer's Disease Dataset for Machine Learning

DISCLAIMER

This report analyses a generated synthetic dataset. No data used is real patient data. Neither the created classification model or represented data should be considered applicable to real Alzheimer's Disease datasets. Instead, the purpose of this investigation is to determine the purpose and validity of synthetic datasets in epidemiological modelling, and their capability in allowing users to independently practice and demonstrate machine learning and data analysis.

ABSTRACT

The primary aim of this report is to analyse the capacity of a synthetic epidemiological Alzheimer's Disease dataset in order to examine potential findings and assess it's quality, ultimately determining if it is fit for purpose and representative of it's advertised goals. This report finds that only five of the thirty-two included variables are significantly predictive in machine learning classification, and that dataset quality impairs any potential findings. Justification for this includes lack of peer-reviewed citation, error inclusion, limited variance in predictive variables, similarity of key variable data and unrepresentative data distributions. As such, this undermines the quality of models that can be generated using the dataset, and undermines potential learning applications which is in direct contrast to the advertised aims. Individuals should be wary of synthetic dataset use for epidemiological models, and in future thorough citation should be provided by model developers in order to develop stronger learning tools and models.

INTRODUCTION

Alzheimer's Disease and Synthetic Datasets

Alzheimer's Disease (AD) is the most prevalent dementia-causing disease and is characterised by the accumulation of various protein plaques within the brain that cause inflammation and neurodegeneration. It was first classified by Alois Alzheimer in 1901 (*De-Paula et al, 2012*), but dementia-related symptoms increasing with age reach back to Greco-Roman periods (*Karenberg and Förstl, 2006*). Globally, 416 million individuals have been estimated to suffer from different forms of AD in 2023 (*Gustavsson et al, 2023*), with it being the sole cause of death for 121,499 individuals in the United States alone (*The Alzheimers Association, 2023*). With a comparatively low mortality rate to other global disease focuses such as heart disease and cancer, the real concern for those affected by AD is palliative care and the subsequent impact of severe dementia symptoms. This leads to global economic strain and severe quality of life reduction for both patients and carers.

The primary symptom of AD is dementia caused by neurodegeneration. This dementia can be extensive and affect different cognitive functions across different patients and lead to progressive

memory loss, aphasia, apraxia, and agnosia. Other psychological symptoms can also manifest such as disorientation and depression. Whilst these symptoms are key clinical markers, AD can only officially be diagnosed post-mortem through presence of amyloid plaques in the brain (*Breijyeh and Karaman, 2020*). This leads to clinical underdiagnosis and difficulty estimating AD prevalence. Risk factors for AD include viral infection, head injury, environmental factors, general lifestyle, and cardiovascular disease. Amongst other risk factors, however, two stand out: age - purely as it is the most influential factor in AD development, and genetic risk (*Armstrong, 2019*) – as this is where much of the research surrounding AD since development of the amyloid cascade hypothesis has been focused. Genes that have long been a focus of Alzheimer's research include APP, PSEN1, PSEN2, and APOE (*Vilatela et al, 2012*). Because of the widespread nature of these risk factors, there is no singularly agreed upon pathology that wholly explains the pathogenesis of AD.

Due to the multitude of issues regarding diagnosis, the development of new diagnostic tools and aggregation of datasets surrounding AD has the potential to change both disease research and clinical response. As such, it is an especially relevant disease for epidemiological model development. The primary issue is that publicly available data is severely (and necessarily) limited due to patient confidentiality, leading to the creation of synthetic datasets for a variety of purposes such as data analysis and machine learning. Here, we examine one of the most popular datasets used for this purpose, and assess both the quality of the machine learning models that are producible from it, and perhaps most importantly the validity and purpose of the model.

METHODS

Data Cleaning

The initial synthetic dataset was sourced from El Kharoua, 2024. This dataset was selected due to its widespread use and representation of the wider publicly available synthetic epidemiology datasets.

The dataset required cleaning and editing to present in a suitable format for subsequent analysis. Column headers were edited to a human readable format, and correct units were added where applicable within SQL. The data was then imported into PowerBI, where coded binary values were substituted for Booleans for readability and MATLAB compatibility. Several scores that were provided in float format were rounded to integers to better represent real-world data collection methods (Physical Activity Score, Diet Quality Score, Sleep Quality Score, Mini Mental State Exam (MMSE) Score, Functional Assessment Score, Activities of Daily Living (ADL) Score). This cleaned data was then exported in .csv format and made available for further analysis.

Machine Learning Classification Analysis

The cleaned data was then imported into the MATLAB classification learner. Formatting was manually verified, and patient ID was prevented from being used as a predictor for the model. The holdout validation method was selected due to the dataset size. 18% of the data was set aside as test data, and 18% as validation data. This validation/test split was determined as optimal through a brief literature review, whereupon the formula determined by *Guyon, 1997* was used. All default

models contained within the MATLAB R2024b classification learner were then trained and tested. 32 predictors were used in model training, with 1763 training observations and 386 test observations.

Manual Data Analysis and Verification

The top five most predictive variables determined from the prior machine learning analysis were subsequently taken forward for manual verification and testing. PowerBI was used to isolate key values and generate figures for the following variables: Functional Assessment Score, ADL Score, Memory Complaints, MMSE, and Behavioural Problems. All other metrics were deemed comparatively insignificant in contributing to model predictivity, and as such were excluded from manual analysis and verification outside of examining data validity (see Appendices). Data was imported into RStudio for point-biserial analysis.

RESULTS

Dataset Overview

Prior to any in-depth analysis, this section provides a brief overview of the key epidemiological variables contained within the dataset to better frame our understanding of the machine learning model.

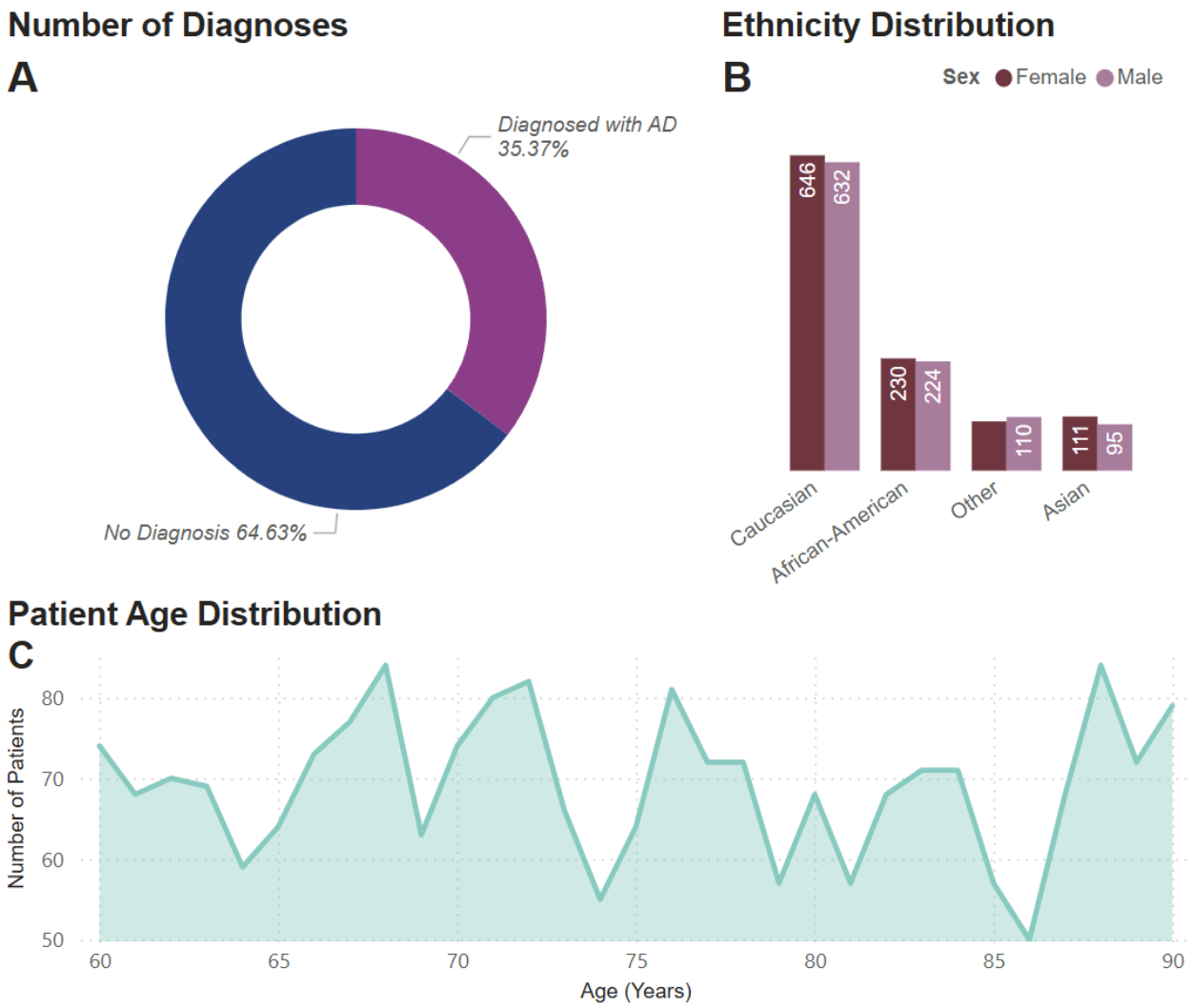

Fig. 1: A – *Elliptical chart showing the split of participants within the dataset diagnosed with AD. 760 participants were diagnosed, whilst 1389 were not.* ***B –*** *Bar graph showing the distribution of sex and ethnicity within the dataset.* ***C –*** *Line chart showing the distribution of all participant ages.*

Of the 2149 participants within the dataset, *Fig. 1A* shows that the majority (1389 participants) were not diagnosed with AD. This is an imbalance in the dataset that is frequently seen reflected in machine learning outcomes, potentially leading to overprediction of the majority group.

It is also observed that in *Fig. 1B* that the majority of the participant cohort consists of Caucasian individuals, which in a real world study may limit the conclusions that can be drawn from such a dataset. This is due to differences in risk allele prevalence across different ethnicities that arises due to the complex genetics involved in AD onset *(Vilatela et al., 2012)*. This dataset does not model differences in onset from factors such as ethnicity or sex, and so all subsequent analyses have been conducted on the entirety of the dataset.

The participant age distribution, as shown in *Fig. 1C* appears random, ranging from 60-90 with a mean of 75 ($\sigma = \pm 8.99$) falling precisely in the middle of the age range. 1061 (49.4%) participants are classed as male, whilst 1088 (50.6%) are classed as female, with no significant differences in the sex of participants occurring in different ethnicity samples.

Machine Learning Analysis

Table 1: *The below table shows each trained model with a Test Accuracy of $\geq$ 90%. Also shown is the cost for model training for both validation and testing methods, alongside test and validation accuracy.*

Model	***Accuracy % (VAL)***	***Total Cost (VAL)***	***Accuracy % (Test)***	***Total Cost (Test)***
RUSBoosted Trees	90.69767	36	91.45078	33
Medium Tree	90.18088	38	91.45078	33
Boosted Trees	89.66408	40	91.19171	34
Bagged Trees	88.63049	44	90.15544	38

Table 1 shows classification model results sorted by Test Accuracy. This data was generated in order to select the optimal model of the group. Note that both the RUSBoosted Trees and Medium Tree models offer identical Test Accuracy. However, given the marginally lower validation training costs, higher validation accuracy, and superior F_1 score, RUSBoosted trees was selected as the optimal model for subsequent analyses.

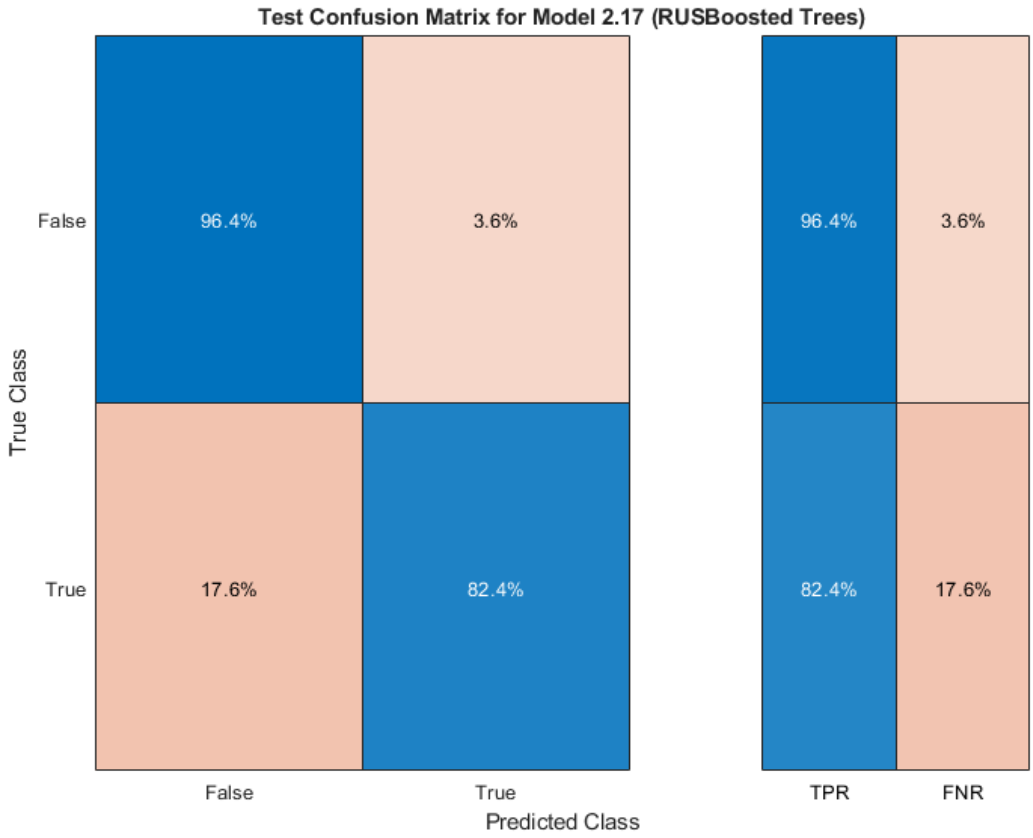

Fig. 2**:** *Confusion matrix for RUSBoosted Trees. Shows true positive rates (TPR) and false negative rates (FNR) for both true and false AD diagnoses.*

The RUSBoosted Trees model in *Fig. 2* has a weighted F_1 score of 91.3%, demonstrating a relatively high accuracy with an overall error rate of 8.5% against test data. Macro and micro F1 scores were 90.4% and 91.5% respectively. Despite this, the model has a high false negative rate (FNR) of 17.6%, resulting in an underprediction of diagnoses. Potential reasoning for this is addressed in the discussion.

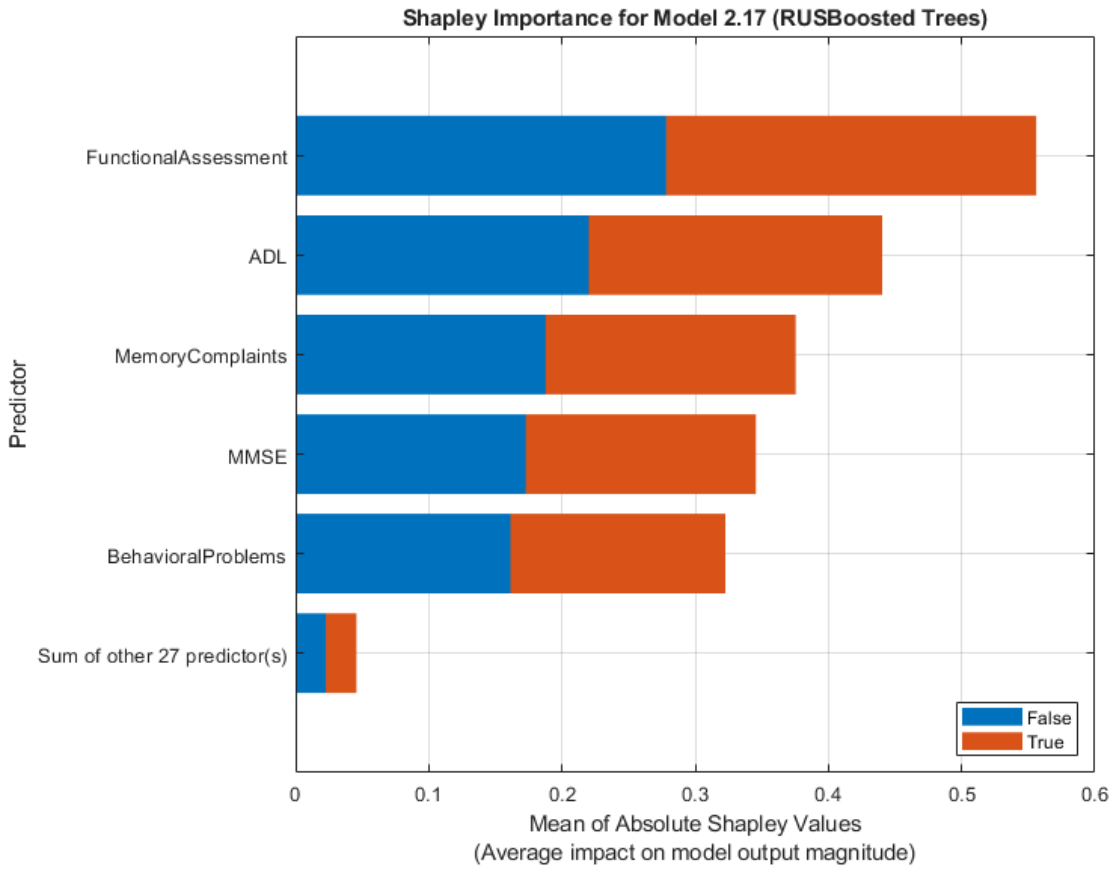

Fig. 3: *Shapley Importance chart showing the average impact of predictors on model output magnitude. The top five contributing predictors are shown individually, whilst the other twenty-seven are summed as the final bar.*

The Shapley Importance chart in *Fig. 3* shows us that the predictors within the dataset are highly unevenly distributed, with the model relying largely on only five of the thirty-two variables. The fact that the sum of the other twenty-seven predictors is so low is unusual, and is likely a result of the method through which the synthetic dataset was generated. The even predictivity between true and false is due to the binary nature of the diagnosis variable.

Manual Analysis

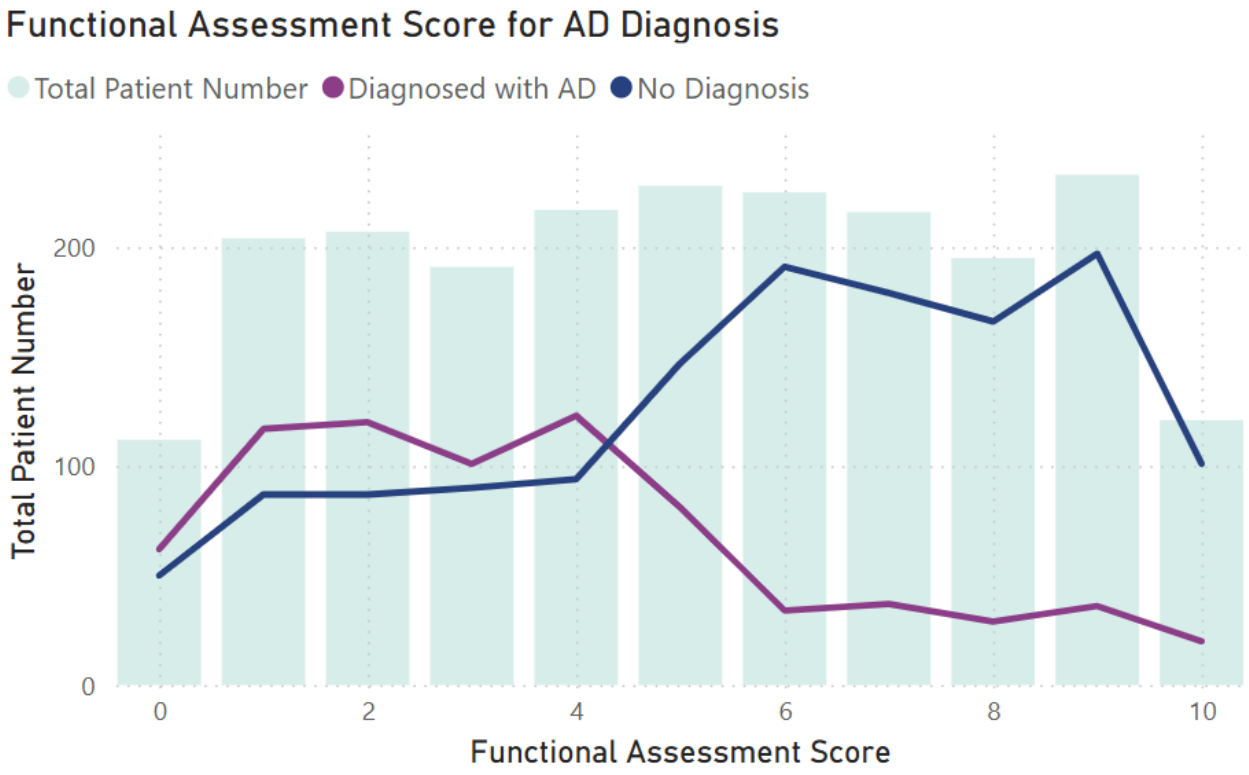

Fig. 3: *Line graph showing the relationship between the Functional Assessment score and the number of participants diagnosed with AD. A higher functional assessment score means fewer AD symptoms present (N.B. this is not the case in real-world FA scores). Columns shows the overall distribution of functional assessment scores across the dataset.*

The mean functional assessment score for all participants is 5.09 ($\sigma = \pm 2.93$), spanning the full range of zero to ten. Those diagnosed with AD had a mean functional assessment score of 4.0 ($\sigma = \pm 3.0$), whilst those without a diagnosis had a mean functional assessment score of 6.0 ($\sigma = \pm 3.0$). The correlation coefficient calculated through point biserial analysis for this variable is -0.361 ($p < 2.2e^{-16}$, *95CI = -0.397, -0.324*).

As shown in *Fig. 3*, the functional assessment score appears predictive of disease outcome at values greater than four. This is expected, and reinforces what was seen in *Fig. 3* of the machine learning Shapley Importance chart. *Fig. 3* also shows the overall distribution of functional assessment scores is fairly even across the participant population, with significant reductions at zero and ten. There is a large problem within the dataset that is easily identifiable here, as in reality, the functional assessment score is given in integers from one to seven, or in certain cases from one to sixteen *(Sclan and Reisberg, 1992)*. Therefore, even if scaling and rounding the data appropriately, none of the population should have a score of zero. The participants showing no diagnosis of AD having the higher score is also incorrect, as the real score measures this with seven (or 7f) as the highest level of cognitive impairment *(Sclan and Reisberg, 1992)*. This, amongst other serious issues within the dataset will be comprehensively addressed in the discussion.

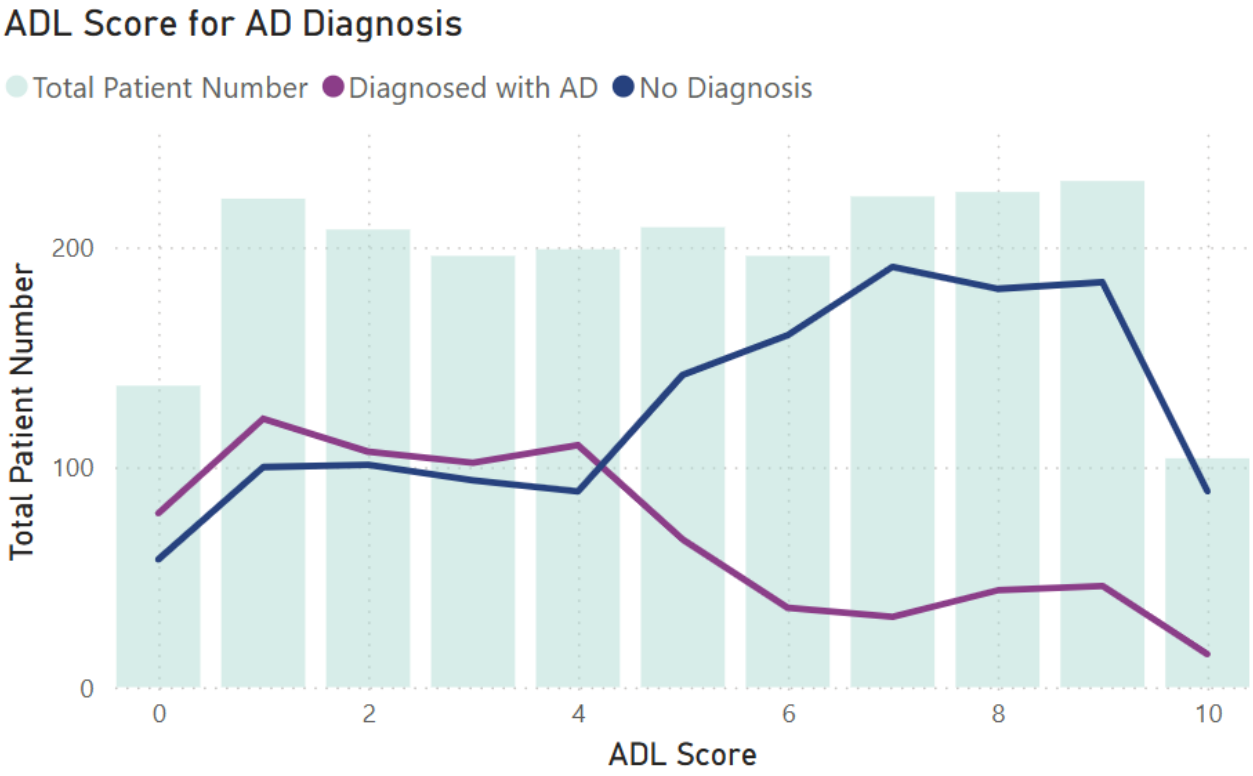

Fig. 4: *Line graph showing relationship between the ADL score and the number of participants diagnosed with AD. A higher ADL score means fewer AD symptoms present. Columns shows the overall distribution of ADL scores within the dataset.*

The mean ADL score for all participants is 4.99 ($\sigma = \pm 2.99$), ranging from zero to ten. Those diagnosed with AD had a mean ADL score of 4.0 ($\sigma = \pm 3.0$), whilst those without a diagnosis had a mean ADL score of 6.0 ($\sigma = \pm 3.0$). Note that the mean and standard deviations are identical for participant groups to the above functional assessment score, which is addressed in the discussion. The correlation coefficient calculated through point biserial analysis for this variable is -0.331 ($p < 2.2e^{-16}$, *95CI = -0.368, -0.293*).

Once again, in *Fig. 4* we see a similar pattern arising as in *Fig. 3*, with scores significantly increasing or reducing when greater than four. The distribution is also relatively even in *Fig. 4,* but with a greater population present with a score of zero, and fewer with a score of ten. This data also shares the same issues as the aforementioned Functional Assessment score, as increased scores within this dataset appear to model reduced AD symptoms when in reality the opposite is true. Another difficulty arises as multiple ADL tests exist, and the dataset fails to specify which was used. It appears that these scores have also been arbitrarily scaled from an unknown range in real-world tests to a zero to ten range *(Eto et al., 1992) (Edemekong et al., 2025).*

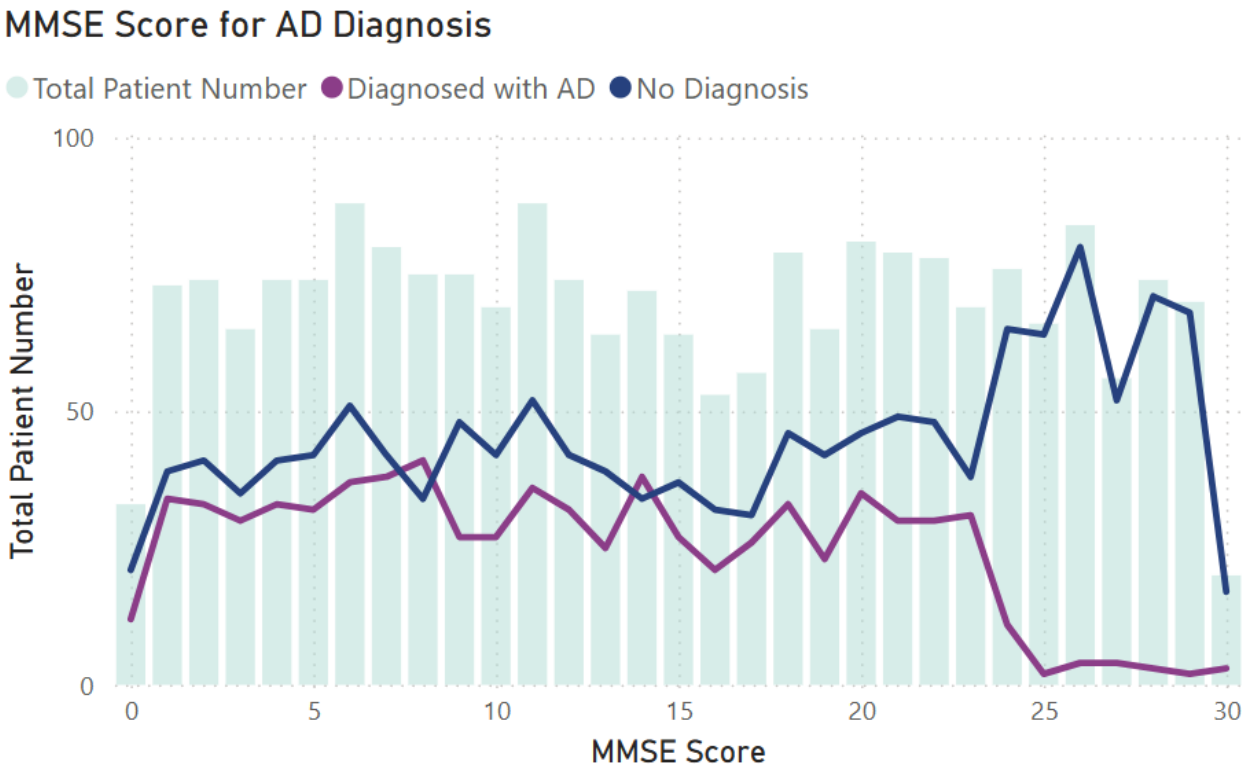

Figure 5: *Line graph showing relationship between MMSE score and the number of participants diagnosed with AD. A MMSE score means fewer AD symptoms present, although this is not the case in reality. Columns shows the overall distribution of MMSE scores within the dataset.*

The mean MMSE score for all participants is 14.67 ($\sigma = \pm 8.61$), ranging from zero to thirty. Those diagnosed with AD had a mean MMSE score of 12.0 ($\sigma = \pm 7.0$), whilst those without a diagnosis had a mean MMSE score of 16.0 ($\sigma \pm 9.0$). The correlation coefficient calculated through point biserial analysis for this variable is -0.236 ($p < 2.2e^{-16}$, *95CI = -0.276, -0.196).*

As seen in *Figure 5*, the MMSE score is scaled correctly – to thirty points instead of ten, and the cutoff point where we see a large uptick in AD diagnoses is largely concurrent with literature *(Arevalo-Rodriguez et al., 2021)* at approximately 26 points. Likely due to the increase in possible variable values, the distribution of MMSE score appears noisier in *Fig. 5* than our previous two examples, but is still largely random. Somewhat unexpectedly, zero participants achieved a score of thirty. In real-world data, this is unlikely but could potentially be accounted for in fringe cases by considering the age of the sample set and related comorbidities.

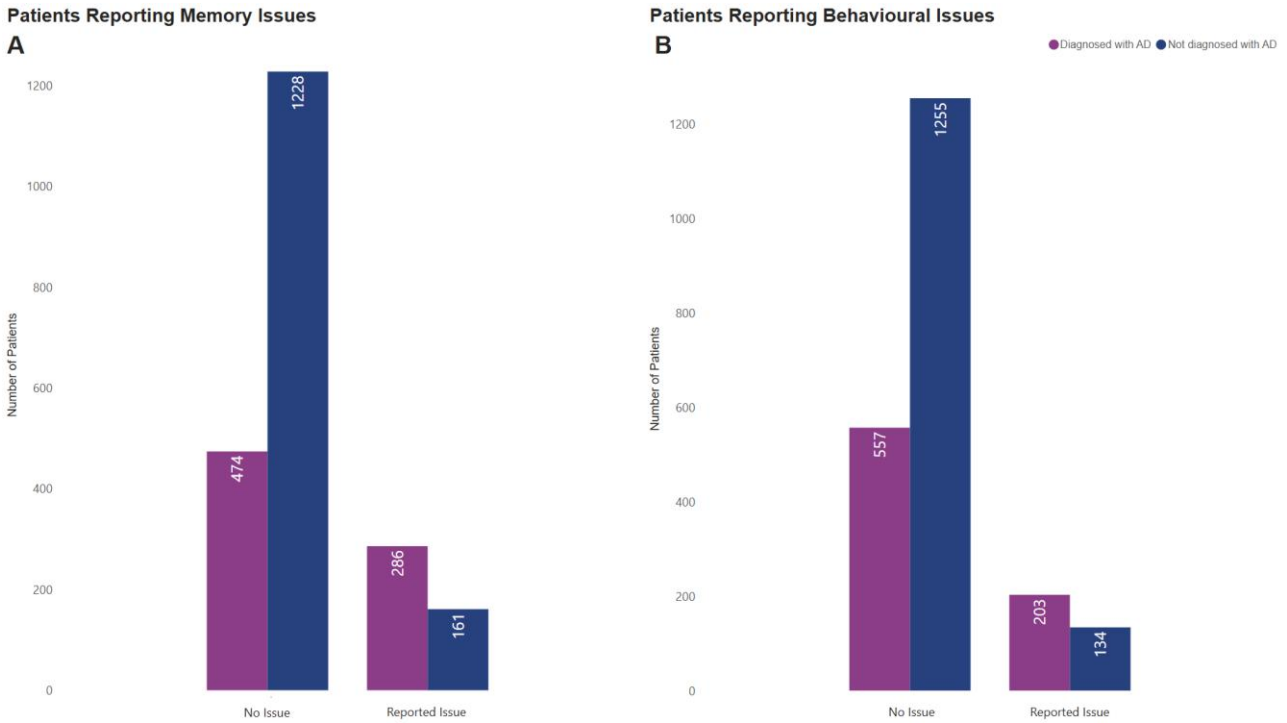

Fig. 6: A – *Bar graph showing the number of participants that report memory issues and their subsequent diagnoses of AD.* ***B –*** *Bar graph showing the number of participants that report behavioural issues and their subsequent diagnoses of AD.*

Both *Fig. 6A* and *Fig. 6B* show very similar patterns regarding the each of the Boolean variables from the top predictors. Each shows that participants reporting memory and behavioural issues are more likely to be diagnosed with AD. It is worth noting that a large proportion of participants diagnosed with AD do not report having these issues, and this is potentially a causative factor in the high false negative rates observed in *Fig. 2*.

DISCUSSION

Model Analysis and Results

The RUS Boosted Trees model appears to be the best model possible given the limitations of the dataset. As the validation and test accuracies are extremely similar, it does not appear as though the model is over- or under-fitting. It also has a relatively strong test accuracy of 91.5%, but the confusion matrix showed a high false negative rate of 17.6% in the diagnosed participant class. This is likely due to several factors, such as the imbalance in the dataset towards participants without a diagnosis and the data shown in *Fig. 6A* and *Fig. 6B*, where a large proportion of individuals diagnosed with AD do not appear to report memory or behavioural issues. It did, however, boast an even higher accuracy when predicting for individuals without an AD diagnosis, with a 96.4% TPR and only a 3.6% FNR.

From manual analysis of the data, we can see that professional assessments for AD initially appear to be indicators for disease onset. Closer examination of correlation coefficients reveals statistically significant ($p < 2.2e^{-16}$) but weak correlations for all three assessments. Despite the relatively low correlations, the extremely low p values could provide a reason that the machine learning model predicts so heavily using these variables.

Regarding limitations, we must also acknowledge that much of this data is derived from the Caucasian ethnicity, which would theoretically limit it's applications to other ethnicities due to the complex genetic factors associated with AD, although those factors are not modelled within this dataset.

Dataset Quality Analysis

This dataset, like many other publicly available synthetic datasets, has a series of weaknesses. The first of these is the lack of accuracy when modelling factors that contribute to AD risk. According to this dataset, only five of the thirty-two factors meaningfully predict for risk of AD. This entirely excludes key factors frequently used by clinicians and researchers such as family disease history, age, and other critical symptoms such as anhedonia and forgetfulness *(Armstrong, 2019)*. This low number of predictive values could also contribute to the high FNR we see in *Fig. 2,* although real-world data would have far more noise and outliers and as such would reduce overall model accuracy.

Unusually, the mean and standard deviation scores for the functional assessment and ADL score data are identical. Whilst this could be an anomaly, I find this highly unlikely given the synthetic origin of the dataset. This also provides an explanation as to why both are ranked highly in *Fig. 3* for machine learning predictivity.

Another significant issue is that the dataset presents all the group of individuals diagnosed with AD as homogenous, with no thought to the multitude of diagnostic subgroups and causative pathologies. At a basic level, genetic versus familial AD should be a consideration when modelling the disease, or a dataset should aim to model only one of these subsets. Whilst the data here infers that familial history has little to no impact on disease outcome, we are not provided any

genetic markers or clarification for which (if any) subtype this dataset aims to model. In addition, there is no modelling of comorbidities or correlation between other variables – for example age has no significant impact on blood cholesterol results (*Fig. A1)*, where typically an upwards trend correlated with age would be expected *(Kreisberg, Kasim., 1987)*.

The distribution of data within each variable also appears even or random, but without the noise you would expect from a real-world dataset. When examining *Fig. 3, 4* and *5,* all of the data is mostly uniformly distributed, with lower participant numbers at the very extremes. Whilst this approximates a normal distribution, you would typically expect to see patients fall into majority groups at certain score cutoff points depending on disease progression *(Bleecker et al., 1988)*.

Perhaps one of the largest issues within the dataset is the choice of variable value ranges. Examples include the above Functional Assessment and ADL scores ranging from zero to ten, when in reality the Functional Assessment score would range from one to seven (S*clan & Reisberg, 1992*), and the ADL score appears derived from an entirely unknown value *(Edemekong et al., 2025)*. The presence of participants who apparently have an ADL score of zero implies that this data has not been scaled, instead randomly generated. More concerning is the fact that the Functional Assessment and ADL scores are inverted, as the dataset purports that a higher score is correlated to a reduction in AD diagnosis when in reality the opposite is true.

If the above issues were clearly stated as limitations on the dataset page, then their impact would be minimal as students and researchers attempting to utilise the data for machine learning would be alerted to the impact they could have on results. At the time of writing, however, this is not the case and could easily lead to data misinterpretation.

Despite these issues with the dataset, the question remains as to it's potential. As demonstrated in this report, it certainly allows the generation of a relatively accurate machine learning model, enabling individuals to demonstrate relevant modelling skills. Given the lack of availability of epidemiological data surrounding AD to those outside of research, many students are limited to datasets such as this. The potential for datasets like this is enormous, and could significantly assist in enabling independent learning, but without correct understanding of the subject matter and lack of cited peer-reviewed material to evidence the model, they are more likely to lead to misunderstanding of the central topic.

Conclusion

Overall, this dataset is concerningly inaccurate for modelling AD diagnosis and onset for the multitude of reasons outlined above. In some cases, this is understandable as the apparent purpose of the data is not to accurately model disease onset, but instead to provide a learning tool for individuals to expand their understanding of data analysis and machine learning. Despite this, there are certain foundational errors present here that I believe undermine that aim, and prevent prospective students and analysts from returning with any significant understanding of the disease, rendering it incredibly difficult to reconcile the data with real-world observations. This is especially important when considering that epidemiological data is not publicly available outside of research groups, and datasets much like this one are all many students have available. This is even further complicated by the dataset being advertised as a way for "researchers and data

scientists […] to explore factors associated with Alzheimer's, develop predictive models, and conduct statistical analyses," and "[explore] the complex interplay of factors contributing to Alzheimer's Disease" (*El Kharoua, 2024*), when it does not reflect a clear understanding of the disease. This dataset was selected as it was the most popular set on the website 'Kaggle' with 19.3k downloads at the time of writing, and has yet to provide any citation or reference for the data source. Upon cursory examination of several other epidemiological datasets, similar problems seem to arise, and those using synthetic datasets for machine learning projects without an understanding of the subject material will ultimately be misled. This has clearly highlighted the need for synthetic datasets and models to exist for learning and development purposes, but for the data within to be thoroughly evidenced from peer-reviewed sources, instead of randomly generated for the purpose of rudimentary model creation.

REFERENCES

Alonso Vilatela ME, López-López M, Yescas-Gómez P. 2012. *Genetics of Alzheimer's disease*. Arch Med Res. **43**(8):622-31. doi: 10.1016/j.arcmed.2012.10.017.

Arevalo-Rodriguez I, et al. 2021. *Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI)*. Cochrane Database Syst Rev. Jul 27;**7**(7):CD010783. doi: 10.1002/14651858.CD010783.pub3.

Armstrong, RA. 2019. *Risk factors for Alzheimer's disease*. Folia Neuropathol. **57**(2):87-105. doi: 10.5114/fn.2019.85929.

Bleecker ML, Bolla-Wilson K, Kawas C, Agnew J. 1988. *Age-specific norms for the Mini-Mental State Exam*. Neurology. Oct;**38**(10):1565-8. doi: 10.1212/wnl.38.10.1565.

Breijyeh, Z., Karaman, R. 2020. *Comprehensive Review on Alzheimer's Disease: Causes and Treatment*. Molecules. **25**(24):5789. doi: 10.3390/molecules25245789.

De-Paula VJ, Radanovic M, Diniz BS, Forlenza OV. 2012. *Alzheimer's disease*. Subcell Biochem. **65**:329-52. doi: 10.1007/978-94-007-5416-4_14.

Edemekong PF, et al. 2025. *Activities of Daily Living*. Treasure Island (FL): StatPearls Publishing. Available at: [https://www.ncbi.nlm.nih.gov/books/NBK470404/].

El Kharoua, R. 2024. *Alzheimer's Disease Dataset*. Kaggle. doi: 10.34740/kaggle/dsv/8668279

Eto F, et al. 1992. *Comprehensive activities of daily living (ADL) index for the elderly*. Nihon Ronen Igakkai Zasshi. Nov;**29**(11):841-8. doi: 10.3143/geriatrics.29.841.

Gustavsson, A., et al. 2023. *Global estimates on the number of persons across the Alzheimer's disease continuum*. Alzheimers Dement. **19**(2):658-670. doi: 10.1002/alz.12694.

Guyon, I. 1997. *A Scaling Law for the Validation-Set Training-Set Size Ratio*. AT&T Bell Lab. [Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=452e6c05d46e061290fefff8b46d0ff161998677].

Karenberg, A., Förstl, H. 2006. *Dementia in the Greco-Roman world*. J Neurol Sci. 244(1-2):5-9. doi: 10.1016/j.jns.2005.12.004.

Kreisberg RA, Kasim S. 1987. *Cholesterol metabolism and aging*. Am. J. Med. Jan 26;**82**(1B):54-60. doi: 10.1016/0002-9343(87)90272-5.

Sclan SG, Reisberg B. 1992. *Functional assessment staging (FAST) in Alzheimer's disease: reliability, validity, and ordinality*. Int Psychogeriatr. **4** Suppl. 1:55-69. doi: 10.1017/s1041610292001157.

The Alzheimer's Association. 2023. *Alzheimer's disease facts and figures*. Alzheimers Dement. **19**(4):1598-1695. doi: 10.1002/alz.13016.

APPENDICES

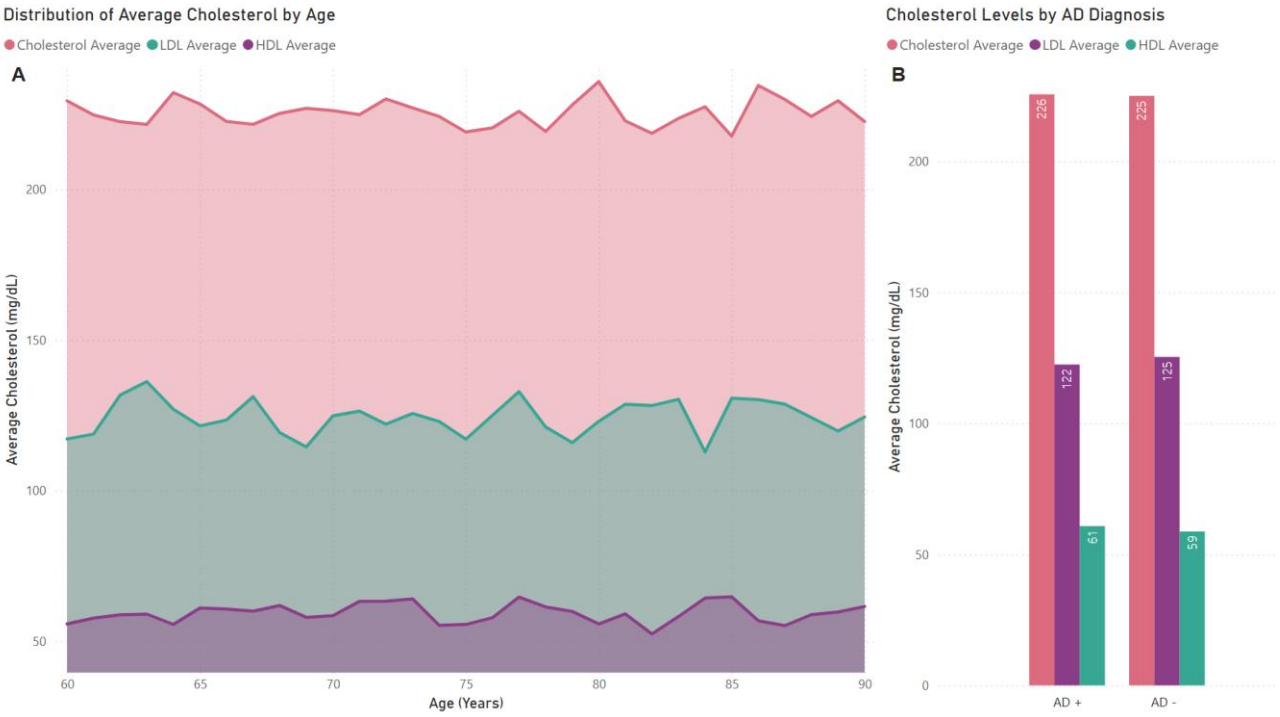

Fig. A1: A - *Line graph illustrating the distribution of average total cholesterol by participant age ($\bar{x}$ = 225.20, σ = 42.53).* ***B –*** *Clustered column chart showing the similarity of cholesterol levels by AD diagnosis.*