Prediction of Parkinson's Disease with Convolutional Neural Networks using Structural T1 Weighted MRIs

Dimitrios Kirbizakis

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science (MCompSc) at

Concordia University

Montréal, Québec, Canada

January 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Dimitrios Kirbizakis**

Entitled: **Prediction of Parkinson's Disease with Convolutional Neural Networks using Structural T1 Weighted MRIs**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science (MCompSc)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Marta Kersten-Oertel

_____ Examiner
Dr. Jean-Baptiste Poline

_____ Supervisor
Dr. Tristan Glatard

Approved by _____
Marta Kersten-Oertel, Chair
Department of Computer Science and Software Engineering

_____ 2026

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Prediction of Parkinson's Disease with Convolutional Neural Networks using Structural T1 Weighted MRIs

Dimitrios Kirbizakis

Parkinson's disease (PD) is a progressive neurodegenerative disorder with no current cure, where early detection and accurate prediction of disease progression would be crucial for patient care and treatment planning. Structural T1-weighted Magnetic Resonance Imaging (MRI) offers a non-invasive way to investigate brain patterns, and Convolutional Neural Networks (CNNs) have shown promise in medical image analysis to detect these patterns automatically. This study evaluates how 3D CNN architectures can be used to predict PD diagnosis using baseline T1-weighted MRIs from the Parkinson's Progression Markers Initiative (PPMI), as diagnosis is the first step to progression. We replicated two published CNN models used for PD classification and followed the preprocessing pipelines found from the original studies. The model's performance was validated using permutation testing and explainability techniques, including Grad-CAM and saliency maps. Results showed that for PD classification, all tested CNNs achieved near-chance performance (ROC AUC around 0.6 at best), with explainability maps revealing no stable or meaningful patterns, suggesting random predictions. To ensure the models were capable of classification in general, we tested the models on sex classification, which is correlated to MRI data. The same models performed substantially better on sex classification tasks (ROC AUC around 0.7-0.8, p-value < 0.01), with consistent spatial focus along the boundaries of the brain, indicating effective pattern recognition in a binary classification setting. These findings suggest that while the evaluated CNNs trained on PPMI data are capable of learning structural features in MRI data, their current configurations and available datasets are insufficient for reliable PD classification.

Acknowledgments

First, I would like to thank my supervisor, Dr. Tristan Glatard, for giving me the opportunity to learn from him and work in his lab. Under his supervision, I have been able to learn more than I had expected and have been able to work with his constant support. I am grateful for everything he has done for me over the years that I have worked with him.

I would like to express my sincere gratitude to the members of my examination committee for taking the time to review my thesis and participate in my defense. Their feedback, questions, and insights have greatly strengthened the quality of this work, and I am truly appreciative of their commitment and expertise.

I would like to thank every member of the Big Data Lab for always being there when I needed support. Without their help, I would not have been able to work as effectively as I have for my thesis. I am glad to have had the opportunity to work with them and learn from them.

Lastly, I would like to thank everyone who is close to me. My parents, my siblings, and my close friends who have always been there for me to keep pushing forward. Without their support, most of my work would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Parkinson's disease (PD) is a neurodegenerative disease that impacts motor functions, mental health, and leads to other serious health problems. Studies showed that approximately 6.1 million people around the world were affected by PD in 2016 [1]. Unfortunately, there is currently no known cure for PD. Although PD cannot currently be cured, early detection and prediction of disease progression could improve quality of life and help understand the disease [2]. Magnetic Resonance Imaging (MRI) has the potential to help with both diagnosis and potentially prognosis of PD. Examination of T1-weighted MRI scans can be helpful in this regard through deep learning to help with pattern recognition from one patient to another without the need for surgery or other invasive procedures.

In this study, our ultimate goal is to predict the progression of PD through the use of 3D Convolutional Neural Networks (CNN). As a first step, we investigate whether CNNs can accurately detect PD from T1-weighted MRI scans.

1.1 Application of Machine Learning in Medical Field

The idea of AI in the medical field in general has existed for many decades, dating back to the 1950s. Throughout this time until the 1970s, medicine was slow to adopt AI. However, the foundation development during this time was useful as a digital resource for the future development of biomedicine [3]. In the 1970s, AI began to take off with more practical medical applications.

Two of the biggest developments from this time were the AI programs INTERNIST-I and MYCIN. INTERNIST-I was a decision-tree based computer program designed to diagnose medical conditions of patients based on patient symptoms. Around the same time, MYCIN was also developed, which was an AI program that helped diagnose and treat bacterial infections with the proper antibiotics [4].

As digital health data, medical imaging, and computational power expanded throughout the 2000s, ML methods became increasingly capable in clinical decision support. The most significant progress emerged with deep learning, particularly CNN-based approaches, which now define state-of-the-art performance in many medical-imaging tasks, including anatomical segmentation, disease classification, and risk prediction [29].

These advancements have been especially impactful in neurology, where early biomarkers of disease can be subtle and difficult to detect visually. MRI-based ML enables extraction of structural and textural patterns that may indicate underlying pathology. As a result, CNNs are now widely used in research, to investigate disorders such as Alzheimer's disease, multiple sclerosis, and PD [30].

1.2 Machine Learning and Parkinson's Disease

Early PD diagnosis on its own has always been challenging, as noticeable differences in motor characteristics and cognitive impairment are difficult to observe [28]. That is why Machine Learning (ML) methods have been implemented and studied over the years to help with better diagnosis and prognosis of PD. The research in the paper by Mei et al. collected 209 different studies from 2020 with machine learning methods that have been used in the diagnosis and differential diagnosis of PD [7]. The datasets, model types, and performance of these models were all written down and put into a table.

Figure 1, reproduced from Mei et al. [7] shows the average accuracies of the models in all the studies found in the article along with the models used. Mei et al. conclude from this study that machine learning approaches have the potential to help with the diagnosis and prognosis of PD [7]. Plenty of machine learning methodology is used with multiple data types in the study, but one of

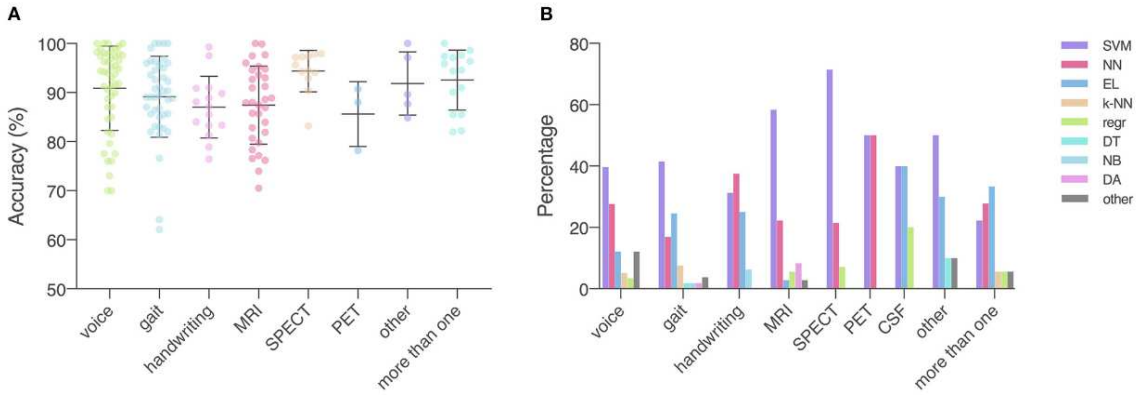


Figure 1: Data type, machine learning models applied, and accuracy. (A) Accuracy achieved in individual studies and average accuracy for each data type. Error bar: standard deviation. (B) Distribution of machine learning models applied per data type. MRI, magnetic resonance imaging; SPECT, single-photon emission computed tomography; PET, positron emission tomography; CSF, cerebrospinal fluid; SVM, support vector machine; NN, neural network; EL, ensemble learning; k-NN, nearest neighbor; regr, regression; DT, decision tree; NB, naïve Bayes; DA, discriminant analysis; other: data/models that do not belong to any of the given categories. Reproduced from Mei et al. [7]

the less explored routes is the diagnosis of PD with magnetic resonance imaging (MRI) as the data type and specifically CNN as the machine learning method, which we would like to explore.

1.3 Convolutional Neural Networks

CNNs are a specialized type of deep learning model that has historically been used for visual applications since the 1980s. After multiple breakthroughs, CNNs have become a very popular model due to their superior performance for image processing applications [8]. CNNs have many layers, but they are primarily composed of three types of layers.

The convolutional layer is arguably the most important layer of the CNN, as this is the part where most of the calculations are required. The most important features of the image, such as edges and textures, are pulled from this layer, which are needed for categorizing [9]. The pooling layer pulls important information from an image and reduces its size, effectively downsampling the data. Common types of pooling used are max and average pooling, where max pooling selects the patches from the input and provides the largest value while ignoring the rest of the information and average pooling takes the average of the filter [8]. The fully connected layer is the final layer, where after the convolutional and pooling layers have extracted the features, this layer produces the required class prediction, similar to how they are used in regular NNs [10].

The purpose of medical imaging is to help visualize the body of a patient, making it easier for clinicians or doctors to make the diagnostic process easier and more efficient. There are many medical images used for this, such as X-rays, computed tomography, positron emission tomography, and, of course, magnetic resonance imaging (MRI) [10]. Of these imaging types, we decided to run MRIs for our CNN input as they are used extensively for neuroimaging. In addition, MRI scans are non-invasive and widely available in large public datasets such as PPMI, making them an ideal choice for reproducible machine-learning studies.

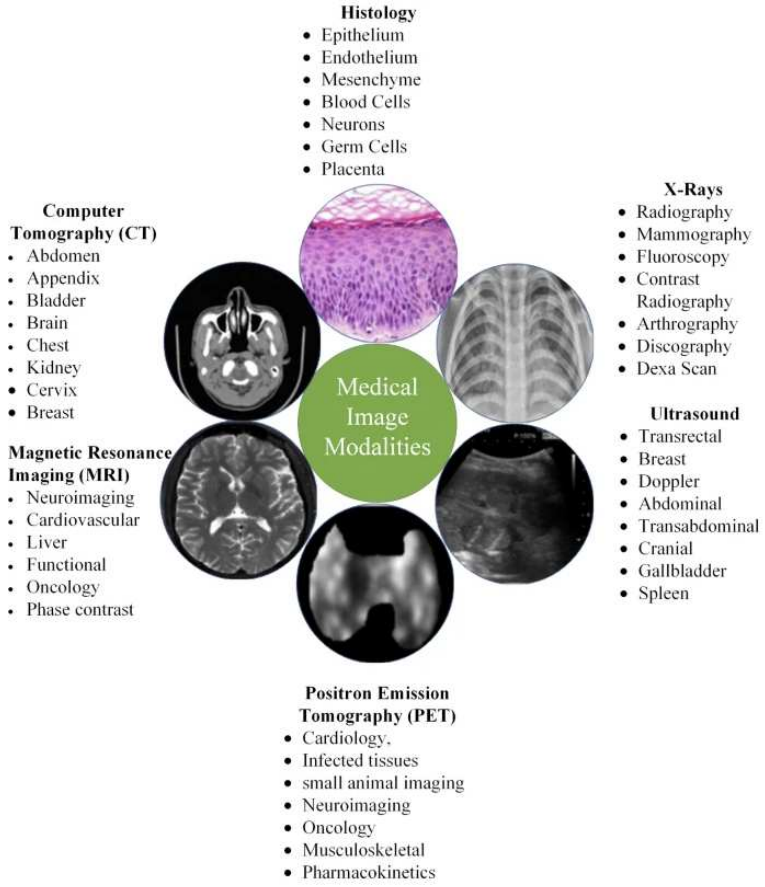


Figure 2: Typology of medical imaging modalities, reproduced from Anwar et al. [10]

1.4 Objectives

The primary objective of this thesis is to evaluate the practical effectiveness of previously published 3D CNN models designed to classify PD using baseline structural T1-weighted MRI scans.

By independently reproducing two established CNN architectures from the literature and validating their performance on the PPMI dataset, we aim to determine whether baseline anatomical MRI contains sufficient discriminative signal to support reliable PD classification.

A secondary objective is to assess whether successful model performance in this setting reflects meaningful disease-related structure rather than artifacts or sampling noise. To address this, we incorporate a permutation-testing framework to quantify the statistical significance of observed performance against chance.

To verify that the models and training pipeline are capable of learning valid structural features when such features exist, we additionally perform a sex-classification experiment using the same dataset and preprocessing steps but with a model explicitly designed for this task. Sex classification serves as a positive control known to yield strong MRI-based separation and is used here to benchmark general binary classification capability.

Chapter 2

Literature Review

In this chapter, we will look at previous work done on classifying PD. For this, we have looked for papers that include similar objectives as this along with as much inclusion details for the input, pre-processing, and model used. We also looked for other related CNNs to see if they could also be applied to PD classification.

2.1 Parkinson's disease classification

The tool we used to look for papers for our study was PubMed [20]. We searched for papers that included keywords such as 3D-CNN and 3D-MRI. For the literature review, the main objective was noted along with other metrics and information shared with the article. The most important parts were what kind of metrics they had recorded in their experiment, if there was any code available with the paper, if the model used for the study was a CNN, what data set the MRIs were taken from and the preprocessing methods performed for the data set. For the dataset, it was important that the data came from Parkinson's Progression Markers Initiative (PPMI), a study whose objective was to establish an early cohort defined by biomarkers to follow to identify progression biomarkers [21]. PPMI is used in many studies related to PD as subject data is collected from various sites in North America, Israel, and Europe, and covers a wide variety of imagine methods, clinical measures, and extensive biological sampling [22]. When it came to the rest of the content in the papers, we encountered some challenges in finding papers that met all of our standards. One of the biggest

challenges we came across was that papers in general did not provide any form of source code to attempt to replicate the structure of the CNN models. Ultimately, we selected the study by Erdas et al. [11] because it provided sufficient methodological detail, used T1-weighted MRIs as input, and shared a CNN architecture compatible with our experimental design.

The study by Erdas et al. [11] was selected because it provided a clear CNN architecture using T1-weighted MRIs as input, making it well suited for comparison with our dataset. Among the configurations reported in their results (Table 1), the All Brain Downsized ×3 condition was the most relevant for our replication, as it corresponded to the 3D CNN setup rather than the 2D slice-based models. The ×4 downsizing condition was not used in our experiments because it produced input dimensions smaller than those compatible with our data. One limitation of the Erdas et al.

	Accuracy	**F_1 Score**	**Precision**	**Recall**
Median Slices	0.9620	0.9452	0.9407	0.9536
Axial Median Slice	0.9319	0.8992	0.9162	0.8825
Coronal Median Slice	0.9381	0.9074	0.9341	0.8866
Sagittal Median Slice	0.9354	0.9036	0.9292	0.8835
All Brain Downsized $\times$ 3	0.9558	0.9046	0.8943	0.9151
All Brain Downsized $\times$ 4	0.9549	0.8953	0.8417	0.9561

Table 1: Classification results for Parkinson's disease detection, reproduced from Erdas et al. [11]

[11] study is that neither the publication nor its supplementary materials reported the area under the curve (AUC), which is an important metric to assess model performance. Because of this omission, it was not possible to verify the degree to which our replicated model matched the original in terms of discriminative accuracy. To address this limitation, we additionally referenced the CNN framework described by Dhinagar et al. [12], which employed a similar input modality, preprocessing approach, and evaluation metrics.

This study had a few different steps compared to Erdas et al. [11]. Instead of classifying PD with 2D and 3D CNNs, this paper classifies both PD and AD using a Random Forest classifier and a 3D CNN. Since our study will only be covering PD, we reviewed all of the steps done in this paper for the PD classification.

As seen in Table 2, we now have an AUC that we can reference to how our model will perform.

Metric	PPMI		UPenn (Average of balanced subsets with the standard deviation)	
	Random Forest	**3D CNN**	**Random Forest**	**3D CNN**
ROC AUC	0.524	**0.667**	0.534 (0.066)	**0.743 (0.111)**
PR AUC	0.815	**0.812**	0.520 (0.067)	**0.777 (0.102)**
Accuracy	0.521	**0.706**	0.567 (0.066)	**0.709 (0.082)**
Precision	0.864	**0.839**	0.563 (0.071)	**0.816 (0.085)**
Recall	0.487	**0.722**	0.673 (0.248)	**0.560 (0.208)**
F1-score	0.623	**0.776**	0.581 (0.148)	**0.631 (0.177)**

Table 2: Performance of Random Forest and 3D CNN for PD classification, reproduced from Dhinagar et al. [12]

There are two AUCs displayed, Receiver Operating Characteristic (ROC) AUC and Precision Recall (PR) AUC. ROC AUC is used to identify a model's general performance while PR AUC is used to identify the positive class' performance in a skewed dataset [23]. Generally, if the ROC AUC is around 0.5, it shows that the model's performance is random. In the paper, it is shown that the random forest classifier used had a random performance with the PPMI dataset, but the 3D CNN performed better than random as shown in Table 2, so it was suitable for our study. There are slight differences when comparing this model to Erdas et al. [11]. The input parameters are somewhat the same, but the preprocessing method is different, and the model has slight numerical changes in the filter sizes. This paper did not have code to replicate it either, so we had to adapt our code to the model description in the paper.

2.2 Sex classification

To validate that the models and input data were functioning as expected, we conducted an additional experiment in which the classification task was changed from Parkinson's disease (PD) status to biological sex. This was done because the ROC AUC values for the PD classification were moderate (Table 2), raising the possibility that the model performance issues could stem from data or preprocessing rather than the task itself.

Since MRI-derived features such as total brain volume and cortical thickness are known to differ significantly between males and [24], we expected the model to achieve a relatively high AUC when

trained for sex classification. A strong performance in this control task would therefore indicate that the model and preprocessing pipeline were operating correctly, whereas poor performance would suggest potential issues in the data or model configuration. For this purpose, we also referenced a previously published 3D CNN designed for sex classification [13].

The model architecture that we replicated from this paper was the BrainNN architecture described in Figure [4]. The code for this model was publicly available for replication, making this process easier to follow. We did not make a new cohort for this, as we figured the model would be able to classify sex using the two cohorts we had already made from the first two papers.

Based on our literature review, three studies were selected for replication: Erdas et al. [11], Dhinagar et al. [12], and Ebel et al. [13]. The studies by Erdas and Dhinagar focused on Parkinson's disease (PD) classification using T1-weighted MRI scans and 3D convolutional neural networks (3D-CNNs), making them directly relevant to our research objective. The study by Ebel et al., which applied a similar CNN framework for sex classification, was included as a methodological control to verify that our models and preprocessing pipeline were functioning correctly.

The primary goal of our study was to replicate the PD classification results reported in Erdas and Dhinagar, particularly the ROC AUC values, using equivalent preprocessing steps and CNN architectures. Following this, we evaluated the same models on the sex-classification task based on the approach described in Ebactuel et al., allowing us to confirm whether our implementation could reliably learn biologically grounded distinctions from MRI data.

2.3 Models and Explainability

To faithfully assess the reproducibility of previously published work, we reimplemented three CNN architectures based on the descriptions in the original papers. Two of these models—proposed by Dhinagar et al. [12] and Erdas et al. [11]—were designed for the classification of PD. The remaining model—introduced by Ebel et al. [13]—was developed to classify biological sex from structural MRI. These two tasks allowed us to evaluate the generalizability of the model.

In the absence of publicly available source code for the PD models, we reconstructed the architectures using detailed layer descriptions, dimension flow diagrams, and hyperparameters provided

in the respective papers. Our implementation of the modular CNN for the classification of PD is illustrated in Figure 3. It consists of four convolutional blocks, each followed by batch normalization, ReLU activation, and max pooling, culminating in a global average pooling layer and two fully connected layers with dropout. This structure reflects a VGG-style hierarchy of progressively abstracted spatial features and is closely aligned with the designs described in [12] and [11].

For sex classification, we reproduced the custom 3D CNN architecture presented by Ebel et al. [13], shown in Figure 4. The model is shallower and employs wider receptive fields—larger convolutional kernels that integrate information over a broader spatial extent of the brain. This design enables the extraction of large-scale morphometric information. Because sex differences in brain structure are generally global—affecting cortical boundaries, total intracranial volume, and ventricular morphology—this architecture is well suited to the problem. We standardized voxel dimensions across both tasks to ensure comparability across preprocessing pipelines.

Overall, care was taken to reproduce each model as faithfully as possible while maintaining consistent training and preprocessing conditions across experiments. This controlled framework enables direct comparison between architectures and ensures that performance differences reflect the underlying data signal rather than implementation artifacts.

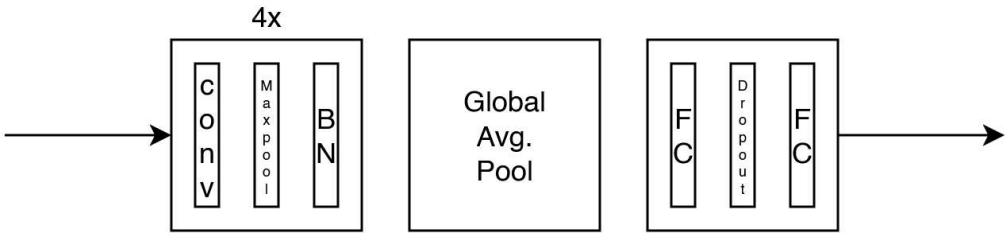


Figure 3: Replicated CNN Structure for classifying PD: Four Modules, Global average pooling layer, Two fully connected layers and a dropout layer. Reproduced from Dhinagar et al. [12]

To improve the interpretability of the CNNs in our study, we implemented two explainability methods: saliency mapping and Gradient-weighted Class Activation Mapping (Grad-CAM). The purpose of these methods is to provide voxel-level visualizations of regions that influence the model's predictions the most.

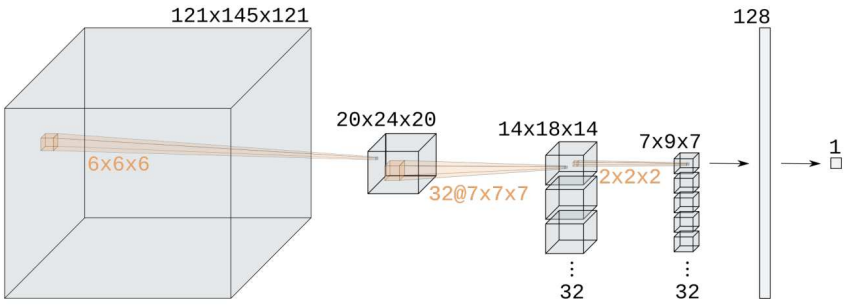


Figure 4: Replicated CNN Structure for classifying sex: The first layer is the input image, a 121 × 145 × 121 voxel MRI scan with one gray value per voxel. It is followed by a 6 × 6 × 6 max pooling (same stride), resulting in a 20 × 24 × 20 layer. To this, a 7 × 7 × 7 convolution with 32 filters is applied, yielding a 32 × 14 × 18 × 14 layer. This is again max-pooled with 2 × 2 × 2 (stride 'same'), resulting in a 32 × 7 × 9 × 7 layer. This isflattened and fully connected to a 128 unit dense layer (left arrow). A dropout layer (not shown) with rate 0.5 is applied before the final dense layer (right arrow). The last layer outputs a single unit - the femaleness probability. Reproduced from Ebel et al. [13].

2.3.1 Saliency mapping

To investigate which regions of the MRI volumes contributed the most to the model predictions, we generated gradient-based saliency maps using guided backpropagation. Saliency maps were computed for individual subjects by taking the gradient of the output class score with respect to the input voxel intensities. The magnitude of this gradient reflects how sensitive the classification decision is to perturbations at each voxel location, allowing us to infer which brain regions the model relies on to make its predictions.

For each subject, the resulting voxel-wise saliency values were normalized and visualized as overlays on anatomical slices in the axial, coronal, and sagittal planes. This representation provides complementary spatial insight by highlighting continuous 3D regions of salient structure while preserving the underlying morphology. Because gradient-based methods can be noisy and sensitive to local perturbations, we visually inspected multiple subjects across both prediction classes to verify that the resulting saliency distributions were consistent and not driven by isolated outliers.

Saliency mapping is particularly valuable in neuroimaging applications because it enables a qualitative assessment of whether CNNs exploit biologically plausible anatomical features or instead rely on spurious patterns such as edges introduced during skull stripping or scanner-specific

artifacts. Although these visualizations are interpretive rather than diagnostic, observing consistency across subjects and models increases confidence that the networks have learned meaningful and generalizable structural patterns.

2.3.2 Gradient-weighted Class Activation Mapping

To further localize the spatial features driving model predictions, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to the final convolutional layer of each network. Grad-CAM computes a weighted combination of the feature maps in the last convolutional layer, where the weights correspond to the gradients of the predicted class score with respect to those feature maps. This enables the extraction of spatially coherent activation regions that contribute the most strongly to the decision of the model.

The resulting Grad-CAM volumes were upsampled to the original MRI resolution using trilinear interpolation and masked to exclude non-brain regions, ensuring anatomically relevant visualization. We examined both raw 3D overlays and 2D slice-wise projections in axial, coronal, and sagittal orientations to assess robustness across views.

Unlike saliency maps—which often highlight fine-grained gradient fluctuations—Grad-CAM emphasizes larger-scale features by taking advantage the receptive fields of deeper convolutional layers. Thus, Grad-CAM provides a more semantically interpretable assessment of model reasoning, particularly in tasks such as sex classification, where global morphometric differences are expected. By comparing Grad-CAM activations across models and preprocessing pipelines, we evaluated whether CNNs consistently attended plausible neuroanatomical regions or exhibited susceptibility to confounding features.

Together, saliency mapping and Grad-CAM provide a complementary framework for interpretability, enabling validation of learned features not only in terms of predictive performance but also anatomical relevance.

Chapter 3

Methodology

The test relevant to our replication experiment in Dhinagar et al. [12] reported a ROC AUC of 0.667 using the PPMI dataset, as shown in Table 2. The other PD classifying CNN found in Erdas et al. [11] did not report any AUCs, but the model was similar. As these papers provided detailed replication parameters for classifying PD, we attempted to replicate these model based on what the papers provided. We started by downloading the MRIs from PPMI with the same settings, followed by using the preprocessing techniques used in both papers. This created two separate cohorts to use in our experiment. With these cohorts, we tested each of them in both CNNs along with the CNN provided in the sex classification paper [13].

Three different CNN models were used for the experiment, one from each paper mentioned in the literature review. Two of the models used for the experiment were similar in structure. Papers [11] and [12] use similar parameters and structure, the minor difference being the number of filters in each module.

3.1 Dataset

All three replicated studies used MRI data obtained from the PPMI database, which provides one of the largest publicly available imaging repositories for Parkinson's disease research. PPMI includes longitudinal clinical, genetic, and imaging data collected across multiple international acquisition sites, making it a widely used benchmark for studies on the early detection and progression

of PD.

For our replication experiments, we restricted the dataset to Baseline (BL) visits and structural T1-weighted 3D MRI scans, matching the inclusion criteria described in Erdas et al. [11], and Dhinagar et al. [12]. Several key demographic and imaging-related variables were extracted to characterize the dataset: age at baseline, biological sex, diagnostic group (Parkinson's disease vs. Healthy Control), and scanner information when available. Because acquisition protocols vary slightly across sites, we applied filtering criteria to ensure consistency across subjects.

3.1.1 PPMI Search and Filtering Parameters

The following filters were applied when querying the PPMI database:

- Group: PD, Control
- Weighting: T1
- Thickness: Between 1 mm and 1.5 mm
- Modality: MRI
- Type: 3D
- Visit: BL
- Model: Any
- Manufacturer: Any

These parameters ensured that all included scans were structurally comparable and matched the specifications reported in the original studies.

3.2 Image Pre-processing

After downloading all available MRI scans that met the inclusion criteria described in the previous section, the next step was pre-processing and quality control. From this dataset, we created two

distinct cohorts: one processed following the preprocessing steps of Erdas et al. [11], and the other following the preprocessing pipeline described by Dhinagar et al. [12].

The pre-processing performed in Erdas et al. [11] involved two main steps. The first step was the spatial normalization of the MRI scans to the Montreal Neurological Institute (MNI) template to ensure consistent anatomical alignment across subjects [14]. The reference image for image registration was the MNI 152 T1w Linear one-mm atlas [15] and the tool used was the FMRIB's Linear Image Registration Tool (FLIRT), which is a tool that is part of the FMRIB Software Library (FSL) [16].

After registering all MRIs, the next step of pre-processing for Erdas et al. [11] was to remove unwanted tissues to eliminate non-brain tissues. For this, we used FSL's Brain Extraction Tool (BET) [17]. The default parameter of BET used a threshold of 0.5 and gradient of 0, but we adjusted these after looking at the images as the default parameters would either extract too much of the brain or not enough. As a final step, we also downsized the images to a size of 92 x 110 x 92 using NiBabel, which is the same downsampled image size used in both cohorts.

The preprocessing performed in Dhinagar et al. [12] was similar to the other cohort, but starts by reordering the images to match the orientation of a standard template image using FSL's fslreorient2std tool. After this, we remove unwanted tissue like the other cohort, but instead of using BET, we use HD-BET [18]. This is a CPU based implementation that did not need manual parameter adjustment like FSL's BET. We then performed non-parametric intensity normalization on the images using ANTs N4 bias field correction [19]. After that, we perform the image registration using FLIRT then downsample the images to a size of 92 x 110 x 92. The number of patients before and after the images were downloaded and the quality control can be seen in the figure below.

Extraction Type	**Total Patients**	**HC**	**PD**	**Sex (Male/Female)**	**Average Age ± Std (years)**
FSL BET	965	189	776	586/379	62.31±10.05
HD-BET	1056	209	847	655/401	62.52±10.13

Table 3: Summary of patients by extraction type.

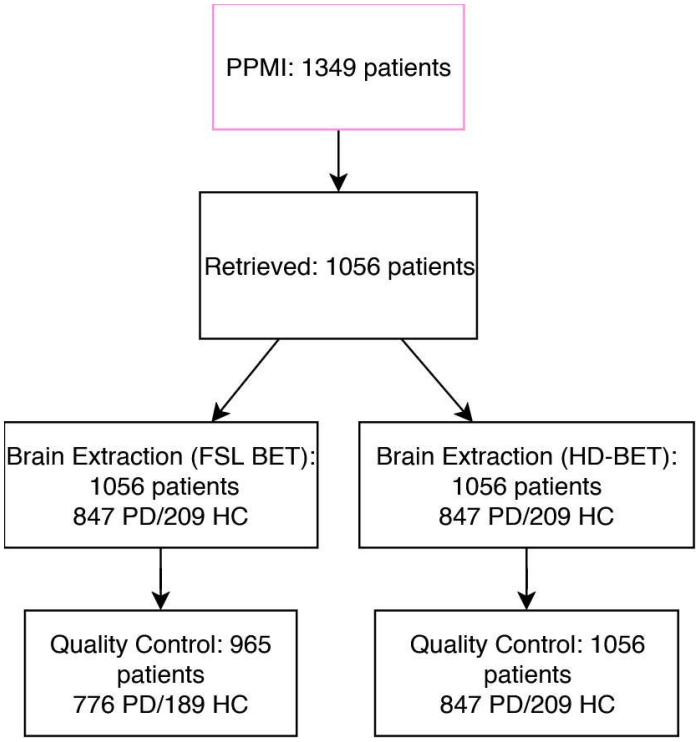


Figure 5: Cohorts created for Erdas et al. [11] (left) and Dhinagar et al. [12] (right)

3.2.1 3D image augmentation

Figure 5 illustrates the two cohorts generated for our experiment. In addition to the standard test set, we introduced a separate test applied to the FSL BET–extracted cohort. The standard test evaluates each model under normal conditions, whereas the additional test incorporates a form of data augmentation applied directly to the images. This augmentation randomly flips 3D volumes along spatial axes and adds subtle Gaussian noise to mimic realistic anatomical and acquisition variability, thereby enhancing model robustness and generalization in 3D medical imaging tasks.

3.3 Convolutional Neural Network Models

In this study, we evaluated three different CNN architectures for two binary classification tasks: (1) Parkinson's disease (PD) versus healthy control (HC), and (2) biological sex classification. Two of the architectures—those from Erdas et al. [11] and Dhinagar et al. [12] were designed specifically for PD classification, whereas the third architecture, proposed by Ebel et al. [13], was originally

developed for sex classification using structural T1-weighted MRIs. Collectively, these models allowed us to examine both the reproducibility of prior PD MRI studies and the general classification capability of CNNs when applied to the PPMI dataset.

All models were implemented in TensorFlow/Keras. Because the original publications did not provide complete source code or full training specifications, we reproduced the architectures and training procedures based on the descriptions available in the papers. In cases where hyperparameters or implementation details were missing, we selected values consistent with the methods typically used in comparable neuroimaging deep-learning studies, ensuring that our replications remained faithful to the intent of the original authors while remaining internally consistent across all experiments.

A notable methodological challenge was that the architectural descriptions in Erdas et al. and Dhinagar et al. were sometimes incomplete or reported at a high level (e.g., "four convolutional modules"), which required careful interpretation and reasonable assumptions. Where possible, we validated these assumptions by examining figure schematics, inferred dimension flows, and consistency with model structures commonly used in 3D CNN literature.

Despite these limitations, all models were trained under a unified experimental framework so that differences in performance could be attributed to architecture and preprocessing rather than to the training setup.

To ensure comparability across experiments, all models were trained using the same set of hyperparameters unless otherwise specified in the original papers. After limited tuning and validation, the following settings were selected:

- **Epochs:** 50 with early stopping
- **Dropout:** 0.2
- **Batch size:** 32
- **Learning rate:** 3×10^{-4} (Adam)
- **Class weights:** computed from the training labels with scikit-learn

These hyperparameters were kept fixed across experiments after light tuning and validation to ensure comparability among the three model implementations.

3.3.1 Model for classifying Parkinson's Disease

The PD classification model used in Erdas et al. [11] and Dhinagar et al. [12] follows a modular 3D CNN architecture. The general structure is shown in Figure 4, where the convolutional layer consists of four modules, followed by a global average pooling layer, and two fully connected layers with a dropout layer in the middle. The main difference between the two models is the size of the filters for the convolutinal layers. The filter size for each of the four filters is 64, 64, 128, 256 filters for the convolutional layers in Dhinagar et al. [12] and 64, 64, 64, 64 filters for the convolutional layers in Erdas et al. [11]. As mentioned earlier, the parameters used for the CNN are consistent throughout every test.

3.3.2 Model for classifying Sex

The sex classification model follows the architecture proposed by Ebel et al. [13]. The model can be seen in Figure 4, the main difference in the model being the input size of the MRIs. To keep the test consistent with the other model, we used the same input parameter of 91 x 110 x 91

3.4 Permutation testing

To evaluate the statistical significance of the performance of our CNN model, we performed permutation tests for each input type and model configuration. Permutation testing is a nonparametric resampling procedure widely recommended in neuroimaging machine-learning studies because it provides an empirical measure of whether a classifier is learning meaningful structure from the data or simply exploiting noise, sampling imbalance, or other spurious correlations. This is especially important for high-dimensional data such as MRI volumes, where models can easily overfit despite apparently high accuracy or AUC.

For each permutation, the class labels in the training set were randomly shuffled while the input MRI images were kept fixed. This destroys any real association between the images and their labels,

but preserves the underlying structure of the covariates (age distribution, scanner noise patterns, brain morphology, etc.). A new model was then trained from scratch on this permuted dataset using the same hyperparameters, architecture, and training schedule as the original experiment. The trained model was evaluated on the same (unpermuted) test set, and the resulting ROC AUC was recorded. The complete implementation is shown in listing 3.1.

This procedure was repeated 150 times for every model–input combination. The same random seed initialization strategy was used across models to ensure that stochastic effects (weight initialization, shuffling, and augmentation) were comparable across experiments. The 150 AUC values produced by the permutations form a null distribution, representing model performance expected if no relationship exists between the MRI data and the class labels.

Unlike the default permutation test implemented in scikit-learn, which trains lightweight estimators and assumes tabular features, our implementation retrains a full 3D CNN for each permutation, preserves all MRI-specific preprocessing and augmentation steps, and computes AUC using TensorFlow's probabilistic outputs. Thus, our test more faithfully captures the true model capacity, stochastic training variability, and imaging-specific correlations inherent in volumetric data.

This nonparametric approach provides a robust estimate of the statistical significance of the predictive performance of the model without assuming any specific distributional form of the data. We will go over the results of each of these models and tests. Each model's code can be found on Github following this link.

Listing 3.1: Permutation test for AUC

```
def permutation_test(X_train, y_train, X_test, y_test, y_test_cat,
                     model_fn, true_auc, num_permutations=150, metric='
                        AUC'):
    results = []

    for i in range(num_permutations):
        print(f"\nPermutation {i+1}/{num_permutations}")

        # Shuffle labels
        rng = np.random.default_rng(seed=SEED + i)
        y_train_shuffled = rng.permutation(y_train)

        # One-hot encode shuffled labels
        y_train_shuffled_cat = tf.keras.utils.to_categorical(
            y_train_shuffled, num_classes=2)
        y_train_shuffled_cat = tf.cast(y_train_shuffled_cat, tf.float32)

        shuffled_train_dataset = tf.data.Dataset.from_tensor_slices(
            (X_train, y_train_shuffled_cat)
        )
        shuffled_train_dataset = shuffled_train_dataset.map(
            augment_3d_image).batch(batch_size)

        model = model_fn()

        model.fit(
            shuffled_train_dataset,
            epochs=20,
            validation_data=val_dataset,  # Use original val set
            verbose=0
        )
```

```
        # Evaluate on the real test set
        preds = model.predict(X_test)
        probs = preds[:, 1]  # class 1 probabilities
        auc = roc_auc_score(y_test, probs)
        print(f"Permutation {i+1} AUC: {auc:.4f}")

        results.append(auc)

    # Compute permutation p-value
    p_value = np.sum(np.array(results) >= true_auc) / len(results)
    print(f"\nPermutation p-value (AUC): {p_value:.4f}")

    return results, p_value
```

Chapter 4

Results

In this chapter, we will focus on the results returned from our multiple tests and experiments. We will first look at the permutation tests performed for each experiment, then follow up with some explainability tests for the model in Dhinagar et al. [12].

We present the permutation test results for each model using different labels, inputs, and augmentations. Each histogram (Figure 6 and 9) illustrates the Receiver Operating Characteristic Area Under the Curve (ROC AUC) values obtained from multiple permutations, along with the corresponding p-value of the model. The blue bars indicate how frequently specific AUC values occurred across the permutations, while the red line marks the observed AUC from the fixed test seed used in every experiment. All histograms are standardized to display the same range of AUCs and frequencies for consistent comparison across models (x-axis: 0.2–0.8; y-axis: 0–30). Each model was evaluated using 150 permutations with an identical test seed to ensure full reproducibility between experiments.

We will identify what our models' predictions are based on. CNNs and other deep learning models in general are considered black box outputs [25], so we have made saliency maps and Grad-CAM maps [26] to help identify and explain CNN decisions. For this approach, we ran both the saliency map and Grad-CAM code through some of our data and selected one image to keep consistent throughout the tests. Because the dataset is different with HD-Bet and with sex labels vs PD labels, we used four different MRIs when the input and labels were changed.

It is important to note that there are limitations to explainability, such as being sensitive to the

model and training data (Adebayo et al. [27]). This means that maps may not always faithfully represent what a model learns or can be inconsistent in highlighting important regions for outliers. Some images in Figures 7 and 10 will be blank as we were keeping the MRIs consistent.

4.1 Sex Classification

Our histogram results from each model can be seen in Figure 6, where the results shown suggest that the models are above chance. Each model type has a set of inputs that are shown to have a high ROC AUC from 0.7 and above with p-values going below 0.01. This shows that the models are capable of classifying sex, and therefore capable of binary classification in general. To support this, we can look at the explainability tests done in Figures 7 and 8. The Grad-CAM images in Figure 7 reveal a recurring emphasis along the outer cortical boundaries. Although the highlighted regions are relatively broad rather than sharply localized, this spatial consistency suggests that the network is learning and utilizing a meaningful pattern for sex classification. For the saliency maps provided in Figure 8, we can see that while the intensity patterns differ from the different models and inputs, the emphasis on the outer edges of the brain are still there, suggesting that the model is making decisions off of the size and structure of the MRIs. Given that the model's ROC AUC performance ins around 0.7-0.8, these results support that the models are capturing stable patterns of the brain for sex classification.

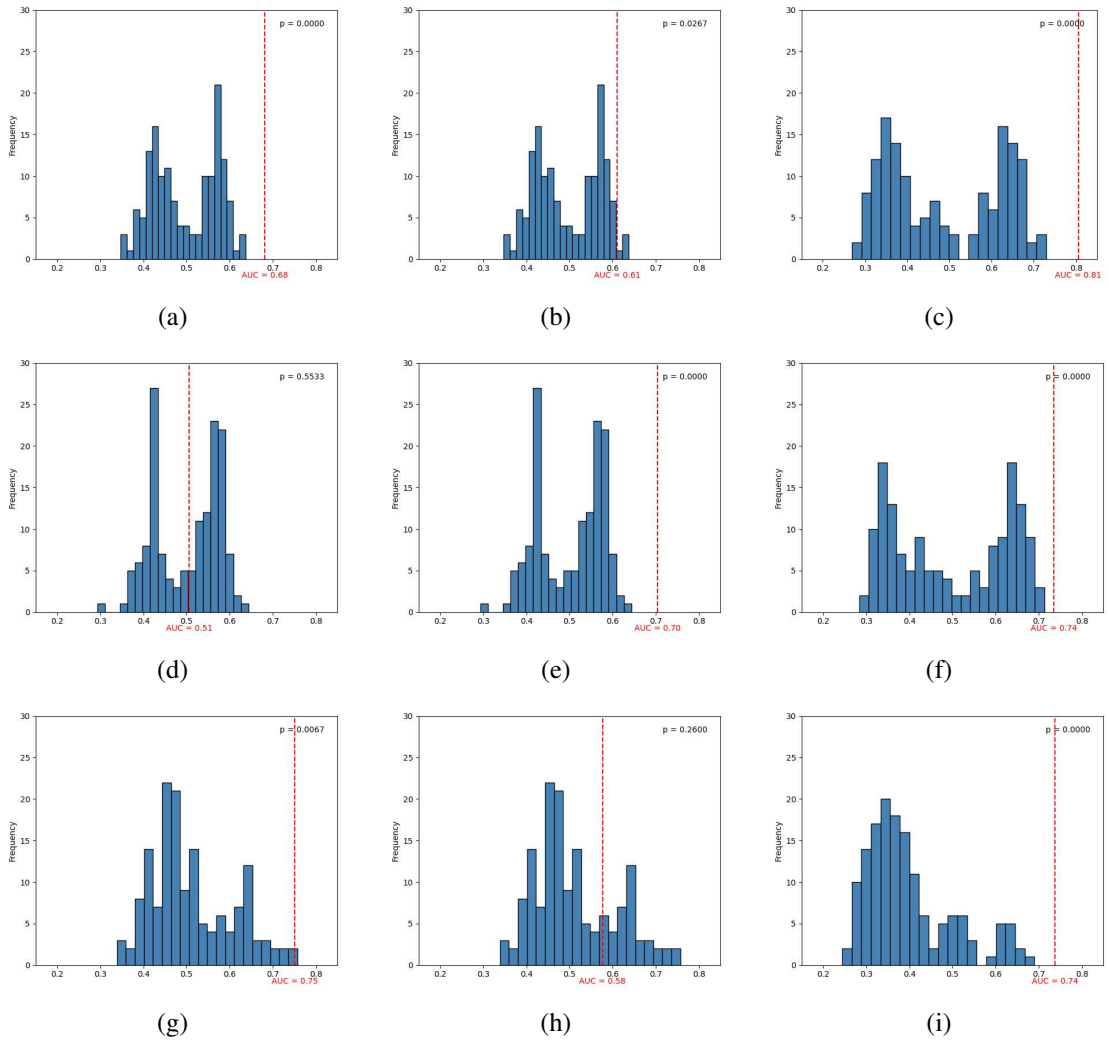
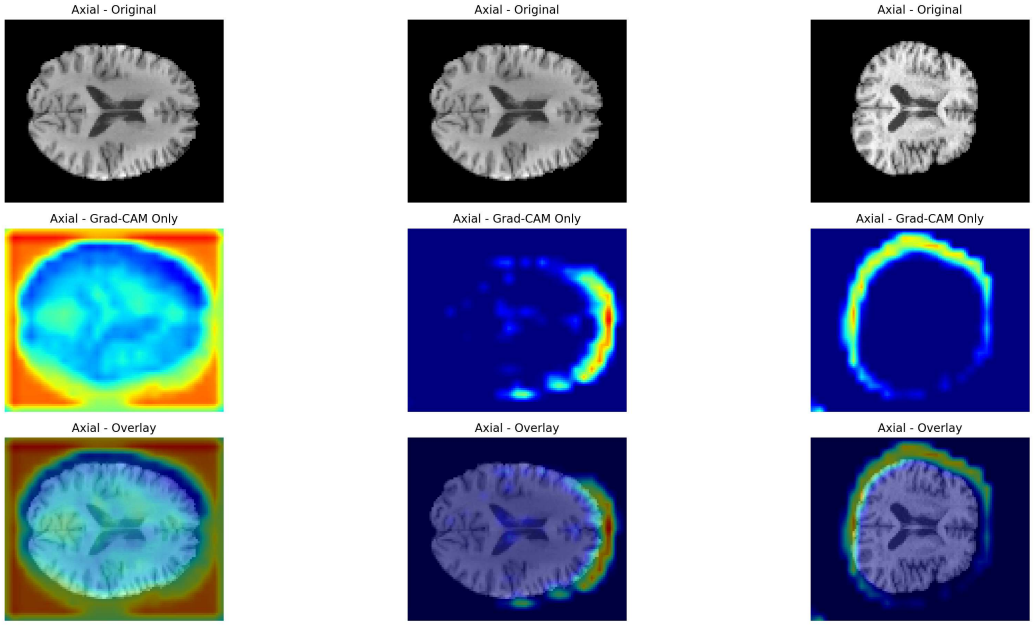
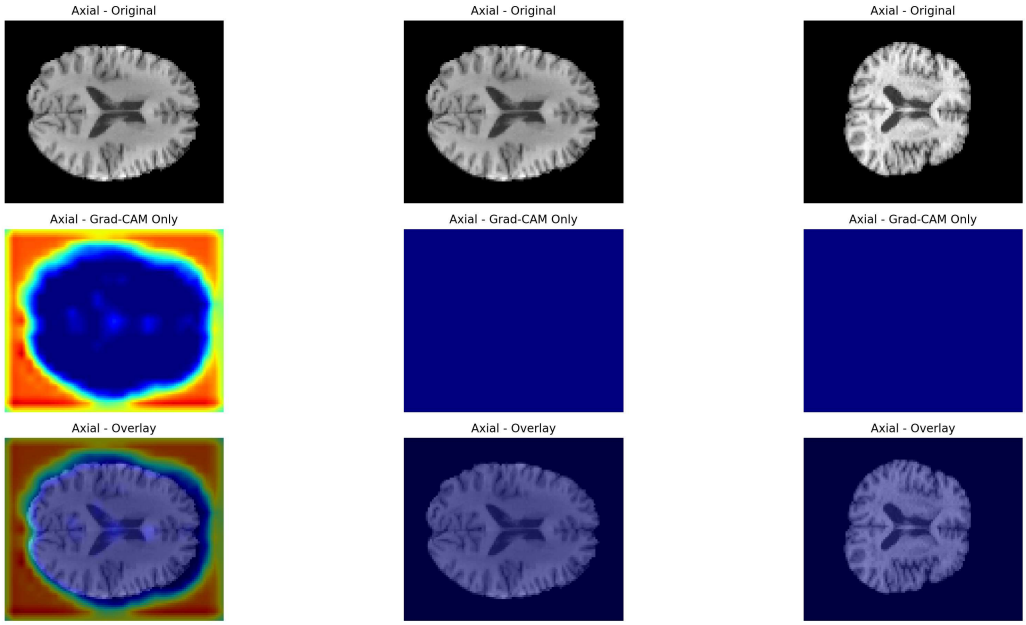


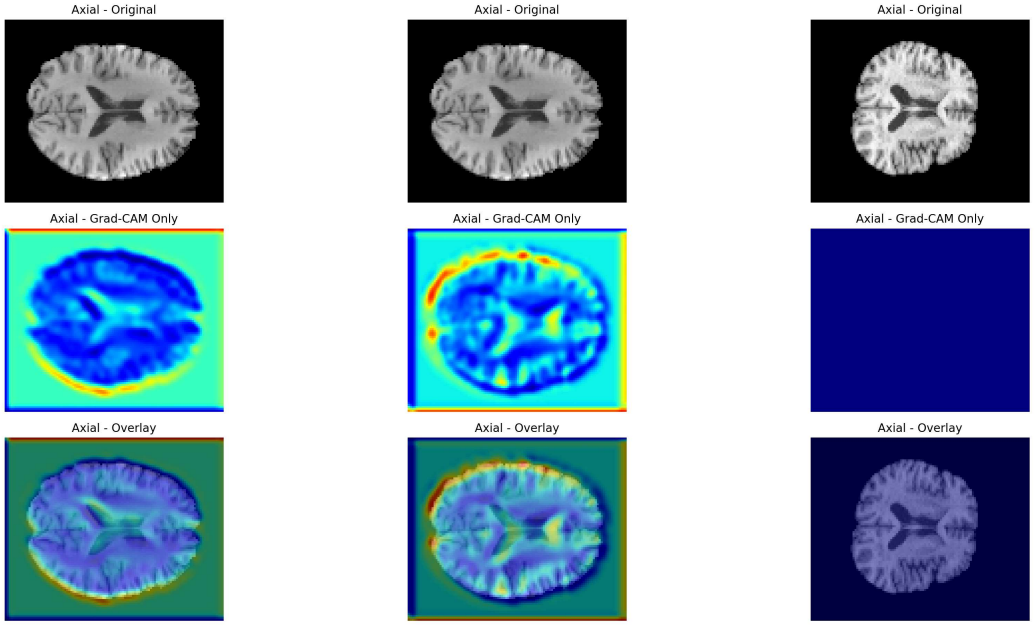
Figure 6: Permutation test results for sex classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Each histogram shows the null distribution of AUC scores with the observed AUC marked in red. Columns: (a ,d ,g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).


Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay

(a) (b) (c)

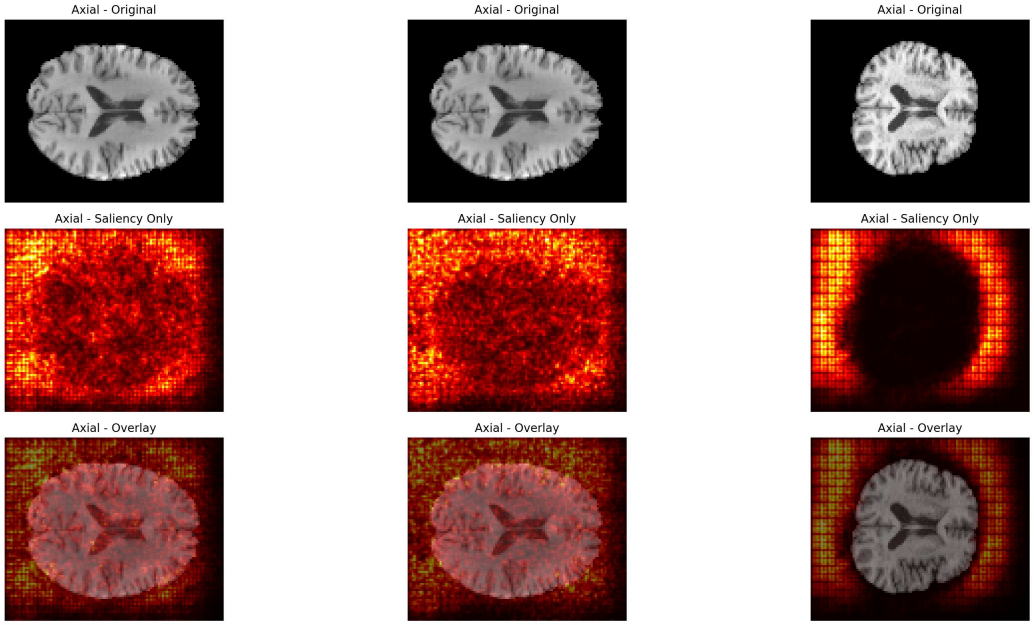

Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay

(d) (e) (f)

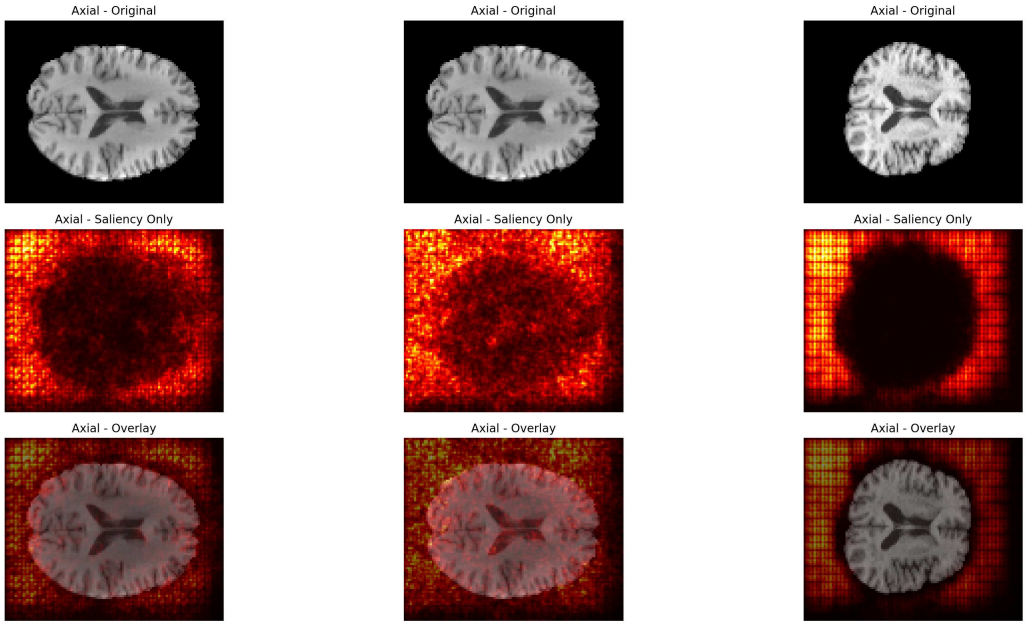


(g) (h) (i)

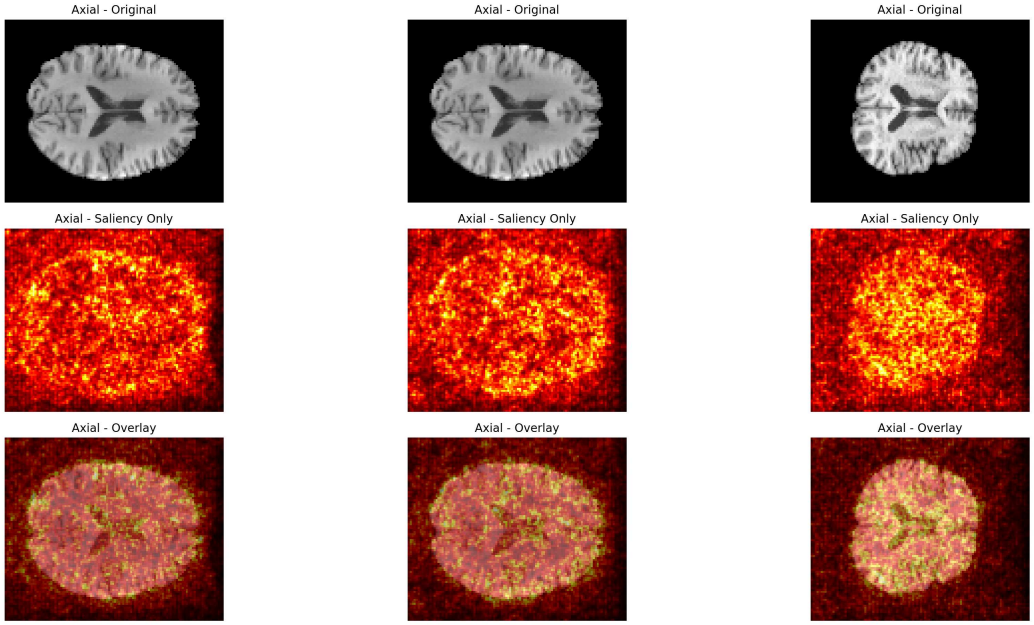
Figure 7: Axial Grad-CAM results for male vs female classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Columns: (a, d, g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).


Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay

(a) (b) (c)


Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay

(d) (e) (f)



(g) (h) (i)

Figure 8: Axial Saliency map results for male vs female classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Columns: (a, d, g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).

4.2 PD vs HC classification

Looking at the histograms presented in Figure 9, histograms (a) to (f) showed a middling observed ROC AUC and high p-value for classifying PD, showing that these models performed as well as randomly guessing, if not slightly better than random guessing. Because these histograms were made with the same fixed seed, it could imply that the observed ROC AUC could be slightly higher or lower with a different seed. These results were expected as Table 2 returns an ROC AUC of 0.667, which would imply that the best test done on that paper was also slightly better than chance. With this in mind, it is possible that their results were not better than chance. Additionally, we performed explainability tests for both models to confirm if our models were randomly guessing. Figures 10 and 11 show that the models are randomly guessing as there is no highlight pattern shown across the images. The Grad-CAM images shown in Figure 10 are extremely variable from the models and inputs, which is not what would be expected from a model that would have learned stable PD features. The saliency maps in Figure 11 are too noisy and cover too many unspecific regions to be finding any patterns. There is very slight recognition of a brain but the noise is too high to be considered an effective pattern recognition.

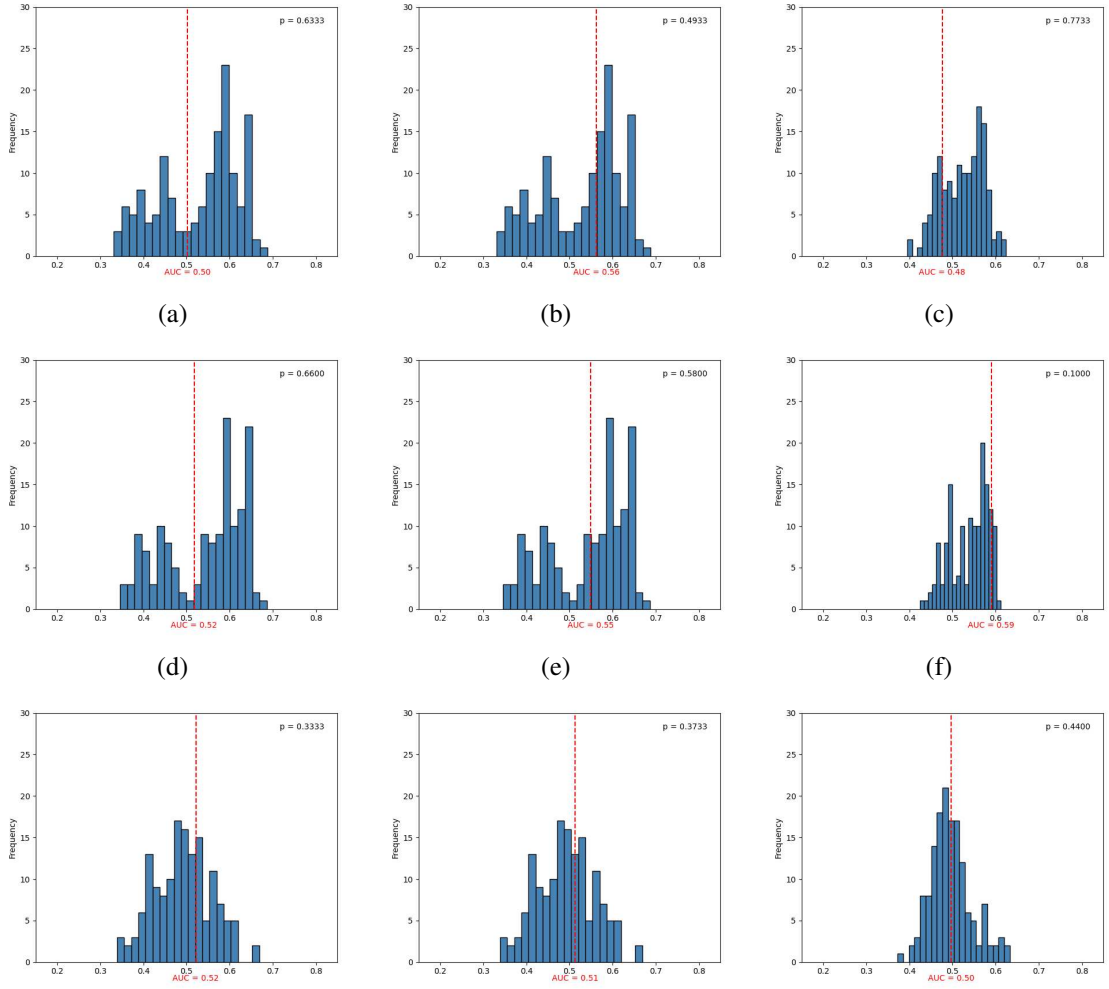
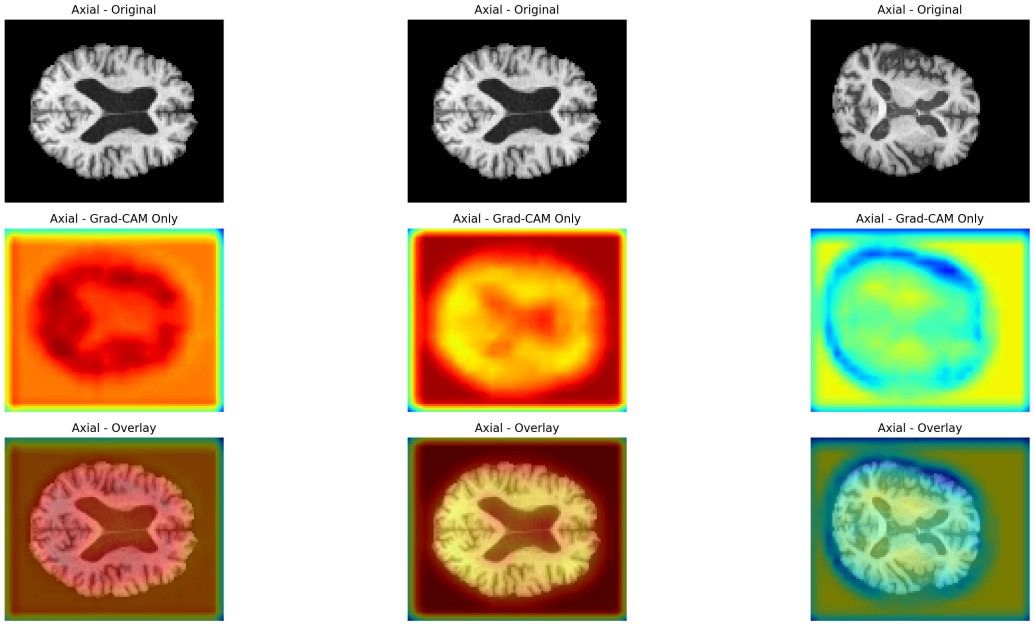
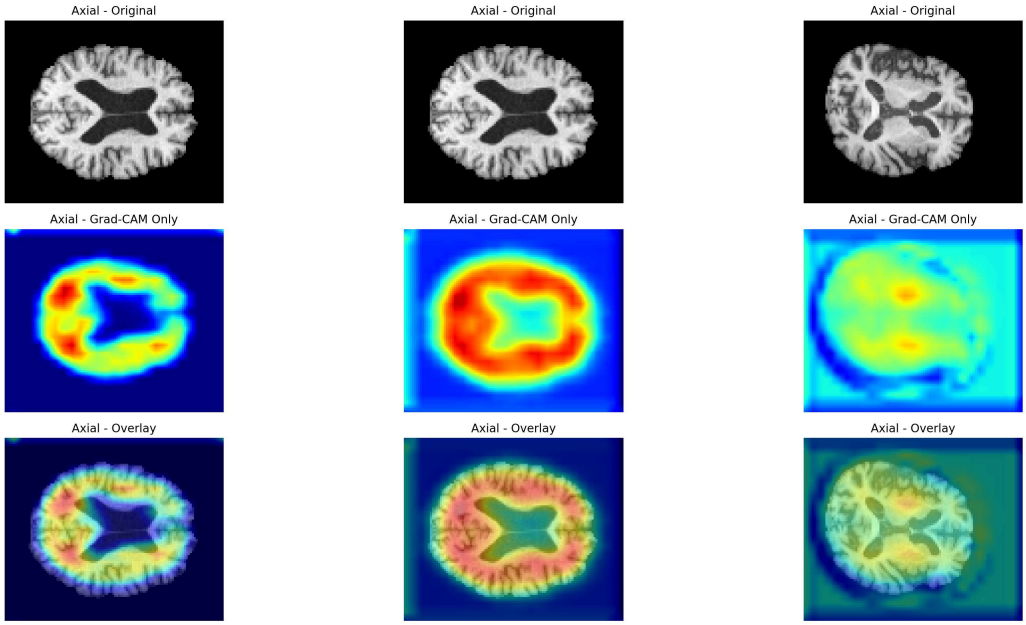


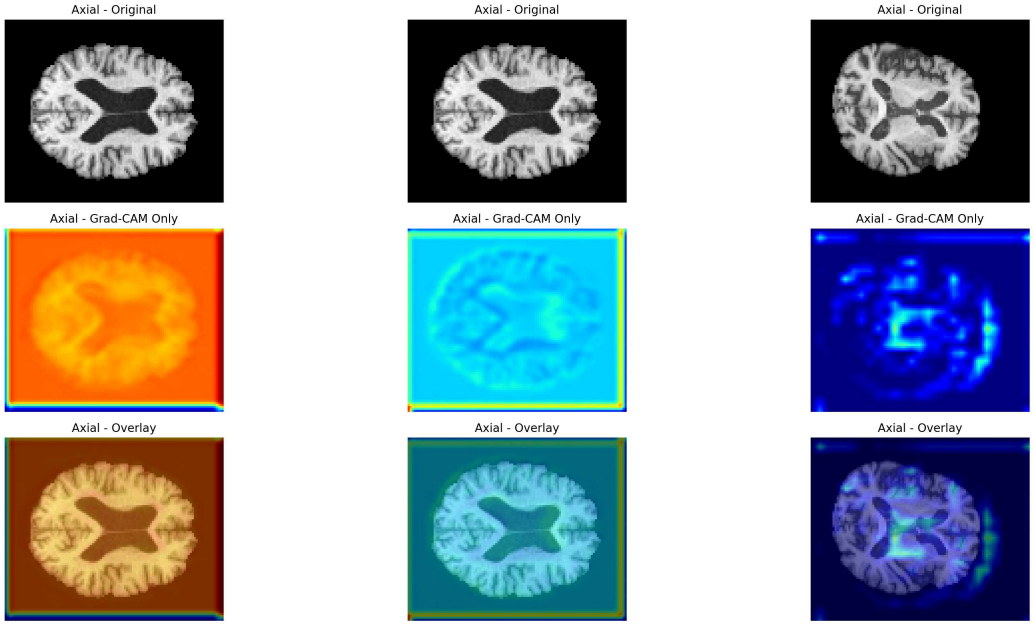
Figure 9: Permutation test results for PD vs HC classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Each histogram shows the null distribution of AUC scores with the observed AUC marked in red. Columns: (a, d, g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).


Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay

(a) (b) (c)

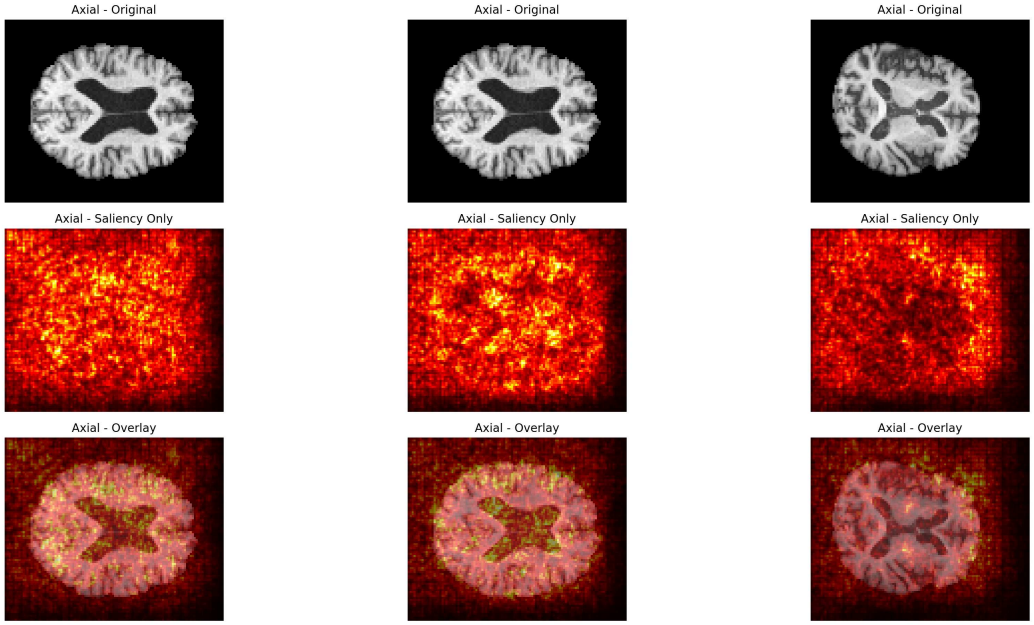

Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay
Axial - Original
Axial - Grad-CAM Only
Axial - Overlay

(d) (e) (f)

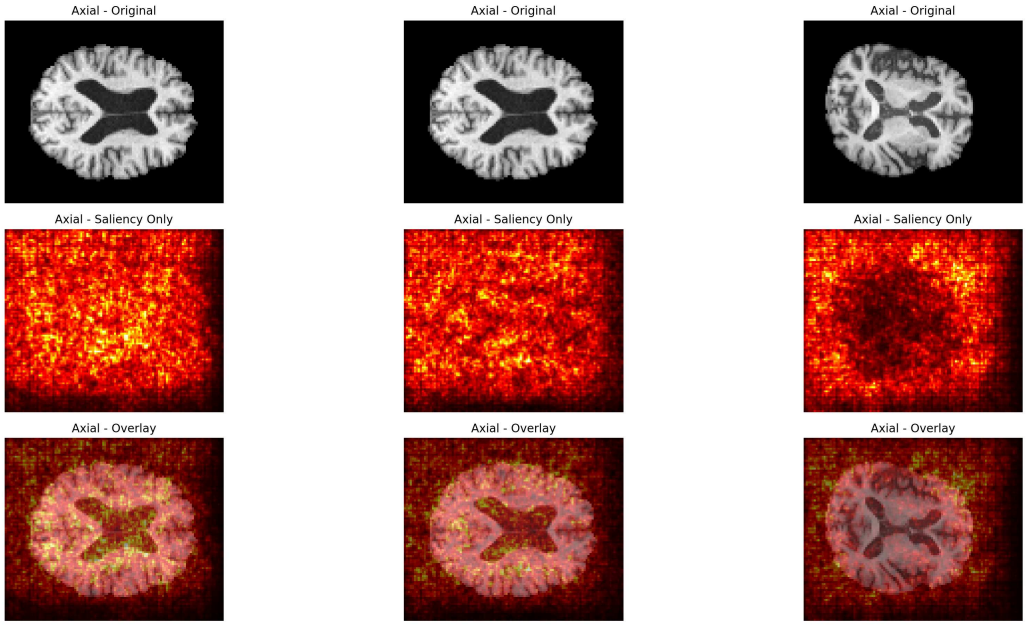


(g) (h) (i)

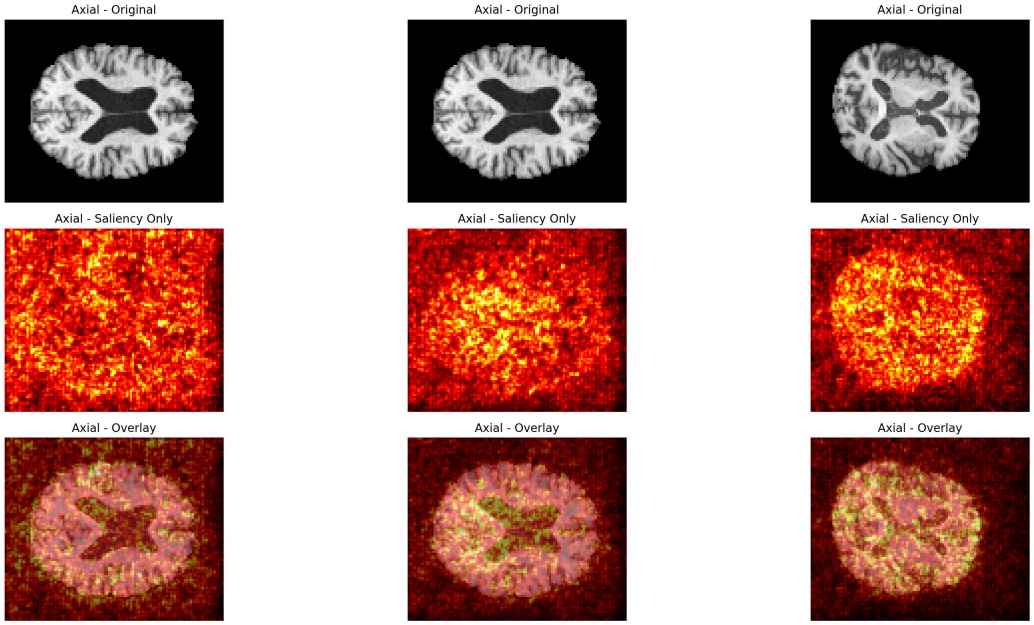
Figure 10: Axial Grad-CAM results for PD vs HC classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Columns: (a, d, g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).


Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay

(a) (b) (c)


Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay
Axial - Original
Axial - Saliency Only
Axial - Overlay

(d) (e) (f)



(g) (h) (i)

Figure 11: Axial Saliency map results for PD vs HC classification models from Dhinagar [12], Erdas [11], and Ebel [13]. Columns: (a, d, g) manual BET; (b, e, h) manual BET with data augmentation; (c, f, i) HD-BET. Rows: authors (Dhinagar, Erdas, Ebel).

Chapter 5

Discussion

We presented all our findings and results for our replication experiments in the last chapter. In this chapter, we discuss our findings and the results that were obtained and what we can conclude based on these results.

In this study, one of our main goals was to compare the two models of Erdas et al. [11] and Dhinagar et al. [12] and to see if we could use them to predict the progression of PD. The models did not perform exceptionally well for classification of PD as they were performing at chance level. To ensure that the models are functional and capable of learning, the second part of the experiment was to do a different classification test. In this case, we wanted to see if these models were capable of classifying sex since we know MRI data is directly correlated to sex [24]. Using the previous models along with the one from Ebel et al. [13], we could see if our models were functional.

In conclusion, our replication experiments showed that the CNN architectures from Dhinagar et al. [12] and Erdas et al. [11] were not sufficient enough for the prediction of PD progression, as their performance for classification hovered around chance level (ROC AUC of 0.667 at best and high p-values) and did not have the consistent pattern detection in the analysis of the explainability maps. Both the Grad-CAM and saliency map visualizations for PD classification showed very variably and noisy activations without any clear structure or pattern, supporting the idea that these models were randomly guessing.

However, using these models for sex classification, all models had a much higher performance,

with ROC AUCs hovering around 0.7-0.8 and significant p-values, usually below 0.01. The corresponding explainability maps revealed a recurring emphasis on the outer boundaries of the brain across the different models and inputs, which suggests that the model was recognizing off of structural differences in the MRIs. These results show that the models are capable of learning patterns for binary classification, although they are not yet suitable for prediction of PD progression. This shows that there is potential for future applications with better targeted training and a more optimized dataset.

Although our replication experiments demonstrated that CNNs can learn meaningful patterns from structural MRI data, the challenge of PD classification remains far from solved. Our results underscore the difficulty of using baseline structural data alone to detect or predict PD progression. Future work should consider incorporating data from different stages of PD, rather than relying solely on baseline scans. Modeling progression dynamically, for example, by comparing early, mid- and late-stage scans or by using longitudinal follow-up data, could help capture subtle morphological or connectivity-based signatures that static datasets fail to reveal.

Another important direction involves expanding beyond structural MRI. Although T1-weighted images capture anatomical features such as cortical thickness and gray matter volume, other modalities could provide complementary information. Diffusion-weighted imaging (DWI) can reveal structural changes in white matter tracts, while functional magnetic resonance imaging (fMRI) can reveal alterations in brain connectivity and activity patterns associated with symptoms and progression of PD. Integrating these modalities through multimodal learning approaches could improve model generalization and sensitivity to disease-relevant signals, potentially overcoming some of the limitations we observed when relying solely on anatomical information.

Our findings emphasize the importance of statistical validation in the study of neuroimaging machine learning. In particular, permutation testing offers a reliable method to assess whether a model's performance truly exceeds chance. This method involves randomizing class labels and recalculating performance metrics to generate a null distribution, providing an empirical estimate of how likely the observed accuracy or AUC could have occurred by random chance. This is especially crucial when dealing with small datasets, high-dimensional inputs, or models with high capacity, all

of which can inflate apparent performance through overfitting. Even models that achieve high metrics (e.g., > 0.8 ROC AUC) should be evaluated against permutation-based baselines to ensure that their predictive power reflects genuine learning rather than noise or sampling bias. Therefore, we recommend that permutation-based validation become a standard component of future neuroimaging machine learning pipelines, particularly when reporting new predictive biomarkers for PD.

Appendix A

Axial, Coronal, and Sagittal Views

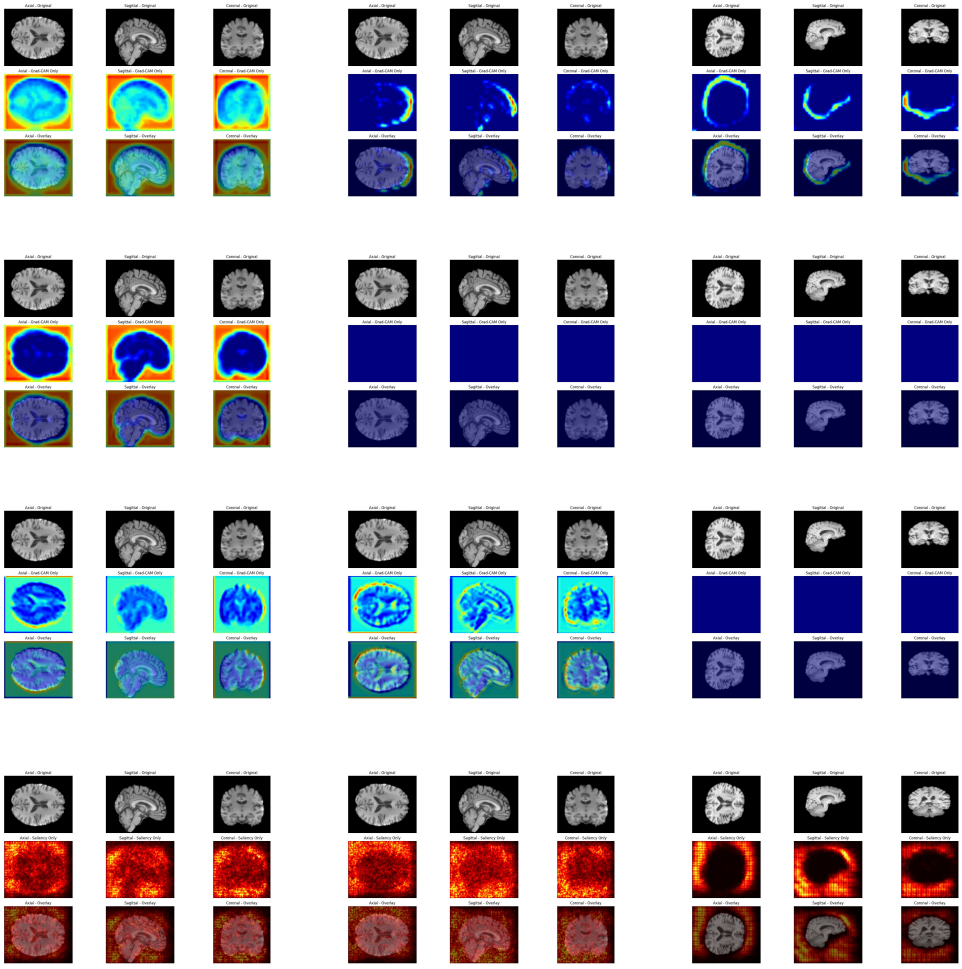


Figure A.1: Grad-CAM and saliency maps across MRI planes (part 1).

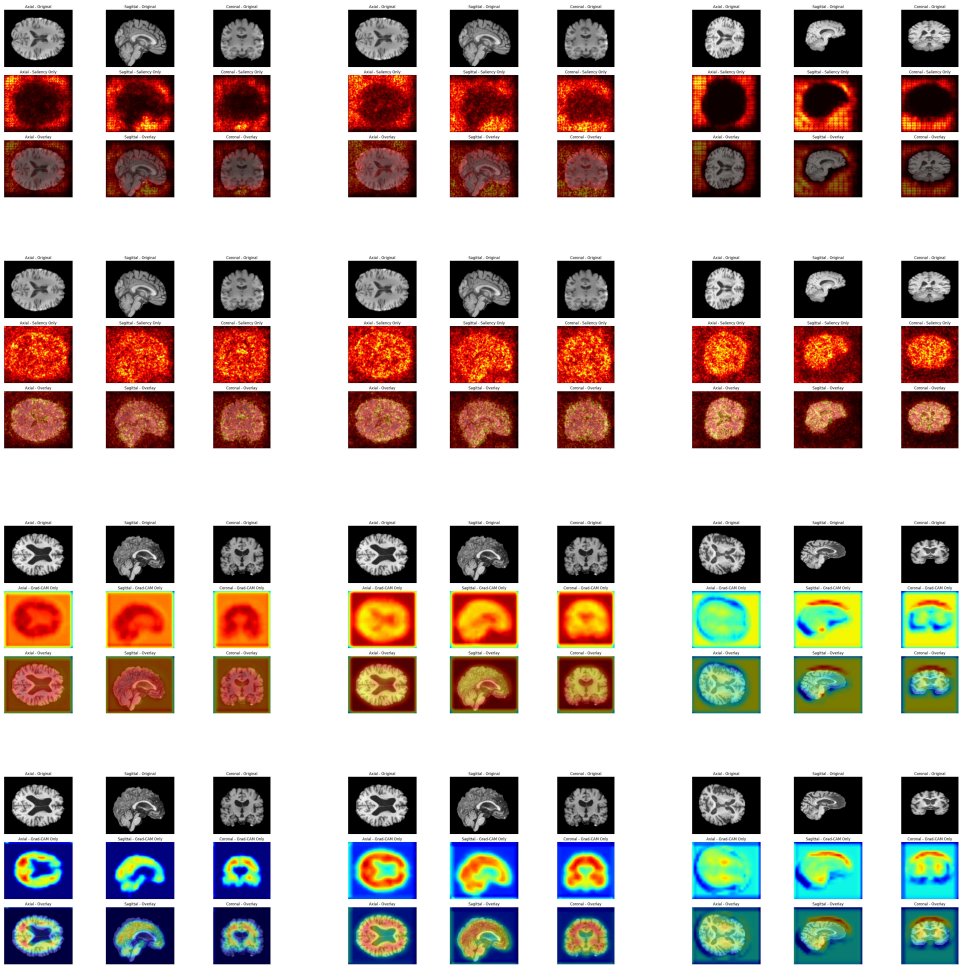


Figure A.1: Grad-CAM and saliency maps across MRI planes (part 2).

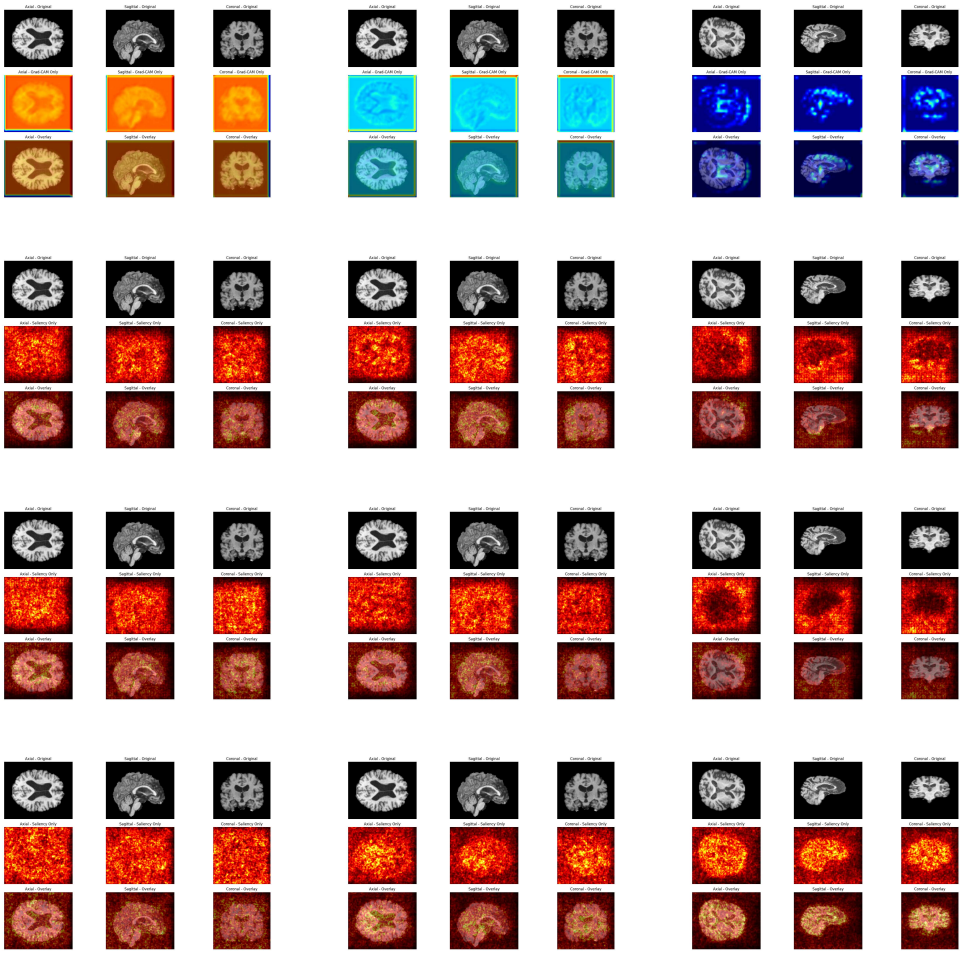


Figure A.1: Grad-CAM and saliency maps across MRI planes (part 3).

References

(1) E. R. Dorsey et al., “Global, regional, and national burden of Parkinson’s disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016,” The Lancet Neurology, vol. 17, no. 11, pp. 939–953, Nov. 2018, doi: 10.1016/S1474-4422(18)30295-3.

(2) A. Govindu and S. Palwe, “Early detection of Parkinson’s disease using machine learning,” Procedia Computer Science, vol. 218, pp. 249–261, Jan. 2023, doi: 10.1016/j.procs.2023.01.007.

(3) V. Kaul, S. Enslin, and S. A. Gross, “History of artificial intelligence in medicine,” Gastrointestinal Endoscopy, vol. 92, no. 4, pp. 807–812, Oct. 2020, doi: 10.1016/j.gie.2020.06.040.

(4) R. Hirani et al., “Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities,” Life (Basel), vol. 14, no. 5, p. 557, Apr. 2024, doi: 10.3390/life14050557.

(5) D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, “Watson: Beyond Jeopardy!,” Artificial Intelligence, vol. 199–200, pp. 93–105, Jun. 2013, doi: 10.1016/j.artint.2012.06.009.

(6) N. Bakkar et al., “Artificial intelligence in neurodegenerative disease research: use of IBM Watson to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis,” Acta Neuropathol, vol. 135, no. 2, pp. 227–247, Feb. 2018, doi: 10.1007/s00401-017-1785-8.

(7) J. Mei, C. Desrosiers, and J. Frasnelli, “Machine Learning for the Diagnosis of Parkinson’s Disease: A Review of Literature,” Front. Aging Neurosci., vol. 13, May 2021, doi: 10.3389/fnagi.2021.633752.

(8) S. S. Kshatri and D. Singh, “Convolutional Neural Network in Medical Image Analysis: A Review,” Arch Computat Methods Eng, vol. 30, no. 4, pp. 2793–2810, May 2023, doi: 10.1007/s11831-023-09898-w.

(9) A. Das, G. R. Patra, and M. N. Mohanty, “LSTM based Odia Handwritten Numeral Recognition,” in 2020 International Conference on Communication and Signal Processing (ICCSP), Jul. 2020, pp. 0538–0541. doi: 10.1109/ICCSP48568.2020.9182218.

(10) S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, “Medical Image Analysis using Convolutional Neural Networks: A Review,” J Med Syst, vol. 42, no. 11, pp. 1–13, Nov. 2018, doi: 10.1007/s10916-018-1088-1.

(11) Ç. B. Erdaş and E. Sümer, “A fully automated approach involving neuroimaging and deep learning for Parkinson’s disease detection and severity prediction,” PeerJ Computer Science, vol. 9, p. e1485, Jul. 2023, doi: 10.7717/peerj-cs.1485.

(12) N. J. Dhinagar et al., “3D convolutional neural networks for classification of Alzheimer’s and Parkinson’s disease with T1-weighted brain MRI,” in 17th International Symposium on Medical Information Processing and Analysis, SPIE, Dec. 2021, pp. 277–286. doi: 10.1117/12.2606297.

(13) M. Ebel, M. Lotze, M. Domin, N. Neumann, and M. Stanke, “Classifying sex with MRI,” Apr. 28, 2022. doi: 10.1101/2022.04.27.22274355.

(14) V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, “Unbiased average age-appropriate atlases for pediatric studies,” NeuroImage, vol. 54, no. 1, pp. 313–327, Jan. 2011, doi: 10.1016/j.neuroimage.2010.07.033.

(15) J. Mazziotta et al., “A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM),” Philos Trans R Soc Lond B Biol Sci, vol. 356, no. 1412, pp. 1293–1322, Aug. 2001, doi: 10.1098/rstb.2001.0915.

(16) S. M. Smith et al., “Advances in functional and structural MR image analysis and implementation as FSL,” Neuroimage, vol. 23 Suppl 1, pp. S208-219, 2004, doi: 10.1016/j.neuroimage.2004.07.051.

(17) S. M. Smith, “Fast robust automated brain extraction,” Hum Brain Mapp, vol. 17, no. 3, pp. 143–155, Nov. 2002, doi: 10.1002/hbm.10062.

(18) F. Isensee et al., “Automated brain extraction of multisequence MRI using artificial neural networks,” Human Brain Mapping, vol. 40, no. 17, pp. 4952–4964, 2019, doi: 10.1002/hbm.24750.

(19) N. J. Tustison et al., “N4ITK: Improved N3 Bias Correction,” IEEE Transactions on Medical Imaging, vol. 29, no. 6, pp. 1310–1320, Jun. 2010, doi: 10.1109/TMI.2010.2046908.

(20) K. Canese and S. Weis, “PubMed: The Bibliographic Database”.

(21) K. Marek et al., “The Parkinson’s progression markers initiative (PPMI) – establishing a PD biomarker cohort,” Annals of Clinical and Translational Neurology, vol. 5, no. 12, pp. 1460–1477, 2018, doi: 10.1002/acn3.644.

(22) M. A. Nalls et al., “Baseline genetic associations in the Parkinson’s Progression Markers Initiative (PPMI),” Movement Disorders, vol. 31, no. 1, pp. 79–85, 2016, doi: 10.1002/mds.26374.

(23) J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in Proceedings of the 23rd international conference on Machine learning - ICML ’06, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

(24) A. N. V. Ruigrok et al., “A meta-analysis of sex differences in human brain structure,” Neuroscience & Biobehavioral Reviews, vol. 39, pp. 34–50, Feb. 2014, doi: 10.1016/j.neubiorev.2013.12.004.

(25) S. Konate et al., “A Comparison of Saliency Methods for Deep Learning Explainability,” in 2021 Digital Image Computing: Techniques and Applications (DICTA), Nov. 2021, pp. 01–08. doi: 10.1109/DICTA52665.2021.9647419.

(26) R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

(27) J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2018.

(28) S. Bhat, U. R. Acharya, Y. Hagiwara, N. Dadmehr, and H. Adeli, "Parkinson's disease: Cause factors, measurable indicators, and early diagnosis," Computers in Biology and Medicine, vol. 102, pp. 234–241, Nov. 2018, doi: 10.1016/j.compbiomed.2018.09.008.

(29) T. Mahmood, A. Rehman, T. Saba, L. Nadeem, and S. A. O. Bahaj, "Recent Advancements and Future Prospects in Active Deep Learning for Medical Image Segmentation and Classification," IEEE Access, vol. 11, pp. 113623–113652, 2023, doi: 10.1109/ACCESS.2023.3313977.

(30) M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia," Brain Inf., vol. 7, no. 1, p. 11, Oct. 2020, doi: 10.1186/s40708-020-00112-2.