Exploring Risk Factor Interactions across the Development Stages of Dementia using an Explainable Machine Learning Model

Zeqi Luo
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
z.luo.3@research.gla.ac.uk

Samuel O. Danso
School of Computer Science
University of Sunderland
Sunderland, United Kingdom
sam.danso@sunderland.ac.uk

***Abstract*—Early prediction of dementia, a long-term progressive disease, has always been a challenge. In recent years, advances in artificial intelligence have led to new computer-aided diagnostic tools. However, these methods often offer limited interpretability due to their simplistic binary outputs and black-box algorithms, restricting their use. In this work, we addressed aforementioned shortcomings by assigning clinically meaningful categories to a longitudinal cohort dataset and using an interpretable random forest algorithm to train the prediction model. Our results show that the model predicts various categories effectively. We further applied an advanced machine learning explanation framework to analyse the predictions, revealing the impact of some key risk factors on the prediction and varying interaction patterns between these factors when predicting different development stages of dementia.**

***Index Terms*—brain health, dementia, risk factors, machine learning**

I. Introduction

Dementia is a complex neurodegenerative disorder known to be associated with a variety of symptoms including cognitive decline, which results in loss of independence and function [1]. Alzheimer's disease (AD) is the most common type of dementia and it is estimated to account for over 70 percent of dementia cases [2]. While AD has no cure, there is adequate evidence that points to the fact that the pathophysiological process a result of accumulation of amyloid in the brain begins up to 20 years before the clinical manifestation of the disease. This evidence has led to categorisation of the stages of dementia based clinical manifestation to facilities early detection.

This categorisation is important is it provides a framework for early detection. The framework enables clear characterisation of the various stages of the disease process for early detection and intervention. While clinical trials involving the stages of mild to moderate dementia have failed to demonstrate clinical usefulness, it is believed that early detection of dementia risk from the preclinical stage will have clinical benefit.

Computational approaches such as Machine Learning have demonstrated to have the potential to accurately predict the dementia risk and assist in the diagnosis of AD [3]. However, majority of these approaches have focused on the later stages of the disease, missing the critical window. A recent research by Danso et al. [4] focused on early detection up to 14 years of dementia onset. The authors defined dementia risk based on family history and apolipoprotein E type 4 allele status.

In this work, we develop an explainable early detection machine learning model based on the Amyloid/Tau/Neurodegeneration (ATN) classification framework [5]. The ATN approach offers the possibility to define the stages of the dementia based on the pathophysiological process, and this accounts for the accumulation of amyloid in the brain. We further apply a framework to explore the interaction between the factors responsible for driving the risk at each stage. We believe that this approach offers a robust and reliable alternative to the early detection of dementia and to understanding the behaviour of risk factors to develop possible intervention strategies.

II. Methodology

A. Data Description and Preprocessing

The data we used in this research was obtained from a multicenter, longitudinal cohort study (LCS) established by the European Prevention of Alzheimer's Dementia Consortium (EPAD) [6]. Data used in this work were obtained from the EPAD LCS V.IMI dataset, which contains a total of 2,096 participants and 5 times of visit. The statistics of participants' demographic data are listed in Table I.

In the dataset, a wide range of demographic, cognitive, clinical, neuroimaging, and biomarker data is collected. As the main focus of our research is to explore how the different risk factors contribute to the prediction of AD and how these factors interact with each other, we selected features used for modelling according to domain knowledge about risk factors. In detail, we extracted the following features from the dataset:

- Sociodemographic: age, sex, handedness, years of education, marital status
- Vital: body mass index (BMI), systolic blood pressure

TABLE I: Statistics of the demographic data

Feature	Category	Count	Proportion (%)
Sex			
	Female	1175	56.06
	Male	920	43.89
	Unknown	1	0.05
Age			
	50-55	150	7.16
	56-65	753	35.92
	66-75	931	44.42
	76-85	251	11.97
	86-90	10	0.48
	Unknown	1	0.05
Ethnicity			
	Caucasian/white	1595	76.10
	Hispanic	10	0.48
	Asian	8	0.38
	Others	11	0.52
	Unknown	472	22.52
Years of Education			
	0-10	360	17.18
	11-15	834	39.79
	16-20	816	38.93
	21-32	75	3.58
	Unknown	11	0.52

- Medical history: family dementia history, hypercholesterolemia, hypertension, anaesthesia, depression, head injury, cancer, diabetes, mild cognitive impairment (MCI), dyslipidemia, anxiety, stroke, insomnia
- Genetic: apolipoprotein E (APOE) gene
- Lifestyle: smoking, drug abuse, current health, physical activity, physical fitness

We constructed the dataset for this study by extracting essential features from various tables within the extensive EPAD dataset, using subject IDs and visit counts as primary keys to merge them into a cohesive set of records. We compute certain features, such as BMI from height and weight, from the original data. After isolating these necessary features and removing irrelevant ones, we address any missing values in the data. Considering that each column has less than 5% missing values, we simply use the statistic from the corresponding column for missing value imputation. We filled missing data in continuous columns with column means and in categorical columns with the most frequent category. For features that has specific medical meaning and could not be imputed in any reasonable approach (for example, the APOE gene), we discarded records that missed the values of these features.

In order to simplify the features and enable the machine learning algorithm to produce better classifiers [7], we applied a series of discretisation to the features. We represented "sex" as a binary feature (1 for female, 0 for male). Similarly, "family dementia history" was coded as 1 for yes and 0 for no. We binarized all medical history variables to indicate the presence (1) or absence (0) of each condition. Additionally, "smoking" and "drug abuse" were encoded with binary values where 0 signifies never and 1 indicates past or current use.

We streamlined the "physical activity" variable, which originally had six categories, into a binary feature indicating whether an individual engages in physical exercise at least twice per week; a value of 1 signifies a potential risk of insufficient activity [8]. For "physical fitness" and "current health," initially rated on a five-point scale from 'very good' to 'poor,' we adopted binary coding: 0 for 'satisfactory' or better and 1 for 'relatively poor' or worse. Additionally, we condensed the four-category "marital status" feature—comprising 'married or cohabiting,' 'single,' 'divorced,' and 'widowed'—into a single binary indicator where 0 represents individuals without partners and 1 those with partners.

We binarized "years of education" (1 for $\leqslant 10$ years and 0 for > 10 years), "BMI" (1 for $\geqslant 25$ and 0 for < 25), and "systolic blood pressure" (1 for $\geqslant 140$ mmHg and 0 for < 140 mmHg) to denote a higher or lower risk for dementia according to [8], [9]. Regarding the APOE gene, we focused on the impact of the type 4 allele ($\epsilon 4$), a primary genetic risk factor for Alzheimer's disease (AD), as identified by [10]. Consistent with findings from [11], which indicated an increased AD risk in individuals carrying one or two $\epsilon 4$ alleles—with a greater risk associated with two copies—we coded the presence of APOE $\epsilon 4$ alleles as 0 (no copies), 1 (one copy), and 2 (two copies) to reflect their respective risks.

B. Class Labelling and Problem Formulation

We used the ATN framework to label each record in the dataset. In line with [5], we divide the records into different groups based on the normal (+) or abnormal (-) CSF Aβ_{42} values (labelled "A"), CSF p-tau values (labelled "T") and neurodegeneration or neuronal injury (labelled "N"). We considered CSF Aβ_{42} < 1000 pg/mL and CSF p-tau > 27 pg/mL as A+ and T+ respectively, which had been validated by [12] right on the EPAD LCS data set. To determine neurodegeneration (N), we used the average medial temporal lobe atrophy (MTA) since the experiment in [12] showed that the average MTA score has good distinguishability in the data. For participants < 65 years of age, an average MTA score $\geqslant 1$ was considered positive; for participants $\geqslant 65$ years of age, an average MTA score $\geqslant 1.5$ was considered positive [13]. According to [5], we defined A-T-N- as healthy controls (HC), A+ samples with any combination of T and N types as AD continuum, and all of the other samples as suspected non-AD pathologic change (SNAP). We removed records that contain missing values in the features used in this process.

After the preprocessing and labelling procedure, there are 2,180 records in the dataset for this research. Now the research problem can be formulated as: given a dataset $\mathcal{D} = [(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)]$, where $\mathbf{x}_i$ is a record of features and y_i is the label, we want to train a classification model $f(\mathbf{x}) \rightarrow \mathbb{R}^C$, which predicts the class $c \in C = \{\text{HC}, \text{SNAP}, \text{AD Continuum}\}$ of a given record $\mathbf{x}$.

C. Model Implementation

We choose the random forest [14] technique for our classification task. It includes an ensemble of decision trees and incorporates feature selection and interactions naturally in

the learning process [15], which provides unique benefits for handling our dataset, characterised by a small sample size and a relatively high-dimensional feature space. Previous work [16] has shown the random forest outperformed many more advanced algorithms like the neural network on tabular data.

We applied cross-validation on the training set to get the optimal hyperparameter settings of the random forest, used machine learning algorithms to train the model, and reported the model's evaluation performance on the test set. For this purpose, we split the preprocessed data into a training set (records from visit 1, 75% of the whole dataset) and a test set (records from visits 2-5, 25% of the whole dataset). The distributions of classes of the training and test sets are similar to the distribution of the original dataset. Table II displays the category distributions pre- and post-split.

The distribution of different classes in the training data is not balanced, which may harm the classifier's generalisation performance. Two techniques are often used to solve data imbalances: re-sampling and cost-sensitive learning. For the latter, a cost matrix provided by a domain expert is required and this is not easy to achieve [17]. In this case, we balanced the training data using resampling methods. We developed four prediction models employing distinct resampling strategies: Model A utilises original training data without modification; Model B is trained on the random oversampled data; Model C incorporates the SMOTE-NC algorithm [18] for resampling; and Model D adopts the balanced random forest approach [19], which balances each bootstrap sample through random undersampling. We trained Models A, B, and C with the standard random forest algorithm on resampled datasets. For Model D, we applied an 'minority' undersampling strategy that excludes only the minority class.

D. Evaluation Metrics

As we were dealing with imbalanced data, simply using accuracy tends to provide misleading evaluation because when computing accuracy in multi-class classification, accuracy is simply the fraction of correct classifications and is mainly affected by the classification performance of the majority class. We applied a series of metrics for the evaluation and comparison of our models based on the number of True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) in the prediction of the classifiers. Although these are defined in the binary classification scenario, we could easily apply them to our multi-class classification problem: when considering the classification performance of the classifiers in one class, we could treat this class as the positive class and all of the other classes as the negative class. In this case, a k-class problem turns into k binary classification problems, and all widely used metrics are available.

TABLE II: Data distribution in the preprocessed dataset

Class	Total (%)	Train (%)	Test (%)
HC	1175 (53.90)	891 (53.93)	284 (53.79)
SNAP	281 (12.89)	223 (13.50)	58 (10.98)
AD Continuum	724 (33.21)	538 (32.57)	186 (35.23)
Total	2180	1652 (75.78)	528 (24.22)

Sensitivity and specificity (1) are a pair of metrics widely used in medical diagnosis [20]. When considering an example of a medical test for diagnosing a condition, sensitivity refers to the test's ability to correctly detect ill patients who do have the condition, while specificity refers to the test's ability to correctly reject healthy patients without a condition.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \ \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1}$$

We also calculated the F1 score (3) for each class, which measures the comprehensive performance in positive class detection by considering both accuracy (precision) and coverage (recall). For the overall performance, we calculated the macro-average F1 score and area under the receiver operating curve (AUROC) which is ideal for imbalanced data [21]. The macro-average is not sensitive to data imbalance while each class has equal contribution to the metric [22].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{Recall} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}$$

E. Explainability

The interpretability of machine learning models is treated as important as the prediction accuracy for most life science problems. For dementia prediction, clinicians and researchers are curious about how risk factors contribute to the prediction and the interactions between factors. We applied the SHapley Additive exPlanation (SHAP) framework [23] to enhance interpretability, building on the inherent clarity of random forest predictions. SHAP is a method for interpreting predictive models that is based on Shapley values from cooperative game theory. The Shapley value is a method of fairly distributing the total profit of the game, ensuring that the contribution of each player (feature) is fairly evaluated. For each feature, its Shapley value is the average of its contribution to the predicted outcome of all possible feature combinations.

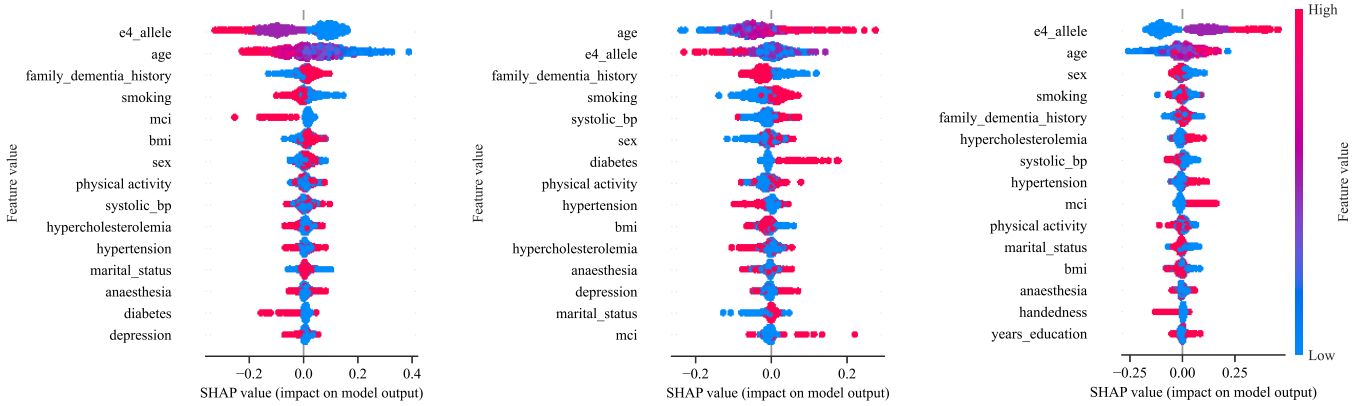
In the case of disease modelling, each risk factor is considered as a 'player' and a 'team' of the risk factors 'cooperatives' to drive the overall risk of disease. In practice, the contribution of the risk factor to the prediction (so-called SHAP value) is the mean marginal contribution of a risk factor value across all possible risk factor coalitions. The framework could show the positive or negative relationship for each variable with the target which surpasses simple feature importance methods. SHAP interaction values extend on this by breaking down the contributions into features' main and interaction effects. Specifically, we used TreeExplainer [24] to interpret our random forest classifiers.

III. Results

Table III summarises the performances obtained when we applied four different models to the unseen test set. In terms of

TABLE III: Classification metrics comparison

	Sensitivity (%)			Specificity (%)			F1 Score (%)			F1 Score (%)	AUROC
Model	HC	SNAP	AD	HC	SNAP	AD	HC	SNAP	AD	**macro-average**	**macro-average**
A	**85.56**	22.41	59.14	57.79	**99.15**	83.92	77.14	34.67	62.68	69.32	0.795
B	83.45	58.62	68.81	74.59	95.32	**86.84**	**81.30**	**59.65**	**71.31**	**75.57**	0.842
C	71.48	65.52	**69.89**	**79.91**	88.09	84.80	75.75	50.00	70.65	70.27	**0.845**
D	61.27	**67.24**	61.29	78.69	82.77	80.12	67.84	45.05	61.54	61.74	0.802



(a) Healthy Control (b) Suspected non-AD pathological change (c) AD Continuum

Fig. 1: SHAP summary plots

sensitivity, Model A achieved an optimal sensitivity of 85.56% in categorising the HC category. Whereas, Model D and C outperformed in classifying SNAP and AD Continuum with sensitivities of 67.24% and 69.89%, respectively.. In terms of specificity, Model A excelled in the classification of SNAP with 99.15%. When it comes to the F1 score, Model B shows the best balanced and strongest performance in all categories, culminating in the highest macro-average F1 score of 75.57%. In terms of the macro-average AUROC, Model B and Model C stand out for their high overall performance. Although a permutation test [25] performed shows statistical significance ($P < 0.05$) of the difference in AUROC between Models B and C, the difference is relatively minor.

Model A is trained using the original dataset and performs good in most of the metrics, indicating that the original dataset is a good representative for this classification task.

Model B uses random oversampling to increase the numbers of samples of SNAP and AD Continuum classes, leading to significant improvements in most of the metrics compared to Model A. The highest F1 score shows this model is more reliable and balanced, effectively avoiding misdiagnoses and missed diagnoses.

Model C employs the SMOTE-NC algorithm for oversampling by generating minority class samples and enhancing model performance across all metrics. However, this comes at the cost of reduced performance of the majority category, suggesting that while oversampling enhances detection of less frequent classes, it may also introduce noise that hampers recognition of the majority class.

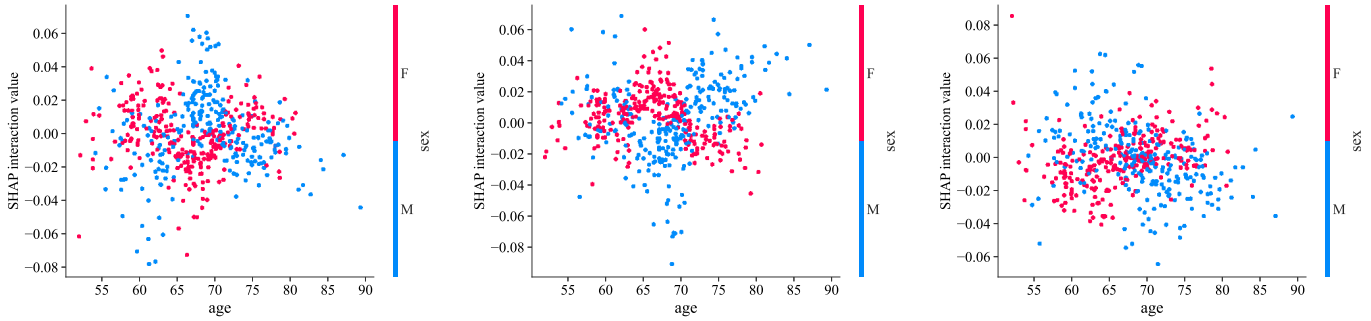
Model D uses random undersampling to balance the sample sizes across categories, matching the count of minority categories. While this improves the sensitivity in SNAP category, it leads to lower metrics in other categories, suggesting that valuable information loss from undersampling diminishes model performance.

Despite variations in performance across categories, Model B consistently achieves high scores in most metrics, particularly excelling in per-class and overall F1 score. It also shows superior sensitivity for the AD Continuum, which is our primary concern, and achieves the highest specificity in identifying Healthy Control, accurately differentiating them from those with potential dementia. We use the predictions from Model B in the interpretation and discussion below.

IV. Discussion

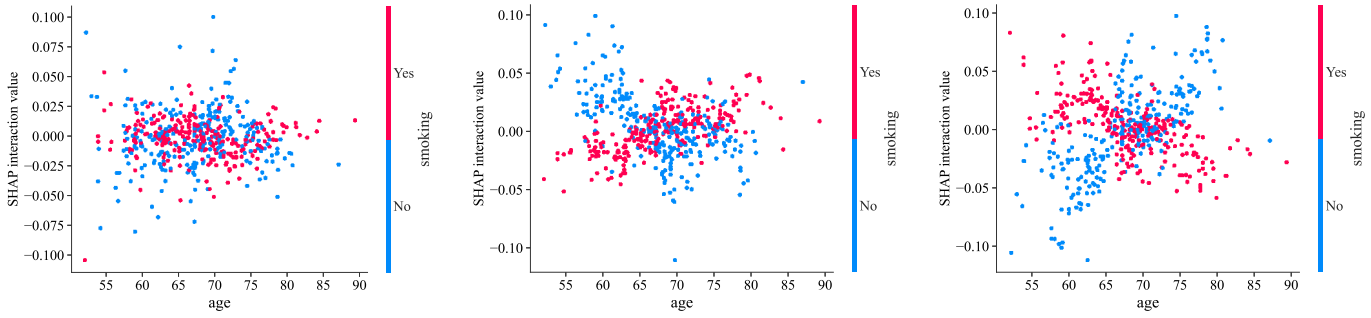
Observing the risk factors known as model features as ranked by the best performing model, Fig. 1 shows the ranking of the features in terms of importance and impact. The figures illustrate the varying importance and impact of features across the three categories. While APOE $\epsilon 4$ alleles, age, smoking and family dementia history are consistently ranked among the top 5 important risk factors, the position of these features varies between the classes. APOE $\epsilon 4$ alleles and age are always in the top 2 position. Smoking is ranked higher than family dementia history in AD Continuum category, whereas the ranking of these factors is different and remains the same in the Healthy Control and SNAP classes.

On the other hand, the impact values of these risk factors on different categories also appear to differ, with APOE $\epsilon 4$ having relatively high impact on AD Continuum class. Furthermore,



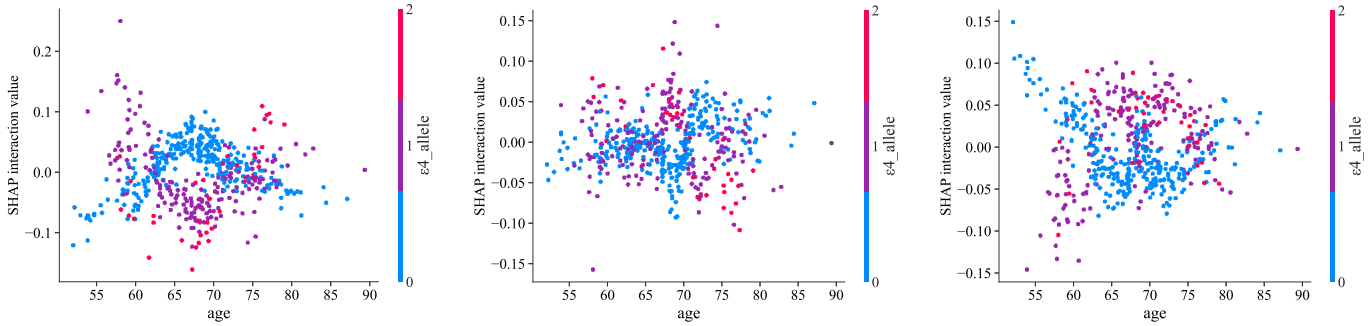
(a) Healthy Control (b) Suspected non-AD pathologic change (c) AD Continuum

Fig. 2: SHAP dependence plots for age and sex



(a) Healthy Control (b) Suspected non-AD pathologic change (c) AD Continuum

Fig. 3: SHAP dependence plots for age and smoking



(a) Healthy Control (b) Suspected non-AD pathologic change (c) AD-Continuum

Fig. 4: SHAP dependence plots for age and ϵ4 allele

while age seems to have uniform distribution of impact among the HC and AD Continuum, the impact seems to be higher in the AD Continuum class. Also, focusing on how these features interact with age, a variety patterns of behaviours emerge. For example, it can be seen from Fig. 2 the interaction between sex and age among the Healthy Controls suggests high impact on males with over 70 years, but the reverse is observed in the AD Continuum group, where females have higher impact, suggesting females have high risk compared to men after age of 70 years. While this is consistent with several findings in the literature [26], [27], our work has highlighted the age at which the risk is elevated in females.

Similarly, the interaction and impact of smoking and age on risk (Fig. 3) appears to be heterogeneous and diverse among the Health Control group. However, different pattern is observed in the SNAP and AD Continuum groups. While non-smoking appears to have higher impact on risk among the SNAP group before the age of 70 years, there is significant impact of age over 70 years. However, the reverse of this phenomenon is observed among the AD Continuum groups (Fig. 3c), where smoking appears to have a higher impact up to the age of 70 years but the impact of this is reduced after the age of 70 years.

Fig. 4 shows the interaction between age and number of

$\epsilon 4$ alleles during the prediction. Compared to the distribution shown in Fig. 4c, the SHAP interaction values are more densely clustered and centred around zero in Fig. 4b, suggesting that the interaction between age and the number of $\epsilon 4$ alleles does not have a strong discriminatory power in predicting non-AD pathologic change. The plot for AD Continuum shows a more dispersed distribution of SHAP values with a slightly higher density of positive values, indicating that the age-$\epsilon 4$ allele interaction might have a more pronounced effect in this group compared to the other two. This aligns with the understanding that APOE $\epsilon 4$ is a significant risk factor for AD [10], while its significance in other dementias is less well-defined.

V. Conclusion

This paper discussed work which focused on development of an explainable machine learning model based on Random Forest algorithm and the ATN dementia classification framework. The ATN framework offered the opportunity to explore pathological process dementia and the influence of this process on the interaction of risk factors. several models we developed and evaluated with the best model achieving high performance prediction accuracy comparable to the best models reported in the literature. Furthermore, employing the the SHAP explainable framework, we explored the interaction of risk factors and their impact of dementia risk. While some of the interactions observed are consistent with the body of knowledge, new patterns of interaction and impact on the various risk groups emerged from this work, which require further investigation.

Acknowledgment

We sincerely thank the participants of the EPAD study for donating their data for this and other research. We also appreciate the EPAD researchers, the EPAD Consortium for approving our request and making data available to carry out this work, and the EPAD research funders for their financial support, which made data collection possible.

References

[1] K. Ritchie and S. Lovestone, "The dementias," *The Lancet*, vol. 360, no. 9347, pp. 1759–1766, 2002.

[2] M. V. F. Silva, C. d. M. G. Loures, L. C. V. Alves, L. C. de Souza, K. B. G. Borges, and M. d. G. Carvalho, "Alzheimer's disease: risk factors and potentially protective measures," *Journal of biomedical science*, vol. 26, pp. 1–11, 2019.

[3] E. Pellegrini, L. Ballerini, M. d. C. V. Hernandez, F. M. Chappell, V. González-Castro, D. Anblagan, S. Danso, S. Muñoz-Maniega, D. Job, C. Pernet *et al.*, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 519–535, 2018.

[4] S. O. Danso, Z. Zeng, G. Muniz-Terrera, and C. W. Ritchie, "Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms," *Frontiers in big Data*, vol. 4, p. 613047, 2021.

[5] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish *et al.*, "Nia-aa research framework: toward a biological definition of alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[6] A. Solomon, M. Kivipelto, J. L. Molinuevo, B. Tom, and C. W. Ritchie, "European prevention of alzheimer's dementia longitudinal cohort study (epad lcs): study protocol," *BMJ open*, vol. 8, no. 12, p. e021017, 2018.

[7] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.

[8] M. Kivipelto, T. Ngandu, T. Laatikainen, B. Winblad, H. Soininen, and J. Tuomilehto, "Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study," *The Lancet Neurology*, vol. 5, no. 9, pp. 735–741, 2006.

[9] J.-H. Chen, K.-P. Lin, and Y.-C. Chen, "Risk factors for dementia," *Journal of the Formosan Medical Association*, vol. 108, no. 10, pp. 754–764, 2009.

[10] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. Roses, J. Haines, and M. A. Pericak-Vance, "Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[11] L. A. Farrer, L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. Van Duijn, "Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis," *Jama*, vol. 278, no. 16, pp. 1349–1356, 1997.

[12] S. Ingala, C. De Boer, L. A. Masselink, I. Vergari, L. Lorenzini, K. Blennow, G. Chételat, C. Di Perri, M. Ewers, W. M. van der Flier *et al.*, "Application of the atn classification scheme in a population without dementia: Findings from the epad cohort," *Alzheimer's & Dementia*, vol. 17, no. 7, pp. 1189–1204, 2021.

[13] H. F. Rhodius-Meester, M. R. Benedictus, M. P. Wattjes, F. Barkhof, P. Scheltens, M. Muller, and W. M. van der Flier, "Mri visual ratings of brain atrophy and white matter hyperintensities across the spectrum of cognitive decline are differently affected by age and diagnosis," *Frontiers in aging neuroscience*, p. 117, 2017.

[14] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[15] Y. Qi, "Random forest for bioinformatics," *Ensemble machine learning: Methods and applications*, pp. 307–323, 2012.

[16] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" *arXiv preprint arXiv:2207.08815*, 2022.

[17] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[19] C. Chen, A. Liaw, L. Breiman *et al.*, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[20] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, 1994.

[21] Y. Ma and H. He, "Imbalanced learning: foundations, algorithms, and applications," 2013.

[22] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171–186, 2001.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[25] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance." *Journal of machine learning research*, vol. 11, no. 6, 2010.

[26] P. Gilsanz, E. R. Mayeda, M. M. Glymour, C. P. Quesenberry, D. M. Mungas, C. DeCarli, A. Dean, and R. A. Whitmer, "Female sex, early-onset hypertension, and risk of dementia," *Neurology*, vol. 89, no. 18, pp. 1886–1893, 2017.

[27] J. Luo, C. R. Beam, I. K. Karlsson, C. J. Pike, C. A. Reynolds, and M. Gatz, "Dementia risk in women higher in same-sex than opposite-sex twins," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 12, no. 1, p. e12049, 2020.