School of Computer Science

University of Auckland

New Zealand

Using deidentified CT brain scan data to classify dementia in older people.

Research Project: SCI077

Supervisor: Professor Gillian Dobbie

Elijah Hayward

February 2025

1 Table of Contents

2 Introduction

2.1 Background

Dementia is a major global health challenge with millions of people affected worldwide. According to the world health organisation more than 55 million people are currently living with dementia, with every year there being nearly a further 10 million new cases [4]. Dementia is a neurocognitive mental disorder that is a general term covering several diseases causing disability and dependency in adults. Early diagnosis of dementia is crucial for managing the disease and planning long-term support strategies. Traditional diagnosis relies on clinical assessments, psychological tests and imaging studies such as CT brain scans. However, these methods can be time-consuming and subject to human interpretation. In recent years machine learning techniques have emerged as powerful tools for improving diagnostic accuracy and efficiency in various medical fields. This project investigates the performance of machine learning models in diagnosing dementia, with a particular focus on the impact of including CT brain scan data and assisted residential care data in the classification process.

This research utilised an extensive dataset containing medical records from Middlemore hospital. The goal is to ascertain the importance of CT brain scan data and assisted residential care data in identifying people living with dementia. We do this by measuring the performance of a classification model both with and without the CT brain scan data and assisted residential care data on diagnosing dementia in patients.

The study used an open sourced software library called Extreme Gradient Boosting (XGBoost) in python to create the classification model. This was selected due to its robustness and accuracy in classification tasks. XGBoost is particularly well suited for medical data analysis due to its ability to handle large datasets with complex interactions between features. The performance of the model was evaluated across a number of metrics such as accuracy, sensitivity, specificity, area under curve (AUC) and f1 score.

By comparing the results of various models created throughout this study, valuable insights are gained into the impact of CT brain scan data and assisted residential care data on the performance of machine learning models in diagnosing dementia. Ultimately, this project aims to contribute to ongoing efforts to enhance the accuracy and efficiency of early dementia diagnosis through the application of machine learning.

3 Related Work

3.1 Machine Learning for Medical Diagnosis

With a swiftly aging world population the rising prevalence of dementia has driven a vast amount of research into computational approaches for early detection and diagnosis. As traditional methods of diagnosis are time consuming and prone to human interpretation, various machine learning (ML) methods have been investigated as a means to streamline and assist in the fast and accurate diagnosis of dementia.

The three main machine learning algorithms for medical diagnosis and prediction to date are support vector machine (SVM), Ensemble methods, and Convolutional Neural Networks (CNN) [4]. Convolutional Neural Networks have been shown to be able to obtain a high accuracy in diagnosis through the use of MRI (Magnetic Resonance Imaging), however it

requires a large amount of image data for the model to be trained and learned deeply [5] and MRI image data wasn't in our dataset. Another study with similar methodology and dataset which investigated kidney disease diagnosis found an Ensemble method called XGBoost (Extreme Gradient Boosting) to perform the best when compared across a number of metrics against SVM [3], hence giving us confidence to move forward with XGBoost as our machine learning model for diagnosing dementia for our given dataset.

3.2 Applications of XGBoost in Healthcare

XGBoost has emerged as a leading algorithm for structured data analysis in healthcare due to its computational efficiency, scalability, and superior performance in classification tasks. Its ensemble learning approach minimises overfitting and adapts flexibly to diverse datasets. Many studies have used XGBoost to assist in medical diagnosis, with applications ranging from kidney disease diagnosis [3] to dementia risk prediction [2].

One study investigated a dementia risk prediction model based on XGBoost, incorporating derived variables to enhance prediction accuracy. The study highlighted the importance of feature selection and hyperparameter tuning for the functionality of an XGBoost model and achieved an accuracy of 85.61% and an F1-score of 79.28% [2]. Another study utilised XGBoost to create a classification model for diagnosing Alzheimer's dementia and again found that with an optimised set of hyperparameters that the model was able to gain a high accuracy of 81% [6].

3.3 Feature Selection and Hyperparameter Tuning

The two main ways an XGBoost model can be optimised for best performance are through feature selection and hyperparameter tuning. Effective feature selection can be crucial for reducing model complexity and computational time, enabling the model to not be clustered by features which have little to no impact on performance. While hyperparameter tuning enables the best set of hyperparameters to be chosen for the specific problem allowing for the model to achieve as high an accuracy as possible.

One commonly used feature selection method is Spearman Rank Correlation which seeks to find the correlation between a feature and the target feature, by calculating a correlation coefficient between the two and only using features which meet a specific correlation threshold [7]. Another widely adopted feature selection technique is Recursive Feature Elimination (RFE). RFE works by iteratively training the model and ranking features based on their importance scores. At each iteration, the least important features are removed, and the model is re-trained with the reduced feature set. This process continues until the desired number of features is reached. This feature selection method can reduce the dimensionality of the feature set while retaining the most important features for classification [8]. In this study, models were developed using both Spearman Rank Correlation and Recursive Feature Elimination to analyse the impact of these feature selection methods on model accuracy.

Hyperparameter tuning optimises XGBoost models performance relative to some metric, often being accuracy. The most important hyperparameters include the number of estimators (n_estimators), maximum tree depth (max_depth), learning rate (learning_rate), column sampling (colsample_bytree) and row sampling (subsample). Methods like grid search and random search are commonly used for tuning, though extensive optimisation can be computationally expensive [2], [3]. In this project, we focused on the most key parameters for

XGBoost to ensure efficient optimisation without excessive computational costs, and employed the grid search method to do this.

3.4 Summary

XGBoost has proven to be a powerful tool in medical diagnostics, with a wide range of applications across the industry. Feature selection methods streamline the dataset by retaining only the most relevant information, while hyperparameter tuning optimises the model's parameters for the specific problem. This research builds on these techniques by investigating the impact of CT brain scan and assisted residential care data on dementia prediction models, providing insight into its potential to enhance dementia diagnosis through machine learning.

4 Methodology

4.1 Pre-processing

For this project the provided dataset had already undergone an initial cleaning process, and there were minimal missing data points. The majority of the missing data was 13 of the 165 patients had no CT data, along with some other patients not having data for things such as a specific blood test. A list of the features which were used before any feature selection was conducted can be seen in table 1.

However the provided data from the hospital was not given in this format, it was instead categorised by the type of dataset it belonged to with its own excel sheet for each dataset. Hence a python script was developed to extract the necessary features from the various sheets and get them into a format such that each patient had one unique row in the data frame with all the associated feature values. This script provided an automated and repeatable extraction process which enabled the data to be manipulated into a compatible form to be used with the XGBoost functions.

Dataset	Features	Definition
CT	TBV	Total Brain Volume
CT	TIV	Total Intracranial Volume
CT	TBV/TIV	TBV divided by TIV
CT	L_HC	Left Hippocampus
CT	R_HC	Right Hippocampus
CT	Combined_HC	L_HC plus R_HC
CT	HC/TIV	HC divided by TIV
Inpatient	Total length of stay	Sum of all time that the patient was in the hospital
Inpatient	Number of admissions	Total number of admissions to the hospital
ED	Number of ED visits	Total number of visits to the emergency department
Outpatient	Number of appointments per specialist	Sum of appointments per specialist
Outpatient	Attendance Data	Sum of attendance statuses from appointments
Contact	Contact Status	Sum of contact statuses from scheduled contacts
Delirium	Total number of CAMs	Sum of confusion assessment methods performed
Delirium	Number of admissions with CAMs	Sum of admissions where a CAM was performed
Delirium	Positive delirium cases	Sum of CAMs where they were positive for delirium
Blood Tests	Number of blood tests	Total number of blood tests

Blood Tests	Number of abnormal blood tests	Total number of blood tests with abnormal results
Blood Tests	Blood Test result	Most recent result for each type of blood test
Pharmacy	Therapeutic Group	Number of prescriptions received per therapeutic group
ARC	Service category day count	Total number of service days spent in each of the different service categories

TABLE 1: FULL LIST OF FEATURES

4.2 Feature Selection

Two types of feature selections were performed and then models created with both sets of selected feature sets, these two methods were the Spearman Rank Correlation and Recursive Feature Elimination.

4.2.1 Spearman Rank Correlation

Spearman rank correlation works to calculate a correlation coefficient between the target variable and each feature in the dataset, then a minimum correlation value is chosen and features which don't meet this requirement are removed from the dataset. In this study the features were found to have relatively low correlations when compared to the target variable (dementia) on their own and hence a threshold coefficient value of 0.15 was chosen. This lead to 29 features of the original 166 being chosen, which can be seen in table 2.

4.2.2 Recursive Feature Elimination

Recursive feature elimination works iteratively by training the model and at each iteration removing the least important feature until you have gotten down to some predetermined number of features to select. In this study several different values for the number of features to select were trialled, but using 20 features was decided on as it was around the point at which any fewer features started to harm the models accuracy. Again the 20 selected features from this method can be seen in table 2.

Spearman Rank Correlation Features		**Recursive Feature Elimination Features**	
TBV/TIV - CT	L_HC - CT	TIV - CT	TBV/TIV - CT
Combined_HC - CT	HC/TIV - CT	L_HC - CT	R_HC - CT
Occurred - Contact	Hospital Stay Days - ARC	Combined_HC - CT	Total Length of Stay - Inpatient
Rest home Stay Days - ARC	Attended - Outpatient	Occurred - Contact	Eosinophils - Blood Tests
Other - Outpatient	Respiratory Medicine - Outpatient	Neutrophils - Blood Tests	TSH - Blood Tests
Dermatology - Outpatient	Gynaecology - Outpatient	Vitamin B12 - Blood Tests	Attended - Outpatient
Acute Allied Health - Outpatient	Stroke - Outpatient	Other - Outpatient	Plastic Surgery - Outpatient
Rheumatology - Outpatient	Muscle Relaxants and Related Agents - Pharmacy	Acute Allied Health - Outpatient	Health Older Person - Outpatient
Viral Vaccines - Pharmacy	Treatments for Dementia - Pharmacy	Vitamins - Pharmacy	Treatments for dementia - pharmacy
Chemotherapeutic Agents - Pharmacy	Antiulcerants - Pharmacy	Non-Steroidal Anti-Inflammatory Drugs - Pharmacy	Long-Acting Beta-Adrenoceptor Agonists - Pharmacy

Inhaled Corticosteroids - Pharmacy	Non-Steroidal Anti-Inflammatory Drugs - Pharmacy	-	-
Antispasmodics and Other Agents Altering Gut Motility - Pharmacy	Antipsychotic - Pharmacy	-	-
Oestrogens - Pharmacy	Antifungal - Pharmacy	-	-
Antidiarrheals and - Pharmacy	Other Ontological Preparations - Pharmacy	-	-
Other Skin Preparations - Pharmacy	-	-	-

TABLE 2: SELECTED FEATURES FROM FEATURE SELECTION

4.3 Models

The two main objectives from this study were to investigate the effect of removing the CT (Computed Tomography) and ARC (Assisted Residential Care) data from the model. In order to do this it led to 16 sub models being created which can be seen in table 3. Each of these models had an underlying base model and then some additional features removed from the model. The four base models were Original, Cleaned, Spearman and RFE, and then they either had nothing removed, CT data removed, ARC data removed, or CT and ARC data removed. All the descriptions for this can be seen in table 3.

Model	Description
Original	Full dataset
Original no CT	Full dataset, CT data removed
Original no ARC	Full dataset, ARC data removed
Original no CT or ARC	Full dataset, CT and ARC data removed
Cleaned	Missing data patients removed
Cleaned no CT	Missing data patients and CT data removed
Cleaned no ARC	Missing data patients and ARC data removed
Cleaned no CT or ARC	Missing data patients and CT and ARC data removed
Spearman	Spearman feature selected, missing data patients removed
Spearman no CT	Spearman feature selected, missing data patients and CT data removed
Spearman no ARC	Spearman feature selected, missing data patients and ARC data removed
Spearman no CT or ARC	Spearman feature selected, missing data patients and CT and ARC data removed
RFE	RFE feature selected, missing data patients removed
RFE no CT	RFE feature selected, missing data patients and CT data removed
RFE no ARC	RFE feature selected, missing data patients and ARC data removed
RFE no CT or ARC	RFE feature selected, missing data patients and CT and ARC data removed

TABLE 3: LIST OF ALL MODELS

4.4 Hyperparameter Tuning

Hyperparameter tuning is when you find the set of hyperparameters for your model which allow for the best performance by some given metric, in this studies case we chose that metric to be accuracy. The chosen method to conduct this hyperparameter tuning was grid search, this works by taking possible combinations of parameters and running the model many times successively to see which combination of parameters gave the best accuracy. As we had 4 base models 4 different sets of hyperparameters were found using the grid search method, these can be seen in table 4.

Models	n_estimators	max_depth	learning_rate	Colsample_bytree	subsample
Original	75	5	0.08	0.50	0.55
Cleaned	80	6	0.08	0.70	0.75
Spearman	105	7	0.09	0.55	0.65
RFE	95	5	0.06	0.65	0.75

TABLE 4: HYPERPARAMETERS FOR EACH MODEL

These 5 hyperparameters in table 4 were deemed to be the most important to be optimised [2], [3]. We chose not to optimise additional hyperparameters, as including more in the optimization process significantly increases computational costs. These hyperparameters were applied to all models which related to that specific sub model.

4.5 K-Fold Cross Validation

In this study we used k-fold cross-validation as part of the model's learning process to enhance its generalisability. K-fold cross validation is a data segmentation technique that divides the entire dataset into k subsets (folds). In each iteration k-1 subsets are used as the training set, while the remaining subset as the testing set. This process is repeated k times, ensuring each subset is used as the testing set exactly once. By testing the model across all subsets of the dataset, k-fold cross validation provides a more comprehensive assessment of its performance and helps reduce overfitting to a particular subset of the data. For this study we selected 5 folds to balance the needs of our dataset with the available computational resources [2].

5 Results

Following our initial results, we identified that the feature Treatments for Dementia from the Pharmacy dataset had a particularly strong impact on predicting whether a patient was diagnosed with dementia. To evaluate its influence further, we conducted another round of modelling with the same setup, excluding the Treatments for Dementia feature. As a result, the analysis has been divided into two sections, one including the Treatments for Dementia feature and another without it.

5.1 Evaluation of Metrics – With Treatments for Dementia

After the 16 different models were set up with their associated feature sets and hyperparameters they were run using XGBoost and the following metrics calculated. For each model and metric the average value across the 5 folds is given along with the standard deviation across the 5 folds. Table 5 summarises the performance considering accuracy, sensitivity, specificity, area under the curve (AUC), and F1-score.

	Accuracy	Sensitivity	Specificity	AUC	F1
Original	0.733 ± 0.059	0.835 ± 0.072	0.609 ± 0.116	0.791 ± 0.074	0.781 ± 0.043
Original no CT	0.648 ± 0.071	0.769 ± 0.085	0.492 ± 0.073	0.701 ± 0.093	0.712 ± 0.057
Original no ARC	0.709 ± 0.078	0.793 ± 0.125	0.608 ± 0.139	0.750 ± 0.087	0.755 ± 0.065
Original no CT or ARC	0.648 ± 0.031	0.771 ± 0.094	0.494 ± 0.031	0.728 ± 0.071	0.712 ± 0.030
Cleaned	0.677 ± 0.146	0.791 ± 0.185	0.499 ± 0.137	0.711 ± 0.131	0.737 ± 0.135
Cleaned no CT	0.706 ± 0.071	0.791 ± 0.162	0.572 ± 0.144	0.733 ± 0.089	0.755 ± 0.087
Cleaned no ARC	0.685 ± 0.080	0.779 ± 0.134	0.538 ± 0.145	0.721 ± 0.111	0.742 ± 0.085
Cleaned no CT or ARC	0.665 ± 0.083	0.768 ± 0.170	0.488 ± 0.146	0.721 ± 0.085	0.725 ± 0.095
Spearman	0.720 ± 0.098	0.740 ± 0.135	0.685 ± 0.169	0.724 ± 0.100	0.753 ± 0.107
Spearman no CT	0.692 ± 0.060	0.746 ± 0.138	0.627 ± 0.125	0.745 ± 0.036	0.741 ± 0.060
Spearman no ARC	0.720 ± 0.098	0.781 ± 0.121	0.643 ± 0.192	0.746 ± 0.081	0.766 ± 0.094
Spearman no CT or ARC	0.692 ± 0.046	0.757 ± 0.102	0.605 ± 0.126	0.732 ± 0.040	0.747 ± 0.031
RFE	0.713 ± 0.087	0.814 ± 0.085	0.556 ± 0.174	0.813 ± 0.086	0.772 ± 0.079
RFE no CT	0.769 ± 0.077	0.866 ± 0.094	0.627 ± 0.103	0.835 ± 0.060	0.817 ± 0.066
RFE no ARC	0.713 ± 0.087	0.814 ± 0.085	0.556 ± 0.174	0.813 ± 0.086	0.772 ± 0.079
RFE no CT or ARC	0.769 ± 0.077	0.866 ± 0.094	0.627 ± 0.103	0.835 ± 0.060	0.817 ± 0.066

TABLE 5: MODEL RESULTS MEAN AND STANDARD DEVIATION

The model which stands out as having the best performance across all metrics is the model with the RFE feature selection and no CT data. The models RFE without CT and RFE without CT and ARC are the same as the RFE feature selection did not select any of the ARC data.

The RFE no CT model had an accuracy of 0.769 ± 0.077, sensitivity of 0.866 ± 0.094, specificity of 0.627 ± 0.103, AUC of 0.835 ± 0.060 and F1-score of 0.817 ± 0.066. This was the best performance for all the models across all metrics other than specificity where the Spearman model had a specificity of 0.685 ± 0.169. This demonstrates that the RFE feature selection was effective in selecting the most relevant features for our particular dataset and removing less impactful noisy features.

Comparing the results to the baseline Original model, which achieved an accuracy of 0.733 ± 0.059 and an F1-score of 0.781 ± 0.043, it is evident that RFE improved model performance. Removing CT data while using RFE further enhanced the results, suggesting that CT data might introduce complexity or noise rather than adding value to the predictive capability for this specific dataset. However it is noteworthy that when the CT data was removed from both the Original and Spearman models that there was a reduction in model performance. The ARC data seems to have had little impact with neither of the feature selection methods choosing any features belonging to this data. The Original model saw a slight worsening in performance after the ARC data removal, whereas the Cleaned model saw a slight improvement after the ARC data removal.

5.2 SHAP Plots – With Treatments for Dementia

The SHAP plots in the figures below give an indication of how impactful each feature was in a specific model for determining whether or not a patient had dementia, with the most impactful feature shown at the top of the plot and then the rest shown in descending order. These SHAP plots were found by considering models made across each fold [9]. Figure 1 shows the top 20 most impact features for the Original model which contained every feature.

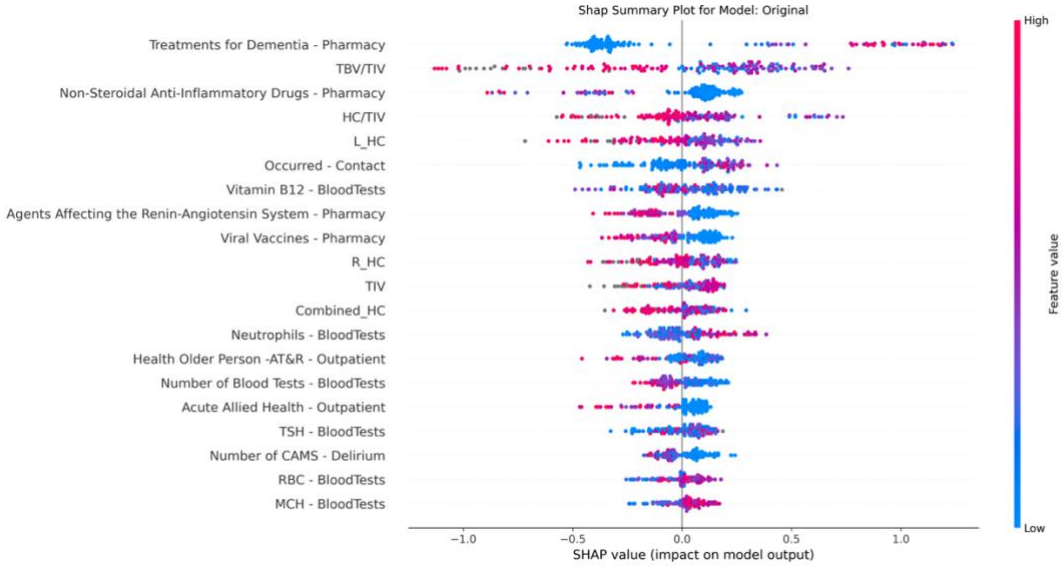

FIGURE 1: SHAP PLOT FOR ORIGINAL MODEL

We can see from this plot that Treatments for Dementia was the most influential feature for the Original model, with several of the CT data features and Pharmacy features also being highly influential in the models dementia diagnosis.

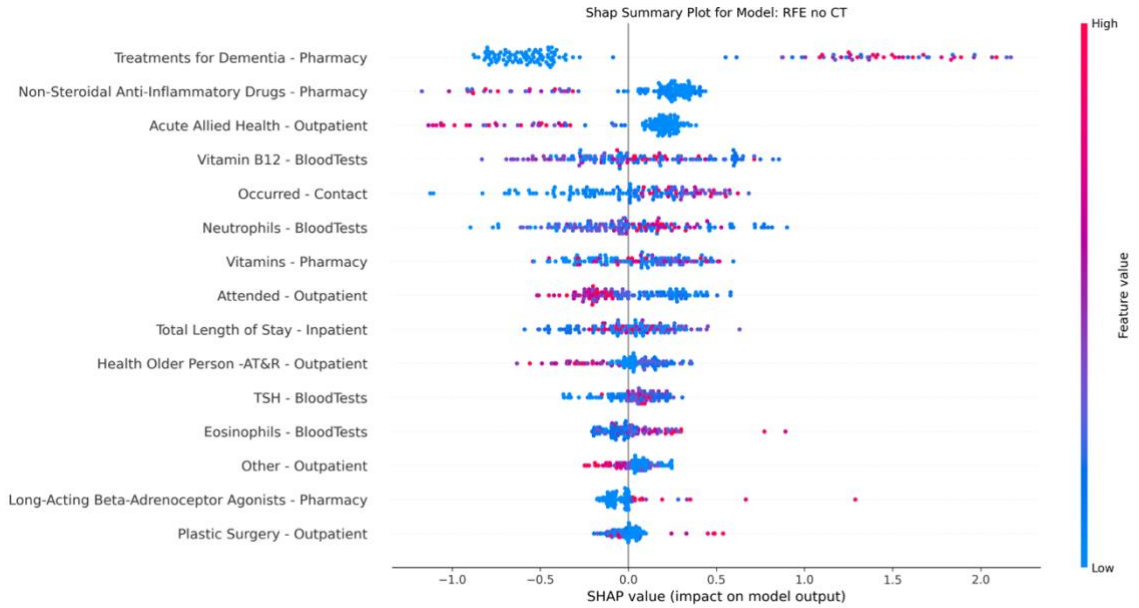

FIGURE 2: SHAP PLOT FOR RFE NO CT MODEL

In figure 2 the SHAP plot of the features for the best performing model RFE no CT can be seen. Again we see that Treatments for Dementia was the most influential feature, with many of the Pharmacy and BloodTests features being impactful. This model after having the RFE feature selection done and CT data removed was left with only 15 features, again indicating that the model's good performance could be due to a lack of overcrowding of other noisy features. We can also see that the max SHAP value for Treatments for Dementia in this more simplified model is almost twice that of in the Original model, indicating that this model

relied more heavily on this feature to determine whether or not a patient had dementia and could be the reason for the models improved performance over the Original model.

5.3 Evaluation of Metrics – Without Treatments for Dementia

Table 6 below has the results and associated metrics for the same models as earlier, but with the Treatments for Dementia feature removed from the data.

	Accuracy	Sensitivity	Specificity	AUC	F1
Original	0.697 ± 0.084	0.803 ± 0.093	0.567 ± 0.128	0.712 ± 0.056	0.751 ± 0.065
Original no CT	0.618 ± 0.036	0.791 ± 0.088	0.398 ± 0.053	0.633 ± 0.075	0.700 ± 0.036
Original no ARC	0.703 ± 0.089	0.845 ± 0.112	0.523 ± 0.093	0.739 ± 0.078	0.763 ± 0.072
Original no CT or ARC	0.630 ± 0.067	0.755 ± 0.107	0.470 ± 0.107	0.627 ± 0.070	0.696 ± 0.065
Cleaned	0.643 ± 0.099	0.792 ± 0.102	0.417 ± 0.102	0.625 ± 0.102	0.725 ± 0.092
Cleaned no CT	0.615 ± 0.089	0.808 ± 0.123	0.313 ± 0.057	0.607 ± 0.101	0.713 ± 0.079
Cleaned no ARC	0.628 ± 0.122	0.778 ± 0.179	0.417 ± 0.102	0.615 ± 0.135	0.707 ± 0.120
Cleaned no CT or ARC	0.643 ± 0.037	0.836 ± 0.133	0.359 ± 0.117	0.661 ± 0.101	0.734 ± 0.048
Spearman	0.699 ± 0.084	0.774 ± 0.133	0.594 ± 0.114	0.714 ± 0.105	0.751 ± 0.084
Spearman no CT	0.664 ± 0.035	0.762 ± 0.074	0.517 ± 0.156	0.692 ± 0.054	0.731 ± 0.041
Spearman no ARC	0.706 ± 0.097	0.795 ± 0.120	0.576 ± 0.128	0.704 ± 0.099	0.761 ± 0.089
Spearman no CT or ARC	0.671 ± 0.029	0.774 ± 0.066	0.519 ± 0.120	0.696 ± 0.046	0.738 ± 0.034
RFE	0.734 ± 0.051	0.845 ± 0.097	0.563 ± 0.134	0.759 ± 0.112	0.792 ± 0.042
RFE no CT	0.727 ± 0.081	0.856 ± 0.077	0.536 ± 0.149	0.768 ± 0.090	0.791 ± 0.065
RFE no ARC	0.734 ± 0.051	0.845 ± 0.097	0.563 ± 0.134	0.759 ± 0.112	0.792 ± 0.042
RFE no CT or ARC	0.727 ± 0.081	0.856 ± 0.077	0.536 ± 0.149	0.768 ± 0.090	0.791 ± 0.065

TABLE 6: MODEL RESULTS WITHOUT TREATMENT FOR DEMENTIA FEATURE

The model with the best overall performance across the metrics was the RFE model. Again the RFE and RFE no ARC model are the same as the RFE feature selection didn't select any of the ARC data. The RFE model consistently outperformed the baseline Original model across most metrics.

The RFE model achieved an Accuracy of 0.734 ± 0.051, Sensitivity of 0.845 ± 0.097, Specificity of 0.563 ± 0.134, AUC of 0.759 ± 0.112, and F1-score of 0.792 ± 0.042. This outperformed the Original model on all metrics other than the Specificity.

Now that the Treatments for Dementia feature has been removed we can see that the accuracy of all the baseline models is now reduced when removing the CT data, unlike when the Treatments for Dementia feature was included and some models saw an increase in accuracy after the removal of the CT data. This indicates that without the Treatments for Dementia feature which has a very strong correlation to whether or not a patient has dementia, that the CT data is useful in the models decision making for determining if a patient has dementia. The ARC data is still seemingly relatively unimpactful, with the Original and Spearman models showing a slight increase in accuracy after its removal and Cleaned a slight decrease.

These results again however highlight that the RFE feature selection process was able to yield the best results, suggesting that by removing noisy and less impactful features the model was able to more accurately diagnose patients with dementia.

5.4 SHAP Plots – Without Treatments for Dementia

The SHAP plot for the Original model without the Treatments for Dementia feature can be seen below in figure 3.

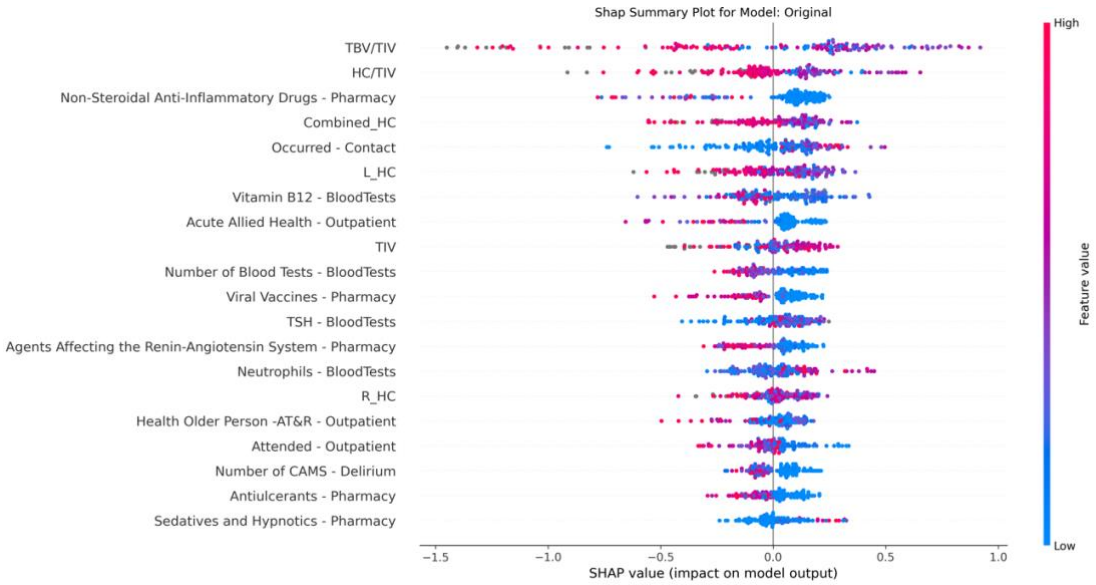

FIGURE 3: SHAP PLOT FOR ORIGINAL MODEL WITHOUT TREATMENT FOR DEMENTIA

The top two most influential features are both from the CT data which indicates that the CT data played a big role in this model for the dementia diagnosis. The rest of the features are similar to the model which had the Treatments for Dementia feature, except some of them seem to have had an increase in impact on the model which can be seen from their higher max SHAP values.

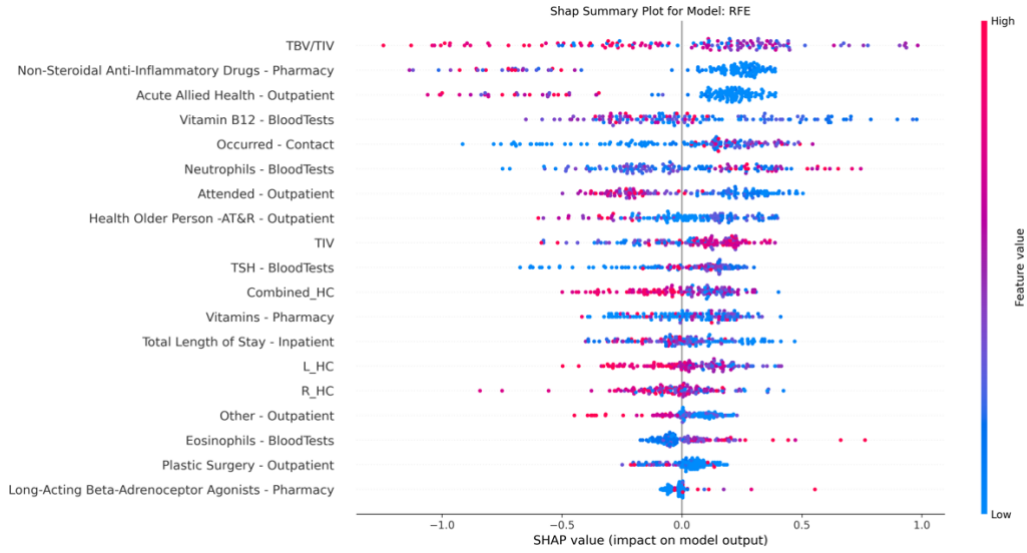

FIGURE 4: SHAP PLOT FOR RFE MODEL WITHOUT TREATMENTS FOR DEMENTIA

Figure 4 shows the SHAP plot for the best performing model when the Treatment for Dementia feature was removed, which was the RFE model. We can see from this plot that the

most influential feature was TBV/TIV which again belongs to the CT data, indicating that the CT data is impactful in dementia diagnosis.

The distribution of the SHAP values from the RFE model is also notably different to that shown in the Original model. We can see that the spread of SHAP values among the different features seems to be more compact rather than being dominated highly by the most important few features.

6 Conclusion

This study aimed to develop and evaluate machine learning models for diagnosing dementia using a combination of clinical, pharmacy, and CT data. Through several feature selection techniques, the analysis revealed key insights into model performance and feature importance.

The inclusion of the feature "Treatments for Dementia" significantly influenced model performance, often dominating the predictive capabilities of the models. While this feature served as a strong indicator of dementia, it overshadowed the contribution of other features. This feature is also likely unavailable in many cases when trying to diagnose a patient with dementia, as if they're already receiving treatment for dementia they have likely already been diagnosed. Hence more weight should be placed on the results obtained from the analysis done with the "Treatments for Dementia" feature removed.

The studies best performing model with "Treatments for Dementia" removed was the RFE feature selected model. It achieved an accuracy of (0.734 ± 0.051) and sensitivity of (0.845 ± 0.097), outperforming the baseline Original model in nearly all metrics except specificity. This highlights the value of feature selection in optimising model performance as well as reducing model complexity and run time, with the RFE model containing just 19 of the Original models 166 features.

The inclusion of CT and ARC data had varying impacts on the models. When ARC data was included, it often seemed to introduce complexity to the models which had varying impacts on model performance. The ARC data was not seen to be impactful when looking at the SHAP plots, and often wasn't a selected feature by the feature selection methods. When "Treatments for Dementia" was removed as a feature the CT data became seemingly very impactful, with all the models worsening in performance after its removal. As well as this it was shown to have the most impactful feature ("TBV/TIV") across all baseline models in the SHAP plots. Hence showing that the CT data was able to contribute significantly to the models ability to accurately diagnose patients with dementia.

However, the study has several limitations. The analysis was just focused on diagnosing dementia at the current time, and had no metrics for future risk of developing dementia. Additionally, the dates corresponding to data such as how recently treatments were given to patients, and when admissions were to hospital weren't heavily considered. These dates could provide additional insight into an effective diagnosis but would require an increase in model complexity. The dataset was also relatively small, with data on just 165 patients, potentially constraining the models ability to generalise to larger populations.

Finally, the findings highlight the disproportionate impact of the "Treatments for Dementia" feature, and bring into question whether this feature would be realistically available when diagnosing dementia. The ARC data didn't seem to have much of an impact on model performance, whereas CT data showed promise in improving dementia diagnosis through machine learning methods.

7 References

[1] Ji, X. (2024). *Using Machine Learning to Classify Severity of Acute Pancreatitis*. ResearchSpace@Auckland.

[2] Ryu, S.-E., Shin, D.-H., & Chung, K. (2020). Prediction Model of Dementia Risk Based on XGBoost Using Derived Variable Extraction and Hyper Parameter Optimization. *IEEE Access*, *8*, 177708–177720. https://doi.org/10.1109/ACCESS.2020.3025553

[3] Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071

[4] Gauhar Kantayeva, José Lima, Ana I. Pereira. (2023). Application of machine learning in dementia diagnosis: A systematic literature review, Heliyon, Volume 9, Issue 11, 2023, e21626, ISSN 2405-8440. https://doi.org/10.1016/j.heliyon.2023.e21626

[5] Prakasam, P., Gnanavel, A., Rajendran, K., & Suresh, S. (2022). Detection and Classification of the Different Stages of Alzheimer's Disease using Sequential Convolutional Neural Network. The Open Biomedical Engineering Journal, 16(1). https://doi.org/10.2174/18741207-v16-e221227-2022-ht27-3589-1

[6] Dementia Identification for Diagnosing Alzheimer's Disease using XGBoost Algorithm. (2021, February 27). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9396777

[7] Rajab, M. D., Jammeh, E., Taketa, T., Brayne, C., Matthews, F. E., Su, L., Ince, P. G., Wharton, S. B., & Wang, D. (2023). Assessment of Alzheimer-related pathologies of dementia using machine learning feature selection. *Alzheimer S Research & Therapy*, *15*(1). https://doi.org/10.1186/s13195-023-01195-9

[8] Alayba, A. M., Senan, E. M., & Alshudukhi, J. S. (2024). Enhancing early detection of Alzheimer's disease through hybrid models based on feature fusion of multi-CNN and handcrafted features. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-82544-y

[9] Kirk, D. (2022, December 27). Using SHAP with Cross-Validation in Python - Towards Data Science. Medium. https://towardsdatascience.com/using-shap-with-cross-validation-d24af548fadc