International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Building Trust in Artificial Intelligence: An Explainable Deep Learning Framework for Brain Disease Detection

**P. V. Siva Kumar*[1], Gautham Mallipeddi[2], Srinivasa Deepesh Kommineni[3], Tapan Ganesh Naram[4], Akhilesh Kumandan Kottakota[5]**

**Submitted**:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

**Abstract:** Artificial Intelligence (AI) has shown promising results across various research fields. However, there is significant concern about its application in medicine due to the critical need for high accuracy and reliable data in this field. A major issue with many existing machine learning models is their lack of transparency; they do not explain the reasoning behind their outputs. This opacity leads to a lack of trust among healthcare professionals, who are hesitant to rely on such technology for critical decisions. Our research aims to address this concern by developing an Explainable Artificial Intelligence (XAI) model. This model not only classifies MRI images but also provides clear explanations for its predictions. By highlighting the specific regions of the brain that influenced each decision, our XAI model helps bridge the gap between AI and clinical practice. This approach empowers clinicians to identify brain diseases more confidently and accurately, fostering greater trust in AI-driven diagnostic tools.

***Keywords:*** *Alzheimer's disease, Brain Tumor, Convolutional Neural Network, Deep Learning, Explainable AI, Magnetic Resonance Image (MRI), Multiple Sclerosis, Transfer Learning*

## 1. Introduction

In a number of study areas, artificial intelligence (AI) has become a disruptive force with the potential to greatly improve productivity and results. AI's capacity to process and analyze enormous volumes of data in medicine offers a chance to completely transform treatment planning and diagnosis. Nonetheless, there is a great deal of skepticism about the use of AI in therapeutic practice. Because medical decision-making is so vital, it requires the highest standards of accuracy and dependability, which many AI models on the market today find difficult to meet on a regular basis.

The "black box" aspect of many AI and machine learning models is one of the main issues. Because these models frequently offer forecasts without supporting details, it can be challenging for medical practitioners to comprehend and have faith in the decision-making process. A significant obstacle to the application of AI in medicine, where there is little room for error and extremely high stakes, is this lack of transparency.

Our work centers on creating an Explainable Artificial Intelligence (XAI) model specifically designed for medical imaging in order to tackle this problem. Our software specifically classifies different brain disorders based on how Magnetic Resonance Imaging (MRI) data are processed. Our XAI model does more than just produce diagnostic outputs; it also highlights the precise brain regions that impacted each prediction, thereby explaining the reasoning behind it. In order to build confidence and make it easier to incorporate AI into standard healthcare operations, this transparency is essential.

In this study, we classified brain tumors and Alzheimer's disease using the InceptionV3 model with transfer learning. We customized InceptionV3 for medical imaging because of its strong feature extraction skills and deep convolutional architecture. This method makes use of the enormous volume of data that the model has been exposed to, which enables it to effectively recognise and classify intricate patterns in MRI images of Alzheimer's patients and brain tumors.

We developed a Sequential 2D Convolutional Neural Network (CNN) for Multiple Sclerosis (MS). The sequential 2D CNN architecture is a useful tool for identifying the subtle and diffuse lesions typical of multiple sclerosis (MS) because it can handle the spatial hierarchies in MRI data. Because this model was trained on a carefully selected dataset of MRI scans, it was highly accurate in its ability to distinguish between brain tissues that were healthy and those that were damaged.

*[1] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, INDIA - 500090*
*ORCID ID : https://orcid.org/0000-0001-9381-0472*
*[2] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, INDIA - 500090*
*ORCID ID : https://orcid.org/0009-0006-8905-0312*
*[3] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, INDIA - 500090*
*ORCID ID : https://orcid.org/0009-0002-6047-5086*
*[4] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, INDIA - 500090*
*ORCID ID : https://orcid.org/0009-0008-3895-5327*
*[5] VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, INDIA - 500090*
*ORCID ID : https://orcid.org/0009-0004-9080-7752*
** Corresponding Author Email: sivakumarpasupuleti@gmail.com*

Our research attempts to improve interpretability in addition to achieving high classification accuracy by incorporating these sophisticated AI models. By utilizing Explainable AI (XAI) techniques, we offer graphical representations that highlight the regions of the MRI scans that had the greatest impact on the model's judgment. The goal of this dual emphasis on explainability and accuracy is to help doctors make better decisions, which will enhance patient outcomes and diagnostic confidence.

## 2. Related Works

Medical image analysis, particularly utilizing MRI scans for diagnosing various brain-related conditions, has been a focal point of research efforts. Several studies have introduced innovative methodologies to enhance the accuracy and efficiency of classification and detection tasks within this domain.

Marwa EL-Geneedy's work stands out for its proposition of a shallow architecture Convolutional Neural Network (CNN), tailored specifically for analyzing 2D T1-weighted MRI brain images. The CNN model, comprising custom-designed layers alongside input and output layers, aims to categorize images into three significant classes: normal, Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). What's particularly striking is the reported testing accuracy, which soared to an impressive 99.68%. This achievement underscores the robustness of the proposed methodology, which also incorporates adaptive synthetic sampling (ADASYN) to address imbalanced datasets commonly encountered in medical image analysis. Moreover, preprocessing steps such as grayscale normalization and resizing images to 150x150 pixels ensure compatibility with the network architecture, while data augmentation techniques, including horizontal flipping, further enhance the model's ability to generalize.[1]

In a similar vein, Nida Aslam et al. shed light on the superiority of MRI scans over traditional clinical data sources like EEG and retinal scans for disease classification tasks. Their research underscores the efficacy of Convolutional Neural Networks (CNNs) over alternative machine learning algorithms such as Support Vector Machines, Decision Trees, and Random Forests, emphasizing the pivotal role of deep learning in medical image analysis.[2]

Meanwhile, Mostafa Salem et al. delved into the challenges associated with using supervised machine learning algorithms for multiple sclerosis (MS) lesion detection in MRI images. They highlighted the resource-intensive nature of acquiring expert-annotated samples, prompting exploration into data augmentation techniques to mitigate this hurdle. However, caution is advised against the potential limitations of such techniques in accurately capturing real-world image variations.[3]

Azam Soltani et al.'s proposition of a CNN architecture for feature extraction further enriches the landscape of medical image analysis. Their method harnesses convolutional layers to systematically analyze input data, leveraging the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and pooling layers to enhance computational efficiency and data robustness. This comprehensive approach facilitates the extraction of meaningful features, thereby bolstering the model's performance across various tasks, including image recognition and object detection.[4]

In the realm of brain tumor detection, Kasi Tenghongsakul et al. devised a comprehensive methodology that employs contrast-limited adaptive Histogram Equalization (CLAHE) to enhance image quality, thereby improving the accuracy of subsequent classification tasks. Leveraging transfer learning with state-of-the-art models such as InceptionResNet-V2, MobileNet-V2, VGG16, and ResNet50, their approach achieves accuracies of up to 96%, with CLAHE-enhanced images demonstrating superior precision.[5]

Moreover, Soheila Saeedi et al.'s development of a detection algorithm using 2D CNN and autoencoders showcases the versatility of deep learning techniques in medical image analysis. By fine-tuning various parameters and layer configurations, they achieved impressive accuracies, further underscoring the potential of deep learning in enhancing diagnostic capabilities.[6]

Shahrair Hossain et al.'s evaluation of Deep Learning architectures for brain tumor classification provides valuable insights into the performance of various models. Their introduction of IVX16, a multiclass classification model based on transfer learning from top-performing TL models, reflects a concerted effort to leverage pre-existing knowledge and optimize model performance in the medical image analysis domain.[7]

The paper by Aman Patel, Nidumoli Gowthami Priya, and G. Divya focuses on automated brain tumor detection using multi-label images of MRI scans and CNNs. With the increasing reliance on automatic defect detection in medical imaging, the need for trustworthy and accurate classification techniques becomes imperative. The study aims to address this need by developing automated tumor detection methods to alleviate the workload of radiologists and ensure proven accuracy. By utilizing computer vision techniques, such as morphological opening for noise removal and binary thresholding alongside Neural Network segmentation methods for tumor detection, the model assesses the presence of brain tumors. The research also explores enhancing accuracy through experimentation with different models and scaling methods, including Efficient B2, B3, and B6.[8]

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed

Bouridane present a survey of Explainable AI (XAI) techniques in healthcare. With the widespread application of deep learning models in medical imaging and healthcare tasks, the need for transparent and interpretable AI becomes crucial. XAI aims to elucidate the decision-making process of black-box deep learning models, mirroring human judgment and interpretation skills. The paper categorizes and summarizes recent XAI techniques used in healthcare and medical imaging applications, highlighting algorithms that enhance interpretability. Additionally, it addresses challenging XAI problems in medical applications and offers guidelines for developing better interpretations of deep learning models in medical image and text analysis. The survey also outlines future directions for prospective investigations in clinical topics, particularly in applications involving medical imaging.[9]

Sajid Nazir, Diane M. Dickson, and Muhammad Usman Akram present a survey focusing on explainable artificial intelligence (XAI) techniques for biomedical imaging with deep neural networks (DNNs). While deep learning has revolutionized disease diagnosis with its exceptional image classification performance, its adoption in clinical practice is hindered by the lack of understanding behind model predictions. The opacity of DNNs raises concerns in the regulated healthcare domain, where trust in automated diagnosis systems is paramount. XAI techniques aim to address this issue by providing insights into model predictions, thereby increasing trust, accelerating diagnosis, and meeting regulatory requirements. The survey categorizes XAI techniques, discusses open challenges, and suggests future directions beneficial to clinicians, regulators, and model developers.[10]

Zehra Karapinar Senturk presents a study focusing on early diagnosis of Parkinson's disease using machine learning algorithms. Parkinson's disease results from the disruption of brain cells responsible for producing dopamine, crucial for communication between brain cells and motor control. Early detection of non-motor symptoms is crucial for halting disease progression. The proposed diagnosis method involves feature selection and classification processes, utilizing methods like Feature Importance and Recursive Feature Elimination for feature selection. Classification and Regression Trees, Artificial Neural Networks, and Support Vector Machines are employed for patient classification, with Support Vector Machines demonstrating superior performance, achieving 93.84% accuracy with minimal voice features for Parkinson's diagnosis.[11]

Mian Muhammad Sadiq Fareed, et al., present ADD-Net, an effective deep learning model for early detection of Alzheimer's Disease (AD) in MRI scans. Alzheimer's Disease is a debilitating neurological disorder affecting memory, behavior, and reasoning, primarily affecting individuals over 40. Manual evaluation of MRI scans and neuro-psychological examinations are currently used for diagnosis, but deep learning offers new avenues for automation. ADD-Net proposes a CNN architecture optimized for AD classification with fewer parameters, ideal for smaller datasets. It accurately distinguishes early AD stages and provides class activation maps as heat maps on the brain. To address class imbalance in the Kaggle MRI dataset, synthetic oversampling techniques are employed. ADD-Net is compared against DenseNet169, VGG19, and InceptionResNet V2, outperforming them across precision, recall, F1-score, Area Under the Curve (AUC), and loss metrics. With 98.63% accuracy, 99.76% AUC, and 0.0549% loss, ADD-Net demonstrates superior performance, showcasing its potential for early AD detection.[12]

Nagaraj Yamanakkanavar, Jae Young Choi, and Bumshik Lee present a survey focusing on MRI segmentation and classification of the human brain using deep learning for the diagnosis of Alzheimer’s Disease (AD). MRI plays a crucial role in analyzing neurological diseases and brain anatomy, particularly in identifying early signs of AD for preventive measures. Deep learning-based segmentation methods have gained attention for their effectiveness in analyzing brain MRI data and diagnosing AD. The paper outlines current deep learning approaches for brain MRI segmentation, particularly focusing on convolutional neural network architectures. It discusses how segmentation improves AD classification, reviews state-of-the-art approaches, and summarizes results using publicly available datasets. Additionally, the paper addresses current issues and proposes future research directions for building computer-aided diagnostic systems for AD.[13]

Abdulkadir Karacı presents VGGCOV19-NET, a deep learning model for automatic detection of COVID-19 cases from X-ray images using a modified VGG19 CNN architecture and the YOLO algorithm. X-ray images are crucial for diagnosing COVID-19, especially in regions with limited access to specialist doctors. The study aims to achieve high precision in COVID-19 classification using pre-trained VGG19 and YOLOv3 algorithms. Models were evaluated with fivefold cross-validation, considering metrics like recall, specificity, precision, and f1-score. The Cascade VGGCOV19-NET model, fed with lung zone X-ray images detected by YOLOv3, achieved 99.84% accuracy for binary classification (COVID vs. no-findings) and 97.16% for three-class classification (COVID vs. no-findings vs. pneumonia). The model outperforms previous studies and demonstrates increased classification performance and reduced training time by incorporating the YOLOv3 algorithm. The results highlight the success of the proposed Cascade VGGCOV19-NET architecture in detecting COVID-19, contributing to both YOLO-aided deep architecture and classification success in the literature.[14]

Authors K.S. Biju, S.S. Alfa, Kavya Lal, Alvia Antony, and M Kurup Akhil present a software solution for Alzheimer’s disease detection based on segmentation of MRI images. The proposed algorithm aims to detect brain abnormalities by generating a 3D representation of the brain from MRI slices, enhancing accuracy and reliability. The process involves denoising, segmentation, 3D construction using Slice-O-Matic, and calculation of residual brain volume. Grey to white matter ratio is utilized to determine Alzheimer’s disease presence.[15]

## 3. Methods and Material

The current research methodology is encapsulated in Fig 1. below, delineating several crucial steps ranging from data preprocessing to model training, and the integration of explainable AI techniques.

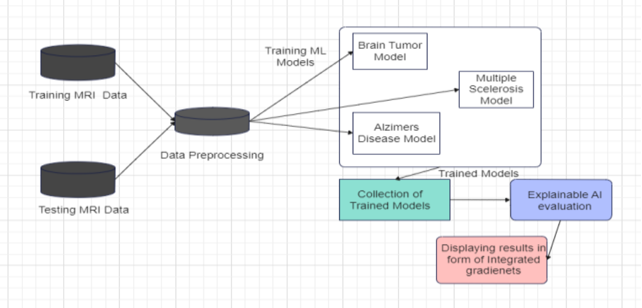
Training MRI  Data
Testing MRI Data
Data Preprocessing
Training ML Models
Brain Tumor Model
Multiple Scelerosis Model
Alzimers Disease Model
Trained Models
Collection of Trained Models
Explainable AI evaluation
Displaying results in form of Integrated gradienets

**Fig 2.** Architecture of the proposed methodology

### 3.1. Dataset

The datasets used for training various disease detection models were primarily composed of FLAIR (Fluid-Attenuated Inversion Recovery) images, a type of MRI sequence known for its sensitivity to pathological changes in the brain.

6400 magnetic resonance images were used for the Alzheimer's disease dataset. 896 photos were classified as mildly demented in this dataset, 2240 as very mildly demented, 3200 as non-demented, and 64 as examples of moderately demented. This comprehensive dataset encompassed a diverse range of Alzheimer's disease progression stages, facilitating robust model training and evaluation.

In the Brain Tumor dataset, a total of 4604 images were included, consisting of both healthy brain scans and those with evidence of brain tumors. Specifically, there were 2089 images representing healthy brain tissues and 2515 images showcasing various manifestations of brain tumors. This dataset provided a valuable resource for developing algorithms capable of accurately distinguishing between healthy brain tissue and tumor-affected regions in MRI scans.

The Multiple Sclerosis (MS) dataset comprised a total of 3427 images, offering insights into the neurological condition marked by an attack on the central nervous system by the immune system. Within this dataset, there were 2016 control images, with 1002 in the Axial plane and 1014 in the Sagittal plane, serving as reference points for normal brain morphology. Additionally, there were 650 images depicting MS lesions in the Axial plane and 761 images capturing MS lesions in the Sagittal plane. This dataset facilitated the development of models capable of identifying and quantifying MS-related abnormalities in MRI scans, aiding in diagnosis and treatment planning for patients with the condition.

### 3.2. Data augmentation and image pre-processing

The MRI images sourced from these datasets exhibited varying dimensions, necessitating preprocessing before input into the neural networks. Specifically, for the Alzheimer's and Brain Tumor disease models, the images were standardized to a resolution of 176x176 pixels to ensure uniformity across the dataset. In contrast, for the Multiple Sclerosis model, a slightly smaller size of 150x150 pixels was deemed sufficient for effective analysis. This resizing procedure served as a crucial step in preparing the data for subsequent feature extraction and classification tasks within the respective disease detection frameworks. Such standardization measures helped optimize the performance and generalizability of the trained models across different MRI datasets and imaging modalities.

### 3.3. Proposed solution

#### 3.3.1. Alzheimer's Disease  - InceptionV3

Utilizing the pre-trained InceptionV3 model as a feature extractor, a convolutional neural network (CNN) architecture is used for picture categorization. Custom classification layers can be added by first loading the base model with its pre-trained weights from the ImageNet dataset, excluding its top layer. The dimensions of the input photos are indicated by the input shape, which is set to (176, 176, 3). The pre-trained model's layers are frozen (base_model.trainable = False), preventing overfitting and saving computing power by blocking weight updates during training.

Next, layers are stacked in a sequential manner to create the sequential model. In order to mitigate overfitting and regularize the network, a dropout layer with a dropout rate of 0.5 is introduced. The feature maps are then spatially aggregated via a global average pooling layer, which lowers the dimensionality of the data without sacrificing significant information. The pooled feature maps are then transformed into a one-dimensional vector by a flatten layer, readying them for input into fully connected layers. After every dense layer, batch normalization layers are added to stabilize and quicken the training process by normalizing the activations.

The classifier portion of the network is made up of dense layers that progressively reduce the dimensionality of the characteristics that the convolutional base extracts. Repaired linear unit (ReLU) activation functions are used in each dense layer, which has 512, 256, 128 and 64 neurons, respectively, to add non-linearity to the model. To further prevent overfitting, dropout layers with a dropout rate of 0.5 are inserted between the dense layers. Lastly, the output layer uses a softmax activation function to output probabilities for each class. It consists of a dense layer with four neurons, which represents the number of classes in the classification task. The network's structure and complexity are summarized, together with the number of parameters in each layer of the model design.

### 3.3.2. Brain Tumor Disease - InceptionV3

Convolutional neural network (CNN) architecture for binary brain tumor picture classification utilizing the pre-trained InceptionV3 model. First, the pre-trained weights of the InceptionV3 model are loaded from the ImageNet dataset, with the top layer excluded to allow for task-specific customisation. The input shape is set to (176, 176, 3), matching the dimensions of the input images. By setting base_model.trainable = False, the layers of the pre-trained model are frozen, ensuring that their weights are not updated during training, which can help prevent overfitting and reduce computational resources.

Then, to regularize the network and reduce overfitting, a dropout layer with a dropout rate of 0.5 is added to the initial InceptionV3 model to create the sequential model. A global average pooling layer is applied to geographically aggregate the feature maps, followed by a flattened layer to transform the pooled feature maps into a one-dimensional vector. Batch normalization layers are inserted after each dense layer to normalize the activations and stabilize the training process.

The dense layers constitute the classifier part of the network, gradually reducing the dimensionality of the features extracted by the convolutional base. Repaired linear unit (ReLU) activation functions are used in each of the four dense layers—which include 512, 256, 128 and 64 neurons, respectively—to add non-linearity to the model. To further prevent overfitting, dropout layers are interspersed between the dense layers, with a dropout rate of 0.5. Lastly, the output layer outputs probability for the two classes (brain tumor present or missing) using a sigmoid activation function and a dense layer with two neurons. The model architecture is summarized, detailing the number of parameters in each layer and providing an overview of the network's structure and complexity.

### 3.3.3. Multiple Sclerosis - Sequential 2D CNN

An image categorization system for multiple sclerosis (MS) using convolutional neural networks (CNNs). It begins with a sequence of convolutional layers, with max-pooling layers for downsampling coming after each one. The rectified linear unit (ReLU) activation function is used to apply the 32 3x3 filters of the first convolutional layer to the input pictures in order to extract different features. Then, 2x2 pool-sized max-pooling layers are used to minimize spatial dimensions and extract the most prominent features.

The process is repeated with deeper layers to capture increasingly complex patterns: the second convolutional layer has 64 filters, and the third has 128. After the final max-pooling layer, the feature maps are flattened into a one-dimensional vector to be processed by fully connected layers.

To further extract high-level information from the flattened representation, a thick layer with 128 neurons and ReLU activation is added. The next layer is a dropout layer with a dropout rate of 0.5, which randomly removes a portion of the neurons during training as a regularization method to avoid overfitting.

The output layer, which uses a softmax activation function to calculate the probability distribution over the classes, consists of a dense layer with as many neurons as there are classes (num_classes) in the classification job. This architecture works well for classifying anomalies associated with multiple sclerosis (MS) in MRI scans because its convolutional and pooling layers efficiently extract pertinent characteristics, which are then classified using fully connected layers.

## 3.4. Performance evaluation metrics

Since they may evaluate the efficacy and dependability of the classification model, performance metrics including accuracy, precision, recall, and F-measure are essential for MRI classification tasks. The model's overall correctness of predictions is measured by accuracy, which offers information about the model's overall performance. Precision measures how many accurate positive predictions there are among all positive predictions, which shows how well the model can detect pertinent cases. Conversely, recall assesses the ratio of accurate positive predictions to all real positive cases, indicating the model's sensitivity in identifying positive examples. An impartial evaluation of the model's performance is provided by the F-measure, which combines recall and precision into a single score and is especially helpful in cases where class imbalance exists. The model's validation, optimisation, and clinical application are facilitated by the comprehensive insights these metrics provide into the model's accuracy in classifying MRI images.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)} \quad (3)$$

$$\text{F1-score} = 2.\frac{\text{Precision. Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

### 3.5. Model evaluations

The Alzheimer's disease detection model is really good at getting it right 95% of the time when looking at MRI images that show signs of Alzheimer's. Similarly, the brain tumor detection model is even better, hitting a 96.95% accuracy rate in telling apart healthy brain tissue from areas affected by tumors. When it comes to spotting multiple sclerosis, the model does a solid job with a 96.21% accuracy rate, spotting the telltale signs of the condition across different MRI scans. These high accuracy scores for these different health issues show that the models are dependable and helpful tools for doctors in diagnosing and planning treatment for patients.

## 4. XAI and Model Outputs

Following the construction and training of Convolutional Neural Network (CNN) models for predicting various outputs, each model was enhanced by integrating with the eXplainable AI (XAI) library LIME. This integration allowed for the identification of specific regions within images that influenced the prediction for a particular class. In these images, green regions indicate positive impacts on the prediction for that class, while red regions signify negative impacts.

Our approach utilizes integrated gradient images to simplify the process of disease identification for doctors. By highlighting the areas within medical images that significantly contribute to the prediction of a specific disease class, we aim to provide healthcare professionals with a clear and interpretable visual aid. This enables them to quickly and accurately identify the presence of a particular disease, thereby facilitating prompt diagnosis and treatment planning for patients.

LIME, which stands for Local Interpretable Model-agnostic Explanations, is an XAI (eXplainable Artificial Intelligence) technique crafted to offer understandable interpretations for machine learning model predictions, especially in image classification scenarios. Its operation involves altering input features, like individual pixels in an image, and observing how these modifications impact the model's predictions. By scrutinizing these alterations, LIME generates explanations that are easy to grasp, shedding light on the features or areas within the input data that significantly influence the model's prediction for a specific category. This methodology aids users, such as physicians or decision-makers, in comprehending the rationale behind a model's specific prediction, thus fostering trust, transparency, and the usability of the AI system. Moreover, LIME is model-agnostic, implying that it can be utilized across a wide spectrum of machine learning models without necessitating access to their internal mechanisms.

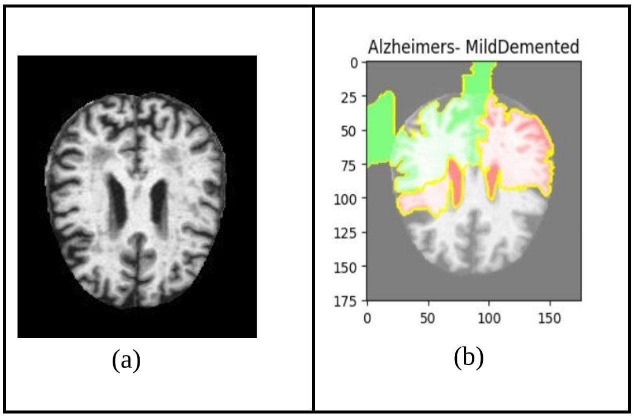

**Fig 2.** Input MRI image for Alzheimer's prediction,(b) Output with prediction of disease and region identified by XAI

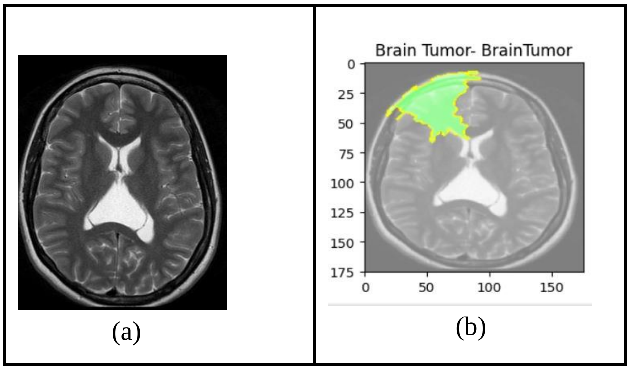

**Fig 3.** Input MRI image for Brain Tumor prediction,(b) Output with prediction of disease and region identified by XAI

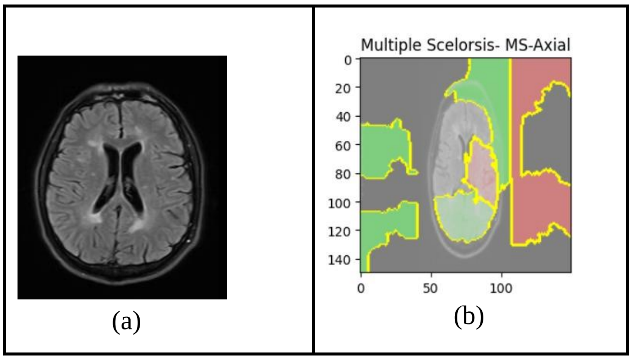

**Fig 4.** Input MRI image for Multiple Sclerosis prediction,(b) Output with prediction of disease and region identified by XAI

## 5. Conclusion

The adoption of the InceptionV3 model for this application serves as a robust foundation for our deep learning framework. By leveraging transfer learning, the pre-trained model is fine-tuned on disease-specific datasets, optimizing its performance and generalization capabilities. InceptionV3's ability to learn intricate features from brain imaging data enables accurate predictions, enhancing diagnostic accuracy.

However, the commitment of this project extends beyond accuracy alone. We recognize the importance of transparency and interpretability in AI systems. To address this, Explainable AI techniques were used in the project. These methods shed light on the model's decision-making process, making it more understandable for clinicians. Feature importance analysis, activation maximization, and model-agnostic interpretability (such as Local Interpretable Model-agnostic Explanations) empower healthcare professionals to gain insights into salient features and regions influencing diagnostic outcomes.

The transparent nature of the system not only fosters collaboration between AI and clinicians but also ensures regulatory compliance and mitigates concerns related to bias or unintended consequences. By bridging the gap between advanced technology and clinical practice, we aim to provide informed treatment decisions and improve patient outcomes.

## 6. References and Footnotes

### 6.1. References

### Acknowledgements

Gautham Mallipeddi, Srinivasa Deepesh Kommineni, Tapan Ganesh Naram, Akhilesh Kumandan Kottakota would like to thank P. V. Siva Kumar for his inputs throughout the project and in the enhancement of this manuscript which have been very beneficial.

### Author contributions

**P. V. Siva Kumar:** Conceptualization, Emphasizing the importance of transparency of AI in medical fields, Suggestions for improving the manuscript. **Gautham Mallipeddi:** Curating the brain tumors dataset, First few model prototypes for the Brain Tumor disease, Prototyping the model for Parkinson's disease, Drafting the final manuscript, **Srinivasa Deepesh Kommineni:** Worked on the model of Multiple Sclerosis disease, Wrote the first draft of the manuscript. **Tapan Ganesh Naram:** Worked on the model of Alzheimer's disease, Building the Explainable AI model **Akhilesh Kumandan:** Worked on the Brain Tumor model.

### Conflicts of interest

The authors declare no conflicts of interest.

## References

[1] Marwa, E.G., Moustafa, H.E.D., Khalifa, F., Khater, H. and AbdElhalim, E., 2023. An MRI-based deep learning approach for accurate detection of Alzheimer's disease. Alexandria Engineering Journal, 63, pp.211-221.

[2] Aslam, N., Khan, I.U., Bashamakh, A., Alghool, F.A., Aboulnour, M., Alsuwayan, N.M., Alturaif, R.A.K., Brahimi, S., Aljameel, S.S. and Al Ghamdi, K., 2022. Multiple sclerosis diagnosis using machine learning and deep learning: Challenges and opportunities. Sensors, 22(20), p.7856.

[3] Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, À. and Lladó, X., 2019. Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET. IEEE Access, 7, pp.25171-25184.

[4] Soltani, A. and Nasri, S., 2020. Improved algorithm for multiple sclerosis diagnosis in MRI using convolutional neural network. IET Image Processing, 14(17), pp.4507-4512.

[5] Tenghongsakul, K., Kanjanasurat, I., Archevapanich, T., Purahong, B. and Lasakul, A., 2023, May. Deep transfer learning for brain tumor detection based on MRI images. In Journal of Physics: Conference Series (Vol. 2497, No. 1, p. 012015).

[6] Saeedi, S., Rezayi, S., Keshavarz, H. and R. Niakan Kalhori, S., 2023. MRI-based brain tumor detection using convolutional deep learning methods and chosen machine learning techniques. BMC Medical Informatics and Decision Making, 23(1), p.16.

[7] Hossain, S., Chakrabarty, A., Gadekallu, T.R., Alazab, M. and Piran, M.J., 2023. Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. IEEE Journal of Biomedical and Health Informatics.

[8] Patel, A., Priya, N.G. and Divya, G., 2023, May. Automated Brain Tumor detection using multi-label images of MRI scans and CNNs. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) (pp. 1-5). IEEE.

[9] Chaddad, A., Peng, J., Xu, J. and Bouridane, A., 2023. Survey of explainable AI techniques in healthcare. Sensors, 23(2), p.634.

[10] Nazir, S., Dickson, D.M. and Akram, M.U., 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. Computers in Biology and Medicine, p.106668.

[11] Senturk, Z.K., 2020. Early diagnosis of Parkinson's disease using machine learning algorithms. Medical hypotheses, 138, p.109603.

[12] Fareed, M.M.S., Zikria, S., Ahmed, G., Mahmood, S., Aslam, M., Jillani, S.F., Moustafa, A. and Asad, M., 2022. ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans. IEEE Access, 10, pp.96930-96951.

[13] Yamanakkanavar, N., Choi, J.Y. and Lee, B., 2020. MRI segmentation and classification of human brain using deep learning for diagnosis of Alzheimer's disease: a survey. Sensors, 20(11), p.3243.

[14] Karacı, A., 2022. VGGCOV19-NET: automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm. Neural Computing and Applications, 34(10), pp.8253-8274.

[15] Biju, K.S., Alfa, S.S., Lal, K., Antony, A. and Akhil, M.K., 2017. Alzheimer's detection based on segmentation of MRI image. Procedia computer science, 115, pp.474-481.