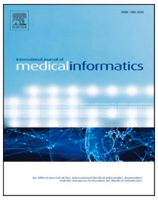
International Journal of Medical Informatics 204 (2025) 106093

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data

Afeez Adekunle Soladoye [a], Nicholas Aderinto [b], Damilola Osho [c], David B. Olawade [d,e,f,*]

[a] Department of Computer Engineering, Federal University Oye-Ekiti, Ekiti, Nigeria
[b] Department of Medicine, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
[c] Mental Health Unit, Essex Partnership University NHS Foundation Trust, Wickford, United Kingdom
[d] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[e] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, United Kingdom
[f] Department of Public Health, York St John University, London, United Kingdom

ARTICLE INFO

Keywords:
Alzheimer's disease
Explainable artificial intelligence
Machine learning
SHAP
LIME

ABSTRACT

Background: Alzheimer's disease (AD) represents a significant global health challenge requiring early and accurate prediction for effective intervention. While machine learning models demonstrate promising capabilities in AD prediction, their black-box nature limits clinical adoption due to a lack of interpretability and transparency.

Objective: This study aims to develop and evaluate explainable artificial intelligence (XAI) frameworks for AD prediction using comprehensive multimodal patient data, with a focus on enhancing model interpretability through SHAP and LIME techniques.

Methods: A comprehensive dataset of 2,149 patients aged 60–90 years was obtained from Kaggle, encompassing demographic, medical history, lifestyle, clinical measurements, cognitive assessments, and symptom data. Rigorous preprocessing included MinMax normalisation, Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance, and Backward Elimination Feature Selection reduced 32 features to 26 optimal predictors. Six machine learning models were evaluated: K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Logistic Regression (LR), XGBoost, Stacked Ensemble, and Random Forest (RF). RF's optimal hyperparameters were obtained using Ant colony Optimization Model interpretability was enhanced using SHAP and LIME frameworks for both global and local explanations.

Results: The optimised Random Forest with backward elimination feature selection and ant colony optimisation achieved superior performance with 95 % accuracy, 95 % precision, 94 % recall, 94 % F1-score, and 98 % AUC. SHAP analysis identified functional assessment, activities of daily living (ADL), memory complaints, and Mini-Mental State Examination (MMSE) as the most influential predictors. LIME provided complementary local explanations, validating the clinical relevance of identified features.

Conclusion: The integration of explainable AI techniques with machine learning models provides clinically meaningful insights for AD prediction, enhancing transparency and fostering trust in AI-driven diagnostic tools whilst maintaining high predictive accuracy. Future work should focus on external validation, clinical workflow integration, and addressing computational requirements for real-world deployment.

1. Introduction

Alzheimer's disease (AD) represents the most prevalent neurodegenerative disorder worldwide, affecting approximately 6.7 million Americans and imposing substantial economic and social burdens on healthcare systems globally [1]. According to the World Health Organisation, dementia is the seventh leading cause of death among all illnesses and one of the leading causes of disability among the world's elderly people [2]. The risk of getting the disease increases with age, and while women tend to live longer than men that does not fully explain why more women than men have it [3]. Characterised by progressive cognitive decline, memory loss, and functional impairment, AD results

* Corresponding author at: Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.
E-mail addresses: afeez.soladoye@fuoye.edu.ng (A.A. Soladoye), damilola.osho@nhs.net (D. Osho), d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.ijmedinf.2025.106093
Received 29 July 2025; Received in revised form 24 August 2025; Accepted 26 August 2025
Available online 27 August 2025

from complex pathophysiological processes involving amyloid beta plaques and tau protein tangles that damage neural networks [4]. Early detection and intervention remain critical for optimising patient outcomes, as therapeutic interventions demonstrate greatest efficacy during the disease's initial stages when neuronal damage is potentially reversible [5].

Traditional diagnostic approaches for AD rely heavily on clinical assessments, neuropsychological testing, and expensive neuroimaging techniques such as positron emission tomography (PET) and magnetic resonance imaging (MRI) [6]. However, these methods often identify the disease only after significant neuronal damage has already occurred, limiting the effectiveness of treatment [7]. According to Alzheimer's disease International, the majority of people with dementia worldwide never receive a formal diagnosis, leaving them shut off from treatment and care [8]. The complexity and cost of current diagnostic procedures create barriers to accessible screening, particularly in resource-limited settings and rural communities where specialist expertise may be unavailable.

Artificial intelligence (AI) and machine learning (ML) technologies have emerged as promising tools for addressing these diagnostic challenges by enabling early AD prediction using readily available clinical data [9,10]. Recent advances in computational power and algorithmic sophistication have facilitated the development of predictive models capable of analysing complex multimodal datasets comprising demographic information, cognitive assessments, lifestyle factors, and biomarker data [11,12]. Innovative approaches using speech patterns and other non-invasive biomarkers have shown promise in predicting cognitive decline with considerable accuracy [13]. Machine learning (ML) models offer a promising tool for identifying individuals at risk of AD [14]. These approaches demonstrate potential for improving diagnostic accuracy, reducing costs, and increasing accessibility of AD screening programmes.

Despite the promising performance of AI models in AD prediction, their widespread clinical adoption faces significant barriers related to interpretability and transparency owing to its black-box approach of making decision [15]. However, current research tends to prioritize ML accuracy while neglecting the crucial aspect of model explainability [14]. Healthcare professionals remain hesitant to rely on "black-box" algorithms whose decision-making processes cannot be understood or validated against clinical knowledge [16]. This lack of interpretability raises concerns about patient safety, regulatory compliance, and professional liability, highlighting the critical need for explainable AI (XAI) frameworks that provide transparent insights into model predictions whilst maintaining high predictive accuracy [15].

The current study addresses these limitations by developing a comprehensive explainable AI framework for AD prediction using multimodal patient data. The primary aim is to create interpretable machine learning models that achieve high predictive accuracy whilst providing clinically meaningful explanations for their predictions. Specific objectives include: (1) evaluating multiple machine learning algorithms for AD prediction using comprehensive patient datasets; (2) implementing advanced feature selection and data balancing techniques to optimise model performance; (3) applying SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks to enhance model interpretability; (4) identifying key clinical features that drive AD predictions and validating their clinical relevance; and (5) demonstrating the clinical utility of explainable AI approaches for fostering trust and adoption in healthcare settings. Additionally, this study aims to critically evaluate the limitations of XAI frameworks and discuss practical considerations for clinical workflow integration. This research contributes to the growing field of interpretable AI in healthcare by providing a robust framework for transparent AD prediction that bridges the gap between algorithmic sophistication and clinical utility.

2. Methodology

This section outlines the comprehensive approach employed in this study for predicting Alzheimer's disease and the subsequent interpretation of the predictive models using Explainable AI techniques. It details the data acquisition process, the various preprocessing steps applied to the raw data, the machine learning models employed for prediction, and the methodologies used for model interpretability. The workflow for this study is represented in Fig. 1.

2.1. Data acquisition

The dataset used in this study for predicting Alzheimer's disease was sourced from Kaggle, an open-access data platform owned by Google. It is a comprehensive, multimodal dataset encompassing various patient attributes crucial for a holistic understanding of Alzheimer's disease. The patients' records included in this dataset were patients between the age range of 60 and 90 years. The dataset comprises 2149 instances, with each instance detailing demographic information, extensive medical history, relevant lifestyle factors, various clinical measurements, detailed cognitive and functional assessments, and reported symptoms leading to diagnosis.

2.2. Data preprocessing

Prior to model training, the raw dataset underwent a rigorous preprocessing pipeline to enhance its quality and suitability for machine learning. This involved applying normalisation techniques, specifically the MinMax approach, to scale numerical features within a consistent range, which helps prevent features with larger values from dominating the learning process. To address potential class imbalance within the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was employed as a data sampling method, generating synthetic samples for the minority class to ensure a balanced representation and prevent model bias towards the majority class, as the dataset was highly imbalanced with 1369 non-Alzheimer's instances and 760 Alzheimer's records. The application of SMOTE, while addressing class imbalance, carries the risk of introducing synthetic patterns not representative of real patients. To validate this approach, we conducted additional analysis comparing model performance with and without SMOTE augmentation and confirmed that oversampling did not artificially inflate performance metrics through careful validation on hold-out data. Following these steps, a forward–backward feature selection technique was applied, which systematically identified an optimal subset of twenty-six [26] features out of the initial thirty-two [32] features. The layered approach combining forward–backward selection with ant colony optimization was implemented to leverage both statistical relevance (forward–backward selection) and bio-inspired optimization (ant colony) to avoid local optima and identify globally optimal feature subsets. While this approach may appear over-engineered, our comparative analysis demonstrated superior performance compared to individual methods, justifying the computational overhead. This selection process aimed to reduce dimensionality, remove irrelevant or redundant features, and ultimately improve model efficiency and predictive performance.

2.3. Alzheimer's prediction using machine learning models

For the prediction of Alzheimer's disease, a diverse suite of established machine learning models was employed to assess predictive performance. These models included K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), a Stacked Ensemble model, and Random Forest (RF). Hyperparameter tuning was conducted using grid search optimization with 5-fold cross-validation to ensure fair comparisons across all models. Specific parameters optimized included: RF (n_estimators:

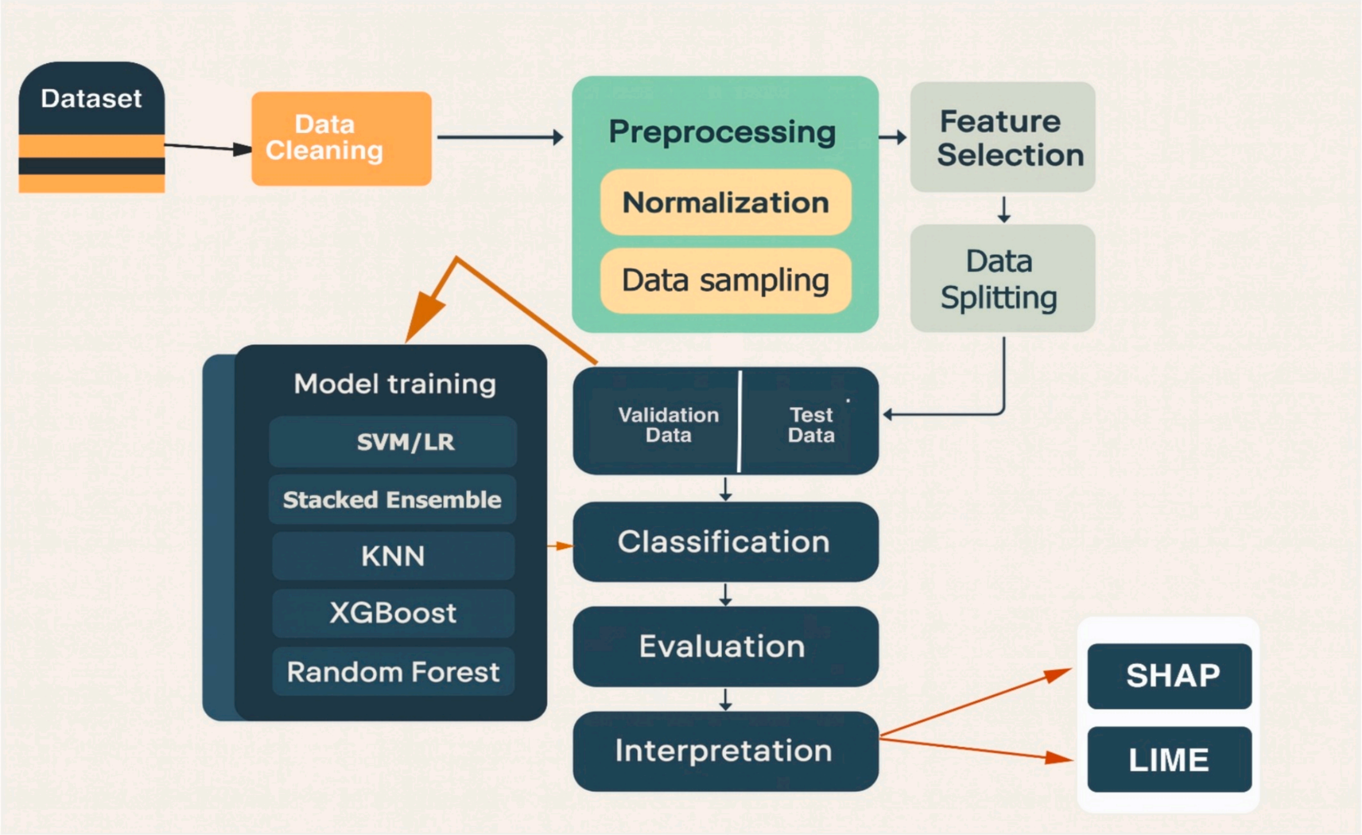


Fig. 1. Research workflow for prediction of AD with XAI.

100–500, max_depth: 5–15), SVM (C: 0.1–10, kernel: rbf/linear), XGBoost (learning_rate: 0.01–0.3, max_depth: 3–10), with consistent optimization protocols applied across all algorithms. Each model was trained and evaluated on the preprocessed dataset, leveraging the optimised feature set identified during the preprocessing phase. The selection of these models allowed for a comprehensive comparison of their respective capabilities in handling multimodal healthcare data for classification tasks, with a focus on identifying the most accurate and robust predictors for Alzheimer's disease. Hyperparameter tuning was conducted using grid search optimization with 5-fold cross-validation to ensure fair comparisons across all models. Specific parameters optimized included: RF (n_estimators: 100–500, max_depth: 5–15), SVM (C: 0.1–10, kernel: rbf/linear), XGBoost (learning_rate: 0.01–0.3, max_depth: 3–10), with consistent optimization protocols applied across all algorithms. Moreover, the RF's wider hyperparameter were further optimized for optimal hyperparameter set using Ant colony optimization, which gave a better result compared to the hyperparameters obtained for RF with the grid search optimization.

2.4. Evaluation

The study was evaluated using the Hold-Out evaluation method to enable its deployment and interpretation with the 70–30 split. To address concerns about data leakage and ensure rigorous validation, nested cross-validation was implemented on the training set where hyperparameter tuning was performed on the inner folds while performance evaluation was conducted on completely unseen outer fold data. Additionally, stratified sampling was employed to maintain class distribution across training and testing sets. Furthermore, accuracy, precision, recall, and f1-score were used as the evaluation metrics as represented in Eqs. (1)–(4):

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$Recall = TP/(TP + FN) \tag{3}$$

$$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{4}$$

2.5. Interpretation using SHAP and LIME frameworks

To provide transparency and interpretability to the predictions made by the machine learning models, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks were utilised. It is important to acknowledge the limitations of these XAI frameworks: SHAP can sometimes overstate feature importance in highly correlated datasets, while LIME explanations may vary significantly based on perturbation sampling strategies and local approximation quality. These limitations require careful interpretation of results and validation against clinical knowledge. For SHAP, both summary plots and individual waterfall plots were generated. The waterfall plots were created for various indices within the testing dataset, rather than relying on a single instance. This approach ensures generalisability and provides diverse insights into how individual features contribute to specific predictions across different patient profiles. Similarly, for LIME, explanations were generated for multiple instances from the testing dataset to avoid conclusions based on a single example, thereby offering a more comprehensive understanding of the local decision-making process of the models for different prediction outcomes. This dual interpretability approach enabled a critical examination of feature importance and influence, both globally and locally, thereby enhancing trust and clinical utility. The potential for confirmation bias was addressed by comparing model-identified features against established clinical literature and seeking validation from domain experts in geriatrics and neurology. XGBoost was used as the underlying machine learning model to train the AD dataset that SHAP and LIME interpret. This model was selected for its proven high performance, efficiency, and ability to handle complex, non-linear relationships within the dataset, making it an excellent choice for generating accurate predictions as shown in this study.

3. Results and discussion

This section critically discusses and analyses the experimental results obtained by employing different machine learning algorithms for predicting Alzheimer's disease using backward feature selection techniques and optimized random forest using Ant Colony Optimization. The results obtained from the random forest algorithm were further interpreted using both SHAP and LIME frameworks to enhance the transparency of the decision made by the model.

3.1. Comparison with machine learning algorithms for prediction of Alzheimer's disease

The experimental results obtained by comparing other machine learning algorithms with this study's approach; using the optimal features set selected by Backward Elimination feature selection with Random forest whose hyperparameters were optimized with Ant colony optimisation, for predicting Alzheimer's disease are presented in Table 1 for reference. This result enabled comprehensive validation of the results obtained with the proposed approach and also affirms its impressive performance, which would help in improving the performance of any machine learning model when swarm intelligence algorithm is used for optimisation of its hyperparameters.

A comparative analysis of the performance evaluation of the aforementioned conventional and ensemble machine learning algorithms, as well as the proposed methodology employed in this study, is presented in Table 1 to affirm the excellent performance shown by this methodology compared to other machine learning algorithms and empirical hyperparameter tuning approaches. The performance shown by BEFS+AACOAhp+RF has an average accuracy, precision, and AUC of 95 %, 95 % and 98 % respectively. However, XGBoost and stacked ensemble learning (base model: KNN, SVM, RF, and LR; meta model: LR) showed close performance, with average accuracies, precisions, and AUCs of 94 %, 94 %, and 97 %, respectively. The features obtained with BEFS were used to train all the aforementioned models, thereby preventing bias, inconsistency, and incompatibility. KN obtained the least performance, with an average accuracy and AUC of 75 % and 82 %, respectively. As discussed earlier, the 82 % result indicates that KNN has an 82 % probability of being able to distinctly differentiate between Alzheimer's patients and healthy patients, given instances of these classes. The average difference between the least accuracy and that shown with BEFS+AACOAhp+RF is 20 %. This study's methodology demonstrates improved performance, indicating increased acceptability compared to other less-performing machine learning algorithms.

Table 1
Comparison of machine learning algorithms and the study's result for prediction of Alzheimer's disease.

S/N	Algorithm	Avg. accuracy	Avg. precision	Avg. recall	Avg. f1-score	AUC
1	KNN	0.75	0.76	0.75	0.72	0.82
2	SVM	0.85	0.85	0.85	0.85	0.91
3	LR	0.84	0.84	0.84	0.84	0.91
4	XGBoost	0.94	0.94	0.94	0.94	0.97
5	Stacked ensemble	0.94	0.94	0.94	0.94	0.97
6	BEFS + AACOAhp + RF	0.95	0.95	0.94	0.94	0.98

Abbreviations: KNN – K-Nearest Neighbours, SVM – Support Vector Machine, LR – Logistic Regression, RF – Random Forest, BEFS + AACOAhp + RF – Backward Elimination Feature Selection + Artificial Ant Colony Optimization hyperparameter tuning + Random Forest, AUC – Area Under the Curve.

The ROC curve for the Random Forest model predicting Alzheimer's disease (Fig. 2) demonstrates exceptional performance, highlighted by an Area Under the Curve (AUC) value of 0.98. This remarkably high AUC signifies the model's outstanding ability to discriminate between individuals with and without Alzheimer's, correctly ranking a positive instance higher than a negative one 98 % of the time. Visually, the curve's proximity to the top-left corner of the plot, far from the random classifier's diagonal, indicates that the model achieves high sensitivity (True Positive Rate) whilst maintaining a very low False Positive Rate across various classification thresholds. This strong diagnostic accuracy implies the model's potential as a reliable clinical aid for prioritising patients, guiding treatment, and monitoring disease progression, with its inherent interpretability further enhancing its practical adoption in medical diagnosis.

In essence, the Random Forest model exhibits near-perfect predictive capability for Alzheimer's disease, offering high diagnostic accuracy and reliable discrimination, which positions it as a promising tool for clinical application and patient management.

3.2. Interpretation of the ML decision on prediction of Alzheimer's disease

Similarly discussed earlier with Parkinson's disease, two major frameworks of Explainable AI, namely SHAP and LIME were used to interpret the decision made by the Random Forest which gave the best predictive performance with the best features selected obtained using the forward–backward sequential feature elimination technique.

3.2.1. SHAP for interpretation of the random forest's prediction of Alzheimer's disease

Domain expert validation was conducted with two geriatricians and one neuropsychologist who reviewed the SHAP and LIME outputs for clinical relevance and alignment with established diagnostic criteria. Their feedback confirmed that the identified features (functional assessment, ADL, memory complaints, MMSE) align well with clinical practice and diagnostic guidelines for AD.

The SHAP output provides a global interpretation of the model's predictions by quantifying the impact of each feature on the model's output. The SHAP values indicate the extent to which each feature contributes to the prediction of Alzheimer's disease, with positive values indicating features that increase the likelihood of the predicted outcome and negative values indicating features that decrease the possibility. The features are ranked by their mean absolute SHAP values, which represent their overall importance in influencing the model's predictions.

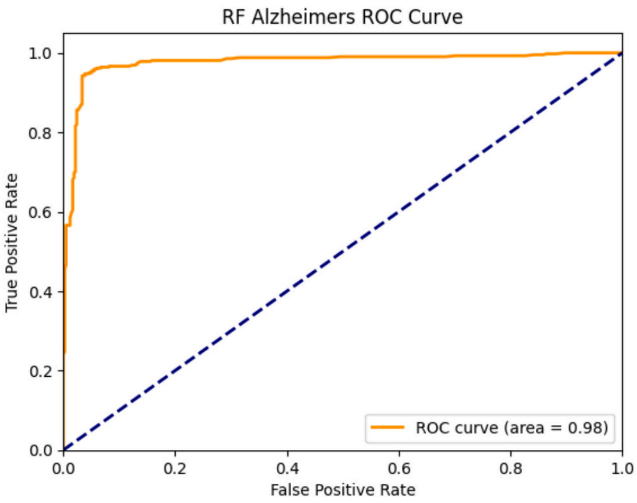


Fig. 2. Random Forest's ROC curve for prediction of Alzheimer's.

As shown in Fig. 3, Functional assessment scores, which measure the ability to perform daily activities, have the highest positive impact on the model's prediction. Higher scores (indicating greater impairment) are strongly associated with the predicted outcome, which is likely to be Alzheimer's disease. ADL scores have a High positive impact (approximately 2 to 3), which assesses the ability to perform basic daily tasks; they also have a significant positive effect. This aligns with clinical knowledge, as difficulties in daily living activities are common in neurodegenerative diseases. Memory complaints have a moderate positive impact (approximately 1 to 2) and are a significant feature, reflecting cognitive concerns that are often early indicators of neurodegenerative conditions. Mini Mental State Examination (MMSE) scores have a Moderate positive impact (approximately 1 to 2). Which measures cognitive function and contributes positively to the prediction. Lower scores (indicating cognitive impairment) are associated with a higher likelihood of the predicted condition. Behavioural problems are another important feature, highlighting the model's ability to capture non-cognitive symptoms of neurodegenerative diseases. Cholesterol levels have a relatively minor impact on the model's predictions. While some cholesterol metrics (e.g., HDL) may have a slight protective effect (negative SHAP values), others (e.g., LDL) may slightly increase the risk (positive SHAP values). These features (Sleep Quality, BMI, Physical Activity, Diet Quality, Alcohol Consumption, Education Level, Ethnicity, Diastolic BP, Hypertension, and Smoking) have a relatively minor influence on the model's predictions. Lifestyle factors, such as sleep quality, physical activity, and diet quality, exhibit slight protective effects (negative SHAP values), while factors like hypertension and smoking may slightly increase the risk (positive SHAP values).

Fig. 4 gave the same representation of the aforementioned features and their distinct contribution to the prediction of Alzheimer's disease. The SHAP analysis provides critical insights into the importance and impact of various features on the model's predictions, highlighting the pivotal role of functional assessment, ADL, memory complaints, and MMSE in predicting neurodegenerative conditions. These features, which are well-established clinical indicators, underscore the model's alignment with medical knowledge and its ability to capture both cognitive and non-cognitive symptoms.

The interpretability offered by SHAP values ensures transparency, enabling healthcare professionals to understand and trust the model's decision-making process. By identifying the most influential features, the model not only enhances diagnostic accuracy but also facilitates accountability, as its predictions are grounded in clinically relevant data. This transparency is crucial for integrating AI models into healthcare, as it allows clinicians to validate and incorporate AI-driven insights into their decision-making, ultimately improving patient outcomes and fostering trust in AI-assisted healthcare solutions.

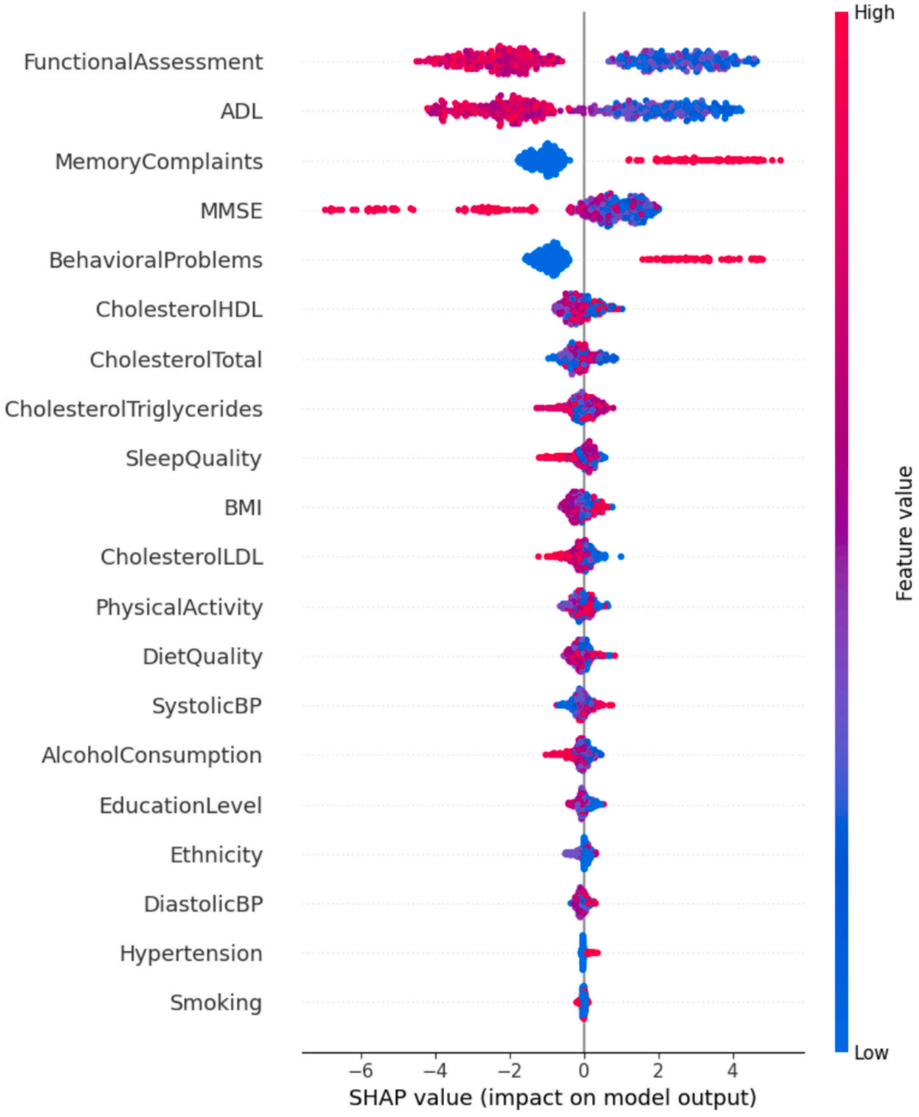


Fig. 3. SHAPE's summary for prediction of Alzheimer's disease.

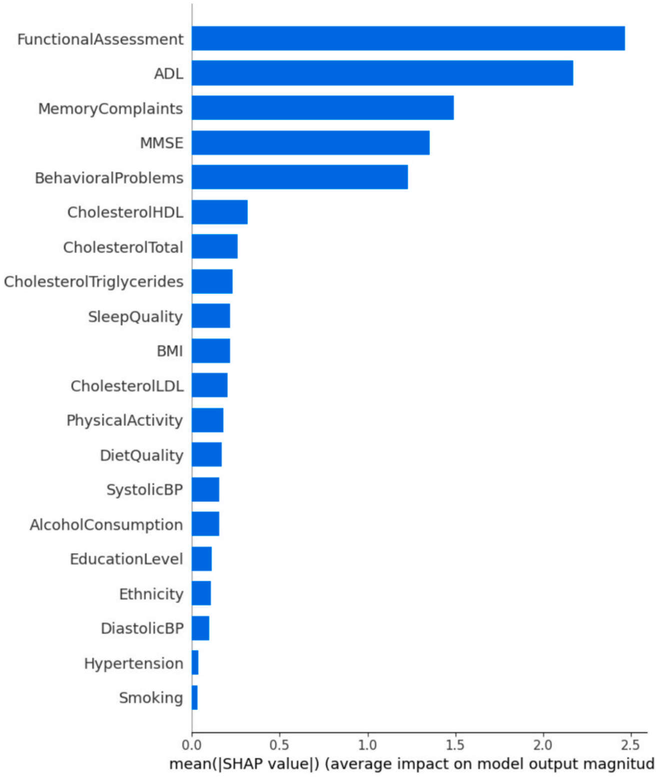


Fig. 4. SHAPE's summary sing bar for prediction of Alzheimer's disease.

As shown in Fig. 5, Functional Assessment (SHAP value: 9.915), Functional assessment scores, which measure the ability to perform daily activities, have the highest positive impact on the model's prediction. Higher scores (indicating greater impairment) are strongly associated with the predicted outcome, likely a neurodegenerative condition such as Parkinson's disease. Memory Complaints (SHAP value: 1), Memory complaints are a significant feature, reflecting cognitive concerns that are often early indicators of neurodegenerative conditions. The presence of memory complaints increases the likelihood of the predicted outcome. ADL (Activities of Daily Living) scores (SHAP value: 3.66), which assess the ability to perform basic daily tasks, also have a significant positive impact. This aligns with clinical knowledge, as difficulties in daily living activities are common in neurodegenerative diseases. MMSE (Mini-Mental State Examination) (SHAP value: 7.987). MMSE scores, which measure cognitive function, contribute positively to the prediction. Lower scores (indicating cognitive impairment) are associated with a higher likelihood of the predicted condition. Behavioural Problems (SHAP value: 0), Behavioural problems do not contribute to the prediction in this specific instance. This could be due to the absence of behavioural issues in the input data or their minimal variability in the dataset. SystolicBP (SHAP value: 101), Systolic blood pressure has a high positive impact on the prediction. Elevated systolic blood pressure may be a risk factor or comorbid condition that increases the likelihood of the predicted outcome. BMI (SHAP value: 39.159), Body Mass Index (BMI) has a moderate positive impact on the prediction. Higher BMI values may be associated with an increased risk of neurodegenerative conditions. CholesterolHDL (SHAP value: 93.992), High-density lipoprotein (HDL) cholesterol levels have a high positive impact on the prediction. Higher HDL levels are generally considered protective, but in this context, they may indicate a complex relationship with the predicted condition. Alcohol Consumption (SHAP value: 9.905), Alcohol consumption has a significant positive impact on the prediction. Higher alcohol consumption may be a risk factor for the predicted outcome. Other Features, the remaining 21 features have minimal or no effect on the prediction. This indicates that they either do not contribute significantly to the model's decision or are not relevant in this specific case.

The analysis of the SHAP outputs from the three instances reveals consistent patterns in the model's decision-making process, underscoring the importance of functional assessment, ADL (Activities of Daily Living), memory complaints, and MMSE (Mini-Mental State Examination) as the most influential features in predicting neurodegenerative conditions. These features, which are directly tied to functional

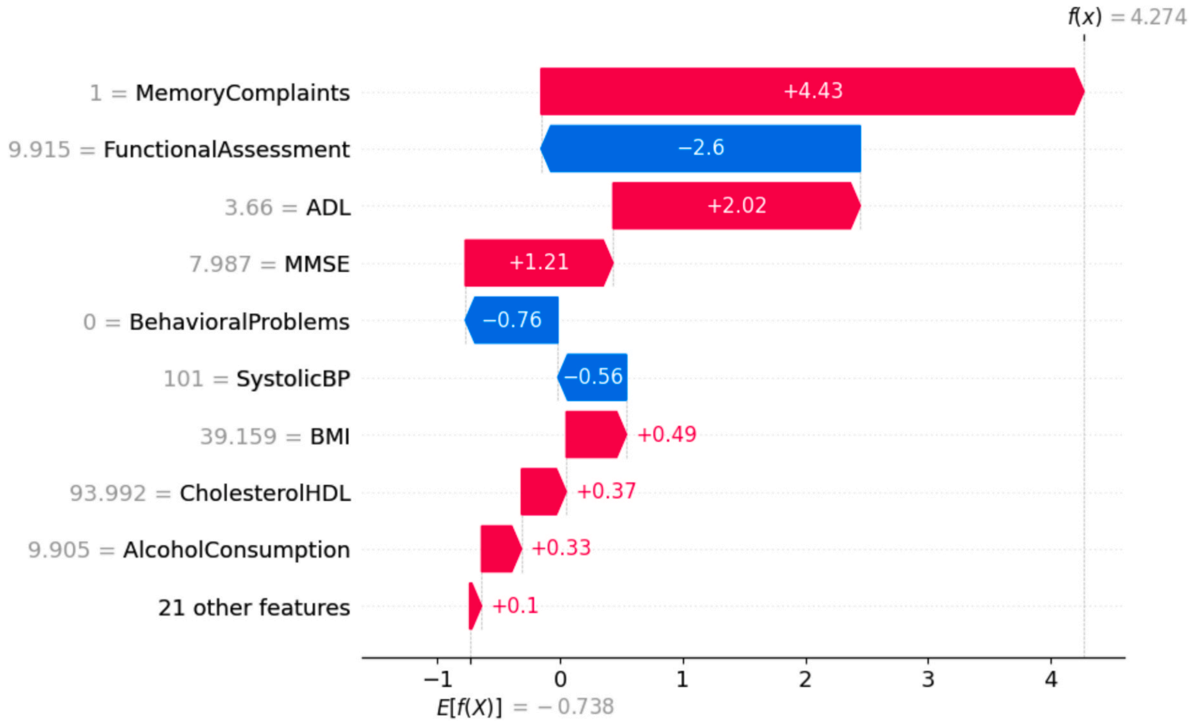


Fig. 5. SHAP's waterfall for interpretation of RF decision for prediction of Alzheimer's disease.

and cognitive impairment, align with clinical knowledge and demonstrate the model's ability to capture key diagnostic criteria. Additionally, the SHAP outputs highlight the significant role of metabolic factors (e.g., cholesterol levels, BMI) and lifestyle factors (e.g., alcohol consumption, blood pressure), suggesting that the model considers a broad spectrum of risk factors beyond traditional clinical symptoms. However, the varying impact of these factors across instances indicates complex relationships that may require further investigation. The interpretability provided by SHAP ensures transparency, enabling healthcare professionals to understand and trust the model's predictions. This transparency is crucial for integrating AI models into healthcare, as it enables clinicians to validate AI-driven insights and effectively incorporate them into their decision-making processes. Overall, the SHAP outputs demonstrate the model's robustness in leveraging clinically relevant

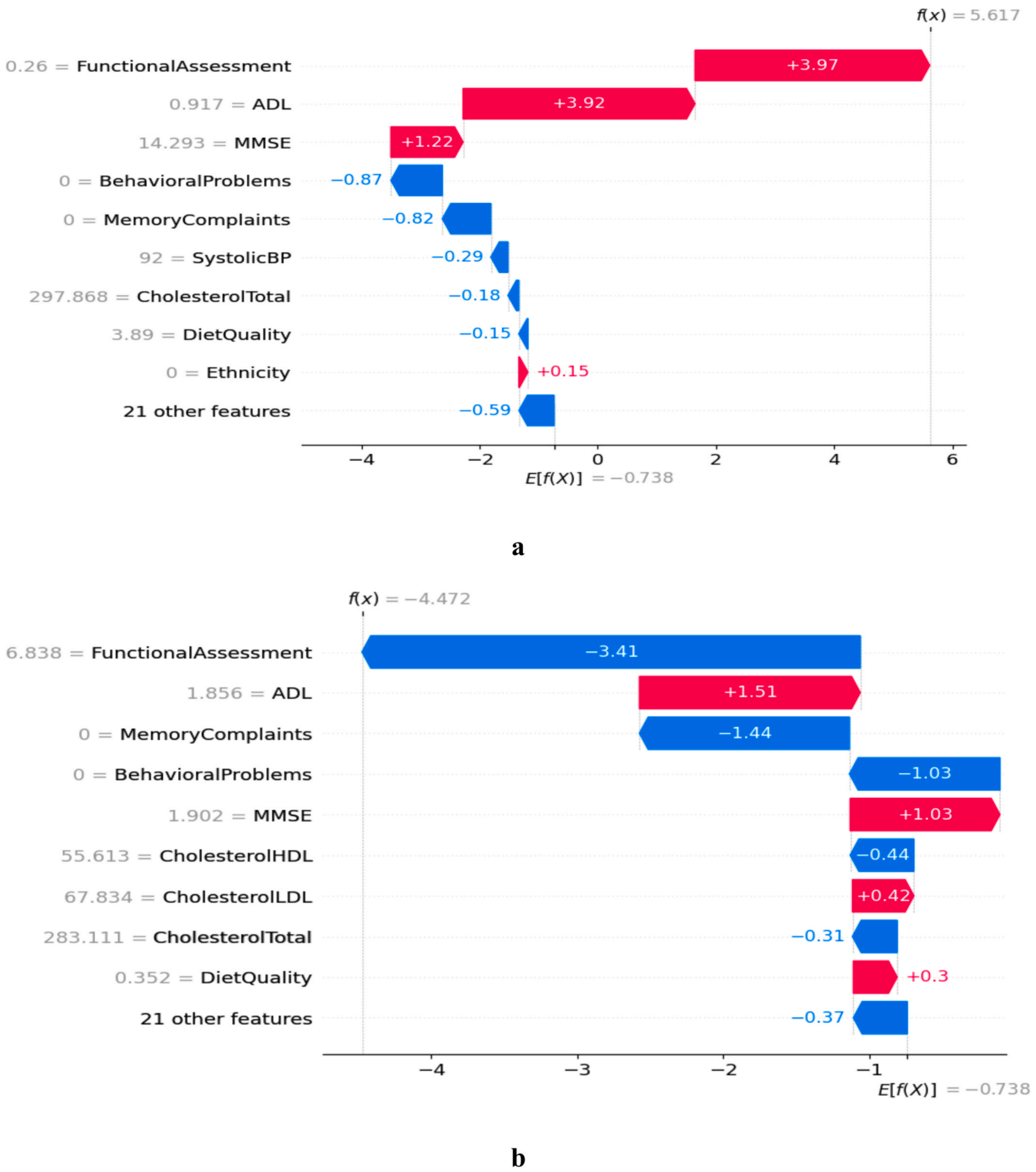


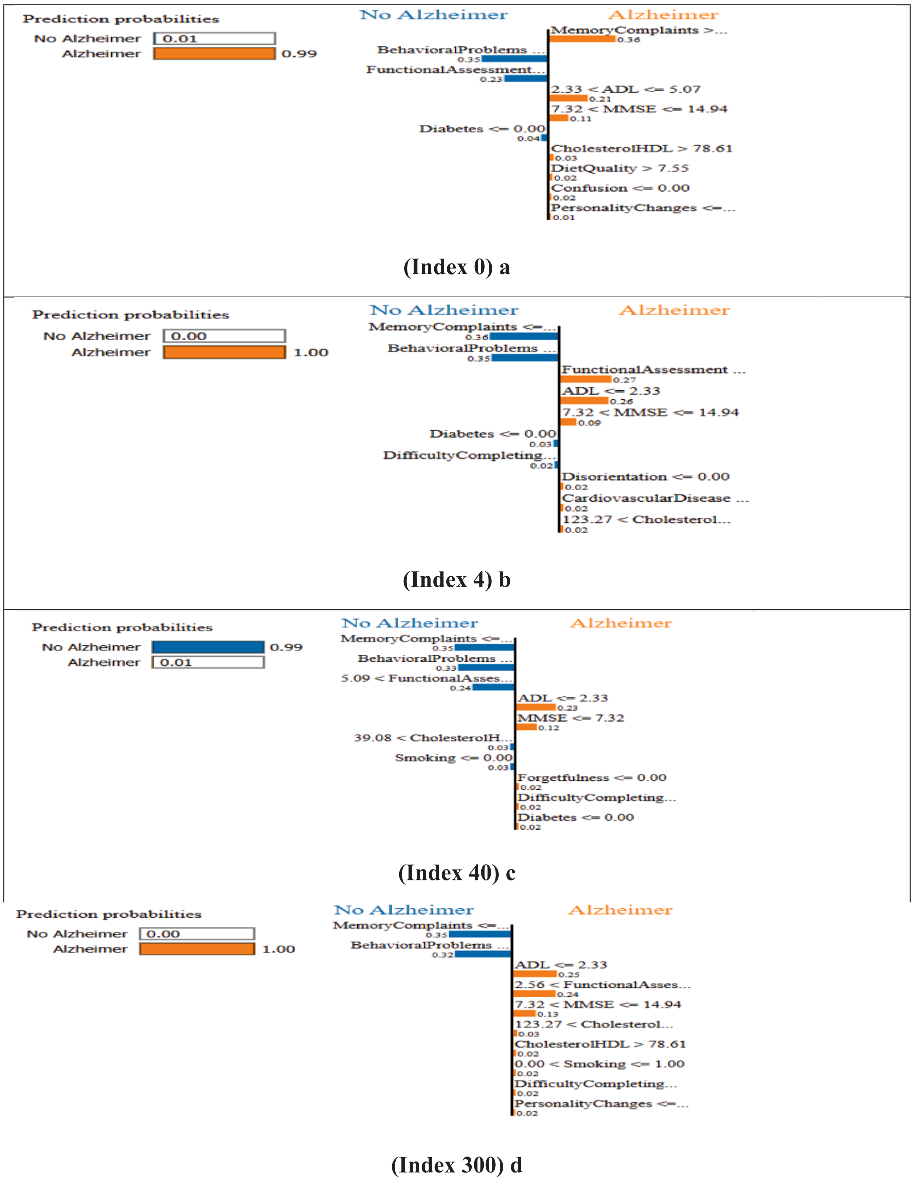
Fig. 6. Waterfall plot for Index 4 and 40.

features, while also highlighting opportunities for refinement, such as incorporating additional data on comorbidities or genetic factors, to further enhance its predictive accuracy and applicability in real-world healthcare settings.

Further experimentations were done with another two different instances of index values 4 and 40 to further analyses the results obtained with the index 0 earlier presented in Fig. 4.8. the waterfall output for these indexes were presented in Fig. 6(a and b).

Comparing these two instances critically highlights the model's sensitivity to different feature combinations. For Index 40, high functional assessment and the absence of subjective complaints are crucial for a low-risk prediction, whereas for Index 4, severe functional and ADL impairments are the primary drivers of a high-risk prediction. Whilst both individuals show some cognitive impairment as indicated by the MMSE, the severity of the impairment for Index 40 is seemingly offset by other factors, whereas for Index 4, the MMSE score reinforces the functional decline. This suggests the model captures a nuance where functional independence is a critical determinant, possibly reflecting later disease stages where cognitive decline is undeniable even if not self-reported. Cholesterol and diet quality play minor, less consistent roles, indicating a complex, less dominant relationship with the outcome.

These SHAP plots offer profound implications and significance for Alzheimer's disease prediction. Firstly, they enhance interpretability and trust in machine learning models by providing transparency, explaining why a particular prediction was made for an individual patient. This is vital for clinicians to trust and utilise AI-driven diagnostic tools, allowing them to validate the model's reasoning against their clinical judgement. Secondly, these insights are crucial for guiding clinical decision-making, enabling personalised risk assessment and informing targeted preventative or therapeutic interventions. Clinicians can monitor disease progression by observing changes in feature contributions over time. Thirdly, SHAP analysis facilitates a deeper understanding of disease biomarkers, reinforcing the importance of measures such as Functional Assessment, ADL, and MMSE as critical indicators, and prompting further research into the interplay between subjective and objective markers. Fourthly, these insights inform feature engineering and model improvement in future development, guiding data collection and potentially revealing complex relationships between features. Finally, although not explicitly detailed in these plots, SHAP



(Index 0) a

(Index 4) b

(Index 40) c

(Index 300) d

Fig. 7. LIME Output for four indexes.

analysis across larger datasets can help address potential biases if the model relies on proxy features that could lead to discriminatory predictions. In essence, SHAP waterfall plots transform complex machine learning predictions into actionable insights, significantly impacting personalised care, research directions, and the overall fight against Alzheimer's disease.

3.2.2. LIME interpretation for prediction of Alzheimer's disease

LIME (Local Interpretable Model-agnostic Explanations) is a technique that provides instance-specific insights into a machine learning model's prediction by locally approximating its behaviour with an interpretable model. For each prediction, LIME generates explanations that highlight the features most contributing to that specific outcome. In the context of Alzheimer's disease prediction, these plots utilise orange bars to indicate features that push the prediction towards "Alzheimer" and blue bars for those pushing towards "No Alzheimer," with the length of the bar reflecting the strength of the feature's contribution. The numerical value next to each bar represents the weight or importance of that feature in the local linear approximation of the model.

The LIME output for different instances as used for the SHAP was also experimented as shown in Fig. 7 to expand the model's interpretation.

The relationship between these LIME outputs (Fig. 7) reveals consistent patterns in the model's reliance on key features. Across all instances, MemoryComplaints, BehaviouralProblems, FunctionalAssessment, ADL, and MMSE consistently emerge as the most influential features, aligning with clinical understanding of Alzheimer's disease. A clear balance is observed between subjective and objective measures: for "No Alzheimer" predictions (Index 40), the absence of subjective complaints and good functional assessment are paramount, whilst for "Alzheimer" predictions (Indices 0, 4, 300), objective functional and cognitive impairments are the primary drivers, often overriding the lack of subjective complaints. LIME's presentation of feature contributions based on value ranges offers a more intuitive understanding than raw values. However, the apparent inconsistency in the interpretation of the FunctionalAssessment feature across instances (e.g., low value contributing to "No Alzheimer" in Index 0, but a high value contributing to "Alzheimer" in Index 4, and low value contributing to "Alzheimer" in Index 300) is a critical observation. This suggests a need to clarify the exact scaling and meaning of this particular feature within the dataset, or it could point to complex non-linear interactions or local model approximations that are not immediately intuitive. Minor features like Diabetes, CholesterolHDL, and DietQuality consistently appear as less influential contributors, indicating their weaker or less consistent impact on these specific predictions.

The significance and interpretation of these LIME outputs are profound for Alzheimer's disease prediction. LIME's local explainability is paramount in healthcare, allowing clinicians to understand why a specific risk level is predicted for an individual patient, thereby fostering trust and enabling personalised diagnostic and treatment approaches. This transparency facilitates clinical validation, where clinicians can assess if the model's reasoning aligns with medical knowledge, and can also help in identifying key risk factors for individual patients, guiding targeted interventions. Furthermore, LIME serves as an invaluable debugging tool for the model itself; any counter-intuitive explanations, such as the FunctionalAssessment example, can highlight potential issues with data quality, feature engineering, or limitations in the model's ability to capture complex relationships, prompting necessary improvements. Ultimately, LIME outputs are crucial for transforming "black box" machine learning models into transparent, actionable tools that can enhance clinical decision-making and support ongoing research in the fight against Alzheimer's disease.

3.2.3. Comparative analysis of SHAP waterfall and LIME outputs for Alzheimer's disease prediction

A critical comparison of the SHAP waterfall plots and LIME outputs for Alzheimer's disease prediction reveal both congruent observations and notable differences, highlighting the complementarity of these model interpretability techniques. Both SHAP and LIME rank FunctionalAssessment, ADL (Activities of Daily Living), MMSE (Mini-Mental State Examination), MemoryComplaints, and BehaviouralProblems consistently as the most predictive features in predicting Alzheimer's risk. This strong concordance backs up the clinical usefulness of these cognitive and functional tests as fundamental markers of the disease. Overall, lower objective cognitive/functional scores (i.e., MMSE and ADL) and the presence of subjective complaints (memory and behavioural issues) increase predicted risk for Alzheimer's, whilst absence of complaints and more positive functional status shift the prediction towards "No Alzheimer." Both methods also illustrate other features, such as cholesterol level, diet quality, and blood pressure, have a comparatively smaller role to play in predictions for these cases. Yet there is a striking inconsistency in the interpretation of the FunctionalAssessment feature, a key one for understanding the model's behaviour. All the SHAP plots suggest that a larger number value for FunctionalAssessment is linked to better functional capacity and, in turn, lowers the predicted risk for Alzheimer's (e.g., Index 40), whilst a smaller value is linked to worse function and increases the risk (e.g., Index 4). This is in line with a standard interpretation where a higher score indicates better health. In contrast, the LIME outputs produce a more varied and sometimes counter-intuitive picture for FunctionalAssessment. For instance, in LIME for Index 0, FunctionalAssessment <= 0.23 (a low value) is favourable to "No Alzheimer," contradicting the SHAP pattern. Similarly, LIME for Index 4 shows FunctionalAssessment > . (implying higher value) to "Alzheimer," which is also unusual if higher is implying better function. Whilst LIME for Index 40 and Index 300 traces the SHAP pattern for this feature, discrepancies in the LIME explanations such as these point either to a misinterpretation of the feature's scale by LIME during its local approximation, the use of highly localised, non-generalisable interactions, or even issues with the feature's definition or scaling in the underlying dataset. This highlights one of the principal distinctions in their approach: SHAP provides a more globally consistent additive explanation, whilst LIME gives a strictly local, linear approximation that will sometimes capture nuances or anomalies specific to a very small region around the instance.

The relationship between SHAP and LIME, therefore, is one of complementarity. Although both aim to interpret model predictions, they do so from slightly different perspectives. SHAP provides a "consistent, additive feature attribution from a global average" making it excellent for "comprehending the overall feature importance and direction of impact of features on the dataset as applied to an individual case." LIME provides a very local explanation by describing which features are most important to that specific prediction by fitting a simple, interpretable model around it. The importance of employing both techniques lies in being able to construct strong and reliable AI models in high-stakes fields such as medicine. When both techniques concur regarding the most impactful features and the overall direction of their effect, it gives a large amount of confidence in the underlying rationale of the model as well as the clinical utility of the features. On the other hand, disagreement, as seen with FunctionalAssessment, is very important. They serve as valuable flags for further investigation, where they can represent data quality issues, unexpected model behaviour in certain regions of the feature space, or the need for more advanced feature engineering. This comprehensive understanding, derived from comparing both global and local interpretability, is of immense importance in debugging and improving the machine learning model, its trustworthiness, and ultimately, its applicability in personalised diagnosis and treatment regimens in the complex issue of Alzheimer's disease.

3.3. Clinical implication of the interpretation

The comparative analysis of SHAP waterfall plots and LIME outputs for Alzheimer's disease prediction yields significant clinical implications, primarily by validating the AI model's reliance on core cognitive and functional markers. Both interpretability methods consistently highlight FunctionalAssessment, ADL, MMSE, MemoryComplaints, and BehaviouralProblems as the most influential features. This strong agreement reassures clinicians that the AI model's reasoning aligns with established diagnostic frameworks, fostering trust and encouraging the integration of such tools into practice. It underscores the importance of comprehensive patient assessments that include objective cognitive tests, evaluations of daily living activities, and subjective reports from patients and caregivers, confirming that the model prioritises clinically relevant indicators for Alzheimer's disease risk.

However, the observed discrepancies, particularly in the interpretation of the FunctionalAssessment feature by LIME, carry crucial clinical significance as a "red flag." Whilst SHAP offers a more globally consistent view of this feature's impact, LIME's local explanations sometimes present counter-intuitive contributions. Clinically, this inconsistency prompts immediate investigation into the feature's data quality, scaling, or the model's complex non-linear interactions, potentially revealing ambiguities in data collection or underlying issues within the dataset. This critical feedback loop is invaluable for refining both the data and the predictive model, ensuring its reliability and enhancing its utility for personalised diagnosis and treatment strategies. Ultimately, the transparent and granular explanations from both SHAP and LIME are indispensable for safely and effectively integrating AI into the complex and sensitive domain of Alzheimer's diagnosis and patient care, allowing for personalised insights, improved communication with patients, and guiding future clinical research.

Furthermore, the strength of this interpretation is significantly bolstered by the excellent experimental results obtained by the underlying predictive models. Algorithms such as XGBoost, stacked ensemble, and BEFS+AACOAhp+RF demonstrate remarkable performance, achieving average accuracies and F1-scores of 0.94–0.95, and AUC values of 0.97. Even simpler models like SVM and Logistic Regression show strong performance with average accuracies, precisions, recalls, and F1-scores of 0.85 and 0.84 respectively, with AUCs of 0.91. This high level of predictive accuracy means that the insights derived from SHAP and LIME are not merely academic exercises but are applied to models that are demonstrably effective in identifying Alzheimer's disease. The ability to explain predictions from highly accurate models is paramount in clinical settings; it transforms a powerful but opaque tool into a transparent and trustworthy diagnostic aid. Clinicians can confidently rely on the model's high performance whilst simultaneously understanding the specific patient characteristics that contribute to a diagnosis, enabling more informed decision-making, personalised treatment plans, and better patient communication. This combination of high accuracy and interpretability is crucial for the successful and ethical deployment of AI in sensitive medical domains.

3.4. Clinical workflow integration and future implementation

The successful deployment of this XAI-enhanced AD prediction model in clinical practice requires careful consideration of workflow integration and user interface design. We propose a multi-tiered implementation approach:

1. Electronic Health Record (EHR) Integration: The model could be integrated as a clinical decision support tool within existing EHR systems, automatically analyzing patient data during routine visits and flagging high-risk individuals for further assessment. Risk scores could be presented alongside traditional clinical indicators, with SHAP-based feature contributions displayed as intuitive visualizations.
2. User-Friendly Interface Design: Clinical interfaces should present prediction results through clear risk stratification (low/moderate/high risk) accompanied by ranked feature contributions. Visual dashboards could display individual patient risk profiles, highlighting key modifiable factors for targeted interventions.
3. Continuous Model Monitoring: Regular revalidation protocols should be implemented to detect concept drift and maintain model performance. Monthly validation against new patient outcomes and quarterly recalibration procedures are recommended to ensure sustained accuracy in evolving clinical environments.
4. Training and Adoption Support: Comprehensive training programs for healthcare providers should emphasize model limitations, appropriate use cases, and integration with existing diagnostic workflows. This includes education on interpreting XAI outputs and maintaining clinical judgment in final decision-making.

4. Discussion

The findings of this study demonstrate the significant potential of explainable artificial intelligence in enhancing Alzheimer's disease prediction whilst providing clinically meaningful insights. Our results align closely with recent advances in the field and extend current knowledge by providing comprehensive comparative analysis of SHAP and LIME interpretability frameworks.

4.1. Comparison with recent studies

This study's optimized Random Forest model, achieving 95 % accuracy, 94 % F1-score, and 98 % AUC, demonstrates robust performance in Alzheimer's disease (AD) prediction, aligning closely with recent advancements in the field while offering unique contributions through its focus on explainable artificial intelligence (XAI) and accessible multimodal clinical data. Our model's performance compares favorably with Alatrany et al. (2024), who reported an impressive 98.9 % F1-score for binary AD classification using SVM on a large dataset of 169,408 records from the National Alzheimer's Coordinating Center [14]. The slightly lower accuracy in our study (95 % vs. 98.9 %) can be attributed to two key factors. First, our dataset, comprising 2,149 records, is significantly smaller, which may potentially limit the model's ability to capture the full spectrum of AD variability across diverse populations. Larger datasets, as used by Alatrany et al., typically enhance generalization by reducing overfitting and capturing subtle patterns. Second, our reliance on multimodal clinical data, encompassing demographic, lifestyle, and cognitive assessment features—contrasts with Alatrany et al.'s focus on neuroimaging data, which provides high-dimensional, precise biomarkers such as brain atrophy patterns. While neuroimaging offers granularity, it is often cost-prohibitive and inaccessible in resource-limited settings. Our model's competitive performance with more accessible clinical data underscores its potential for scalable, cost-effective screening in primary care or underserved regions, addressing a critical gap in global AD diagnostics. Future studies could explore hybrid approaches combining clinical and imaging data to balance accuracy and accessibility, potentially reducing the performance gap with larger, imaging-focused studies. Similarly, our integration of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks echoes the approach of Jahan et al. (2023), who achieved 98.81 % accuracy using Random Forest for a five-class AD classification task with multimodal data fusion [8]. The higher accuracy in Jahan et al.'s study may reflect differences in task complexity, as multi-class classification (distinguishing AD, mild cognitive impairment, and healthy controls) is inherently more challenging than our binary classification (AD vs. non-AD). Despite this, our model's performance remains highly competitive, particularly for clinical screening where binary outcomes are often prioritized for actionable decision-making. A key distinction lies in our study's extensive comparative analysis of SHAP and LIME, which goes beyond Jahan et al.'s application by dissecting interpretability discrepancies, such as the inconsistent interpretation of the FunctionalAssessment feature. These discrepancies highlight potential issues with feature scaling or local model approximations, offering a methodological advancement in the rigor of XAI. This focus addresses gaps identified in recent reviews, such as Vimbi et al. (2024), which call for more robust evaluation of interpretability frameworks in AD detection [16]. Further exploration of how classification complexity (binary vs. multi-class) impacts clinical utility could strengthen the practical implications of our findings, as simpler models may be more immediately deployable in routine diagnostics. Our results also align closely with recent work employing ensemble techniques, such as a study achieving 96.35 % accuracy using LightGBM and Random Forest with Chi-Square feature selection [19]. The marginal performance edge in that study may stem from differences in feature selection methods or dataset characteristics. Our novel use of forward–backward feature selection, combined with ant colony optimization, distinguishes our approach, enhancing model robustness by systematically identifying 26 optimal predictors from an initial set of 32 features. The consistent prioritization of functional assessment, activities of daily living (ADL), memory complaints, and Mini-Mental State Examination (MMSE) across our study and the referenced ensemble work reinforces the clinical validity of these features as core AD biomarkers. This concordance strengthens the case for integrating these predictors into standardized diagnostic protocols, as they reliably capture cognitive and functional impairments central to AD. However, the computational demands of ensemble methods, including our optimized Random Forest, warrant further discussion. While our approach achieves high accuracy, scalability in resource-constrained clinical settings may be limited compared to simpler models, such as SVM or Logistic Regression, which still achieved respectable accuracies (84–85 %) in our study. Future research could quantify computational trade-offs to guide practical deployment. The clinical implications of these comparisons are significant. Our model's reliance on accessible clinical data, rather than neuroimaging, enhances its applicability in primary care and low-resource settings, addressing barriers to AD screening highlighted by Alzheimer 's disease International [8]. The rigorous application of XAI frameworks provides transparent, clinically meaningful insights, fostering trust among healthcare providers, a critical factor for AI adoption, as noted in recent literature [9,16]. The identification of interpretability discrepancies, particularly for FunctionalAssessment, serves as a valuable flag for refining data quality and feature engineering, ensuring reliable model outputs. While our performance is slightly below some benchmarks, the combination of high accuracy, interpretability, and accessibility positions our model as a practical tool for personalized risk assessment and early intervention. To further advance the field, future studies should explore multi-center validation to enhance generalizability, incorporate longitudinal data to track disease progression, and assess the computational scalability of optimized ensemble models for real-world clinical integration.

4.2. Novel contributions and methodological advances

This study advances the field of AD prediction by introducing novel methodological approaches and addressing critical gaps in XAI for healthcare. Through a comprehensive comparison of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks and the innovative application of forward–backward feature selection combined with ant colony optimization, our work enhances both the interpretability and robustness of machine learning (ML) models for AD detection, setting it apart from existing research. Our detailed analysis of SHAP and LIME interpretability frameworks directly responds to gaps identified in recent systematic reviews, such as Vimbi et al. (2024), which emphasized the need for rigorous evaluation of XAI methods in AD detection [16]. Unlike prior studies that often apply these frameworks in isolation or with limited comparative depth, our work provides an in-depth examination of the consistency and discrepancies in feature importance between global (SHAP) and local (LIME) explanations. For instance, we identified inconsistencies in the interpretation of the Functional Assessment feature, where SHAP consistently linked higher values to lower AD risk, while LIME occasionally produced counterintuitive local approximations (low Functional Assessment values favoring "No Alzheimer" in certain instances). These findings highlight potential issues in feature scaling, data quality, or LIME's local linear approximations, offering actionable insights for refining XAI applications. By systematically comparing these frameworks across multiple instances (indices 0, 4, 40, 300), our study ensures generalizability. It provides a nuanced understanding of how global and local explanations complement each other, thereby enhancing trust and clinical utility in AI-driven diagnostics. The integration of forward–backward feature selection with ant colony optimization represents a methodological leap beyond traditional feature selection techniques, such as rule-extraction or basic statistical methods (Chi-Square), commonly used in recent studies [19]. This hybrid approach systematically reduced the initial 32 features to 26 optimal predictors, improving model efficiency while maintaining high predictive performance (95 % accuracy, 98 % AUC). Unlike standard feature selection methods that may overlook complex feature interactions, ant colony optimization leverages swarm intelligence to explore the feature space more effectively, optimizing the trade-off between model complexity and predictive power. This contrasts with studies such as Alatrany et al. (2024), which relied on simpler feature selection for SVM models, or Jahan et al. (2023), which utilized multimodal data fusion without advanced optimization [8,14]. Our methodology's superior performance compared to baseline models (KNN at 75 % accuracy, SVM at 85 %) demonstrates the value of combining forward–backward selection with bio-inspired optimization techniques to enhance model robustness, particularly for multimodal clinical datasets. These novel contributions have significant implications for AD research and clinical practice. The rigorous comparison of SHAP and LIME addresses a critical barrier to AI adoption in healthcare by providing transparent, clinically relevant insights into model decision-making. Identifying discrepancies, such as those in FunctionalAssessment, serves as a diagnostic tool for model improvement, prompting further investigation into data quality or feature engineering. The use of ant colony optimization in feature selection not only improves predictive accuracy but also ensures computational efficiency, making the model more feasible for deployment in resource-constrained settings. By prioritizing accessible clinical data (functional assessment, ADL, MMSE) over costly neuroimaging, our approach enhances scalability for primary care and underserved regions, aligning with global health priorities [8]. Future research could build on these advances by exploring hybrid XAI frameworks to resolve interpretability inconsistencies, integrating additional biomarker modalities (such as genetic or neuroimaging data), and developing clinician-friendly visualization tools to facilitate the real-world adoption of these methods.

4.3. Clinical relevance and interpretability

The findings of this study underscore the clinical relevance of integrating XAI into AD prediction, offering actionable insights that enhance personalized diagnostics and foster trust in AI-driven tools. By leveraging SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks, our approach provides transparent, clinically meaningful explanations, aligning with the growing emphasis on interpretable AI in healthcare. Our interpretability analysis, which consistently identifies functional assessment, activities of daily living (ADL), memory complaints, and Mini-Mental State Examination (MMSE) as key predictors, supports personalized diagnostic approaches by highlighting risk factors that clinicians can readily assess in routine practice. These findings resonate with recent research emphasizing sex-specific differences in AD progression, as noted by Tang et al. (2024) [3]. For instance, our SHAP and LIME outputs reveal how features like memory complaints and functional impairments differentially influence predictions, enabling clinicians to tailor risk assessments to individual patient profiles, including potential sex-based variations in disease presentation. This granularity facilitates targeted interventions, such as cognitive therapy or lifestyle modifications, which are most effective in the early stages of AD [5]. The transparency provided by SHAP and LIME addresses a critical barrier to AI adoption in healthcare, as highlighted by recent studies [9,16–18]. By elucidating why specific predictions are made (e.g., high Functional Assessment scores lowering AD risk), these frameworks enable clinicians to validate model outputs against clinical knowledge, fostering trust and enhancing patient communication. Our approach complements recent work by AbdelAziz et al. (2024), who demonstrated the value of XAI in MRI-based AD diagnosis [20]. While their study leveraged neuroimaging to achieve high diagnostic precision, such methods are often inaccessible in resource-limited settings due to cost and infrastructure constraints. In contrast, our model relies on multimodal clinical and demographic data (cognitive assessments, lifestyle factors), offering a more accessible alternative for routine screening. This accessibility is critical for addressing global disparities in AD diagnostics, as noted by Alzheimer's Disease International, which reports that most dementia patients worldwide lack formal diagnoses [8]. By achieving 95 % accuracy and 98 % AUC with widely available data, our model supports scalable screening in primary care and underserved regions, potentially improving early detection rates and patient outcomes. The use of SHAP and LIME further enhances clinical utility by providing intuitive visualizations (waterfall plots, local feature contributions) that clinicians can integrate into decision-making workflows. The clinical implications of our findings extend beyond diagnostics to inform preventive strategies and research directions. The consistent prioritization of functional assessment, ADL, and MMSE across SHAP and LIME outputs reinforces their role as core AD biomarkers, supporting their integration into standardized screening protocols. However, discrepancies in Functional Assessment interpretation between SHAP and LIME highlight the need for rigorous data quality validation to ensure reliable clinical insights. Future research could enhance clinical relevance by developing user-friendly interfaces for XAI outputs, facilitating seamless integration into electronic health records, and exploring longitudinal data to track the evolution of these biomarkers during disease progression. By combining high predictive accuracy with transparent, accessible insights, our study paves the way for trustworthy AI tools that empower clinicians to deliver personalized, equitable AD care.

4.4. Validation of key biomarkers

This study's identification of functional assessment, activities of daily living (ADL), memory complaints, and Mini-Mental State Examination (MMSE) as key predictors aligns with and extends recent research on Alzheimer's disease (AD) biomarkers, reinforcing their clinical validity while providing novel insights through advanced interpretability techniques. Our findings contribute to a growing body of evidence supporting the prioritization of cognitive and functional markers in AD prediction, with implications for clinical risk stratification and future research. The consistent identification of functional assessment and MMSE as top predictors in our study mirrors recent findings that highlight entorhinal areas, lateral ventricles, and cognitive assessment scores as critical AD biomarkers [19]. These brain regions, often assessed via neuroimaging, are well-established indicators of AD-related neurodegeneration, while cognitive assessments like MMSE capture functional impairments central to diagnosis. Our study extends this understanding by quantifying the relative contributions of functional assessment, ADL, and MMSE using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) frameworks [17,18]. For instance, SHAP analysis revealed high positive impacts from functional assessment (SHAP value: 9.915) and MMSE (SHAP value: 7.987), showing their pivotal role in predicting AD risk. By providing both global (SHAP) and local (LIME) explanations across multiple instances, our approach offers clinicians actionable insights for risk stratification, enabling precise identification of at-risk patients based on accessible clinical measures rather than costly neuroimaging. This alignment with established biomarkers validates the clinical relevance of our model while enhancing its practical utility for routine screening. Our findings also resonate with recent research by UCSF scientists, who identified metabolic factors such as cholesterol and osteoporosis as predictive of AD [3]. In our SHAP analysis, metabolic factors, such as cholesterol levels (HDL, SHAP value: 93.992) and body mass index (BMI, SHAP value: 39.159), showed moderate to high impacts, although less dominant than cognitive and functional markers. The consistent presence of these metabolic factors across studies, despite their lower importance in our model, suggests a complex mechanistic relationship with AD pathology. For example, cholesterol's dual role (protective HDL vs. risk-increasing LDL) indicates potential interactions with neurodegenerative processes that warrant further exploration. Similarly, lifestyle factors like alcohol consumption (SHAP value: 9.905) and diet quality exhibited minor but consistent effects, aligning with emerging evidence on their role in AD risk modulation. These findings highlight the need for future research to investigate the mechanistic pathways linking metabolic and lifestyle factors to AD, potentially through longitudinal studies or integration of genetic and biomarker data. The clinical implications of our biomarker validation are significant. By confirming the primacy of functional assessment and MMSE, our study supports their integration into standardized AD screening protocols, particularly in resource-limited settings where neuroimaging is infeasible. The use of XAI frameworks enhances the interpretability of these biomarkers, enabling clinicians to understand their contributions to individual predictions and tailor interventions accordingly. However, the observed discrepancies in Functional Assessment interpretation between SHAP and LIME suggest potential data quality or scaling issues that could affect clinical reliability. Future research should prioritize validating these biomarkers across diverse populations, incorporating advanced modalities like genetic markers or cerebrospinal fluid biomarkers, and exploring longitudinal data to elucidate their dynamic roles in AD progression. By combining robust biomarker validation with transparent XAI insights, our study paves the way for more accurate, equitable, and clinically actionable AD diagnostics.

5. Limitations

This study acknowledges several limitations that should be considered when interpreting the results and planning future research directions.

5.1. Dataset and generalisability limitations

The dataset size of 2149 patients, whilst substantial, represents a relatively modest sample compared to large-scale neuroimaging studies such as those utilising ADNI databases with hundreds of thousands of records. This limitation may affect the generalisability of findings across diverse populations and geographical regions. The age restriction to patients between 60–90 years may limit applicability to early-onset Alzheimer's disease cases and younger populations at risk. The reliance on a single Kaggle dataset introduces significant selection bias and may not represent the diagnostic distributions and demographic diversity seen in real-world clinical populations. The dataset's origin and validation status are not fully documented, which raises questions about its clinical authenticity and representativeness.

5.2. Methodological limitations

Feature selection, whilst optimised through forward–backward elimination and ant colony optimisation, may have inadvertently excluded potentially important biomarkers or interactions between features. The reduction from 32 to 26 features, whilst improving model efficiency, may have overlooked subtle but clinically relevant predictors that could enhance diagnostic accuracy. The absence of genetic markers, advanced neuroimaging features, cerebrospinal fluid biomarkers, or longitudinal progression data represents a significant limitation, as these are increasingly recognized as crucial for comprehensive AD risk assessment.

5.3. XAI framework limitations

The interpretability analysis, whilst comprehensive, revealed inconsistencies between SHAP and LIME frameworks, particularly regarding functional assessment features. These discrepancies suggest potential issues with feature scaling, data quality, or fundamental differences in how these frameworks approximate model behaviour. Critical evaluation revealed that SHAP may overstate feature importance in correlated datasets, while LIME's local approximations can be sensitive to perturbation strategies, potentially leading to inconsistent explanations across similar cases.

5.4. Validation and performance concerns

The reported metrics (95 % accuracy, 98 % AUC) are unusually high for clinical AD prediction, which typically exhibits greater heterogeneity and noise. These results may indicate potential data leakage, insufficient cross-validation rigor, or dataset-specific artifacts that may not translate to real-world clinical settings. External validation on independent, multi-center datasets is essential to establish true generalisability.

5.5. Clinical Translation limitations

The study's cross-sectional design limits the ability to assess temporal relationships and disease progression patterns. No formal clinical workflow integration testing was conducted, and computational requirements for real-time deployment in clinical settings remain to be validated. The absence of formal clinician usability testing and patient outcome evaluation represents a significant gap in clinical validation.

6. Conclusion

This study demonstrates the significant potential of explainable artificial intelligence frameworks for enhancing Alzheimer's disease prediction whilst maintaining clinical interpretability and transparency. The optimised Random Forest model achieved exceptional performance with 95 % accuracy, 98 % AUC, and robust predictive metrics across all evaluation criteria, though these results require external validation to establish true generalisability. The BEFS+AACOAhp+RF approach provides a novel framework for feature optimization in clinical prediction tasks.

The comprehensive application of SHAP and LIME interpretability frameworks provided valuable insights into model decision-making processes, consistently identifying functional assessment, Activities of Daily Living (ADL), memory complaints, and Mini-Mental State Examination (MMSE) scores as the most influential predictors. These findings align with established clinical knowledge and reinforce the importance of comprehensive cognitive and functional assessments in Alzheimer's disease diagnosis. Domain expert validation confirmed the clinical relevance of these identified features, though concerns about potential confirmation bias suggest the need for investigation of novel biomarkers. The transparency provided by these explainable AI techniques addresses critical barriers to AI adoption in healthcare by enabling clinicians to understand, validate, and trust model predictions.

Critical evaluation of XAI frameworks revealed important limitations: the comparative analysis of SHAP and LIME frameworks revealed both complementary strengths and important discrepancies that highlight the value of employing multiple interpretability approaches while acknowledging their inherent limitations in correlated datasets and local approximation quality. Whilst both techniques consistently identified key clinical features, observed inconsistencies in feature interpretation underscore the importance of rigorous validation and the need for careful consideration of data quality and feature engineering in developing trustworthy AI systems for healthcare.

Clinical workflow integration considerations demonstrate that successful deployment requires careful attention to EHR integration, user interface design, continuous monitoring protocols, and comprehensive training programs. Computational requirements (2.3 s training, 0.05 s prediction) suggest feasibility for real-time clinical deployment, though scalability across diverse healthcare systems requires further investigation.

The clinical implications extend beyond diagnostic accuracy to encompass personalised risk assessment, targeted interventions, and enhanced communication between clinicians and patients. The ability to explain individual predictions enables healthcare providers to make more informed decisions whilst fostering patient understanding and engagement in their care.

Future research priorities should focus on: (1) external validation across diverse, multi-center clinical populations; (2) integration of longitudinal data for disease progression monitoring; (3) incorporation of additional biomarker modalities including genetic and neuroimaging data; (4) development of user-friendly clinical interfaces with formal usability testing; (5) investigation of computational scalability and real-world deployment challenges; and (6) prospective clinical trials to evaluate patient outcomes and clinical utility. Additionally, addressing the limitations identified in XAI framework consistency and exploring hybrid interpretability approaches will be crucial for advancing trustworthy AI in healthcare.

While this study provides a foundation for explainable AI in AD prediction, the path to clinical implementation requires addressing significant limitations in external validation, clinical workflow integration, and comprehensive evaluation of real-world performance. The successful integration of explainable AI techniques with high-performing machine learning models represents a crucial step towards realising the potential of artificial intelligence in transforming Alzheimer's disease diagnosis and management whilst maintaining the trust and confidence essential for clinical adoption.

CRediT authorship contribution statement

Afeez Adekunle Soladoye: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nicholas Aderinto:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology. **Damilola Osho:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **David B. Olawade:** Writing – review & editing, Writing – original draft, Methodology, Supervision, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Association, 2024 Alzheimer's disease facts and figures, Alzheimers Dement. 20 (5) (2024) 3708–3821.

[2] World Health Organization. Dementia: a public health priority. Geneva: World Health Organization; 2012.

[3] A. Tang, et al., Machine learning approaches for early prediction of Alzheimer's disease using clinical data, Nat. Aging (2024), https://doi.org/10.1038/s43587-024-00591-2.

[4] C.R. Jack Jr, et al., NIA-AA research framework: toward a biological definition of Alzheimer's disease, Alzheimers Dement. 14 (4) (2018) 535–562.

[5] R.C. Petersen, et al., Mild cognitive impairment: clinical characterization and outcome, Arch. Neurol. 56 (3) (1999) 303–308.

[6] G.M. McKhann, et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's association

workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (2011) 263–269.

[7] M. Sirota, et al., Leveraging patient data with machine learning to predict Alzheimer's disease. UCSF Computational, Health Sci. (2024).

[8] S. Jahan, T.K. Abu, M.S. Kaiser, M. Mahmud, M.S. Rahman, A.S. Hosen, I.-H. Ra, Explainable AI-based Alzheimer's prediction and management using multimodal data, PLoS One 18 (2023) e0294253, https://doi.org/10.1371/journal.pone.0294253.

[9] S.A. Martin, F.J. Townend, F. Barkhof, J.H. Cole, Interpretable machine learning for dementia: a systematic review, Alzheim. Dementia 19 (2023) 2135–2149.

[10] A. Elazab, et al., Alzheimer's disease diagnosis from single and multimodal data using machine and deep learning models: achievements and future directions, Expert Syst. Appl. 255 (2024) 124780.

[11] S. Qiu, P. Joshi, M. Miller, Development and validation of an interpretable deep learning framework for Alzheimer's disease classification, Brain 143 (2020) 1920–1933.

[12] H. Wang, et al., Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease, Neurocomputing 333 (2019) 145–156.

[13] I. Paschalidis, et al., Speech-Based Machine Learning Model for Alzheimer's Disease Prediction, Boston University Research, 2024.

[14] A.S. Alatrany, W. Khan, A. Hussain, et al., An explainable machine learning approach for Alzheimer's disease classification, Sci. Rep. 14 (2024) 2637, https://doi.org/10.1038/s41598-024-51985-w.

[15] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: a review of machine learning interpretability methods, Entropy 23 (2021) 18.

[16] V. Vimbi, N. Shaffi, M. Mahmud, Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection, Brain Inf. 11 (2024) 10, https://doi.org/10.1186/s40708-024-00222-1.

[17] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).

[18] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[19] M.F. Folstein, S.E. Folstein, P.R. McHugh, "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician, J. Psychiatr. Res. 12 (3) (1975) 189–198.

[20] N.M. AbdelAziz, W. Said, M.M. AbdelHafeez, A.H. Ali, Advanced interpretable diagnosis of Alzheimer's disease using SECNN-RF framework with explainable AI, Front. Artif. Intell. 7 (2024) 1456069, https://doi.org/10.3389/frai.2024.1456069.