Original Research Article

Classifying Alzheimer's Disease, Parkinson's Disease, and Control Cases Using Transfer Learning, Ensemble Learning, and Explainable AI

Joshua P. Asirvatham

Troy Athens High School, 4333 John R Rd, Troy, MI 48085, United States

ABSTRACT

Early detection of neurodegenerative diseases can be challenging, where Deep Learning (DL) techniques have shown promise. Most DL techniques provide a robust and accurate classification of performance. However, due to the complex architectures of the DL models, the classification results are difficult to interpret, causing challenges for their adoption in the healthcare industry. To help improve the adoption of AI in healthcare, this study incorporates Transfer Learning, Ensemble Learning, and XAI techniques to propose an effective and interpretable model. This work compares the performances of pre-trained models for the early detection of Alzheimer's Disease (AD) and Parkinson's Disease (PD). Specifically, the XAI technique Saliency Map was used to overlay gradients on the MRI scan, elucidating the regions on the MRI scan that led the model to its diagnosis. The Kaggle dataset used in the study has three classes: Parkinson's disease (PD), Alzheimer's disease (AD), and control (healthy). This study compares the performance of various pretrained models. Additionally, the diagnoses produced by the pretrained models were ensembled to produce a final diagnosis. Combining the predictions of multiple pretrained models can boost the performance of the model because it combines the strengths of multiple pretrained models, achieving a higher performance than the pretrained models. The best pretrained model EfficientNetB7 received an accuracy of 94.58% with an F1-score of 95.81%. The proposed model of this study is the ensemble learning model with an accuracy of 97.04% and an F1-score of 97.69%.

Keywords: Deep Learning; Alzheimer's Disease; Parkinson's Disease; Explainability Artificial Intelligence; Saliency Map; Transfer Learning; Ensemble Learning

INTRODUCTION

Alzheimer's Disease (AD) is a neurological disorder that causes gradual and irreversible damage to cognitive ability and memory, hindering the patient's ability to perform daily tasks. AD is a neurodegenerative disease, meaning their condition will only get worse over time (1). The disease, ultimately, results in death (2).

As their conditions worsen over time, their quality-of-life declines (1). AD patients may lose their sense of self as the disease progresses. As patients lose their ability to carry out daily activities like dressing, or cooking, they may feel frustration and helplessness because they must rely on others. Furthermore, as the disease progresses, the loss of memories and skills can result in the loss of

Corresponding author: Joshua P. Asirvatham, E-mail: jasirvatham2602@gmail.com.

Accepted December 4, 2025
https://doi.org/10.70251/HYJR2348.36791806

identity (1).

Parkinson's Disease (PD) is a progressive disorder that mainly affects movement and motor control, which is due to the loss of brain cells in the Substantia Nigra—the region in the brain responsible for transmitting movement signals (1). Patients also face trouble with essential basic skills, including speaking, walking, writing, and many tasks that require fine motor control (2).

Like the neurological disorder AD, PD can have negative emotional effects on people. The loss of motor functions results in slow movement and difficulty with balance. This makes everyday tasks such as walking downstairs and making breakfast a very challenging task. These symptoms can lead to depression and physiological distress, ultimately decreasing to quality of life until death. Early detection and treatment can help "manage symptoms, slow the disease progression, and improve the patient's quality of life" (1).

The "improvement of the quality of life for the impacted individuals largely depends on the timely and precise diagnosis of these illnesses" (2). Clinical diagnoses can take 3 to 5 years to classify PD (3). Similarly, clinical diagnoses typically take 5.5 years to reach a final diagnosis of AD (4). Such a long period causes symptoms to worsen over time, hindering their quality of life. Furthermore, clinical diagnoses using blood tests to detect AD have an accuracy of 88% to 92% (5). Similarly, clinical diagnoses can detect PD with an accuracy of 90.3% (6). While clinical diagnoses are fairly accurate, they take a very long time to make their diagnosis. AI techniques can help reduce the time and increase the precision of diagnosis of neurological diseases, helping improve impacted patients' quality of life.

It is common for medical datasets to be small due to the time and cost of collecting each MRI scan. Transfer Learning can be used to increase the accuracy of the model even with a small dataset (7). Pre-trained models are machine learning models that have already been trained on a large dataset and can be used to solve new problems utilized in Transfer Learning (7). Specifically, the pre-trained models used in this study VGG16, VGG19, ResNet50, InceptionV3, Xception, EfficientNetB0, and EfficientNetB7 have been trained on the ImageNet dataset, which contains millions of images classified into thousands of classes (1). Since these pre-trained models are originally trained on a large dataset with millions of images for a different task, it can reach higher accuracy in a short period when tailored to a new task since it doesn't have to start from scratch. Furthermore, by training multiple pretrained models, there will be multiple predictions made by the pretrained models; these predictions can be ensembled or combined to make a final prediction. This is called Ensemble Learning. Ensemble Learning combines the strengths of multiple models, achieving higher performance. However, using Transfer Learning and Ensemble Learning alone creates a black-box experience. Since the model does not illuminate the reason why it made its diagnosis, healthcare workers do not know when to trust AI models. This is where visual eXplainable Artificial Intelligence (XAI) techniques come in. Visual XAI techniques like Grad-CAM and Saliency maps highlight features on the MRI scan, elucidating why the AI model made its decision (8). This also builds trust between healthcare workers and AI models; since the XAI techniques like Saliency Maps highlight regions on the MRI scan that led it to its diagnosis, healthcare workers know why the model makes its diagnosis and when to trust the model. Without XAI, the AI model would provide a diagnosis without providing any explanation for why it made its diagnosis, leading healthcare workers to feel unsure of when to trust and when not to trust the AI model. The usage of XAI technique helps increase interpretability and build trust between healthcare workers and AI. However, the XAI technique Saliency Maps has not been done before while classifying AD and PD with Transfer Learning models. Overall, XAI techniques build trust between healthcare workers and AI technology in the medical field. Integrating XAI techniques with Transfer Learning and Ensemble Learning allows for both high accuracy and high interpretability, improving patients' quality of life. This study aims to classify AD and PD with Transfer Learning, Ensemble Learning, and the XAI technique Saliency maps to provide interpretable and precise predictions.

LITERATURE REVIEW

XAI in medical imaging

To help improve the interpretability of AI models, it is important to understand how others utilized Transfer Learning and/or XAI techniques to classify neurological disorders. Viswan *et al* used Transfer Learning with the pre-trained models: ResNet50, ResNet101, InceptionV3, Inception, ResNetV2, and EfficientNetB0 to classify between AD, PD, and control cases (1). He also created heatmaps from Grad-CAM, an XAI technique, which highlights regions in the MRI scan that led the model to its diagnosis. He then calculated Pearson's correlation

coefficient, which is a number between –1 and 1 that determines the correlation between the original MRI scan and the heatmap. This study's highest performing model ResNet152 received 99.9% accuracy on the NTUA dataset.

Mansouri *et al* used the pre-trained models: CNN, VGG16, ResNet50, and AlexNet to classify stages of AD: No impairment, Moderate Impairment, Mild Impairment, and Very Mild Impairment (9). To address the limited number of MRI Scans of AD, it created a synthetic MRI scan using Generative Adversarial Networks (GANs), specifically WGAN-CP. They then trained the models with real and synthetic MRI scans and a model with only real MRI scans. They then used the XAI technique Grad-CAM to extract gradients from the last Convolutional layer from each image to overlay heatmaps onto the image. The CNN model trained by both real and synthetic MRI scans received an accuracy of 97.50% and an F1-score of 98.25—higher than the model trained by real MRI scans alone, receiving an accuracy of 88.98% and an F1-score of 92.75%.

Mahmud *et al* utilized Transfer learning with the pre-trained models: VGG16, VGG19, DenseNet169, and DenseNet201 to classify stages of Alzheimer's Disease (10). They used the explainable AI technique (XAI) including saliency maps and Grad-CAM to increase the interpretability of the model. They performed many image-preprocessing techniques such as resizing them to 224 × 244, normalization, removing noise by median filter, and more. Their model received an accuracy of 96%.

Dentamaro *et al* classified PD with XAI techniques such as Integrated Gradients and Attention Heatmaps for Vision Transformer (ViT) after using Transfer Learning to train the pre-trained models: ResNet and DenseNet (11). This paper used 3D MRI scans instead of 2D MRI scans. Their DenseNet model performed the best with an accuracy of 96.6%.

Transfer Learning for neurological disorders

Al-Zharani *et al* also used Transfer Learning with the pre-trained models: VGG16, ResNet50, and DenseNet121 to classify stages of Alzheimer's Disease (8). It performed data augmentation techniques such as horizontal flipping of images, rotation of image by 5°, and shifting the images to increase the number of images in the training dataset. The best pre-trained model was DenseNet121, which had an accuracy of 97.33%.

Siddiqua *et al* first performed data cleaning techniques such as sharpening and denoising (2). It used a separate pre-trained denoising model to clean the data. Cleaning the data removes biases and allows the models to train with higher quality images, it improves the model's accuracy in the test dataset. They then used Transfer Learning with pre-trained models: EfficientNetB0, InceptionV3 model, ResNet50, and Xception to classify between AD, PD, and Control Cases with their highest performing model EfficientNetB0, receiving 99% accuracy.

Rama *et al* used transfer learning to train the two pre-trained models: CNN and VGG-19 to classify PD with CONTROL (12). They performed data preprocessing steps on the MRI scans. Each model had convolutional layers for feature extraction, pooling layers (e.g., MaxPooling) to reduce spatial dimensions, and dense and dropout layers for classification and regularization. Ultimately, the CNN model received an accuracy of 98.04%; while the VGG-19 model received 92.04%.

There has been research using transfer learning on classifying PD with Control Cases (12), AD with Control Cases (8) and between AD, PD, and Control Cases (2). There has also been research integrating XAI methods with Transfer Learning to increase both the accuracy and interpretability of the model. Dentamaro *et al* classified PD with Control Cases with the XAI techniques Integrated Gradients and Attention Heatmaps with the pre-trained models: ResNet and DenseNet (11). Mahmud *et al* classified AD with Control Cases with the XAI techniques: saliency maps and Grad-CAM with the pre-trained models: VGG16, VGG19, DenseNet169, and DenseNet201 (10). Mr. Mansouri *et al* also classified AD with Control Cases with the XAI technique: Grad-CAM (9). However, he used Transfer Learning with the pre-trained models: CNN, VGG16, ResNet50, and AlexNet. It also created a WGAN-CP model to create synthetic MRI scans to address the limited dataset size, increasing the accuracy. Lastly, Viswan *et al* used the Grad-CAM technique with the pre-trained models: ResNet50, ResNet101, InceptionV3, InceptionResNetV2, and EfficientNetB0 (1).

However, other XAI methods have not been used when classifying between AD and PD. Only the XAI technique Grad-CAM has been used to classify AD and PD with Transfer Learning. Other studies that utilized XAI techniques and Transfer Learning only classify between one neurological disease (AD or PD) with Control cases—not both. Other studies also don't focus on interpretability by using XAI techniques; they use Transfer Learning alone to classify neurological disorders. Additionally, Ensemble Learning can be

used to combine the predictions of pretrained models, further increasing the performance of the model. This led to the research question: How does the integration of XAI, specifically Saliency Maps, Ensemble Learning and Transfer Learning improve the interpretability and the performance of classification of AD, PD, and Control Cases?

This paper will focus on using Saliency Maps combined with the use of Transfer Learning and Ensemble Learning with pre-trained models. The research aims to build trust between healthcare workers and AI models. XAI techniques provide insights into the features and regions crucial for its decision (9). This information from the model can be used by clinicians to make their diagnosis of the patients, instilling confidence in healthcare workers. XAI techniques create a collaborative experience between healthcare workers and AI, where humans and AI technology work together to produce accurate and reliable diagnoses, saving lives. Moreover, XAI techniques also help improve the model's accuracy through the debugging process. If the model makes a wrong diagnosis, the programmer can view the Saliency Map to see why the model made the wrong diagnosis. Because the programmer knows where and why the model made the wrong diagnosis, it becomes easier to fix the problem. Overall, the implementation of XAI techniques not only helps improve the performance of the model but also helps create trust between the AI models and healthcare workers. The collaboration of AI models and healthcare workers helps create fast and reliable diagnoses, helping patients receive treatment as fast as possible, and improving their quality of life.

METHODS AND MATERIALS

Acknowledging that there is an existing gap between classifying AD and PD with XAI techniques and Saliency Maps as it has not been done before; there has been various studies classifying AD and PD alone with XAI techniques and Transfer Learning while there has only been one study that classified AD and PD with XAI techniques, specifically they used the GRAD-CAM technique, with Transfer Learning. Since there are no studies that utilize Saliency Maps to classify AD and PD with Transfer Learning, it is an existing gap. This study also incorporated Ensemble Learning to help improve the performance of the models. Design-based research methodology was used since it focuses on designing models that can be used in the real world, and it requires iterative improvements to the model. Iterative changes were made to the methodology to help improve the performance of the model.

Choosing the Data and Data Preparation

A good dataset must include MRI images of neurological disorders supported by the type of neurological disorder; but they don't need to be separated by train, validation, and test since the code can separate them. Kaggle, the world's largest data science company, has a dataset with 2D MRI scans of Parkinson's, Alzheimer's, and Control Cases. Specifically, the dataset "Alzheimer Parkinson Diseases 3 Class," has a size of 51.88 MB with unknown demographics (13). The MRI scans show an axial view of the brain. The Kaggle's built-in IDE was used along with its NVIDIA TESLA P100 GPU to run the code for the model with 12 GB of RAM, and 20 GB of disk space. The training dataset has 6477 MRI images across the three classes. Many of the MRI images had noise, so they were removed from the dataset to improve performance of the model. Specifically, Noisy images were defined as MRI scans with either too small or large ventricle or had an extremely dark background, making it hard to see the features of the scan; these noisy MRI scans were removed from the dataset, as it would reduce performance of the model. Specifically, 2200 AD scans, 2905 CONTROL scans, and 713 MRI scans were removed from the dataset. The Kaggle dataset does not have a license in its dataset page. Therefore, the research was treated as publicly accessible if it was cited and used for non-commercial purposes. research methodology was exempt from approval by the Institutional Review Board (IRB) because this research does not contain identifiable human subjects. The publicly available Kaggle dataset contains unidentifiable human subjects and therefore does not need to be approved by the IRB.

Kaggle originally split the dataset into AD, PD, and CONTROL MRI images in both test and train folders. These folders were combined and then randomly split into 70% training, 10% validation, and a 20% testing folder. This ensures that training, validation, and testing MRI scans have similar difficulty to predict. All of the images provided by Kaggle had the same format: Portable Network Graphics (Png) of size 176 pixels by 208 pixels.

In this study, an AI model was developed to classify MRI scans Each pretrained model requires the MRI scan to be in a certain format. One might expect the pixel values to be 0 and 1, and another might expect it to be –1 and 1. Therefore, the images were preprocessed with their own build in function from TensorFlow, converting

the image from 0 to 255 to any pixel values the pretrained model requires. This also allows the pretrained model to apply any other transformations they want to the MRI scan. Additionally, the dataset for images was already split into training and testing datasets; all the testing MRI scans were moved into the corresponding training class. Then, the MRI scans were stratified randomly split into 70% training, 10% validation, and 20% test. Stratified random sampling was used so that the distribution of AD, PD, and CONTROL images in the validation and test folder remained the same as the training folders. Combining the images into one folder and then randomly splitting the MRI scans into train, validation, and test ensures that each folder has a similar difficulty to MRI scans, helping the model perform on the test dataset.

Data augmentation techniques are a good approach to increase the amount of training data, generally improving the performance of the model. However, these transformations changed the MRI scans too much, resulting in lower performance. This may be because without rotations or translations of the MRI scans, the pretrained model may become better at identifying important features on the MRI scan, resulting in better diagnoses. For example, rotating the MRI scan may result in certain features to be in different locations, making it harder for the model to train and thus, resulting in worse performance. Additionally, data augmentation techniques may lead to overfitting on the training data. Therefore, data augmentation techniques were omitted from the methodology because of the risk of overfitting and causing different location of features.

To address the imbalance of images across the class, for instance, there are significantly more AD and CONTROL MRI images than PD images in both training and testing datasets, class weights were used to inform the model of the percentage of AD, PD, and CONTROL images in the dataset. This allows the model to perform better despite the data imbalance of the dataset.

Training the model

Transfer Learning techniques were used to train the model. Specifically, the pre-trained models VGG16, ResNet50, InceptionV3, Xception, EfficientNetB0, and EfficientNetB7 models were used. These models were pre-trained on millions of images from ImageNet, which is why when trained to classify neurological disorders they will have a higher accuracy because the model has already been trained on a different task before. Furthermore, the first few blocks (chosen based on the architecture of the model) of the pretrained model were frozen during the first phase of training (20 epochs). This allows the model time to identify which small regions on the MRI scan are important first. Then the model returns the model with the lowest loss during the first 20 epochs. After 20 epochs, the initially unfrozen layers become frozen, and the initially frozen layers become unfrozen for the next 35 epochs. This allows it to train to find important fundamental regions of the brain, which is why it takes more time. Similarly, the model returns to the model with the lowest loss during the first 50 layers. Furthermore, Early Stopping was implemented, monitoring validation loss with a patience of 12 epochs. This was implemented to stop training the model if it did not improve after 12 epochs. Additionally, Reduce Learning Rate on Plateau was implemented, monitoring validation loss with a factor of 0.3, patience of 2 epochs, and a minimum learning rate of . This was implemented to reduce learning rate at plateau. The last layer of the model was changed to have 3 output layers: one for AD, one for PD, and one for control cases. Additionally, the XAI technique Saliency Maps overlayed the weights of the model onto the image, highlighting the most important regions on the MRI scan that led it to its diagnosis. Unlike Grad-CAM, which overlays the gradience from the last layer of the CNN model, Saliency Maps uses the weights and features identified from all of the layers of the CNN. XAI techniques can be useful to help debug the model and help assist healthcare workers see why the model made its decision. If the model made a diagnosis based on the right region on the MRI scan, the healthcare worker can trust its diagnosis; knowing when to and when not to trust the AI model because of XAI techniques helps create a collaborative experience between healthcare workers and AI models, assisting healthcare workers produce fast and reliable diagnosis. As a result, Patients begin to trust AI models. Ensemble Learning can help boost the performance of the pretrained models, creating more reliable diagnoses. After each pretrained model makes a prediction, the predictions are now fed as inputs into a Logistic Regression model with a max iteration of 1000 and a random seed of 5—chosen arbitrary. The Ensemble Learning model was only fed the final diagnosis without the probabilities from the pretrained models. This was done for simplicity and faster training time.

Architecture of VGG16 and VGG19

VGG16 is a pre-trained model with 16 layers while VGG19 is a pre-trained model with 19 layers. Both VGG16 and VGG19 have similar architectures. VGG19

is slightly bigger and may require a slightly longer training time. Both VGG16 and VGG19 have an image input size of 224 by 224 pixels. The images from the dataset were resized and preprocessed with their corresponding preprocessing function from TensorFlow before being placed in the model (14). Both architectures have a 3*3 small kernel that can shift right/left/up/down by 1 pixel each time it moves in the convolutional layer. The ReLU activation function was used in the hidden layers without Visual Geometry Graph since it consumes too much memory (14). The output layer was resized to have 3 outputs for 3 different classes: AD, CONTROL, and PD. The SoftMax activation function was used for the output layer, which makes the sum of the probabilities equal to 1.

Architecture of EfficientNetB0 and EfficientNetB7

EfficientNet is also a architecture family that uses transfer learning. The EfficientNetB0 model has 237 total layers in its architecture. EfficientNetB0 is the smallest and the most efficient in its architecture family. Similar to VGG16 and VGG19 architecture, EfficientNetB0 utilizes 224 * 224 input image size. However, EfficientNetB7 requires a bigger input size of 600 * 600 pixels (15). The EfficientNet family uses Mobile inverted Bottleneck Convolution layer, or MBConv, which is a key component of the EfficientNet family of neural network architectures, which is useful when computational resources are limited (15). The activation of the hidden layers and output layers are the same as VGG16 and VGG19. EfficientNetB7 is a later version of EfficientNetB0, prioritizing performance over efficiency, which is why it has more total layers in its architecture with 813 total layers in its architecture.

Architecture of InceptionV3

The InceptionV3 model, which belongs to the Inception family of architectures, "is a widely used pretrained model particularly designed for image - classification and detecting different object as a task. It incorporates inception modules, which are comprised of parallel convolutional layers that possess varying filter sizes. InceptionV3, which has 48 layers total, also introduces factorized convolutions as a means to reduce the computational cost associated with the model. Moreover, it utilizes batch normalization, auxiliary classifiers, and global average pooling. By employing these architectural elements, the model becomes capable of effectively capturing both local and global features" (2).

Architecture of ResNet50

ResNet50 is a modified version of the Residual Network (ResNet) design. ResNet "pioneered the notion of residual learning, which involves utilizing skip connections to circumvent one or more layers, thereby facilitating the training of exceedingly deep networks. ResNet50 encompasses a total of 50 layers and utilizes residual blocks with shortcut connections. The incorporation of skip connections aids in alleviating the vanishing gradient problem, thereby enabling the model to acquire more effective representations" (2).

Architecture of Xception

Xception, also known as Extreme Inception, "represents an expansion of the Inception framework, albeit with a distinct methodology for dealing with convolutional operations. Instead of relying on conventional convolutions, Xception employs depth wise separable convolutions, decompose a standard-convolution into depth wise convolution and pointwise convolution. This segregation enhances the efficiency of the model and diminishes the number of parameters. In terms of design, Xception achieves commendable performance across various computer vision tasks while adopting a comparatively simpler approach" (2).

Hyperparameters of the models

The same hyperparameters were used for all pre-trained models except for batch size; the optimizer used was Adam because it is a popular optimizer used and because of its adaptive learning rates. A callback used to increase performance during training reduced the learning rate on a plateau. This method reduces the learning rate when a certain metric is not improving anymore; the metric chosen was validation loss. If the validation loss doesn't decrease after a couple epochs, then the learning rate decreases, which helps continue to increase in performance. Additionally, the loss function used was categorical cross-entropy, which is a loss function used during multi-class AI models that penalizes incorrect diagnoses based on the probability distribution of the class prediction, encouraging the model to predict the right class with a high probability. Additionally, the AI model chooses the best performing pretrained model after finished training (e.g. it may be the 55 epoch or an earlier version). This helps choose the best performance of the model. The only hyperparameter that was different for certain pretrained models was batch size. For all pretrained models other than EfficientNetB7, the batch size was 16; however, because EfficientNetB7

has a bigger architecture and takes bigger images, requiring more computational resources, a batch size of 8 was used to save computation resources.

Hypotheses

a. The performance of the EfficientNetB7 model will be better than EfficientNetB0 because EfficientNetB7 has more layers than EfficientNetB0. EfficientNetB7 has 813 layers while EfficientNetB0 has only 237 layers (16).
b. The Parkinson's MRI scan will highlight the motor cortex region of the MRI scan since Parkinson's disease affects motor control, causing a change in the motor cortex region of the brain, which the model may be able to pick up on.
c. The ResNet50 model will perform better than VGG16 model since ResNet50 has more layers than VGG16; ResNet50 has 50 total layers while VGG16 only has 16 (2, 13). A more complex architecture is better suited for classification of neurological disorders since it is also complex.
d. The Ensemble Learning model will perform better than any of the pretrained models, specifically the top 2 performing pretrained models because it ensembles the pretrained models' predictions into a final diagnosis and therefore, combining the strengths of the pretrained models, mitigating the weaknesses, ultimately receiving higher performance. For instance, if a model has a class precision of 99% for AD, whenever the model makes an AD prediction, it is correct 99% of the time. In that case, the ensemble learning model can use that model's prediction alone. Ensemble learning can help combine strengths of the model in this way. Additionally, if one model makes the wrong diagnosis and all of the other models make the correct diagnosis, the ensemble learning model will choose the correct diagnosis since it was more common, thus boosting performance.

Measuring the performance of the model

The following metrics are used to measure the performance of the AI model. After training each pre-trained model and implementing Saliency Maps, these metrics were calculated to see how well the model performed.

Accuracy is an important metric to evaluate the overall performance of the model measuring the quality of positives and negative predictions. Eq. 1 shows how to calculate accuracy. It is based on the ratio of the correct observations to the total observations (10). Precision is the amount of correct positive predictions out of total predictions (10) shown by Eq. 2. Recall or sensitivity is the amount of correct positive predictions out of actual positive predictions shown by Eq. 3. F1-score is also an important metric to analyze the performance of AI model. F1-score is a harmonic mean of precision and recall, two important metrics for performance. This metric, calculated below by Eq. 4, is useful when both recall and precision are important.

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

$$PRECISION = \frac{TP}{TP + FP} \quad \text{(Eq. 2)}$$

$$RECALL = \frac{TP}{TP + FN} \quad \text{(Eq. 3)}$$

$$F1 - score = 2 \cdot \frac{PRECISIOIN \cdot RECALL}{PRECISION + RECALL} \quad \text{(Eq. 4)}$$

TP denotes true positive, TN denotes true negative, FP denotes fake positive, and FN denotes false negative.

Comparing the performances of the models

Because it is hard to say if a pretrained model performed better than another pretrained model purely by chance, statistical methods can be employed to determine whether there is convincing evidence if one pretrained model is better than the other. Specifically, the McNemar's Test is useful when comparing the performance of 2 Machine Learning models, where represents the number of images that model 1 got correct and model 2 was incorrect. Similarly, represents the number of images that model 1 got incorrect and model 2 got correct. The degrees of freedom used was one. The McNemar's Test was applied to the test dataset to measure difference in performance on unseen data. The total test size is 202 MRI scans. Using the value and comparing with , the p-value can be calculated to determine statistical significance, indicating whether there is convincing evidence that one model is better than the other or the difference in performance occurred purely by chance; if the p value is less than 0.05, there is convincing evidence that one model architecture is better than the other. Conversely, if the p-value is greater than 0.05, there is no convincing evidence that one

model architecture is better than the other, difference in performance were likely due to chance.

$$\chi^2 = \frac{(|n_{10} - n_{01}| - 1)^2}{n_{10} + n_{01}} \quad (17)$$

RESULTS

The pre-trained models VGG16, VGG19, ResNet50, InceptionV3, Xception, EfficientNetB0, and EfficientNetB7 had an accuracy of 92.12%, 91.13%, 93.60%, 90.64%, 94.09%, 90.15%, and 94.58%. These models had an F1-score of 93.73%, 93.01%, 94.98%, 92.57%, 95.39%, 91.99%, and 95.81%. The best pretrained model was EfficientNetB7 with an accuracy of 94.58% and an F1-score of 95.81%; however, the ensemble model performed much better. All the pretrained models performed between 90% and 95% accuracy; however, putting these models' predictions together—essentially working together—allowed the model to receive 97.04% accuracy, higher than any of the pretrained models. Ensemble learning combines these models' strengths together to boost performance of the model, thus receiving higher performance than any of the pretrained models. Therefore, the proposed model for this study is the Ensemble Learning model since it had the highest accuracy and F1-score. Interestingly, the F1-score is slightly higher than accuracy for all pretrained models and the Ensemble Learning model. This indicates that the precision and recall values are both higher than accuracy, resulting in a higher F1-score because it is a harmonic mean of the two metrics. This study demonstrates the potential of Explainable AI (XAI) and Deep Transfer Learning by accurately classifying AD and PD with the Ensemble Learning model with an accuracy of 97.04% and an F1-score of 97.69%. These findings build trust between AI-based diagnoses and healthcare workers, making it a valuable tool for early detection. The findings highlight the importance of explainability in medical AI, ensuring that healthcare professionals can understand and rely on model predictions (Table 1).

A confusion matrix provides information about how well the model performed in each class. First, the EfficientNetB7 model received 100% class accuracy (recall) when predicting PD, 92% when predicting AD, and 96% when classifying CONTROL. EfficientNetB7 did slightly worse classifying AD with a class accuracy of 92%. The model mainly struggled to classify AD and Control. The precision values seen on the confusion matrix normalized by column indicate the precision of the class. For instance, 0.97 means that out of all the times the model predicted AD, it was correct 97% of the time; similarly, out of the times the model predicted CONTROL, it was correct 90% of the time. This is useful in clinics because if the model predicts AD or PD,

Table 1. Performance of the pre-trained models

Model Name	Accuracy	F1-score
VGG16	92.12%	93.73%
VGG19	91.13%	93.01%
ResNet50	93.60%	94.98%
InceptionV3	90.64%	92.57%
Xception	94.09%	95.39%
EfficientNetB0	90.15%	91.99%
EfficientNetB7	94.58%	95.81%
Ensemble Learning model	97.04%	97.69%

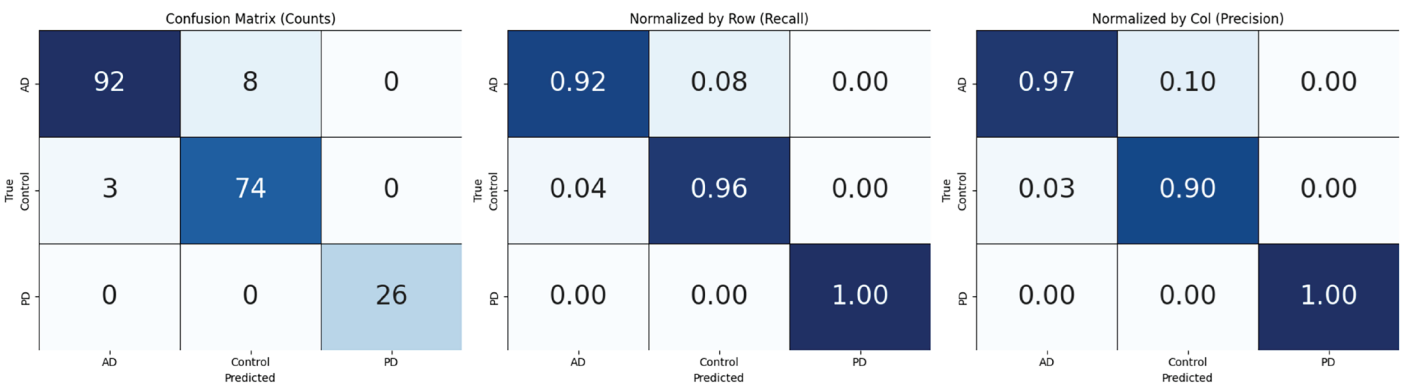

Figure 1. Confusion Matrices of the pretrained model EfficientNetB7 (Original figure created by the author).

it is most likely correct; however, if the model predicts CONTROL, more tests may be needed before confirming the diagnosis.

The ensemble learning model received slightly lower accuracy (recall) on CONTROL cases with 95% accuracy; however, it received higher class accuracy on AD with 98% accuracy. The ensemble learning model was correct 96% of the time when it made an AD prediction; similarly, out of the times the model predicted CONTROL, it was correct 97% and predicted 100% of PD cases. When the ensemble learning model makes a prediction, it is almost certainly correct, achieving high performance—an essential necessity for being applicable to healthcare facilities. When a patient is in critical condition and needs treatment immediately, it is imperative that the diagnosis is accurate so that they can receive the correct treatment.

Notably, the model was able to reach 100% accuracy on the training dataset early on even before phase 2 of training. The model's validation accuracy plateaued at 95% because its training accuracy converged to 100% approximately epoch 8. In general, both the validation and training accuracy of the model generally increased as the epochs increased.

All of the gradients produced by EfficientNetB7 are lighter and less bright than the other pretrained models, which may be due to the larger input size of 600 by 600 pixels compared to the 224 by 224 pixels in the other models. This may cause the gradients to be more spread out, resulting in less brightness.

The model made a prediction that the MRI scan was an AD MRI scan, which is correct. It predicted AD with confidence with over a 99.9% probability of being AD. A higher maximum probability in the distribution indicates

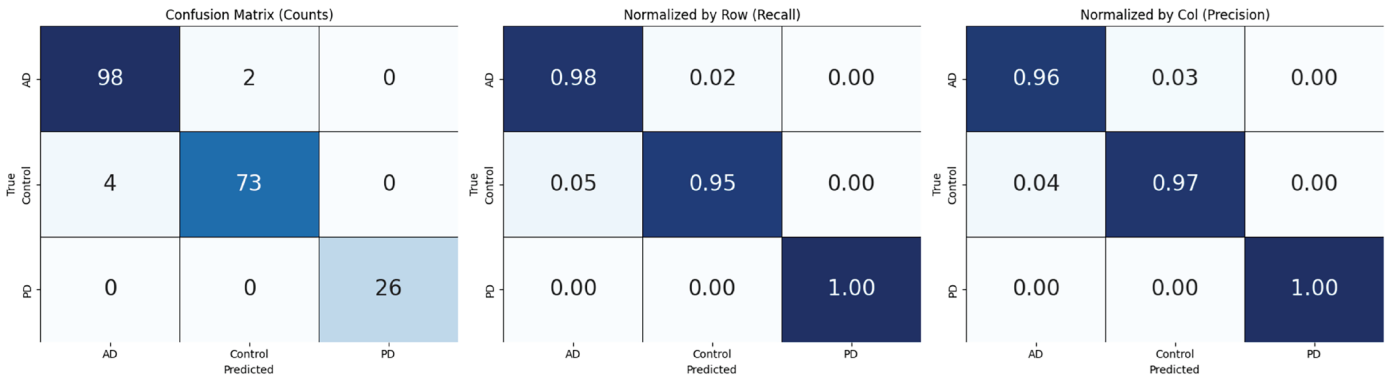

Figure 2. Confusion Matrices of the ensemble learning model (Original figure created by the author).

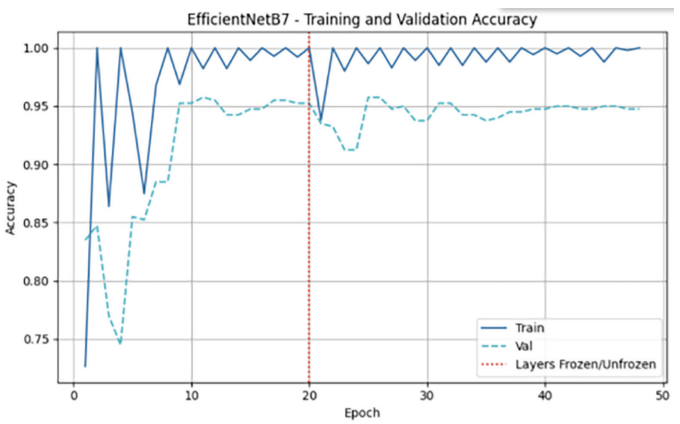

Figure 3. Training and validation accuracy of the pretrained model EfficientNetB7 vs. Epochs (Original figure created by the author).

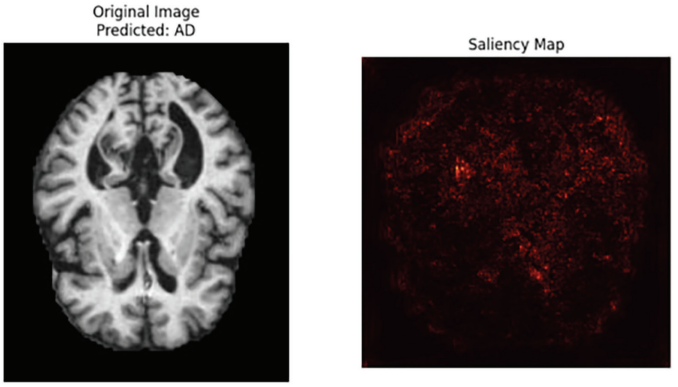

Figure 4. Original MRI scan of AD and the Saliency Map from the pretrained model EfficientNetB7 (Original figure created by the author).
Predicted class: AD (index: 0), Probabilities: [9.9997950e-01, 1.8500439e-05, 1.9888994e-06]

higher confidence because if the model had no idea of what the diagnosis is, the probability distribution could be [0.333, 0.333, 0.334]. The model would still be able to make a diagnosis and may even be correct sometimes, but it still would not be confident. It is not favorable to have a probability distribution like this since the model is equally likely to pick AD as PD, meaning that it may get the wrong answer about a 1/3 of the time. A higher probability increases the likelihood of the model being correct, so a 99.9% probability is favorable. Looking at the Saliency Map produced by the model in Figure 4, the upper middle region is highlighted in yellow dots. This region is the reason why the EfficientNetB7 pre-trained model made its diagnosis.

The model made a prediction that the MRI scan was a Control Case in Figure 5, which is correct with over 99.9% probability. Looking at the Saliency Map produced by the model, the upper middle region of the brain is highlighted in yellow, highlighting that the region the MRI scan made its diagnosis was because of the middle region of the MRI. This is the same region of the brain where the model examined for AD, indicating that the upper middle region is responsible for deciding between AD and CONTROL.

The model made a prediction that the MRI scan was a PD scan in Figure 6 with over 99.9% probability, which is correct. Comparing the Saliency Map produced by the model to the axial view of the brain in Figure 7, the right middle region is highlighted in yellow where the motor cortex lies. Figure 7 only shows the motor cortex region of the right side of the brain, but the motor cortex is also located in the left side of the brain. The model's diagnosis was based on the motor cortex region, which is primary affected by being diagnosed with PD.

All of the gradients produced by VGG19 model are brighter than EfficientNetB7's gradients, resulting in more highlighted regions on the MRI scan. This may be due to the larger input size of 600 by 600 pixels on EfficientNetB7's input size compared to the 224 by 224 pixels in the other models, including VGG19. This may cause the gradients to be more spread out, resulting

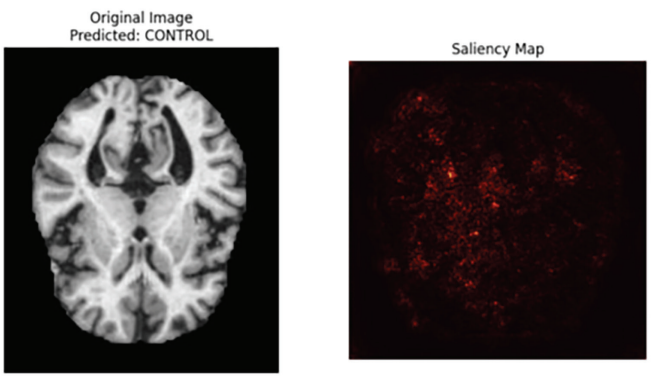

Figure 5. Original MRI scan of CONTROL and the Saliency Map from the pretrained model EfficientNetB7 (Original figure created by the author).
Predicted class: CONTROL (index: 1), Probabilities: [2.0680003e-04, 9.9977845e-01, 1.4799224e-05]

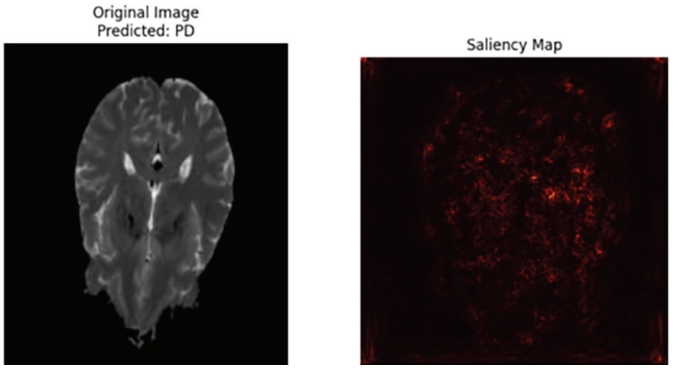

Figure 6. Original MRI scan of PD and the Saliency Map from the pretrained model EfficientNetB7 (Original figure created by the author).
Predicted class: PD (index: 2), Probabilities: [7.3012649e-07, 4.6850533e-07, 9.9999881e-01]

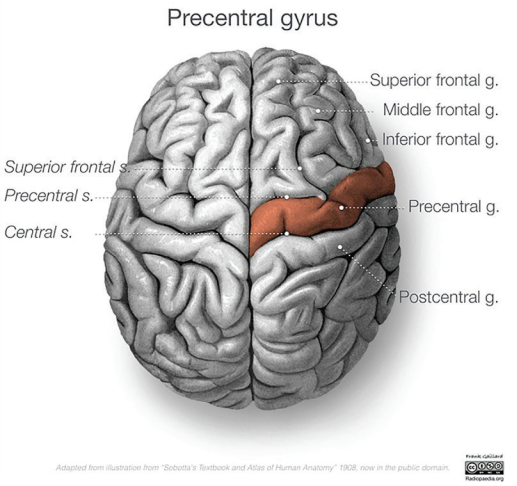

Figure 7. Motor cortex (Precentral gyrus or Postcentral gyrus) and other regions of the brain highlighted in superior view and coronal view of the brain. Adapted from *Radiopaedia.org* under a permissive license: Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) (18).

in less brightness. Additionally, the VGG19 model confidence was lower than EfficientNetB7. This may be due to the lower accuracy of the model, leading to less confidence in its prediction.

The model made a prediction that the MRI scan was an AD MRI scan, which is correct. It predicted AD with confidence with 81.37% probability of being AD, lower than EfficientNetB7's confidence. Looking at the Saliency Map produced by the model in Figure 8, the upper middle region is highlighted in yellow dots. This region is the reason why the VGG19 pre-trained model made its diagnosis. It appears that both the VGG19 and EfficientNetB7 pretrained models highlighted the same regions of the brain, which led to the same diagnosis. This may indicate that the upper region of the brain may be affected by AD. The model also highlighted the lower edge of the brain, which may mean that there is a slight difference in that region when the MRI scan is AD or CONTROL, which may indicate that the lower edge region is affected after receiving AD.

The model made a prediction that the MRI scan was a Control Case in Figure 9, which is correct with 82.69% confidence. Looking at the Saliency Map produced by the model, the upper middle region of the brain is highlighted in yellow, highlighting that the region the MRI scan made its diagnosis was because of the middle region of the MRI. Similar to EfficientNetB7's prediction, this is the same region of the brain where the model looked for classifying AD, indicating that the upper middle region is responsible for classifying AD and CONTROL. Similar to Figure 8, the VGG19 model highlighted the lower edge region of the brain, indicating that the lower edge region may not have been damaged in this CONTROL case. Additionally, this lower edge region is not highlighted by the EfficientNetB7 model, elucidating the idea that different pretrained models can showcase different features of the brain and come to the same conclusion.

Lastly, the model made a prediction that the MRI scan was a PD scan in Figure 10, which is correct with exactly 100% confidence. Comparing the Saliency Map produced by the model to the axial view of the brain seen in Figure 7, the VGG19 model highlighted the

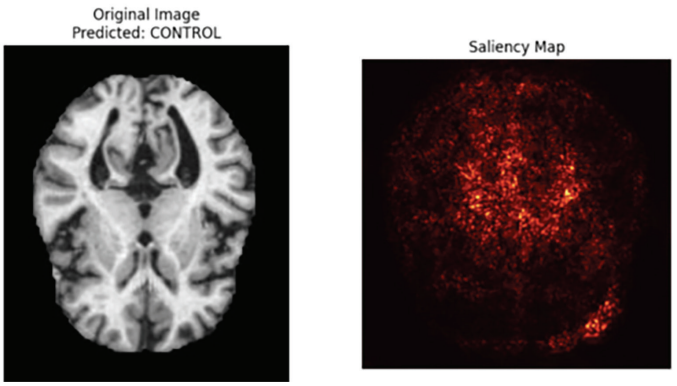

Figure 9. Original MRI scan of CONTROL and the Saliency Map from the Pre-trained model VGG19 (Original figure created by the author).
Predicted class: CONTROL (index: 1), Probabilities: [8.2686730e-02, 9.1704845e-01, 2.6489707e-04]

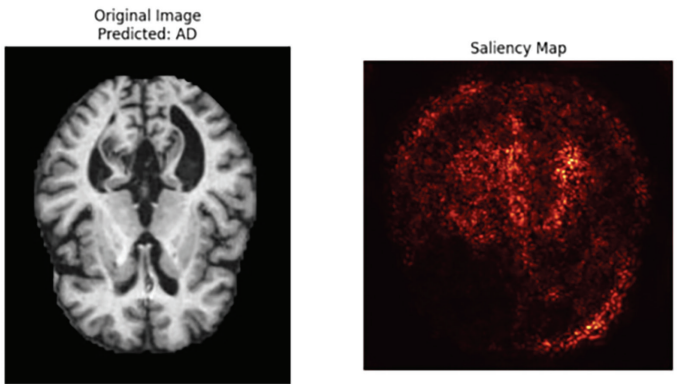

Figure 8. Original MRI scan of AD and the Saliency Map from the pretrained model VGG19 (Original figure created by the author).
Predicted class: AD (index: 0), Probabilities: [0.8137051, 0.18491198, 0.00138297]

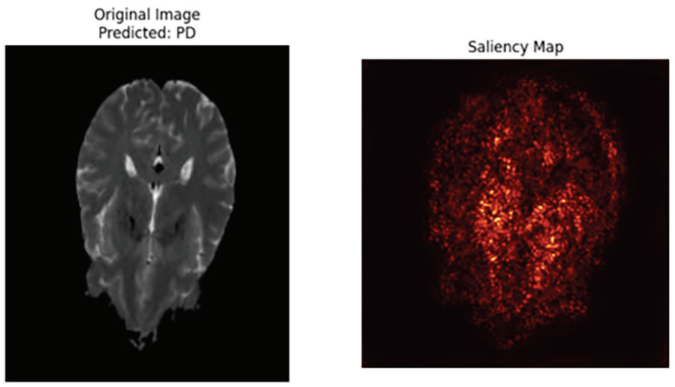

Figure 10. Original MRI scan of PD and the Saliency Map from the Pre-trained model VGG19 (Original figure created by the author).
Predicted class: PD (index: 2), Probabilities: [1.2079941e-11, 8.535346e-11, 1.0000000]

bottom half region of the brain instead of the motor cortex region. This may indicate that other regions of the brain have been affected besides the motor cortex, so the model notices a difference in those regions of the brain. Nonetheless, the pretrained model VGG19 arrives at the right diagnosis, despite identifying a different region of the brain. This showcases the idea that different pretrained models of the brain may identify different features of the same MRI scan. Both of these MRI scans can help physicians observe multiple regions of the brain, building confidence in their diagnosis.

While these saliency maps help elucidate the influential regions of the MRI scan, they become more informative when observing how the outputs correspond to neuropathological features. It helps indicate the damaged regions on the neurodegenerative diseases AD and PD, helping diagnose, as the AI models tend to highlight specific regions in these neurodegenerative diseases. Both VGG19 and EfficientNetB7 highlighted the upper middle region (middle or inferior frontal gyrus) of the MRI scan while diagnosing an AD MRI scan. While this is not the primary site of AD pathology, the frontal gyrus is affected in later stages of AD. This indicates that the models observed the frontal gyrus to make their diagnosis. Moreover, both VGG19 and EfficientNetB7 highlighted the right precentral gyrus or the motor cortex, also a non-primary site of the neurological pathway of the disease. Therefore, both pretrained models highlight regions of the brain that are affected during later stages of the neurological disorder; further analysis is needed to identify which regions of early-stage neurological disorders can be interpreted in the future to help with early diagnosis.

According to Table 2, the p-value for the Xception vs. Ensemble Learning model was 0.04, making it significant. Because $0.04 < 0.05$, this result is statistically significant and there is convincing evidence that the Ensemble Learning model architecture is better than the Xception model architecture. However, for the model pairs EfficientNetB7 vs. EfficientNetB0, ResNet50 vs. VGG16, and EfficientNetB7 vs. Ensemble Learning Model, the p-values were 0.11, 0.75, and 0.13, respectively. These values are not statistically significant because these p-values are above 0.05. Therefore, there is no convincing evidence that one of these model architectures is better than the other. While 0.13 is not statistically significant, these high p-value may be caused by the small test dataset. Comparing the results shown in Table 2, specifically the EfficientNetB7 vs. Ensemble and Xception vs. Ensemble, the n_{10} value was 1 and 0 respectively while the n_{01} values were both 6. The increase in n_{10} by only one caused the p-value to increase

Table 2. Comparing pairs of pretrained models with McNemar's Test

Model Pair (Model 1 vs. Model 2)	Accuracy (%)	F1-Score (%)	n_{10}	n_{01}	χ^2	P value
EfficientNetB7 vs. EfficientNetB0	94.58% / 90.15%	95.81% / 91.99%	17	8	2.56	.11
ResNet50 vs. VGG16	93.60% / 92.12%	94.98% / 93.73%	6	4	.1	.75
EfficientNetB7 vs. Ensemble	94.58% / 97.04%	95.81% / 97.69%	1	6	2.29	.13
Xception vs. Ensemble	94.09% / 97.04%	95.39% / 97.69%	0	6	4.17	.04

Table 3. Highest performing pretrained model of previous studies and this study

Study	Model Type	Dataset	Classes	Accuracy	F1-score
Viswan (1)	ResNet152	NTUA	AD/PD/Control	99.9%	N/A
Mansouri (9)	CNN (Real + Synthetic MRIs)	Kaggle	AD/Control	97.5%	98.25%
Mahmud (10)	XAI Deep TL Ensemble model	OASIS Kaggle	AD/Control	96.0%	91%
Dentamaro (11)	XAI DenseNet11 model	PPMI	PD/Control	96.6%	96.5%
Siddiqua (2)	EfficientNetB0	Kaggle	AD/PD/Control	99.4%	N/A
N/A	Clinical	N/A	AD/Control PD/Control	88-92% 90.3%	N/A
This study	Ensemble Learning Model	Kaggle	AD/PD/Control	97.04%	97.69%

by 0.09. This variation can be reduced with a larger test dataset size, resulting in data that won't vary a lot due to a small change in n_{10}.

DISCUSSION

Compared to existing literature

Compared to prior work, which largely focus on classifying one neurodegenerative (AD or PD) with XAI techniques, there has not been a study that classified AD and PD with Ensemble Learning and XAI techniques. Many of these studies are highly interpretable and have high performance, yet they don't classify multiple neurological disorders. However, Viswan *et al* used the XAI technique Grad-CAM with the pre-trained models: ResNet50, ResNet101, InceptionV3, InceptionResNetV2, and EfficientNetB0 to classify AD and PD (1). Its highest performing pre-trained model ResNet50 has accuracy of 98.8%, 1.76% higher than this study's accuracy. This study's highest performing model was the Ensemble Learning model, which received an accuracy of 97.04% and an F1-score of 97.69%. This study received similar performance as Dr. Viswan with his model performing slightly better than the Ensemble Learning. He used Grad-CAM while this study utilized Saliency Maps (1). Grad-CAM utilizes the gradients of the last convolutional layer of a CNN to generate a heatmap while saliency maps can be applied to any type of neural network architecture with gradients from multiple layers. This study helps improve the interpretability of integrating XAI techniques in the classification of AD and PD.

Dentamaro *et al* classified PD with XAI techniques such as Integrated Gradients and Attention Heatmaps for Vision Transformer (ViT) after using Transfer Learning to train the pre-trained models: ResNet and DenseNet (11). Their DenseNet model performed the best with an accuracy of 96.6%, classifying PD and CONTROL—3.4% lower than the Ensemble Learning Model's 100% PD class accuracy.

Siddiqua *et al* used four pretrained models, which were also used in this study to classify AD, PD, and CONTROL, respectively (2). Its EfficientNetB0 models has a class precision of 1.00, 0.99, 1.00 for AD, PD, and CONTROL with a 99.4% overall accuracy. The overall F1-score of the model was not provide by the study. The Ensemble Learning model has a higher-class precision for PD, but lower-class precision value for AD and CONTROL with 0.98 and 0.95, respectively. The AD class precision is fairly similar to the Siddiqua's model, but the CONTROL model is 5% lower than Siddiqua's model. This study performed 2.36% worse than their study's model. However, the other pretrained models (InceptionV3, ResNet50, and Xception) in Siddiqua's study performed worse than the corresponding pretrained models in this study. In Siddiqua's study the InceptionV3, ResNet50, and Xception pretrained model had an accuracy of 88%, 62%, and 89% respectively while this study had an accuracy of 90.64%, 93.60%, and 94.09%.

Mansouri classified AD and Control cases with an accuracy of 97.5% and an F1-score of 98.25%-- a 0.46% higher accuracy and a 0.56% higher F1-score (9). The study used a GANs to create synthetic MRI scans to help train the model, improving the accuracy of the model. Mahmud classified AD and Control cases with an accuracy of 96.0% and an F1-score of 91.0%--1.04% lower accuracy and 6.69% lower F1-score.

However, this performance difference of models from other studies may have occurred purely by chance, as statistical testing, specifically McNemar's test, wasn't employed to compare the performance of the models. McNemar's test would require the models to be tested on the same test dataset to see the difference in performance. Because, the other studies use different datasets, there is no convincing evidence that one pretrained model is better/worse than the other.

As mentioned in the introduction, clinical diagnoses using blood tests to detect AD have an accuracy of 88% to 92% (5). Similarly, clinical diagnoses can detect PD with an accuracy of 90.3% (6). According to Figure 3, the Ensemble Learning model has a class accuracy of 98% for classification of AD. It also has a class accuracy of 100% class accuracy for classification of PD. These results indicate that AI in classification of neurological disorders can provide faster and more accurate diagnoses than clinical diagnoses. Integrating AI in healthcare can help provide faster and more accurate diagnoses, allowing the patient to receive quicker treatment, improving their lives.

Limitations

Some limitations faced by Kaggle—the coding environment. Specifically, Kaggle's NVIDIA TESLA P100 GPU was used when compiling the code. Using a higher quality GPU can help improve the runtime and performance of the model. Additionally, the model's performance was hindered due to its small dataset size. With either a larger, more balanced dataset or higher quality GPU, the performance and interpretability of the AI model would be improved. Furthermore, the PD class

size was much smaller than the amount of MRI scans in AD and CONTROL. This dataset imbalance may have affected the performance of the pretrained models. Additionally, another limitation is the risk of overfitting to the data, which may result in worse performance when the model sees clinical data; however, this limitation may be able to be resolved by an external validation dataset.

CONCLUSION

Addressing the Hypotheses

In my first hypothesis, it was predicted that EfficientNetB7 would perform better than EfficientNetB0 and all other pretrained because EfficientNetB7 is the biggest architecture out of the pretrained models; it has 813 total layers compared EfficientNetB0's 237 total layers. This study rejects the hypothesis because although EfficientNetB7 performed better with an accuracy of 94.58% and an F1-score of 95.77% while EfficientNetB0 has an accuracy of 90.15% and an F1-score of 91.99%, the p-value according to Table 2 is 0.11. Therefore, there is no convincing evidence that the EfficientNetB0 architecture is better than the EfficientNetB7 architecture. This observed difference likely happed purely by chance alone. Therefore, a bigger architecture may be more complex but may not always perform better.

In my second hypothesis, it was predicted that the Parkinson Disease's MRI scan would highlight the brain region with the motor cortex since PD is responsible for hindering motor function. According to Figure 7, the EfficientNetB7 pretrained model highlighted the right motor cortex, the striatum, and the globus pallidum. Because it highlighted the motor cortex, this study accepts its second hypothesis. The highlighted regions of the brain were the primary reason the pretrained model classified the MRI scan as PD. During the progression of the neurological disease, the disease may have affected these regions, which is why the model was able to notice a difference in those regions of the brain.

In my third hypothesis, it was predicted that the pretrained model ResNet50 would perform better than VGG16 since ResNet50 has more total layers than VGG16. This study rejects the hypothesis because although the VGG16 pre-trained model had an accuracy of 92.12% and an F1-score of 93.73% while the ResNet50 pre-trained model had an accuracy of 93.60% and an F1-score of 94.98%, according to Table 2, the p-value it had to the McNemar's test was 0.75. By the same reasoning as shown in the first hypothesis, the ResNet50 architecture is not better than the VGG16 architecture; the observed difference happened due to chance.

Finally, in my last hypothesis, it was predicted that the Ensemble Learning Model would perform much better than any of the pretrained models, specifically the top 2 performing pretrained models. This study partially accepts this hypothesis because while the top 2 pretrained models had lower accuracy and F1-scores than the Ensemble Learning model, the model pairs (EfficientNetB7 vs. Ensemble and Xception vs. Ensemble) had a p-values of 0.13 and 0.04, respectively. Because $0.04 < 0.05$, there is only enough convincing evidence to conclude that the Ensemble Learning Model architecture is better than the Xception architecture.

The Ensemble Learning model performed better than the EfficientNetB7 model, but not enough to have convincing evidence that the Ensemble Learning architecture is better than EfficientNetB7's architecture. Comparing the precisions and sensitivity (recall) values using confusion matrices shown in Figure 1 and 2 of the EfficientNetB7 model and the Ensemble Learning model, the EfficientNetB7 model had a sensitivity value of 0.92 and 0.96 while the Ensemble Learning model had sensitivity values of 0.98 and 0.95 for AD and CONTROL, respectively. The Ensemble Learning model performs a lot better classifying AD patients while it performs similar to the EfficientNetB7 model when classifying CONTROL. The precision values for EfficientNetB7 was 0.97 and 0.90 while the Ensemble learning model had 0.96 and 0.97 classifying AD and CONTROL. The EfficientNetB7 performs only slightly better when it makes an AD prediction; however, it performed a lot worse precision value at CONTROL. Ensemble Learning is a technique that can combine pretrained models' strengths, mitigating weaknesses and thus, improving the performance of the model. For instance, in a certain MRI scan of CONTROL, many models may misclassify it as AD because it looks similar to a AD case; if one pretrained models is able to classify that MRI scan correctly, the ensemble learning model can put more weight on that diagnosis if it has a higher precision value for CONTROL (e.g. when it makes a prediction of CONTROL, there is a higher chance for that pretrained model to produce its diagnosis). Each pretrained models thinks differently, identifying different features because of their different architectures and different preprocessing steps. This can cause them to identify different features—useful in some MRI scans and not on others; ensembling these models allows the model to reach a much higher performance. This may

be why the Ensemble Learning had a higher precision value for CONTROL. Moreover, both pretrained models received 1.00 sensitivity and precision values, indicating that the models were better at classifying Parkinson's disease. While there the Ensemble Learning model performed better than the EfficentNetB7 model there is no convincing evidence that the EfficientNetB7 architecture is worse than the Ensemble Learning model's architecture, as the p-value of $0.13 > 0.05$.

However, because there is no statistical significance between EfficientNetB7 and the Ensemble Learning model, a larger test dataset size may be needed in order to conclude statistical significance, reduce variance, and provide higher statistical power.

Implications

This study improves the interpretability of AI in diagnosing neurological disorders with excellent accuracy. The model is able to highlight regions on the MRI scan to make its diagnosis. Since models without XAI techniques do not illuminate the reason why it made its diagnosis, healthcare workers do not know when to trust AI models. The XAI technique Saliency Maps can highlight features in the MRI scan, elucidating why the AI model made its decision. The high interpretability of the model created by XAI techniques builds trust between the AI model and the physician since physicians know when to trust the model. Physicians can use the AI model to assist them in making fast and accurate decisions, allowing patients to receive immediate treatment. Immediate treatment improves the patient's quality of life by mitigating symptoms of the disease.

Future Directions

This study developed an AI model that classified AD and PD with Transfer Learning, Ensemble Learning and the XAI technique Saliency Maps. To improve the interpretability of the model, other XAI techniques (e.g., SHAP) can be incorporated or a combined XAI framework can be incorporated to produce a visual gradient map of important features of the MRI scan from the Ensemble Learning model. To improve the performance, studies can utilize generative adversarial networks (GAN) to generate new MRI scans in each class, addressing the small train and test dataset size; this also reduces variability and increases statistical power of McNemar's test. Moreover, integrating multi-modal data such as PET, CT scans, or blood samples with MRI scans can also help enhance the performance of the model since the model has access to more information. Moreover, Stack Ensembling, which includes the probability of the diagnoses for the pretrained models in addition to its diagnosis, can be incorporate to enhance performance since the Stacked Ensemble Learning model have more information on the confident and uncertain predictions of the pretrained models to produce its final diagnosis. Lastly, AI can help classify many more neurological disorders besides just AD and PD. Adding multiple neurological disorders with muti-modal data be useful to detect all neurological disorders with high speed and accuracy, leading to quicker treatment and ultimately aimed at one goal: improving the patient's quality of life.

ACKNOWLEDGEMENTS

I would like to thank Mrs. Webb, my AP Research teacher at Troy Athens High School, for her valuable feedback and guidance during the development of this paper. I also wish to acknowledge Oliver C. Muellerklein for his contributions to the methodology section of the paper. I am grateful for their help.

CONFLICT OF INTEREST

The author declares no conflict of interest, as the data utilized in this study was obtained from Kaggle—a publicly available dataset. Furthermore, no external funding was required for this study.

REFERENCES

1. Viswan V, Shaffi N, Mahmud M, Subramanian K, Hajamohideen F. A Comparative Study of Pretrained Deep Neural Networks for Classifying Alzheimer's and Parkinson's Disease. *2021 IEEE Symposium Series on Computational Intelligence (SSCI).* 2023; 1334–1339, doi:https://doi.org/10.1109/ssci52147.2023.10371843.
2. Siddiqua A, Oni AM, Miah MJ. A Transfer Learning Approach for Neurodegenerative Disease Classification from Brain MRI Images: Distinguishing Alzheimer's, Parkinson's, and Control Cases. *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT).* 2024; 347–351, doi:https://doi.org/10.1109/iceeict62016.2024.10534463.
3. Rossi M, Perez-Lloret S, Merello M. How Much Time Is Needed in Clinical Practice to Reach a Diagnosis of Clinically Established Parkinson's Disease? *Parkinsonism & Related Disorders.* 2021; 92: 53–58, doi:https://doi.org/10.1016/j.parkreldis.2021.10.016.

4. Kvello-Alme M, Bråthen G, White LR, Sando SB. Time to Diagnosis in Young Ons*et al*zheimer's Disease: A Population-Based Study from Central Norway. *Journal of Alzheimer's Disease.* 2021; 82: 965–974, doi:https://doi.org/10.3233/jad-210090.
5. Reynolds S. Accurate Blood Test for Alzheimer's Disease Available online: https://www.nih.gov/news-events/nih-research-matters/accurate-blood-test-alzheimer-s-disease (accessed on 6 January 2025).
6. Virameteekul S, Revesz T, Jaunmuktane Z, Warner TT, De Pablo-Fernández E. Clinical Diagnostic Accuracy of Parkinson's Disease: Where Do We Stand? *Movement Disorders.* 2023; 38. doi:https://doi.org/10.1002/mds.29317.
7. Geeksforgeeks ML | Introduction to Transfer Learning Available online: https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning/ (accessed on 14 January 2025).
8. Al-Zharani M, Ansarullah SI, Al-Eissa MS, Dar GM, Alqahtani RA, Alkahtani S. Exploring the Efficacy of Deep Learning Techniques in Detecting and Diagnosing Alzheimer's Disease: A Comparative Study. *Journal of Disability Research*. 2024; 3: doi:https://doi.org/10.57197/jdr-2024-0064.
9. Mansouri D, Echtioui A, Khemakhem R, Hamida AB. Explainable AI Framework for Alzheimer's Diagnosis Using Convolutional Neural Networks. *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP).* 2024; 93–98, doi:https://doi.org/10.1109/atsip62566.2024.10639037.
10. Mahmud T, Barua K, Umme Habiba S, Sharmen N, *et al*. An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning. *Diagnostics*. 2024; 14: 345–345. doi:https://doi.org/10.3390/diagnostics14030345.
11. Dentamaro V, Impedovo D, Musti L, Pirlo G, Taurisano P. Enhancing Early Parkinson's Disease Detection through Multimodal Deep Learning and Explainable AI: Insights from the PPMI Database. *Scientific Reports.* 2024; 14. doi:https://doi.org/10.1038/s41598-024-70165-4.
12. Rama B, Praveen P, Shaik MA. Machine Learning Model to Detect Parkinson's Disease Using MRI Data. *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA).* 2023; 1516–1521. doi:https://doi.org/10.1109/icscna58489.2023.10370527.
13. Alzheimer Parkinson Diseases 3 Class Available online: https://www.kaggle.com/datasets/farjanakabirsamanta/alzheimer-diseases-3-class/data (accessed on 15 September 2024).
14. Muthamil Sudar K, Nagaraj P, Nithisaa S, Aishwarya R, *et al*. Alzheimer's Disease Analysis Using Explainable Artificial Intelligence (XAI). *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).* 2022. doi:https://doi.org/10.1109/icscds53736.2022.9760858.
15. Shijin Knox GU, Anurenjan PR, Sreeni KG. Detecting Alzheimer's Disease Using Multi-Modal Data: An Approach Combining Transfer Learning and Ensemble Learning. *2023 International Conference on Control, Communication and Computing (ICCC).* 2023; 1–6, doi:https://doi.org/10.1109/iccc57789.2023.10165454.
16. Agarwal V. Complete Architectural Details of All EfficientNet Models Available online: https://medium.com/data-science/complete-architectural-details-of-all-efficientnet-models-5fd5b736142 (accessed on 15 April 2025).
17. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*. 1998; 10: 1895–1923. doi:https://doi.org/10.1162/089976698300017197.
18. Gaillard F. Neuroanatomy: Superior Cortex (Diagrams). Radiopaedia.org. 2018. doi:https://doi.org/10.53347/rid-59317.