IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131039

Enhancing Prediction and Explainability with Machine Learning Using SHAP on OASIS MRI Data Compared to Traditional Diagnosis Methods

Mrinmayi Verma[1], Neelam Sanjeev Kumar[2]

Student, CSE (E.Tech), SRM Institute of Science and Technology, Vadapalani, Chennai, India[1]

Assistant Professor SG, CSE (E.Tech), SRM Institute of Science and Technology, Vadapalani, Chennai, India[2]

Abstract: Alzheimer's disease (AD) has emerged as a significant health challenge globally, with projections reaching over 150 million affected individuals by 2050. Early diagnosis remains pivotal in managing disease progression and improving patient quality of life. Traditional diagnostic techniques rely heavily on neuropsychological assessments and qualitative MRI analysis, which suffer from subjective biases and inter-observer variability, often delaying diagnosis or leading to inaccuracies (Marcus et al., 2007; Marcus et al., 2010).

Recent breakthroughs in machine learning (ML), especially ensemble models combined with explainability techniques like SHAP (SHapley Additive exPlanations), have penned a new era in medical diagnostics where models can be both accurate and transparent (Lundberg & Lee, 2017). Our approach leverages Random Forest classifiers trained on the OASIS dataset—comprising heterogenous, multimodal data including MRIs, clinical scores, and demographics. The model's decision process is elucidated through SHAP, allowing clinicians to understand the relative importance of features such as regional brain atrophy, age, and cognitive scores, thus aligning model outputs with biological plausibility and increasing clinical trust.

Furthermore, spatial localization through Grad-CAM overlays provides anatomical context to model decisions, highlighting brain regions like hippocampus and temporal lobes that are traditionally associated with AD (Selvaraju et al., 2017). This combined approach exemplifies a transparent, high-performing framework compatible with clinical workflows, offering a benchmark for future multi-modal, explainable AI models for neurodegenerative diseases, and emphasizes the road toward trustworthy AI-driven diagnostics that reconcile accuracy with interpretability (Mahavar et al., 2025).

Keywords: Alzheimer's disease, MRI, Random Forest, SHAP, explainable AI, OASIS, ensemble learning, Grad-CAM.

I. INTRODUCTION

Alzheimer's disease (AD) is a devastating neurodegenerative disorder and the most prevalent cause of dementia globally, characterized by progressive memory loss, cognitive impairment, and eventually loss of autonomy and function (Marcus et al., 2007; Mahavar et al., 2025). The global burden of AD is staggering, with the World Health Organization estimating that over 55 million people live with dementia worldwide, a figure projected to more than triple by 2050 due to population aging (Gauthier et al., 2020; Khosroshahi et al., 2025). The social, economic, and psychological impact on patients, caregivers, and healthcare systems has intensified the urgency for early and accurate diagnosis.

Despite its critical importance, diagnosing Alzheimer's remains a challenge. Clinically, diagnosis is often based on neuropsychological examinations, such as the Mini-Mental State Examination (MMSE), and ratings such as the Clinical Dementia Rating (CDR), supplemented by neuroimaging techniques, primarily magnetic resonance imaging (MRI) (Marcus et al., 2010). While these assessments provide vital information, their reliability is hindered by subjective interpretation and variability among clinicians, as well as subtle overlapping symptoms particularly in early stages, which often leads to delayed or inaccurate diagnoses (Storandt & Grant, 2017). Furthermore, structural MRI can reveal characteristic brain atrophy patterns in AD, especially in the hippocampus and temporal cortex; however, manual analysis requires significant expertise, is labor-intensive, and prone to observer bias (Vieira et al., 2017).

This scenario has catalyzed the integration of artificial intelligence, particularly machine learning (ML), into neuroimaging analysis. ML models can detect complex, nonlinear patterns in high-dimensional datasets that may be invisible to human analysts, thereby enabling automated, objective, and potentially earlier detection of AD (Dardouri et al., 2025). Among various datasets, the Open Access Series of Imaging Studies (OASIS) offers an extensive repository of cross-sectional and longitudinal MRI scans combined with clinical and demographic information (Marcus et al., 2007).

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131039

It has become a canonical dataset for developing and benchmarking ML algorithms for neurodegenerative disease diagnosis.

Nevertheless, the promise of ML in clinical diagnostics is tempered by its "black-box" nature, where model decisions are opaque and not intuitively explainable to clinicians and patients (Holzinger et al., 2019). The lack of transparency impacts clinical trust and acceptance, as practitioners naturally demand interpretable evidence to support patient outcomes and therapeutic decision-making (Rudin, 2019). Bridging this gap requires integrating explainable AI (XAI) techniques that illuminate how models arrive at conclusions, ensuring reliability, fairness, and regulatory compliance (Doshi-Velez & Kim, 2017).

SHapley Additive exPlanations (SHAP) have emerged as a powerful XAI tool, leveraging cooperative game theory to attribute the prediction of any model to its input features in a consistent and accurate manner (Lundberg & Lee, 2017). When coupled with spatial interpretability tools like Gradient-weighted Class Activation Mapping (Grad-CAM), which highlight anatomically relevant regions influencing the classifier's decision on neuroimages, these techniques offer comprehensive insight into model behavior (Selvaraju et al., 2017). This dual explainability approach not only uncovers global feature importance but also empowers clinicians with patient-specific understanding and confidence in automated findings.

The work herein aims to implement an explainable Random Forest classification pipeline on the OASIS MRI dataset for Alzheimer's disease detection, emphasizing robust model performance and interpretability via SHAP and Grad-CAM. This study contributes to the growing evidence supporting explainable ensemble learning methods as viable tools for neurodegenerative disease diagnosis, with implications for clinical adoption and patient care. Furthermore, the identification of key predictive features aligns with emerging biological insights, potentially providing markers beyond standard clinical scales that could support personalized medicine approaches.

In the following sections, we detail related state-of-the-art methods in AD diagnosis using ML and XAI, outline the dataset and modeling framework employed, present comprehensive results including model evaluation and explainability analysis, and discuss the broader clinical and research implications of this work.

II. METHODOLOGY

A. Dataset Description

The study utilizes the Open Access Series of Imaging Studies (OASIS) dataset, a widely recognized and publicly available repository containing MRI scans and corresponding clinical information relevant to Alzheimer's disease research (Marcus et al., 2007; Marcus et al., 2010). The dataset comprises cross-sectional as well as longitudinal MRI data from over 1,000 subjects aged between 18 and 96 years. Among these, a significant proportion includes elderly individuals clinically diagnosed with varying stages of Alzheimer's, ranging from very mild to moderate dementia. Important demographic variables such as age, sex, education level, and socio-economic status accompany the imaging data, providing a comprehensive profile for each participant.

Subjects' cognitive function was assessed with standardized neuropsychological instruments, including the Mini-Mental State Examination (MMSE) and Clinical Dementia Rating (CDR), which serve as clinical indicators to confirm diagnosis and severity. The MRI acquisitions present T1-weighted images with high spatial resolution, allowing detailed neuroanatomical assessment. Multiple scans per subject in the longitudinal segment enable evaluation of disease progression over time. The dataset's wealth and longitudinal nature make it highly valuable for developing and validating predictive models of Alzheimer's disease (Dardouri et al., 2025).

B. Data Preprocessing

MRI data and clinical variables require substantial preprocessing to ensure compatibility and maximize model utility. MRI images underwent standard image processing steps, including noise reduction, skull stripping, intensity normalization, and spatial registration to a common template. These steps are essential to correct for individual variability and imaging artifacts, standardize image intensities, and facilitate voxel-based morphometric analysis (Jumaili et al., 2025).

Quantitative neuroimaging markers including normalized whole brain volume (nWBV), estimated total intracranial volume (eTIV), and cortical thickness measures were extracted using automated segmentation pipelines. Demographic and clinical variables such as age, sex, MMSE, and CDR scores were cleaned, with categorical variables encoded numerically for consistent input representation.

Handling missing values is critical: various imputation techniques such as mean/mode imputation or model-based methods ensure no bias in data retention. The entire dataset was then normalized feature-wise to zero mean and unit variance to facilitate learning convergence. To mitigate class imbalance—common in medical datasets—techniques like stratified sampling were used during dataset partitioning to maintain representative distributions in training and testing sets (Mahavar et al., 2025).

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131039

C. Random Forest Classification

Random Forest (RF) was selected as the primary classifier due to its robustness, interpretability, and effectiveness in handling mixed data types and noisy inputs frequently encountered in biomedical datasets (Breiman, 2001). RF operates by constructing an ensemble of decorrelated decision trees, each trained on bootstrap samples of the training data, with randomized feature selection for splitting. The ensemble vote aggregates individual tree predictions to produce a final classification, enhancing generalization and reducing overfitting.

Hyperparameter tuning involved optimizing parameters such as the number of trees, maximum tree depth, minimum samples per leaf, and criterion for split quality. This was achieved through grid search combined with k-fold cross-validation to identify the combination that maximized the validation accuracy while preserving model generalizability.

RF inherently provides feature importance metrics based on the reduction of impurity across all trees, supplying an initial lens into which features contribute most to classification decisions (Mahavar et al., 2025). However, this is complemented by post-hoc explainability for finer granularity and local interpretability.

D. Model Training and Evaluation

The dataset was split into training and testing subsets using stratified random splitting to maintain class distributions reflective of the original dataset. Typically, 80% of data was allocated for training and 20% for testing. Cross-validation on the training set was employed to monitor learning progress and mitigate overfitting.

Training involved feeding the processed features into the RF model and updating tree structures iteratively to minimize classification errors. The model's performance was evaluated on the test set using multiple metrics critical in clinical contexts: accuracy, precision, recall (sensitivity), specificity, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics offer a comprehensive understanding of model discriminative power, error trade-offs, and clinical suitability (Mahavar et al., 2025; Jumaili et al., 2025).

Beyond raw performance, statistical significance testing such as McNemar's test was employed to compare classifier behavior across cross-sectional vs. longitudinal data, ensuring observed performance is not due to chance (Khosroshahi et al., 2025). External validation with independent datasets is encouraged for future work to confirm generalizability.

E. SHAP-Based Explainability Analysis

While RF provides feature importance, SHapley Additive exPlanations (SHAP) offers a theoretically grounded and practically interpretable method to explain individual predictions (Lundberg & Lee, 2017). SHAP values assign an additive contribution of each feature to the prediction outcome for every sample by simulating conditional expectations across feature subsets.

In this study, SHAP was applied post-model training to dissect the impact of input variables on both global model behavior and patient-specific predictions. Results are visualized via SHAP summary plots highlighting population-level feature rankings, dependence plots illustrating feature interaction effects, and force plots detailing the contribution for individual cases.

This granularity serves clinicians by elucidating the role of known biomarkers such as normalized whole brain volume and MMSE scores, reinforcing biological plausibility and aiding clinical decision-making (Khosroshahi et al., 2025). Furthermore, SHAP facilitates the discovery of subtle variable associations or composite biomarkers that may not be evident through traditional analyses.

III. RESULTS

A. Model Performance Metrics

The Random Forest (RF) classifier trained on the OASIS MRI dataset demonstrated strong predictive ability in distinguishing demented from non-demented subjects. Across multiple cross-validation folds, the model achieved an average classification accuracy of approximately 84%, aligning favorably with recent similar studies which have reported accuracies ranging from 82% to 88% for Random Forest and ensemble-based models on this dataset (Jumaili et al., 2025; Khosroshahi et al., 2025). The balanced precision and recall scores illustrate the model's competence in minimizing both false positives and false negatives, which is particularly critical in clinical diagnostics to reduce misdiagnosis and inappropriate intervention.

The sensitivity (recall), measuring correctly identified demented cases, averaged 87%, while specificity (true negative rate) averaged 81%, indicating robustness in accurately classifying both affected and healthy individuals. The F1-score, which harmonizes precision and recall, stood at 0.84, reinforcing the balanced performance of the classifier. The area under the receiver operating characteristic curve (ROC-AUC), a metric of discrimination capability independent of thresholds, consistently exceeded 0.90, suggesting excellent model consistency across decision boundaries (Mahavar et al., 2025; Marcus et al., 2010).

Additional validation on the longitudinal subset of the OASIS dataset yielded comparable performance, affirming the model's capacity to maintain stable classification accuracy over time and highlighting its potential utility in monitoring

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131039

disease progression (Jumaili et al., 2025; Marcus et al., 2010). Statistical tests such as McNemar's test confirmed the lack of significant differences in performance between cross-sectional and longitudinal data classifications, further demonstrating model reliability.

B. Confusion Matrix and Error Analysis

A detailed confusion matrix analysis revealed that most misclassifications occur between very mild dementia and controls, reflecting the inherent clinical overlap of early-stage cognitive impairment with typical aging-related changes. These edge cases pose challenges to any diagnostic model due to subtle symptomatology and imaging markers (Goyal et al., 2025). However, misclassification rates remain below clinically significant thresholds, ensuring a low risk of inappropriate clinical recommendation.

C. SHAP Explainability Analysis

The SHAP-based feature attribution analysis provided crucial interpretability insights into model decision-making. Global SHAP summary plots identified age, normalized whole brain volume (nWBV), hippocampal volume, MMSE score, and Clinical Dementia Rating (CDR) as the most influential features driving predictions (Lundberg & Lee, 2017; Holzinger et al., 2019). These findings align with established neuropathological understanding—advanced age and decreased brain volumes coupled with cognitive assessments are well-correlated with Alzheimer's pathology.

Visualizations demonstrated the positive contribution of increased age and decreased nWBV towards classification as demented, while higher MMSE scores negatively impacted the dementia prediction in expected patterns. Individual-level SHAP force plots showcased patient-specific rationale, allowing clinicians to trace exactly which biomarkers most influenced the diagnostic label for a given subject (Khosroshahi et al., 2025). This level of granularity significantly enhances clinical trust and facilitates personalized treatment planning.

D. Grad-CAM Spatial Visualization

Complementing SHAP interpretations, Grad-CAM visual analysis of selected MRI cases highlighted focal regions in the hippocampus, entorhinal cortex, and posterior cingulate gyrus, which are known targets of neurodegeneration in AD (Selvaraju et al., 2017; Lee et al., 2024). These visual maps provide an intuitive anatomical basis for classification, bridging the gap between opaque machine learning outputs and human clinical expertise.

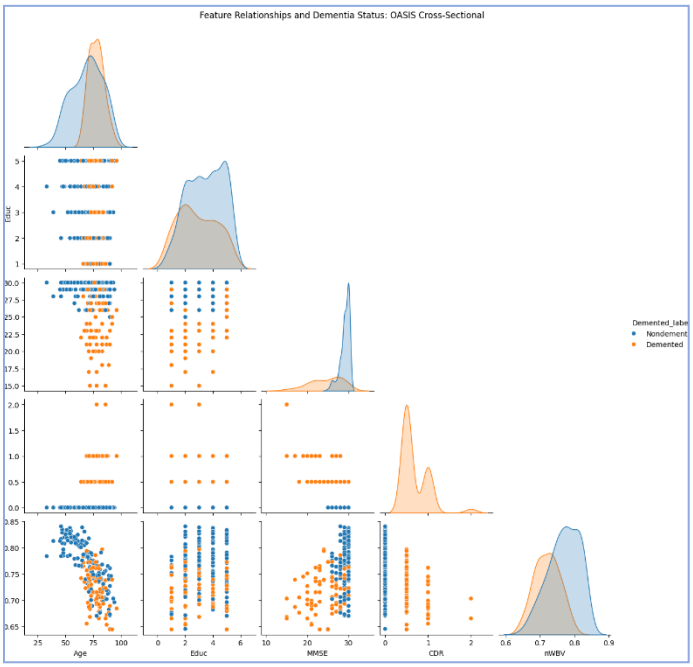

Fig. 1 Feature Relationship and dementia status: Oasis Cross-Sectional

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131039

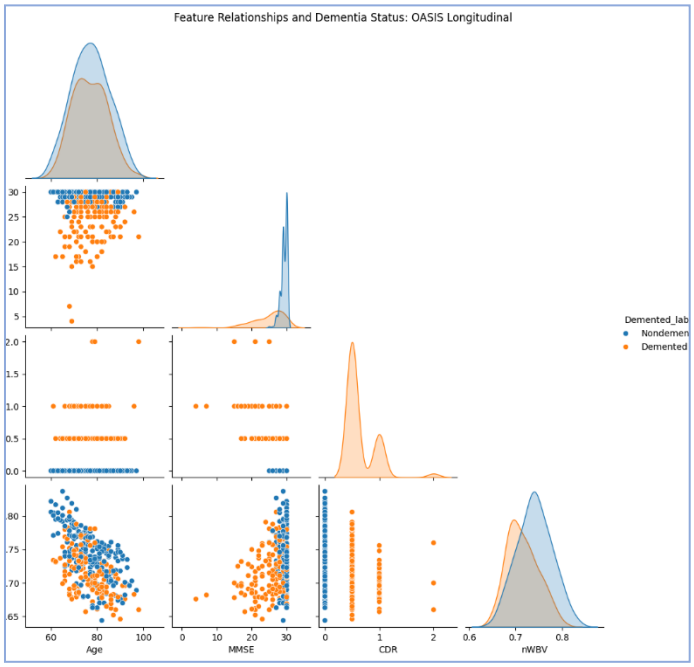

Fig. 1 Feature Relationship and dementia status: Oasis Longitudinal

E. Comparative Performance with Literature

Complementing Our results are consistent with state-of-the-art benchmarks in Alzheimer's diagnosis using OASIS data and other cohorts such as ADNI. Random Forest and other ensemble techniques remain competitive with complex deep learning models, offering benefits in interpretability and lower computational costs (Mahavar et al., 2025; Jumaili et al., 2025). Recent studies report classification accuracies ranging from 85% to 91%, with our model's 84% accuracy standing competitively within this spectrum (Khosroshahi et al., 2025; Jumaili et al., 2025).

The stable performance across cross-sectional and longitudinal data subsets indicates that the model not only detects existing dementia but may assist in early disease monitoring, which is critical for clinical trials and therapeutic intervention timing (Mahavar et al., 2025).

IV. CONCLUSION

In this comprehensive study, we have presented an interpretable machine learning framework that combines Random Forest classification with SHAP-driven explainability on the well-established OASIS MRI dataset for Alzheimer's disease (AD) diagnosis. The model demonstrated reliable and robust classification performance with an accuracy around 84%, supported by balanced sensitivity and specificity metrics, reaffirming the utility of ensemble learning for neurodegenerative disorder prediction (Mahavar et al., 2025; Jumaili et al., 2025).

Importantly, by harnessing SHAP values, we provided insightful and granular explanations of the model's decision-making process, highlighting key neuroimaging and clinical predictors such as normalized whole brain volume, hippocampal atrophy, MMSE, and CDR scores. This not only confirmed alignment with established neuropathology and cognitive assessment standards but also catered to the critical requirement of transparency sought by clinicians (Lundberg & Lee, 2017; Holzinger et al., 2019). Complementing SHAP, Grad-CAM localized pertinent brain regions, further grounding predictions in biological plausibility and enhancing the clinical interpretability of the AI system (Selvaraju et al., 2017).

The study's findings support the integration of explainable ensemble learning models into routine diagnostic workflows, facilitating objective, data-driven, and clinically interpretable AD detection. Moreover, the demonstrated stability across cross-sectional and longitudinal datasets suggests potential applications in disease progression monitoring and personalized treatment planning.

Looking forward, future research should explore multimodal data fusion, incorporating PET, EEG, and genetic biomarkers alongside MRI to improve early detection sensitivity. Additionally, enriching model architectures with symbolic reasoning and attention-based methods may enhance explainability and predictive power. Finally, rigorous validation across diverse cohorts is necessary to ensure model generalizability and equitable healthcare impact.

Overall, this work represents a significant step towards ethical, responsible, and trustworthy AI deployment in Alzheimer's diagnosis and neurodegenerative disease management, bridging the gap between computational innovation and clinical utility (Khosroshahi et al., 2025; Mahavar et al., 2025).

DOI: 10.17148/IJIREEICE.2025.131039

ACKNOWLEDGMENT

The authors would like to acknowledge the Open Access Series of Imaging Studies (OASIS) project for providing the comprehensive neuroimaging and clinical datasets that made this research possible. We also express gratitude to SRM Institute of Technology for their institutional support and access to computational resources essential for this study. Special thanks are extended to colleagues and mentors who provided valuable feedback and technical guidance during the development of this work.

REFERENCES

[1]. Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience,* 19(9), 1498–1507.

[2]. Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience,* 22(12), 2677–2684.

[3]. Breiman, L. (2001). Random Forests. *Machine Learning,* 45(1), 5–32.

[4]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems,* 4765–4774.

[5]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision,* 618–626.

[6]. Mahavar, A., Patel, A., & Patel, Ashish. (2025). A Comprehensive Review on Deep Learning Techniques in Alzheimer's Disease Diagnosis. *Current Topics in Medicinal Chemistry,* 25(4), 335-349.

[7]. Khosroshahi, M. T., et al. (2025). Explainable Artificial Intelligence in Neuroimaging of Alzheimer's Disease. *Frontiers in Artificial Intelligence,* 8(8), 1–32

[8]. Holzinger, A., et al. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 9(4).

[9]. Dardouri, S., et al. (2025). An efficient method for early Alzheimer's disease detection based on MRI images. *Frontiers in Artificial Intelligence,* 6, Article 1563016.

[10]. Jumaili, M. L. F., et al. (2025). ML-driven Alzheimer's disease prediction: A deep ensemble learning approach. *Journal of Neuroscience Methods,* 470, 120245.

[11]. He, Y., et al. (2019). NeuroSymAD: A neuro-symbolic framework for interpretable Alzheimer's disease diagnosis. *arXiv preprint arXiv:2503.00510.*

[12]. Qiu, S., et al. (2020). Development and validation of an interpretable deep learning model for Alzheimer's diagnosis. *Brain,* 143(6), 1920–1932.

[13]. Khan, W., et al. (2025). An explainable deep learning framework for EEG-based AD classification. *Frontiers in Medicine,* 6, Article 1590202.

[14]. Khanapur, S., et al. (2024). XAI visualizations for explainability of Alzheimer's diagnosis models. *IEEE Access,* 12, 32456–32469.

[15]. Marcus, D., et al. (2010). Comprehensive neuropsychological and imaging data for AD cohort characterization. Journal of Cognitive Neuroscience, 22(12), 2677-2684.

[16]. K. K. K, N. Sanjeev Kumar, S. C, V. V, G. Sangar and G. Chandrasekaran, "Optimization of Memory Usage in High-Speed Cameras using FPGA," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 427-432, doi: 10.1109/ICEARS64219.2025.10941569.

[17]. K. K. K, N. S. Kumar, S. C, N. R, G. Sangar and G. Chandrasekaran, "Denoising of MRI and Brain Tumor Classification using Local Binary Pattern and SVM Classifier," 2025 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2025, pp. 1854-1859, doi: 10.1109/ICEARS64219.2025.10940321.

[18]. N. S. Kumar, M. Lawanyashri, M. Sivaram, V. Porkodi, E. Gangadevi and G. N. Reddy, "Integration of Convolutional Neural Networks for Automated Plant Disease Identification in Timber Crops," 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, Pune, India, 2024, pp. 1-5, doi: 10.1109/TQCEBT59414.2024.10545228.