SHAP-Based Explainability for Local and Global Insights in Alzheimer's Detection

Shraddha Khanapur
Department of CSE, B.M.S. College of Engineering, India
shraddha.cs20@bmsce.ac.in (corresponding author)

Jyothi S. Nayak
Department of CSE, B.M.S. College of Engineering, India
jyothinayak.cse@bmsce.ac.in

B. S. Rajeshwari
Department of CSE, B.M.S. College of Engineering, India
rajeshwari.cse@bmsce.ac.in

M. Namratha
Department of CSE, B.M.S. College of Engineering, India
namratham.cse@bmsce.ac.in

Chirag B. Bharadwaj
Department of CSE, B.M.S. College of Engineering, India
chiraggb.cs20@bmsce.ac.in

Raghav Bhardwaj
Department of CSE, B.M.S. College of Engineering, India
raghav.cs20@bmsce.ac.in

Received: 7 August 2025 | Revised: 26 September 2025 and 22 October 2025 | Accepted: 24 October 2025

 DOI: https://doi.org/10.48084/etasr.13932

ABSTRACT

Alzheimer's disease is a progressive neurodegenerative disorder that leads to cognitive decline and loss of independence, making early and accurate diagnosis essential. Recent advances in Machine Learning (ML) have enhanced medical image analysis, but the opaque nature of deep learning models limits their adoption in clinical practice. This study introduces SCR NetX, a CNN model based on the VGG 16 architecture, to classify Alzheimer's disease into four stages: non demented, very mild, mild, and moderate dementia. To improve interpretability, the model integrates Explainable AI (XAI) using SHAP (SHapley Additive eXplanations) for both local and global analyses. Local explanations highlight MRI regions that influence individual predictions, aiding in case-specific evaluation, while global explanations reveal the overall behavior of the model. Two segmentation methods—grid-based for broad region analysis and SLIC (Simple Linear Iterative Clustering) for fine-grained superpixel analysis—are employed to ensure precise and clinically interpretable outputs. This framework combines accurate classification with transparent decision-making, bridging the gap between AI-driven diagnostics and practical clinical application.

Keywords-Alzheimer's disease; convolutional neural networks; medical image; explainable AI; interpretability

I. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia. More than 55 million people were affected in 2023, with numbers expected to be close to 140 million by 2050. AD causes a gradual decline in memory, cognition, and behavior, often leading to full dependence, and contributes to an annual global cost exceeding $1.3 trillion. It is characterized by beta amyloid plaques and tau tangles that damage neurons, with age, the APOE ε4 gene, and certain health conditions as major risk factors. Symptoms advance from mild memory issues to severe

dementia. Although incurable, early diagnosis with biomarkers and neuroimaging enables timely treatment, better care planning, and improved quality of life.

Machine Learning (ML) is transforming medical diagnosis and image analysis, enabling early detection of diseases through modalities such as CT and MRI. In AD detection, ML models provide scalable data-driven insights into early brain degeneration markers. Convolutional Neural Networks (CNNs) effectively identify neurodegenerative changes such as hippocampal atrophy from MRI scans, with advanced 3D CNNs that improve accuracy through volumetric data integration. Random Forests (RFs) combine clinical, cognitive, and imaging features, handling high-dimensional data for risk stratification. Support Vector Machines (SVMs) classify neurodegenerative patterns using kernel methods such as RBF for nonlinear data. Graph Convolutional Networks (GCNs) model brain connectivity from DTI and fMRI, capturing disruptions linked to AD.

The adoption of ML in healthcare remains limited, largely due to the lack of interpretability. Clinicians require AI systems to be transparent, reliable, and comprehensible, but many models—particularly Deep Learning (DL) methods such as CNNs—function as "black boxes," offering little insight into their decision-making. In AD diagnosis, where the results have significant implications, this opacity undermines trust and acceptance. Interpretability enables practitioners to verify predictions against clinical judgment, a necessity reinforced by regulations such as HIPAA and GDPR. Physicians are more likely to trust AI models that provide clear explanations. Explainable AI (XAI) addresses this by creating models that maintain accuracy while offering interpretable outputs, thereby supporting clinical integration and informed patient care.

A. *Contributions of this Study*

This work is divided into two key components: first, developing an accurate image classification model for diagnosing AD, and second, enhancing the interpretability of the model to ensure its practical application in clinical settings. Although building a reliable diagnostic tool is essential, the main focus is on introducing interpretability through XAI. In more detail, this study:

- Develops a reliable diagnostic model for AD. A CNN-based image classification model called SCR-NetX (using VGG-16 as the base model architecture) is developed and trained to accurately diagnose AD by classifying MRI scans into four different stages: non-demented, very mildly demented, mildly demented, and moderately demented. The goal is to ensure high diagnostic accuracy, addressing the need for reliable early detection.
- Enhances the interpretability of the model with SHAP explanations, generating both local and global explanations. Local explanations pinpoint the specific areas of an MRI scan that contribute to a given prediction, while global explanations provide an overview of the model's behavior across all inputs. To achieve this, two segmentation techniques are applied: grid-based segmentation for uniform region analysis, and SLIC (Simple Linear Iterative Clustering) segmentation to capture more refined, superpixel-level regions. This interpretability ensures that the model's output is transparent, enabling medical professionals to confidently integrate it into real-world diagnostic workflows.

B. *Previous Works*

In recent years, ML has become increasingly integral to advances in healthcare, providing novel solutions for the diagnosis, prognosis, and treatment planning of a range of medical conditions. In [1], an in-depth study on ML and DL techniques was presented, showing CNN-based models achieving 98.6% accuracy for brain tumor detection and 96.39% for breast cancer. In addition, this study highlighted the effectiveness of U-Net with a Dice score of 0.85 for brain tumor segmentation, and transfer learning models, such as ResNet-50 and DenseNet-169, achieving 97.2% and 96.8% accuracy, respectively, in COVID-19 detection from chest X-rays. In [2], the strength of CNN-driven approaches was highlighted, with VGG-16 achieving 89.3% accuracy in lung nodule detection. In [3], advances in architecture design were explored, where U-Net variants achieved Dice scores between 0.85 and 0.90 for organ segmentation, and attention-based networks improved classification accuracy by 4-7%.

In the domain of AD detection, an extensive review of ML methods in [4] examined both traditional methods and DL architectures using multimodal neuroimaging data. In [5], a comparative analysis on the ADNI dataset showed that Random Forest (RF) was superior with 97.44% accuracy, followed by SVM and neural networks at 96.15% and 94.87%, respectively. DL approaches have shown particular promise, achieving 96.85% accuracy using DenseNet-161 transfer learning on MRI data [6]. In [7], a whole-brain 3D-CNN approach achieved 94.1% accuracy in binary classification and 87.15% in four-class classification (CN, MCI, LMCI, AD).

Although ML models have shown remarkable performance in medical diagnosis, their black-box nature has led to a growing emphasis on explainability, particularly in healthcare, where transparency and trust are paramount. In [8], a comprehensive review of XAI in healthcare classified approaches into ante-hoc (transparent by design) and post-hoc (explanation after training) methods, highlighting that visualization-based methods were used in 48% of medical XAI studies, followed by example-based methods at 23%. In [9], a systematic analysis of 2,421 papers showed a significant surge in XAI medical publications from 2019-2023, with particular focus on imaging applications (42% of studies) and clinical decision support (35%). In [10], human-computer interaction aspects in medical XAI were analyzed, identifying that local interpretability methods such as LIME and SHAP were employed in 65% of reviewed studies, while global interpretability approaches were used in 35%.

Although ML models have demonstrated encouraging results in AD detection, the integration of XAI methods has emerged as crucial for clinical adoption and trust-building in diagnosis systems. In [11], an analysis of 89 studies from 2018-2023 revealed that Grad-CAM was the most widely adopted XAI method (42% of studies) for AD detection, followed by SHAP (28%) and LIME (21%), showing that XAI methods

predominantly focused on highlighting discriminative brain regions in neuroimaging data. In [12], a targeted review of LIME and SHAP applications in AD detection found that SHAP demonstrated superior consistency in identifying relevant biomarkers across different model architectures. Building on these insights, in [13], a Vision Transformer-GRU hybrid model achieved 90.3% accuracy, using attention visualization to track the model's focus on temporal changes in the brain structure. Recent innovations include the explainable tensor multi-task ensemble learning approach [14], which achieved 88.9% accuracy in dynamic AD prediction while providing interpretable temporal progression patterns. In [15], MAD-Former achieved 92.1% accuracy through multi-patch attention mechanisms, offering trace-based explanations of the decision-making process of the model. In [16], the AMURA model was developed to identify amyloid-β and tau markers, using SHAP values to explain the relationship between microstructural features and AD pathology markers, achieving 87.6% accuracy in classification tasks. In [17], SHAP-based feature selection was combined with ensemble boosting (XGBoost, LightGBM, Gradient Boosting) for AD diagnosis, achieving over 94% accuracy with minimal features and reduced testing time. In [18], a hybrid CNN-VGG16 model with transfer learning and SHAP explainability achieved 94% accuracy for brain tumor detection and 81% for AD across multiple MRI datasets.

AD detection has advanced from traditional ML to deep learning, achieving accuracies above 95%, with XAI improving transparency and clinical trust. Current trends focus on multimodal data, attention-based interpretability, and real-time explanations. This work advances this field by introducing SCR-NetX, a CNN with dual-level SHAP explanations, providing both local and global insights for clinically meaningful diagnosis.

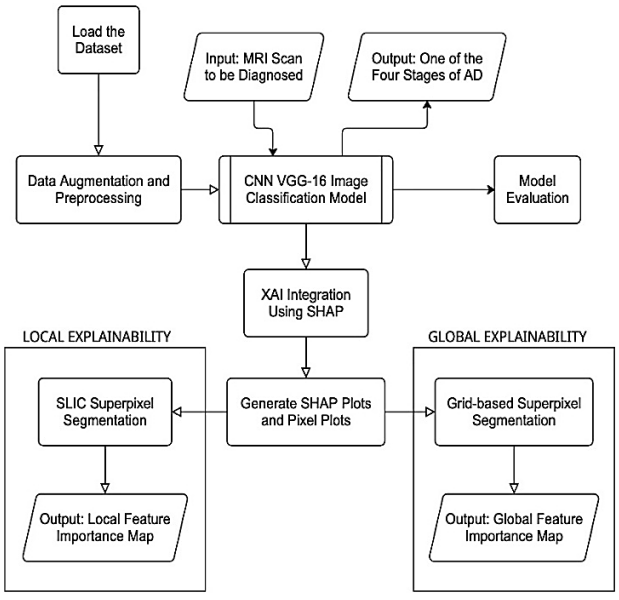

Fig. 1. Architecture of the proposed XAI-based SCR-NetX model for AD classification.

II. METHODOLOGY

A. Proposed Framework

Figure 1 shows a flowchart of the method used. The dataset for CNN-based image classification is first loaded, and data preprocessing techniques are applied before feeding it into the modified CNN VGG-16 model for training. The VGG-16 model, a deep convolutional neural network known for its simplicity and effectiveness, serves as the base architecture for this study. The original VGG-16 architecture was slightly modified to better suit the specific requirements of AD disease classification, named SCR-NetX. This model retains the core structure of 16 layers, including convolutional layers, pooling layers, and fully connected layers, but with some adjustments to enhance performance and adapt it to the nature of MRI scan data. This makes it more suitable for accurately capturing the subtle spatial and structural features necessary to diagnose AD.

Once the model is trained, it can input an MRI scan and classify it into one of the four stages of Alzheimer's disease: non-demented, very mildly demented, mildly demented, or moderately demented. The model's diagnostic capabilities are assessed using evaluation metrics, including accuracy, precision, recall, F1-score, and others. SHAP is integrated to enhance the interpretability of the proposed SCR-NetX model, which is a game-theoretic approach to explain the output of ML models. SHAP computes the contribution of each input feature by evaluating all possible feature subsets and estimating their marginal contributions, making it a reliable and theoretically grounded method for attributing predictions to input features.

B. Dataset and Preprocessing

The dataset used in this study consists of 6,400 MRI scans resized to 128×128 pixels, categorized into four classes: non-demented, very mildly demented, mildly demented, and moderately demented. All images were obtained from [19]. To improve robustness and generalization, data augmentation was applied sequentially using random rotation (±15%), random zoom (-30% to -10%), horizontal and vertical flips, and Random Contrast (±0.2). Class imbalance was addressed by assigning higher weights to underrepresented categories during training, ensuring balanced model performance. Table I shows class distributions and weights.

TABLE I. WEIGHTS FOR EACH CLASS OF THE DATASET

Class / Stage	Number of MRI Scans	Weights
Mildly demented	896	1.75
Moderately demented	64	25.0
Non demented	3200	0.5
Very mildly demented	2240	0.71

C. Architecture of the SCR-NetX Model

CNNs are a class of DL models designed to process structured grid-like data such as images. They automatically learn and extract features through convolutional filters, pooling layers, and non-linear activation functions, enabling the detection of spatial hierarchies such as edges, textures, and complex patterns. This makes them highly effective for medical image classification. Among CNN architectures,

VGG-16—developed by the Visual Geometry Group at Oxford—has been widely recognized for its simplicity, uniform structure, and strong performance. It comprises 16 layers, including 13 convolutional and 3 fully connected layers, using small 3×3 filters throughout. Input images of size 224×224×3 pass through five convolutional blocks, each followed by max pooling, progressively extracting features at different levels of abstraction. The final section includes dense layers of 4096, 4096, and 1000 units, with a softmax layer producing classification probabilities for ImageNet's 1000 categories.

This study adapted and optimized VGG-16 to develop SCR-NetX (Figure 2), a model specifically designed for classifying AD into four stages from MRI scans. The input size is reduced to 128×128×3 to match MRI resolution. After the convolutional layers from VGG-16, the traditional fully connected layers are replaced by a GlobalAveragePooling2D layer, which reduces parameters, minimizes overfitting, and improves computational efficiency by averaging feature maps. A BatchNormalization layer is applied after pooling to standardize activations, accelerating convergence and enhancing training stability. The dense section is restructured to gradually reduce from 2048 to 512, 256, and 64 units, allowing refined feature representation before the final classification layer. The output layer consists of four units with softmax activation, corresponding to the Alzheimer's stages: non-demented, very mild dementia, mild dementia, and moderate dementia. These modifications—particularly the use of global average pooling, batch normalization, and a compact dense architecture—optimize the model for medical image classification, ensuring a balance between computational efficiency, reduced overfitting, and high diagnostic accuracy in clinical applications.

D. SHapley Additive exPlanations (SHAP)

SHAP is a powerful method in XAI that helps understand how ML models arrive at their predictions. Drawing from Shapley values in game theory, SHAP measures the contribution of each input feature to the model's output. This approach provides both local and global explanations, making it particularly useful for understanding complex models like convolutional neural networks. Mathematically, SHAP computes the contribution of each feature by analyzing every possible combination of features and assessing their effect on the model's prediction. For a specific feature i, its contribution, or SHAP value, is computed as:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \left(f(S \cup \{i\}) - f(S)\right) \quad (1)$$

where S represents a subset of features, N denotes the total set of features, and $f(S)$ indicates the model's prediction based solely on the features included in subset S. This formula ensures that the contribution of each feature is calculated fairly by averaging all possible feature subsets, making SHAP a reliable method for understanding how input features drive predictions. In image classification, SHAP calculates pixel-level contributions to identify which areas of an image most impact the model's predictions. By analyzing SHAP values for each pixel, we can pinpoint influential regions and create heatmaps to visualize these contributions.

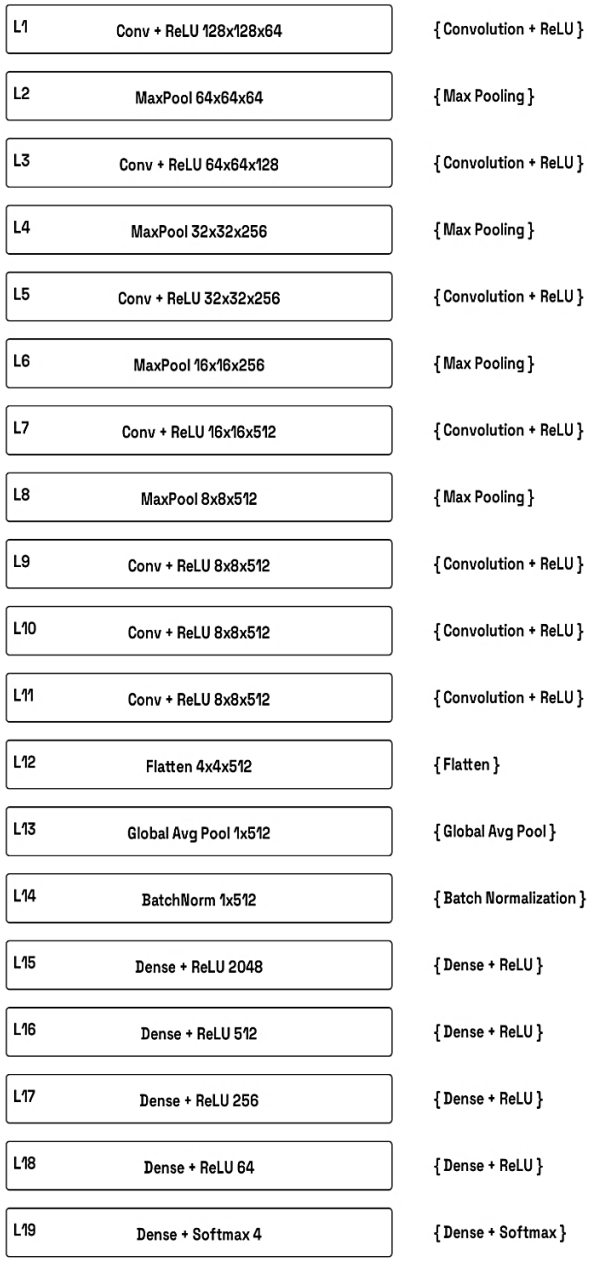

Fig. 2. Architecture of SCR-NetX.

By aggregating SHAP values across multiple images, patterns in feature importance can be identified, providing a holistic understanding of the model's behavior throughout the dataset. This study used SHAP because it can provide both local and global explanations, offering a complete understanding of how the CNN model classifies AD stages. SHAP helps not only in interpreting individual MRI scans but also in understanding the model's broader diagnostic behavior across the dataset. This makes it an ideal tool for improving the transparency and trustworthiness of the proposed model.

1) Local Explanations

For an individual MRI scan, explanations are obtained using SHAP values computed via the GradientExplainer and pixel plot method in Python. This technique assesses the model's sensitivity to each pixel by computing the gradient of the model's output in relation to the input pixels. For a 128×128 input MRI scan, 16,384 SHAP values are obtained, each representing the contribution of a single pixel to the final prediction. Higher SHAP values (depicted in red/pink) indicate pixels with a strong influence on the decision, while lower values (in blue) show less impact.

The SHAP plot shown in Figure 3, accompanied by a pixel plot, illustrates these contributions and helps to visualize how different parts of the MRI scan affect the model's prediction. To simplify the visualization and interpretation of pixel-level contributions, pixels are grouped into segments using SLIC (Simple Linear Iterative Clustering) super-pixel segmentation. In practical applications like medical diagnosis, it is unrealistic to analyze every pixel individually. Instead, SLIC divides the image into meaningful regions called super-pixels. These super-pixels are created by balancing color and spatial information, grouping neighboring pixels that are visually and structurally similar. This makes it easier to understand which larger and more relevant regions of the image influence the model's decisions. SLIC partitions the image into perceptually meaningful super-pixels by minimizing the cost function:

$$E = \sum_{i \in S}(d_c(x_i, \bar{x}) + m \cdot d_s(x_i, \bar{x})) \quad (2)$$

where d_c is the color distance and d_s is the spatial distance between pixels and the super-pixel centroid, with m balancing these distances. This approach captures important structural and textural information in the MRI scans. Thus, by focusing on these larger segments rather than individual pixels, the analysis is simplified to better understand which areas of the MRI scan are most influential in the classification, enhancing interpretability and practical usability.

After obtaining the segments, the SHAP values are aggregated within each superpixel, producing a heatmap that overlays these segments on the MRI scan, as seen in Figure 4. This heatmap highlights the importance of each segment, where higher SHAP values are shown in warm colors (red/pink) and lower values in cooler colors (blue). This segmentation-based heatmap provides a clearer understanding of how different regions of the MRI scan influence the model's results, making the output more relevant and interpretable in a real-world context.

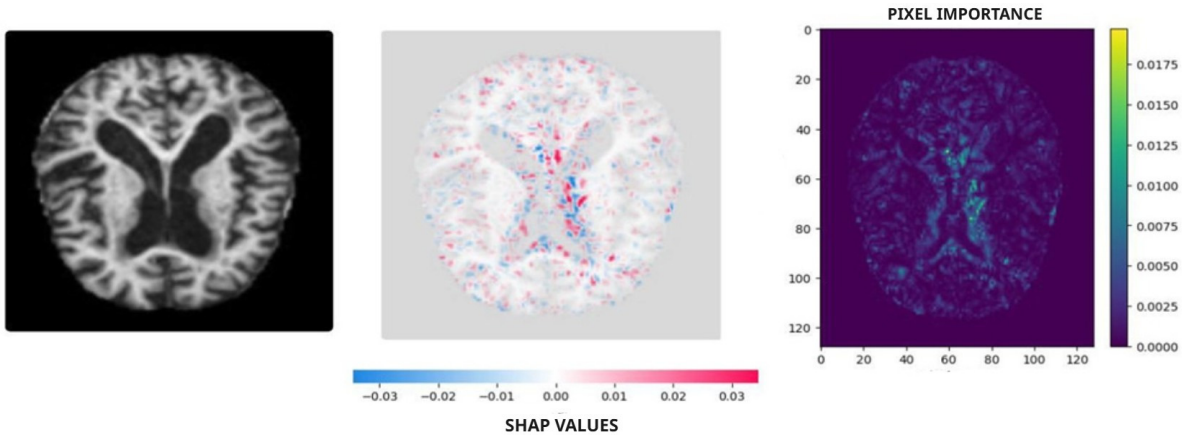

Fig. 3. A mildly demented brain MRI scan with its SHAP plot and pixel plot.

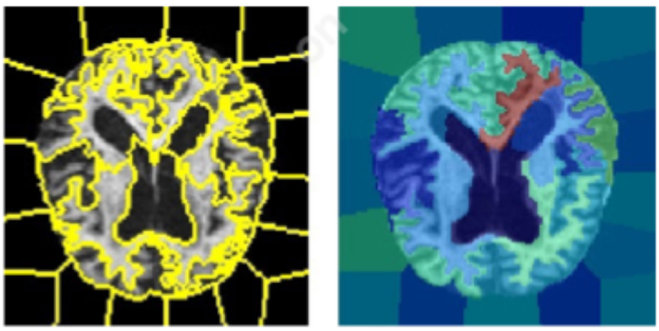

Fig. 4. Left: Same mildly demented brain MRI scan after SLIC segmentation; Right: Local Explanation segment-wise.

2) Global Explanations

For global explanations, SHAP values are computed per pixel and then aggregated across 20 MRI scans (five per class) to reveal overall feature importance. To simplify the 128×128 pixel space, pixels are grouped into larger segments. Unlike SLIC superpixels, which vary between images, a fixed grid segmentation ensures a uniform layout, enabling consistent global comparisons and reliable feature importance analysis across scans.

In the grid-based method, each MRI scan is divided into 20 fixed segments (5×4, F1–F20), with mean SHAP values aggregated into a heatmap (red = high importance, blue = negative). This yields a consistent global view of feature relevance. For comparison, SLIC segmentation provides local, boundary-adaptive explanations, while the grid enables cross-sample uniformity and global aggregation. Despite methodological differences, both approaches highlight overlapping brain regions, confirming the reliability of the proposed interpretability framework and offering complementary individual- and population-level insights.

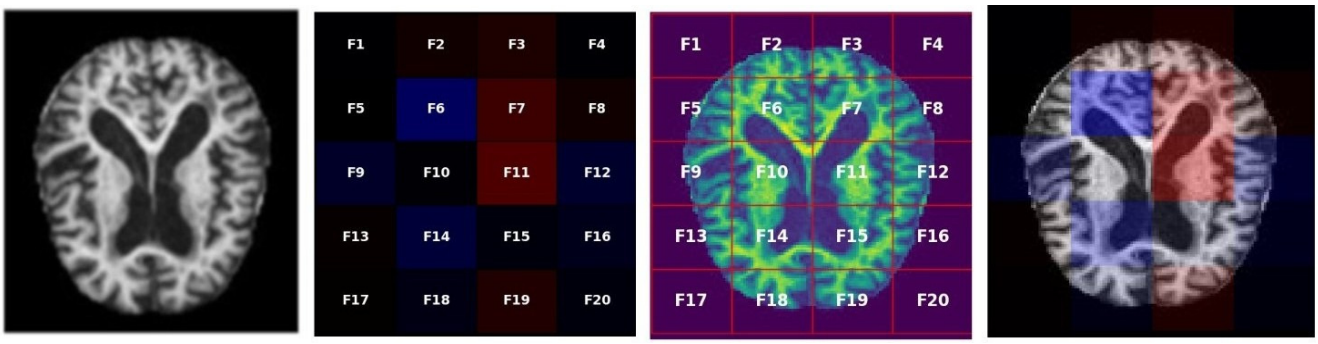

Fig. 5. Grid-based segmentation of the same mildly demented MRI scan.

III. RESULTS

A. SCR-NetX Model

Table II presents an accuracy comparison of existing benchmark models against SCR-NetX. The performance of the proposed SCR-NetX model was evaluated using several widely recognized metrics, including accuracy, loss, AUC (Area Under the Curve), precision, and recall. These metrics offer a thorough evaluation of the model's performance on both the training and testing datasets.

TABLE II. ACCURACY COMPARISON

Model	Accuracy
3D-CNN [7]	87.15%
Tensor Ensemble [14]	88.9%
AMURA [16]	87.6%
SCR-NetX	**91.74%**

Table III presents the evaluation results of the proposed model. Figure 6 compares training and testing performance, showing 95.4% accuracy for training and 91.7% for testing, with slightly higher loss in testing. The results indicate good generalization to unseen data.

TABLE III. PERFORMANCE OF THE MODEL

Metric	Training Data	Testing Data
Accuracy	0.9609	0.9174
Loss	0.1086	0.2296
AUC	0.9975	0.9903
Precision	0.9640	0.9173
Recall	0.9574	0.9159

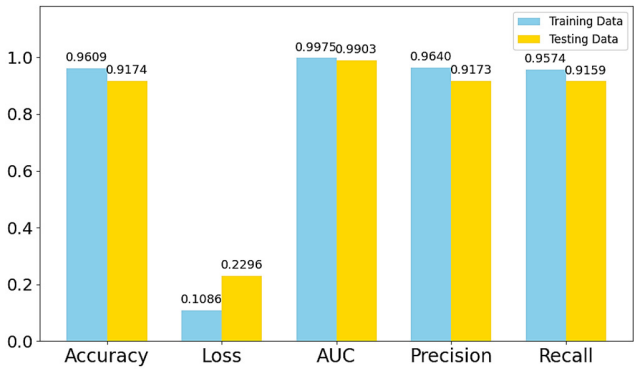

Fig. 6. Model performance: Training vs. Testing results.

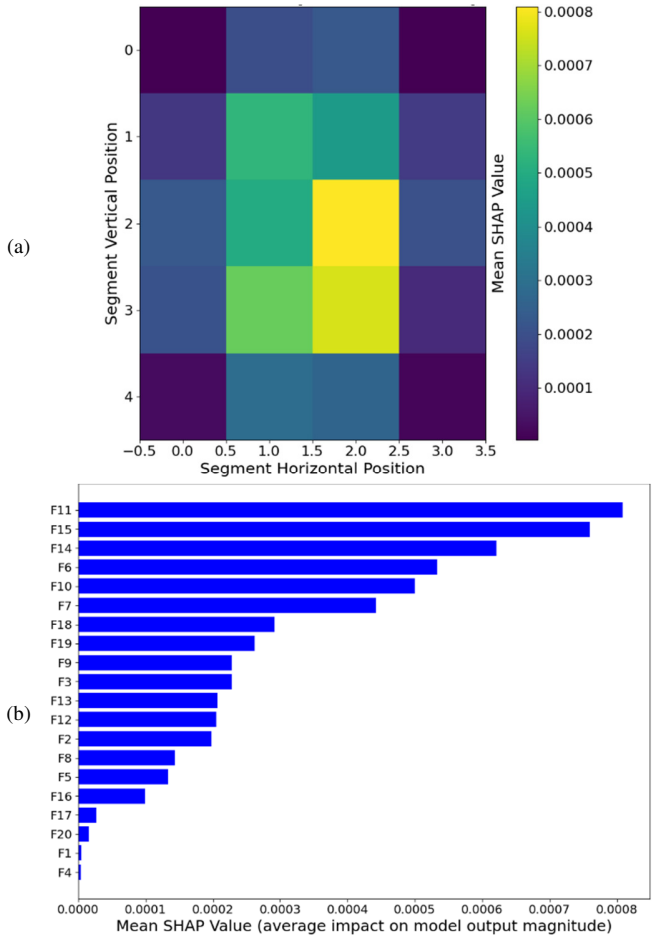

Fig. 7. (a) Global SHAP importance heatmap; (b) Feature importance of each segment.

B. SHAP Analysis

Figure 7 illustrates the global SHAP importance heatmap, where Feature 11 (F11) is the most important segment, followed by F15 and F14. This heatmap visually highlights the relative significance of different regions, with the brightest segments representing the most influential areas. The spatial arrangement of these segments corresponds to anatomically relevant brain regions, suggesting a biologically consistent pattern of feature relevance across the dataset.

Figure 7(b) contains the exact numerical SHAP values for each of the 20 features (F1–F20), enabling precise comparison and quantitative analysis. Ranking the segments by their global SHAP values provides a more straightforward comparison.

TABLE IV. GLOBAL SHAP VALUES

	Column 1	Column 2	Column 3	Column 4
Row 1	4.84e-06	1.98e-04	2.28e-04	3.52e-06
Row 2	1.33e-04	5.33e-04	4.42e-04	1.43e-04
Row 3	2.28e-04	5.00e-04	8.09e-04	2.05e-04
Row 4	2.07e-04	6.21e-04	7.60e-04	9.95e-05
Row 5	2.64e-05	2.92e-04	2.62e-04	1.59e-05

It is important to note that these results could vary when considering different or larger samples. However, the model's interpretability is supported by the fact that corner features, which generally represent blank areas common in all MRI scans, consistently have near-zero global SHAP values. This alignment indicates that the model appropriately identifies less relevant areas, reaffirming the reliability of the SHAP analysis. Figure 8 shows the local explanations for 8 individual test samples. For these samples, the global explanations were estimated, providing a rough general overview of feature importance across the dataset.

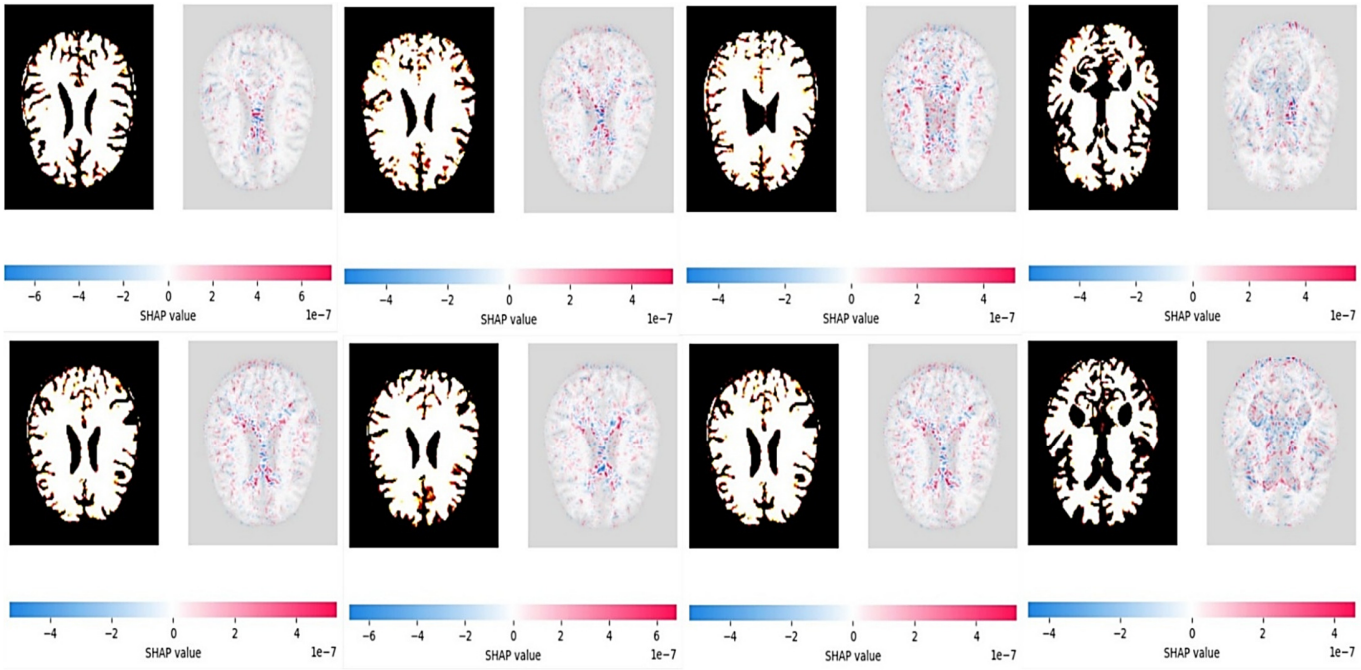

Fig. 8. SHAP plots for non demented, very mildly demented, mildly demented, and moderately demented.

IV. CONCLUSION

This study addressed the challenge of achieving both high accuracy and interpretability in the diagnosis of AD. SCR-NetX is a modified VGG-16 architecture that classifies AD stages into four categories with improved accuracy. Its novelty lies in combining architectural enhancements with SHAP-based explanations at both local and global levels. Using grid-based and SLIC segmentation, the model generates clinically meaningful explanations, offering both patient-specific insights and overall interpretability.

Unlike existing XAI-based techniques that focus on either local or global explanations, this work uniquely combines both through SHAP with grid and SLIC segmentation, ensuring clinically consistent and multi-level interpretability. This dual-layered framework provides patient-specific insights as well as broader interpretability, ensuring that highlighted regions align with clinically meaningful brain structures. The results show that SCR-NetX delivers accurate and transparent predictions, bridging the gap between AI-driven analysis and clinical applicability. This dual emphasis on performance and explainability advances current knowledge in medical imaging and strengthens trust in AI-assisted healthcare. Beyond technical improvements, the framework also emphasizes usability by producing explanations in formats accessible to clinicians. This positions SCR-NetX as a practical step toward integrating explainable DL into everyday diagnostic workflows. Future work will explore deeper architectures such as DenseNet, ResNet, and VGG-19, along with interpretability methods such as LIME and Grad-CAM. Broader validation on diverse datasets will further improve generalizability, while fairness, ethical responsibility, and transparency remain central to fostering trustworthy AI in medical diagnostics.

REFERENCES

[1] M. Rana and M. Bhushan, "Machine learning and deep learning approach for medical image analysis: diagnosis to detection," *Multimedia Tools and Applications*, vol. 82, no. 17, pp. 26731–26769, July 2023, https://doi.org/10.1007/s11042-022-14305-w.

[2] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018, https://doi.org/10.1109/ACCESS.2017.2788044.

[3] S. K. Zhou *et al.*, "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, Feb. 2021, https://doi.org/10.1109/JPROC.2021.3054390.

[4] M. Tanveer *et al.*, "Machine Learning Techniques for the Diagnosis of Alzheimer's Disease: A Review," *ACM Transactions on Multimedia Computing, Communicactions, and Applications*, vol. 16, no. 1s, Dec. 2020, Art. no. 30, https://doi.org/10.1145/3344998.

[5] M. B. Antor *et al.*, "A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease," *Journal of Healthcare Engineering*, vol. 2021, no. 1, 2021, Art. no. 9917919, https://doi.org/10.1155/2021/9917919.

[6] Y. AbdulAzeem, W. M. Bahgat, and M. Badawy, "A CNN based framework for classification of Alzheimer's disease," *Neural Computing and Applications*, vol. 33, no. 16, pp. 10415–10428, Aug. 2021, https://doi.org/10.1007/s00521-021-05799-w.

[7] G. Folego, M. Weiler, R. F. Casseb, R. Pires, and A. Rocha, "Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI," *Frontiers in Bioengineering and Biotechnology*, vol. 8, Oct. 2020, https://doi.org/10.3389/fbioe.2020.534592.

[8] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Aug. 2021, https://doi.org/10.1109/TNNLS.2020.3027314.

[9] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *Discover Artificial Intelligence*, vol. 4, no. 1, Feb. 2024, Art. no. 15, https://doi.org/10.1007/s44163-024-00114-7.

[10] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques," *IEEE Access*, vol. 9, pp. 153316–153348, 2021, https://doi.org/10.1109/ACCESS.2021.3127881.

[11] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, and F. Hajamohideen, "Explainable Artificial Intelligence in Alzheimer's Disease Classification: A Systematic Review," *Cognitive Computation*, vol. 16, no. 1, pp. 1–44, Jan. 2024, https://doi.org/10.1007/s12559-023-10192-x.

[12] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, Apr. 2024, Art. no. 10, https://doi.org/10.1186/s40708-024-00222-1.

[13] S. M. Mahim *et al.*, "Unlocking the Potential of XAI for Improved Alzheimer's Disease Detection and Classification Using a ViT-GRU Model," *IEEE Access*, vol. 12, pp. 8390–8412, 2024, https://doi.org/10.1109/ACCESS.2024.3351809.

[14] Y. Zhang, T. Liu, V. Lanfranchi, and P. Yang, "Explainable Tensor Multi-Task Ensemble Learning Based on Brain Structure Variation for Alzheimer's Disease Dynamic Prediction," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 1–12, 2023, https://doi.org/10.1109/JTEHM.2022.3219775.

[15] J. Ye *et al.*, "MAD-Former: A Traceable Interpretability Model for Alzheimer's Disease Recognition Based on Multi-Patch Attention," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 6, pp. 3637–3648, June 2024, https://doi.org/10.1109/JBHI.2024.3368500.

[16] L. Brusini *et al.*, "XAI-Based Assessment of the AMURA Model for Detecting Amyloid-β and Tau Microstructural Signatures in Alzheimer's Disease," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 12, pp. 569–579, 2024, https://doi.org/10.1109/JTEHM.2024.3430035.

[17] B. K. Raghupathy, M. R. Reddy, P. Theeda, E. Balasubramanian, R. K. Namachivayam, and M. Ganesan, "Harnessing Explainable Artificial Intelligence (XAI) based SHAPLEY Values and Ensemble Techniques for Accurate Alzheimer's Disease Diagnosis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 20743–20747, Apr. 2025, https://doi.org/10.48084/etasr.9619.

[18] S. Alsubai *et al.*, "Transfer deep learning and explainable AI framework for brain tumor and Alzheimer's detection across multiple datasets," *Frontiers in Medicine*, vol. 12, June 2025, Art. no. 1618550, https://doi.org/10.3389/fmed.2025.1618550.

[19] S. Kumar and S. Shastri, "Alzheimer MRI Preprocessed Dataset." Mendeley Data, June 25, 2025, https://doi.org/10.17632/3R8HW8WMMK.1.