Digit. Signal Process. 147 (2024) 104407

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp

An analysis of data leakage and generalizability in MRI based classification of Parkinson's Disease using explainable 2D Convolutional Neural Networks

Iswarya Kannoth Veetil [a], Divi Eswar Chowdary [a], Paleti Nikhil Chowdary [a], Sowmya V. [a], E.A. Gopalakrishnan [b,c,*]

[a] Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[b] Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India
[c] Amrita School of Artificial Intelligence, Bengaluru, Amrita Vishwa Vidyapeetham, India

ARTICLE INFO

Dataset link: www.ppmi-info.org/access-dataspecimens/download-data

Dataset link: www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors

Dataset link: fcon_1000.projects.nitrc.org/indi/retro/parkinsons.html

Keywords:
Parkinson's Disease
Magnetic Resonance Imaging (MRI)
Convolutional Neural Network (CNN)
Explainable Artificial Intelligence (XAI)
Data leakage
Generalizability

ABSTRACT

Background and Objective: Parkinson's Disease (PD) is a progressive neurological disorder caused by the death of dopamine producing neurons. Neuroimaging techniques such as Magnetic Resonance Imaging (MRI) allows the visualization of the structural changes in the brain due to PD. Advances in computer vision has led to a new area of research that combines the expertise of deep learning (DL) tools such as Convolutional Neural Networks (CNN) to detect PD from MRI. Despite the promising results obtained, the clinical integration of the DL models is held back by questions of bias, generalizability and explainability.
Methods: In the present work the identification of bias propagation is carried out through an analysis of data leakage and generalizability of T1 weighted MRI data driven CNN models. For the same, 12 diverse pre-trained CNN models were trained on T1 weighted MRI from the PPMI dataset. Of these, the top 3 models were tested on three different datasets under three simulated cases of data leakage - Subject-wise split, slice-wise split and longitudinal split. A Grad-CAM based visualization was implemented to visualize and explain the output from the CNN without data leakage, and identify regions of importance (ROI) in the brain.
Results: Results from the data leakage simulation revealed that slice level data leakage and longitudinal data leakage can result in over 67% and 30% inflation of accuracy score in hold out test sets. Testing the generalizability of the CNN models to external patient cohorts was able to capture the implicit bias due to data leakage and enable the selection of the most robust CNN architecture. The VGG19 model displayed a consistent performance when tested within the PPMI dataset and the external datasets. The results from the explainable artificial intelligence analysis revealed the identified ROIs were significant with the expected disease progression, validating the proposed method.
Conclusions: The study presents the possible avenues of bias propagation in the MRI data driven classification using CNN models through a simulation of data leakage and by testing the generalizability of the models. The study highlights the need for generalizability and the importance of the testing with heterogeneous populations in ensuring the robustness of the developed models, and in capturing any data mishandling oversights and associated bias. The results suggest that the pre-trained VGG19 model can be used to create a generalizable and explainable model to aid in the detection of PD from T1 weighted MRI.

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disease that has affected over 6.1 million people worldwide [33]. PD is characterized by the degeneration of dopamine producing neurons in the Substantia Nigra region of the mid-brain [90]. Clinically, the disease is diagnosed by its cardinal motor symptoms which include bradykinesia (slowness of movement), muscle rigidity, and rest tremors. In advanced

* Corresponding author at: Amrita School of Artificial Intelligence & Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India.
E-mail addresses: kv_iswarya@cb.students.amrita.edu (I.K. Veetil), eswar.divi.902@gmail.com (D.E. Chowdary), nikhil.paleti@outlook.com (P.N. Chowdary), v_sowmya@cb.amrita.edu (V. Sowmya), ea_gopalakrishnan@blr.amrita.edu (E.A. Gopalakrishnan).

https://doi.org/10.1016/j.dsp.2024.104407

Available online 2 February 2024

stages of the disease, balance issues and postural instability are also seen [96]. The diagnosis of the disease is carried out by a physician based on the above symptoms using a clinical evaluation criteria such as Movement Disorder Society-Sponsored Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [47]. The final confirmation of the diagnosis is achieved through nuclear neuroimaging - DaTScan (Dopamine Transporter Scan). This allows the diagnostician to accurately determine the extent of dopaminergic neuronal degeneration [90]. However, studies show that while the common age of diagnosis of PD based on physical symptoms is from 50 to 60 years of age [128], the person may have had the disease up to 20 years prior to diagnosis [39]. This long prodromal period and variable path of progression means that by the time the motor symptoms of the disease are seen, up to 80% of the dopamine producing neurons could be degenerated [61]. Therefore, there is a pressing need to identify accurate precursors for the early diagnosis of the disease.

The neuroimaging statistics from hospitals in the last quarter show that MRI was taken more than six times the combined total of the nuclear medicine, Positron Emission tomography (PET) and Single-photon emission computed tomography (SPECT) scans combined [88]. With the increased use of MRI in routine neuroimaging diagnostic procedures and the lack of accessibility and high cost of nuclear imaging, a significant part of ongoing research has focused on developing MRI based PD detection tools [107]. MRI display the soft tissue in the brain with high spatial resolution and aid in accurately mapping the structural changes with PD [139]. Identifying the subtle changes using visual discrimination can be quite challenging [19], hence researchers have focused on designing diagnostic decision support systems using artificial intelligence (AI) based tools for the detection of PD [2,5,99,134].

Machine learning (ML) classifiers such as Support Vector Machine [83,119,93,108], Random Forest [122], and Bayesian classifiers [83, 122] have been used by researchers to learn underlying differences in the MRI of normal cohorts (NC) and people with PD for classification. These ML algorithms, however, require strategies for feature extraction. Some research studies focused on using voxel based morphometry (VBM) to extract different features of NC and PD brains such as cortical thickness, gray matter volume, white matter volume, and cortical thickness, [1,83,93,108,122] as features to be given to the ML classifiers. Some other studies focused on using statistical features [148] and graph-based metrics [5]. Further, most studies include an additional feature selection step after feature extraction to reduce the dimensionality of the feature space. These strategies include methods such as the Kohonen Self Organizing Map [119], Joint Feature Sample Selection algorithm [1,2], Radial Basis Function [93], and Principal Component Analysis [108,122]. The machine learning studies are heavily dependent on accurate feature crafting and feature selection which would require a high degree of domain expertise [101,38].

The need for more automation and in a step towards the direction of reducing the amount of explicit feature definition and selection, Convolutional Neural Networks (CNN) were used by researchers to detect PD from MRI in a data-driven manner [5,120,94,133,99]. Previous studies using CNN for the classification of MRI in PD are diverse in their approaches from CNN architecture, MRI input selection, validation procedures and testing. A summary of the previous studies is presented in Table 1 and detailed in Section 2. An analysis of previous research revealed the following gaps in the studies. Most of the studies employed relatively small datasets and hence resorted to using transfer learning (TL) as a strategy to train the CNN models, however the choice of the architectures and its validity in the target domain is not provided by most researchers. Studies also presented a varied choice of MRI selection as input to the CNN models such as using almost all the MRI slice - with or without data augmentation, subset of slices selected using strategies based on the information in the MRI slices, segmented portion of the MRI relevant to PD, or specific feature extracted brain maps. With the aim of developing a first level, diagnostic decision support tool, this wide variation leaves a lot of room for interpretation and some bias in studies where the region of interest (ROI) specific to PD alone is considered. Further, to the knowledge of the authors, there are no published works that compare the slice selection choices listed in the studies. A study of multiple slice selection criteria would allow future researchers in objectively choosing a method.

The studies on MRI data driven approaches paint a very promising picture as a hybrid diagnostic decision support tool. However, wide spread adoption of data driven tools is held back by concerns of bias, a lack of generalizability and non-explainability of the model's predictions [9,32,3,82]. Potential data leakage and over-optimistic results were identified in previous studies. Improper data handling protocols and the effect it has on data driven systems has been analyzed by researchers [106,22,144,143,77], who report above 30% increase in accuracy values due to data leakage. The effect of slice level data leakage with a hold-out independent test set, and the longitudinal data leakage has not been presented in any studies so far. Additionally, most studies in MRI based classification of PD did not present the generalizability of the model to external patient populations. Generalizability provides trustworthiness to the developed models by showcasing the performance on completely heterogeneous data, with characteristics different from what was used for training.

This research aims to address the gaps identified by developing a generalizable and explainable CNN model for the classification of PD from T1-weighted MRI using a TL strategy. The specific goals of the study are detailed below:

1. Investigate the performance of 12 diverse and popular CNN models to identify which architecture is better suited to capture the characteristics of T1-weighted MRI for the task of PD classification.
2. Explore two empirical slice selection strategies and determine a common range of informative slices to be used as input to the CNN.
3. Simulate the effect of subject-level (no data leakage), slice-level and longitudinal data leakage and study their effect on the performance of the developed CNN models using an independent holdout test set.
4. Evaluate the generalization capability of the CNN models (within the context of data leakage) to two external datasets.
5. Identify the regions of importance (ROI) through an explainable AI (XAI) interpretation of the output predictions by the CNN.

The rest of the paper is organized as follows: previous work related to MRI based classification of PD using CNN is presented in section 2. The design of the experiment is detailed in Section 3. Section 4 describes the results of the experiments and its analyses. Section 5 details the conclusions drawn from this research and the future scope of this work.

2. Related work

Shah et al. [114] used a custom CNN to classify a 100x100 segmentation of the Substantia Nigra region of the mid-brain. The authors selected slice 22 from T2 weighted MRI from the PPMI dataset as input to the CNN. Sivaranjini and Sujatha [120] explore the classification of T2-weighted MRI from the PPMI dataset using a transfer learned AlexNet model. The authors used feature maps of the last convolutional layers to show distinct differences in the two classes of data (NC and PD). However, the study does not list a validation strategy during training and the splitting of the dataset at the slice level leads to a possibility of data leakage. Esmaeilzadeh et al. [36] developed a custom architecture that used age and gender as inputs along with MR scans from the PPMI datasets to predict PD. The study used data augmentation strategies of flipping the right and left hemispheres to double the amount of training data. The authors also conducted a sensitivity analysis by occlusion to reveal that the regions of the basal ganglia, Substantia Nigra and the superior parietal regions of the right hemisphere are important.

Yagis et al. [144] used ResNet50 and VGG16 architectures to investigate the effect of slice level data leakage in T2 weighted MRI from the PPMI dataset. The authors used a entropy based slice selection strategy in their study and report up to 26% inflation of the accuracy value due to wrong data selection. An extension of this work to T1 weighted MRI from the PPMI and Versilia datasets was presented in [143]. The authors consider the effect of slice level data leakage during cross validation (CV) and report 48% and 55% inflation of accuracy values on the PPMI and Versilia datasets, respectively. Chakraborty et al. [23] implemented the classification of whole MRI volumes through a custom CNN architecture trained on T1 weighted MRI. Using a gradient-weighted class activation mapping (Grad-CAM) visualization, the authors identify the Substantia Nigra as a region of importance (ROI) for PD. West et al. [142] used T1 weighted MRI from the PPMI and IXI datasets to compare the performance of four CNNs and convolutional auto-encoders(AE). The authors presented the results at a slice level classification with a majority voting scheme and used occlusion based sensitivity analysis to identify the Cerebellum and Occipital lobes as ROI.

Veetil et al. [133] evaluated five pre-trained CNN architectures - VGG16, VGG19, ResNet50, Xception and DenseNet201 on T2 weighted MRI from the PPMI dataset. The authors reported that the VGG19 was most suited for the task of classifying PD using T2 weighted MRI. Classification based on three sub-division within an MRI - the whole brain, smoothed gray and white matter areas, and un-smoothed gray and white matter regions was carried out by Mostafa and Cheng [84]. The authors used T1 weighted MRI from the PPMI and IXI databases and reported that the best performance was obtained using smoothed scans. An occlusion based analysis was performed to identify that the important regions in the classification of PD as the gray matter region of the superior frontal gyrus and white matter region of the postcentral gyrus.

Madan et al. [77] investigated the effect of slice level data leakage on T2 weighted MRI slices from the PPMI database. The authors used an intensity based slice selection strategy and reported a 30% inflation of accuracy values upon wrongly splitting the dataset at the slice level. Yin et al. [147] used T2 weighted MRI in the PD class from the PPMI database and NC MRI from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to classify PD using a custom CNN with Gabor filter convolutional layers. The CNN was trained on the segmentation of the mid-brain region of the MRI and clinical data and reported a 20% increase in accuracy with the custom Gabor filters layers and heterogeneous data. Madan et al. [78] used Variational AE to augment T2 weighted MRI from the PPMI dataset before using it to train a DenseNet201 CNN model. The authors report a 6% increase in accuracy upon augmenting the data using AE. Vyas et al. [135] tested Flair and T2 weighted MRI from the PPMI database and report that the 3D CNN outperforms the 2D CNNs.

Rajanbabu et al. [99] proposed the usage of an ensemble of VGG16, VGG19 and DenseNet201 architectures to classify T2 weighted MRI from the PPMI dataset. The authors used a data augmentation strategy to increase the amount of data available and report up to 10% increase in accuracy due to the usage of an ensemble of CNN models instead of individual architectures. Bhan et al. [13] used a CNN model to classify whether segmented T1 MRI from the PPMI database contained the Substantia Nigra region or not. The segregated scans with the Nigral region was then used to train a second CNN model. The study compared four CNN architectures and reported the highest performance by a transfer learned ResNet50 architecture.

Yang et al. [146] used a ResNet model on 3D volumes of T1 weighted MRI from the PPMI dataset. The authors examined the results of the classification using Grad-CAM and report that the Frontal lobe plays a crucial role. The study verifies the role of the frontal role by correlating the heat maps with the MDS-UPDRS scores of both the classes. Dhinagar et al. [30] trained a random forest (RF) classifier and a custom CNN model on radiomics data and T1 weighted MRI, respectively, from the PPMI dataset. The study demonstrated that the CNN outperformed the ML classifier when tested within the PPMI dataset and on an independent UPenn dataset.

Camacho et al. [19] used T1 weighted MRI from 13 different databases in their work. The study used diffeomorphic transformation to generate Jacobian maps that was used as the input to a custom CNN and evaluated the models on an independent dataset as well. The authors reported several gray matter structures, orbital-frontal regions, and fronto-temporal regions as the ROI through saliency analysis. An improvement of the classification accuracy of T2-weighted MR images from PPMI using median filtering for noise removal during preprocessing and a custom CNN for classification was demonstrated by Sangeetha et al. [109].

A summary of the key points from the background study of previous works is presented in Table 1. A critical analysis of the previous studies that have used CNN for the classification of PD using MRI reveals the following points. An almost equal number of studies considered T1 weighted MRI [143,23,142,84,13,146,30,19] and T2 weighted MRI [114,120,144,133,147,77,78,99,109] as the modality of choice, with [135] considering Flair and T2 weighted MRI, and [36] not having specified the modality used in the study. More than half the studies used transfer learning (TL) as the method of choice for the CNN. Among the architectures used, the most popular choice included variations of the ResNet architectures [144,143,142,84,13,146,30], the VGG architectures [144,143,133,77,99,13], the DenseNet architectures [133,84,77,78,99] and AlexNet [120,13]. A few other CNN models were also considered by other studies. Since using pre-trained CNN architectures help with faster convergence and data scarcity, TL is considered as the method of choice in this study. The diversity of the CNN architectures, each with its own pros and cons leads to the first *research question* of which architecture is most suited for the task of classifying PD using T1 weighted MRI. An investigation of 12 such diverse architectures is undertaken in this study.

Different types of CNN inputs and data selection strategies have been considered by the studies so far. Some studies have used almost all the MRI slices [120,133,99,109], selected slices [144,143,77], segmented mid-brain region containing the Substantia Nigra [114,13], or specific feature extracted brain maps [19,84]. However, not all the studies have explained the process of slice selection and the studies using segmentation requires domain expertise. This led to the formulation of the second research question which looks to address the methods for MRI slice selection and whether a common range of slices can be selected through multiple methods.

Almost all the previous studies report the classification performance by testing within the same dataset. Only two studies reported the generalizability of the developed models on an external independent test set – [30] on the UPenn dataset and [19] on the PD MCI Montreal dataset. As a crucial step towards building trust in data driven systems, there is a need for reporting the generalizability of the CNN model on external patient cohorts [134]. This leads to the third *research question* of how well the developed CNN models trained on a particular dataset generalizes to an independent test set.

As shown in Table 1, the accuracy values reported in previous studies range from 48% to 100%. This wide variation in the accuracy values reported indicated a discrepancy in the fundamental methodology adopted by some of the experiments. An analysis of the possibility of data leakage (see Table 1 and Section 3.6 for more details) in the studies led to the detection of possible slice level data leakage [114,120,133,99,84,13,109] and wrong validation split data leakage [23]. The possibility of such widespread data leakage in almost half the studies led to the fourth *research question* of investigating the effect of slice-level and longitudinal data leakage in hold-out test sets when classifying PD using T1 weighted MRI. In addition to slice level data leakage, which has been investigated by previous studies [144,143,77], this study investigates the effect of longitudinal data leakage. The previous works on data leakage by [144] and [77] focus on T2 weighted MRI. The study by [143] focused on the effect of slice level data leakage on CV using T1 weighted MRI, not an independent hold-out test set.

Table 1
Summary of previous studies on the CNN based classification of PD using structural MRI. The potential for data leakage in the studies is also listed.

Dataset	MRI type	CNN	MRI input	CV	Test set	Results	Data leakage	Study
PPMI	T2	Custom	Single slice	10%	20%	96% accuracy	None detected	Shah et al. [114]
PPMI	T2	AlexNet	2D slices	Not specified	20%	88.9% accuracy	Possible - Slice level leakage	Sivaranjini and Sujatha [120]
PPMI	Not specified	Custom	Age, Gender, 3D volumes	10%	5%	10% accuracy	None detected	Esmaeilzadeh et al. [36]
PPMI	T2	VGG16, ResNet50	Selected slices	Not specified	20%	26% inflation	Investigated	Yagis et al. [144]
Versilia, PPMI	T1	VGG16, ResNet50	8 slices	10 fold	No hold out data	48% and 55% inflation on PPMI, versilia	Investigated	Yagis et al. [143]
PPMI	T1	Custom	Whole volume	5 fold	No hold out data	95.29% accuracy	Possible - Wrong validation split	Chakraborty et al. [23]
IXI, PPMI	T1	Custom	Multiple types	Not specified	20%	75% accuracy	None detected	West et al. [142]
PPMI	T2	VGG16, VGG19, ResNet50, Xception, DenseNet201	2D slices	5 fold	20%	92.6% using VGG-19	Possible - Slice level leakage	Veetil et al. [133]
IXI, PPMI	T1	Resnet101, SqueezeNet, Densenet, MobileNet	GM, WM Maps	20%	20%	94.7% accuracy	Possible - Slice level leakage	Mostafa and Cheng [84]
PPMI	T2	VGG19, DenseNet121, DenseNet201	Selected slices	5 fold	20%	30% inflation	Investigated	Madan et al. [77]
ADNI, PPMI	T2	Custom	Clinical data, MRI slices	Not specified	Not specified	77.9% accuracy	None detected	Yin et al. [147]
PPMI	T2	DenseNet	2D slices	Not specified	20%	6% improvement with augmentation	None detected	Madan et al. [78]
PPMI	Flair, T2	Custom	2D slices, 3D Volumes	Not specified	18 MRI scans	88.9% accuracy	None detected	Vyas et al. [135]
PPMI	T2	Ensemble -VGG16, VGG19, DenseNet201	2D slices	20%	20%	10% increase with ensemble	Possible - Slice level leakage	Rajanbabu et al. [99]
PPMI	T1	VGG19, ResNet34, ResNet50, improved AlexNet	SN region	Not specified	10%	99.80%	Possible - Slice level leakage	Bhan et al. [13]
PPMI	T1	ResNet	3D volume	5 fold	51 MRI scans	94.5% accuracy	None detected	Yang et al. [146]
Upenn, PPMI	T1	ResNet	3D volume	20%	10%	70% accuracy	None detected	Dhinagar et al. [30]
13 datasets	T1	Custom	Jacobian maps, Clinical parameters	5%	10%	79.3% accuracy	None detected	Camacho et al. [19]
PPMI	T2	Custom	2D slices	15%	10%	98% accuracy	Possible - Slice level leakage	Sangeetha et al. [109]

With the increased research focus towards developing hybrid decision support systems, the need for explainable and generalizable data-driven models that is free from any implicit bias is necessary. The focus of this study is towards exploring the appropriate CNN architecture through TL for a generalizable and explainable classification of PD using T1 weighted MRI. Further, the selection of MRI slices, and, the effect of slice level and longitudinal data leakage on the performance of the developed model is also investigated.

3. Design of the experiment

The design of the proposed study is detailed in Fig. 1. The T1 weighted MRI volumes from the PPMI dataset [79] was pre-processed and used to train 12 CNN models. The CNN was trained using TL and evaluated in a five-fold CV setting to determine the best models. The top three CNNs were then selected for additional testing after selecting the MRI slices using two methods. The final models, after tuning for the hyper-parameters, were evaluated for its generalizability to two external datasets. The effect of slice level and longitudinal data leakage is simulated to understand the bias that it can introduce. Finally, the predictions of the CNN was interpreted using a Grad-CAM based visualization of the outputs.

3.1. Data

The data used in this study comes from three different datasets namely PPMI, NEUROCRON and Tao Wu. Among the three, only the PPMI dataset was used for training. The other two datasets – the NEUROCRON and Tao Wu datasets were used exclusively as independent test sets. In all three datasets, the subjects were age-matched to ensure that age related degeneration of the brain was accounted for.

Table 2
Demographics of the subjects across the three databases.

Database	Group	Subjects	Age	Gender
PPMI	NC	213	59.44 ± 11.33	133 M / 80 F
	PD	213	61.27 ± 9.71	125 M / 88 F
NEUROCRON	NC	16	67.63 ± 11.89	4 M / 12 F
	PD	27	68.13 ± 12.86	17 M / 10 F
Tao Wu	NC	18	63.72 ± 4.85	11 M / 7 F
	PD	18	65.61 ± 4.45	10 M / 8 F

3.1.1. Parkinson Progression Marker Initiative (PPMI)

The PPMI database [79] is a publicly available repository that contains imaging data, bio-specimen samples and clinical data. For this study, 213 age-matched T1-weighted MRI scans were selected at random from each class. The images were acquired using MRI scanner equipment by Siemens, GE, Philips and Toshiba, and the magnetic field strength was either 1.5 T or 3.0 T. All the images chosen were acquired in the Sagittal plane. The age of the NC volunteers range from 31 to 85 years, and that of PD range from 34 to 85 years. The demographics of the data are given in Table 2 and the complete details of the PPMI study and the imaging protocols can be found in [79]. The details of the patient ID and the subsequent visits were retained through the pre-processing steps to ensure a controlled assessment of the effect of data leakage.

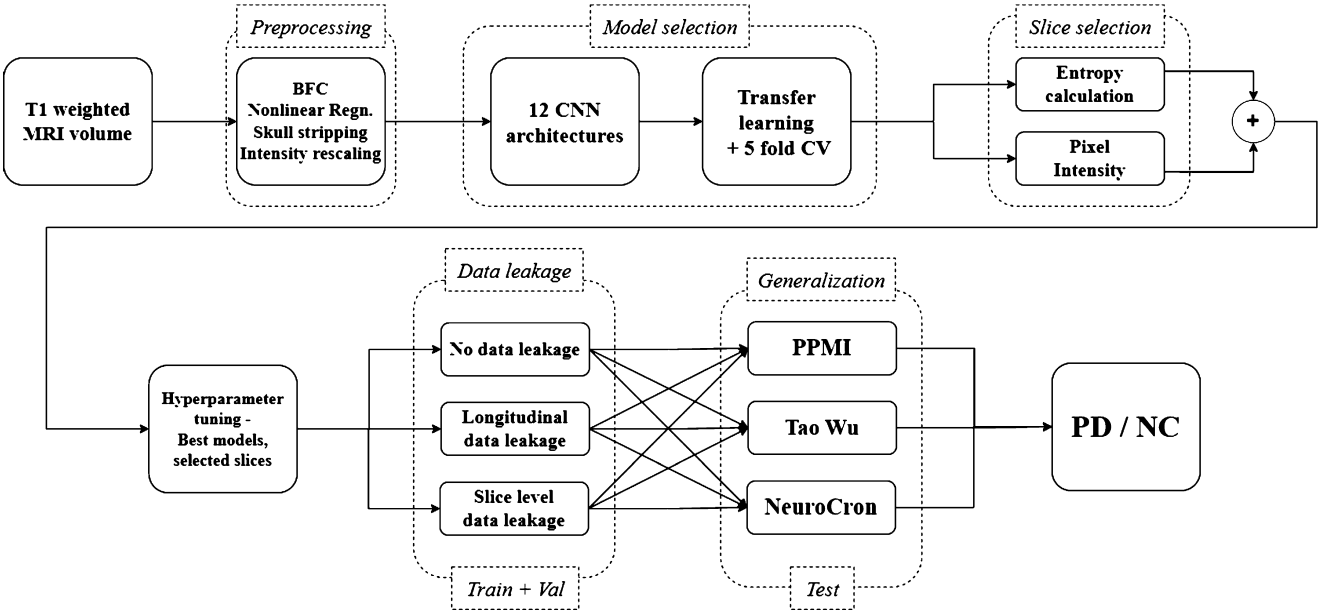


Fig. 1. Overview of the proposed methodology.

3.1.2. NEUROCRON

The NEUROCRON dataset is one of the independent test sets used in this study [10]. It is a public dataset that has been collected by the Neurology Department of the University Emergency Hospital, Bucharest, Romania. The dataset consists of 16 NC and 27 age-matched PD subjects. The T1-weighted MRI was acquired using a Siemens Avanto scanner at 1.5 T. In this dataset, the ages of the NC volunteers range from 46 to 82 and the people with PD range in age from 45 to 86. The details of the demographics of the subjects are detailed in Table 2 and the further details can be found in [10].

3.1.3. Tao Wu dataset

The second test dataset used in this study is the Tao Wu dataset [10]. The dataset was collected and made publicly available by Department of Neurobiology, Beijing Institute of Geriatrics, Xuanwu Hospital, Capital Medical University and Parkinson Disease Centre of Beijing Institute for Brain Disorders, China. This study used T1 MRI from 18 age-matched T1-weighted MRI scans from both each class. The images were acquired using Siemens Magnetom scanner at 3 T. The ages of the PD subjects range from 58 to 74 and the NC control have ages ranging from 57 to 75. The details of the demographics of the data can be found in Table 2 and further details can be found in [10].

3.1.4. Limitations of the datasets

The limitations of the three datasets used in this study are detailed here. The PPMI dataset is a multi-center, longitudinal study that presents study data, biological samples and neuroimaging data for prodromal to mild PD cases [79]. While, at present, the PPMI dataset has data collected from 52 sites[1] across the world, the data collections sites does not include the geographical locations of the Asia-Pacific, South America and Africa (except for one study center in Nigeria). The NEUROCRON and Tao Wu datasets are limited to participants from Romania and China, respectively [10]. There have been studies that indicate a difference in the brain structure due to the difference in the geographical location and cultural variations [127,57]. These differences can be a potential source of bias in the generalizability study. Further, the PPMI dataset had participants enrolled de Novo and within three years of a diagnosis, at the baseline. However, the Tao Wu and NEUROCRON datasets had patients who were already taking medication. Any comparison of the longitudinal data across the three datasets would be unequal as a result. The Tao Wu and NEUROCRON datasets had subjects whose MRI was acquired in the 'eyes closed' position, while the participants of the PPMI study had their MRI taken with 'eyes open'. Since the pre-processing steps involved skull stripping and, hence, removal of extra-cranial tissue, this may not influence the results of the study.

3.2. Preprocessing

In this work, the 2D slice method was followed where each MRI slice was treated as a separate image. The preprocessing steps followed were based on the "Extensive" pre-processing pipeline detailed by Wen et al. [140]. The pre-processing steps involved bias field correction, skull stripping, non-linear registration, brain extraction, intensity normalization, and extracting the MRI slices. The standard MRI pre-processing workflow followed in this work is illustrated in Fig. 2 [50].

Structural MRI can acquire a low frequency intensity gradient known as the bias, illumination non-uniformity, inhomogeneity, or gain field, which corrupts the image [129]. The intensity gradient makes some parts of the image brighter than the other and can erroneously influence image analysis algorithms [140] To reduce the intensity variations due to magnetic field inhomogeneity, bias field correction (BFC) was carried out using the Insight Toolkit (ITK) implementation of the N4 algorithm [129] in the 3D slicer software. After BFC, skull stripping of the MRI was carried out.

Skull stripping or brain extraction is the step used to non-brain tissue from the MRI of the whole head. The process removes extra-cranial tissues not needed for analysis such as the skull and eyeballs. Skull stripping was carried out using the Brain Extraction Tool (BET) [121] in the FSL software. The robust brain extraction option with a fractional intensity threshold (FIT) of 0.4 was used. The value of the FIT was fixed at 0.4 empirically based on a visual inspection of multiple samples on FSLeyes[2] after running BET with different settings. In FSL BET, lower values of FIT results in larger brain estimates and the default value of 0.5 did not give precise brain extraction.

[1] https://www.ppmi-info.org/about-ppmi/ppmi-clinical-sites.

[2] https://git.fmrib.ox.ac.uk/fsl/fsleyes/fsleyes/.

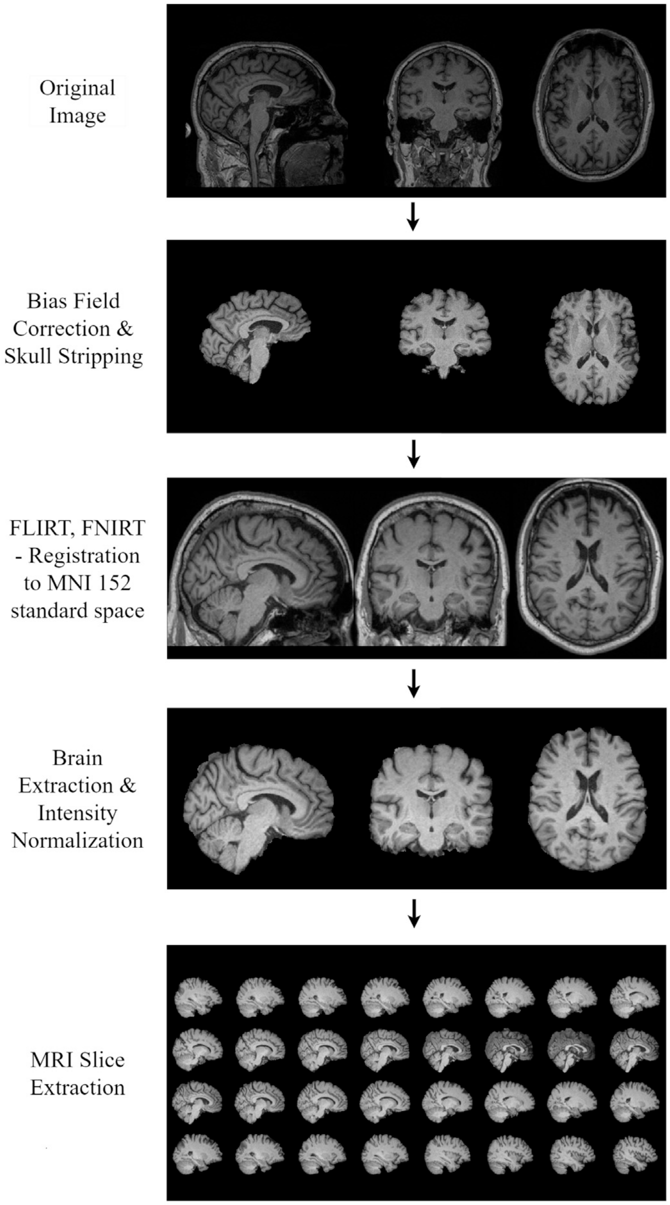

Original Image
Bias Field Correction & Skull Stripping
FLIRT, FNIRT - Registration to MNI 152 standard space
Brain Extraction & Intensity Normalization
MRI Slice Extraction

Fig. 2. Standard MRI pre-processing pipeline followed in this work.

After brain extraction, the MRI scans were then nonlinearly registered to a standard space using the Montreal Neurological Institute and Hospital (MNI) 152 atlas. This allows all the scans to be aligned to a common space such that any group-wise difference between the scans are easier to identify. Further, aligning the MRI to a standard template space facilitates mapping the brain regions from an atlas to get information relevant to specific brain regions. In this study, the registration was carried out in four steps – linear registration, nonlinear registration, warping, and brain extraction. First, a linear registration of the T1-weighted MRI to MNI152-T1 atlas with 12 Degrees of Freedom (DOF) using the FMRIB's Linear Image Registration Tool (FLIRT) in the FSL software was carried out [64,63]. Twelve DOF was chosen to get the size and orientation of the T1-weighted MRI as close to the nonlinear registration template as possible. The second step was a nonlinear registration of the scans using FSL - FMRIB's Nonlinear Image Registration Tool (FNIRT) [49,7]. FNIRT aligns the MRI scan to a T1-weighted MNI 152 - 2mm brain atlas. Finally, the FNIRT output was then warped to a T1-weighted MNI 152 1mm template to generate a high resolution output. Since FNIRT uses the BFC MRI as input, the BET was rerun to get the final aligned and brain-extracted MRI image.

MRI scans are acquired by various scanners with different parameters which results in a wide variation in the intensity of the same tissue types across scans. Since this can adversely affect MRI processing and further analysis, global intensity normalization between 0 and 1 was thus carried out [140]. All the images were checked manually after each processing stage to correct any errors in the parameters chosen for the processing step and to remove damaged files. The final image size after pre-processing was $182 \times 218 \times 182$ which indicated the dimensions of the image along the Sagittal, Coronal and Axial direction, respectively. After viewing the MRI scans in the Sagittal axis, it was seen that the first 12 and the last 30 slices were lacking in information and could be excluded. This resulted in 140 slices for each MRI. From these MR images of size $140 \times 218 \times 182$, the slices along the Sagittal plane (x-axis) were extracted and stacked three times along the z-axis as needed by the CNN. The extracted MRI slices were then saved to a Numpy (.npy) format to be used as input to the CNN. The final size of each input image to the CNN was - $218 \times 182 \times 3$, with 140 slices or images coming from each MRI scan of a person.

3.3. Convolutional Neural Network (CNN)

3.3.1. Transfer learning (TL)

Stemming from cognitive research, TL refers to the process of improving the performance in a new task by transferring knowledge in a related task. An example of humans leveraging such knowledge would be how writing skills in children can be improved by an associated knowledge of hand movement gained by drawing or coloring. Given the scarcity of data and associated high cost of expert annotations, many medical imaging research studies have adopted TL approaches [69]. The most popular choice of source domain are CNN models trained on the ImageNet database which comprises of millions of natural images [105].

A formal definition of TL given by Pan and Yang [92] is as follows: "A domain D is made up of two components - a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, x_2, ...x_n\} \in \chi$. In the learning task of image classification, each image is a feature. χ is the space for all image feature vectors and X is a particular learning sample. In a specific domain represented by $D = \{\chi, P(X)\}$, a task is denoted by $\tau = \{\Upsilon, f(.)\}$. Here Υ refers to the label space and $f(.)$ is the objective predictive function. The task is not learned from training data consisting of pairs (x_i, y_i), where $x_i \in X$ and $y_i \in \Upsilon$."

Specifically, considering the ImageNet source domain with natural images to be denoted as
"$D_s = \{(x_{s_1}, y_{s_1}), ...(x_{s_n}, y_{s_n})\}$,
where the data instance $x_{s_i} \in \chi_s$ and the corresponding label $y_{s_i} \in \upsilon_s$. Similarly, the target domain data of MRI is denoted as $D_t = \{(x_{t_1}, y_{t_1}), ...(x_{t_n}, y_{t_n})\}$,
where the data instance $x_{t_i} \in \chi_t$ and the corresponding label $y_{t_i} \in \upsilon_t$. Given the Image source domain D_s and learning task τ_s, TL aims to help the learning of the target predictive function $f_t(.)$ in D_t, the classification of PD from MRI in this case, using knowledge in the source domain, where $D_s \neq D_t$ or $\tau_s \neq \tau_t$. The TL type followed in this study is *inductive transfer learning* where labels available in target domain *induce* an objective predictive function $f_t(.)$ for PD classification from MRI [92]."

TL from natural images to medical imaging has been shown to allow faster training with lower computational resources, faster convergence with pre-trained weights and performance improvements in large models when pre-trained [48,86,98,69]. And in this study, CNN architectures that have been pre-trained on the ImageNet dataset is used. The 12 CNN models used in this work are DenseNet 121, DenseNet 201, Inception V3, Inception ResNet V2, ResNet 50 V2, ResNet 152 V2, MobileNet, VGG 16, VGG 19, NASNet Large, NASNet Mobile, and Xception.

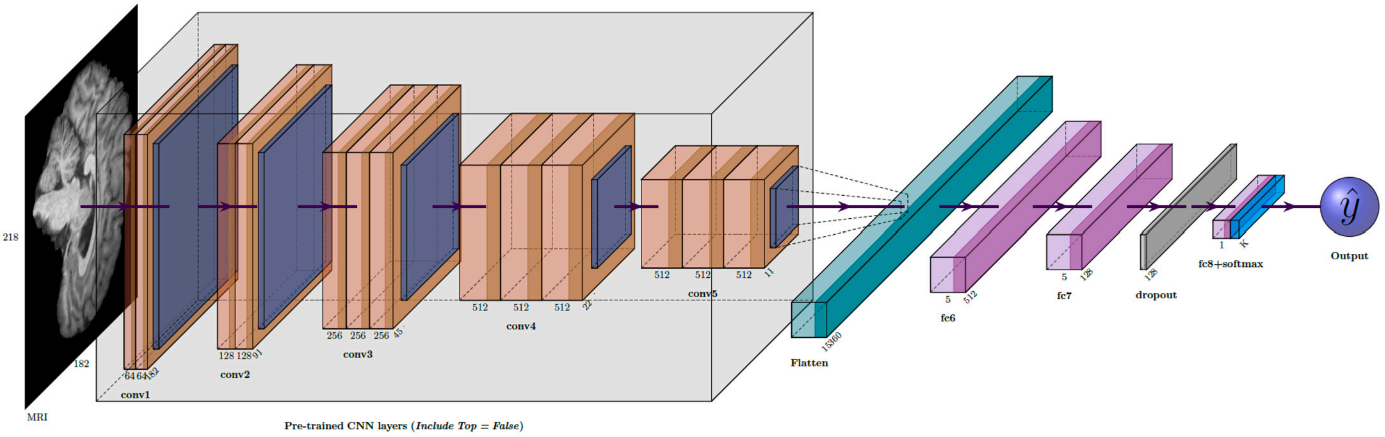


Fig. 3. The modified Convolutional Neural Network architecture.

3.3.2. Description of the chosen CNN architectures

The twelve architecture used in this study were chosen with a specific view of diversity in mind. Each architecture has its own pros and cons and a brief description of the chosen CNN models is given below.

The Inception architecture was the 2014 winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [126]. The model introduced the inception module which reduced the number of model parameters by a large degree. The inception model is adaptable for multi-scale inputs and is known to be versatile across tasks. The Visual Geometry Group (VGG) architecture was the runner-up at the ILSVRC [118]. Both VGG-16 and VGG-19 are deep architectures with a repetitive structure, known for strong performance in general image classification. The core idea of VGG is the usage of small convolutional filters of size 3x3 in all layers with more focus on depth than width.

The ResNet architecture was the winner of ILSVRC 2015 [53]. The ResNet models have a deep architecture with a specific focus on reducing the complexity of the model. ResNet introduced 'additive identity mapping' and 'shortcut connections' in the architecture to deal with the vanishing gradient problem of CNN. DenseNet or Densely Connected Convolutional Network is an architecture that has every layer connected to every other subsequent layer in a feed-forward manner [58]. DenseNet offers feature reuse and deep supervision, while reducing the possibility of over-fitting.

Xception or 'Extreme Inception' is an architecture that performs depthwise convolution or spatial correlation mapping first, followed by pointwise convolution [25]. These modules in Xception are called 'depthwise separable convolutional layers' and the architecture also incorporates residual connection introduced in ResNet to improve its performance on large datasets. MobileNet, is light weight architecture developed for mobile applications [56]. MobileNet uses depthwise separable convolutional layers and is a smaller model offering efficient and fast computation in resource-limited environments.

NASNet is based on the Neural Architecture Search (NAS) framework that uses the concept of reinforcement learning to optimize the efficiency of the model and eliminate lesser performing parts [150]. Known for its automated architecture optimization, NASNet offers high performance in diverse tasks with a complex structure. NASNet Mobile is a version of NASNet Large optimized for performance with fewer resources while maintaining robustness. The Inception-ResNet architecture combines the Inception modules and residual skip connections to create a more efficient neural network [125]. The hybrid model mirrors the performance of two top models with reduced complexity. The 12 pre-trained CNN models were modified to function as feature extractors and then retrained on the PPMI dataset. The description of the modifications carried out are given below.

3.3.3. Modifications to the CNN architecture

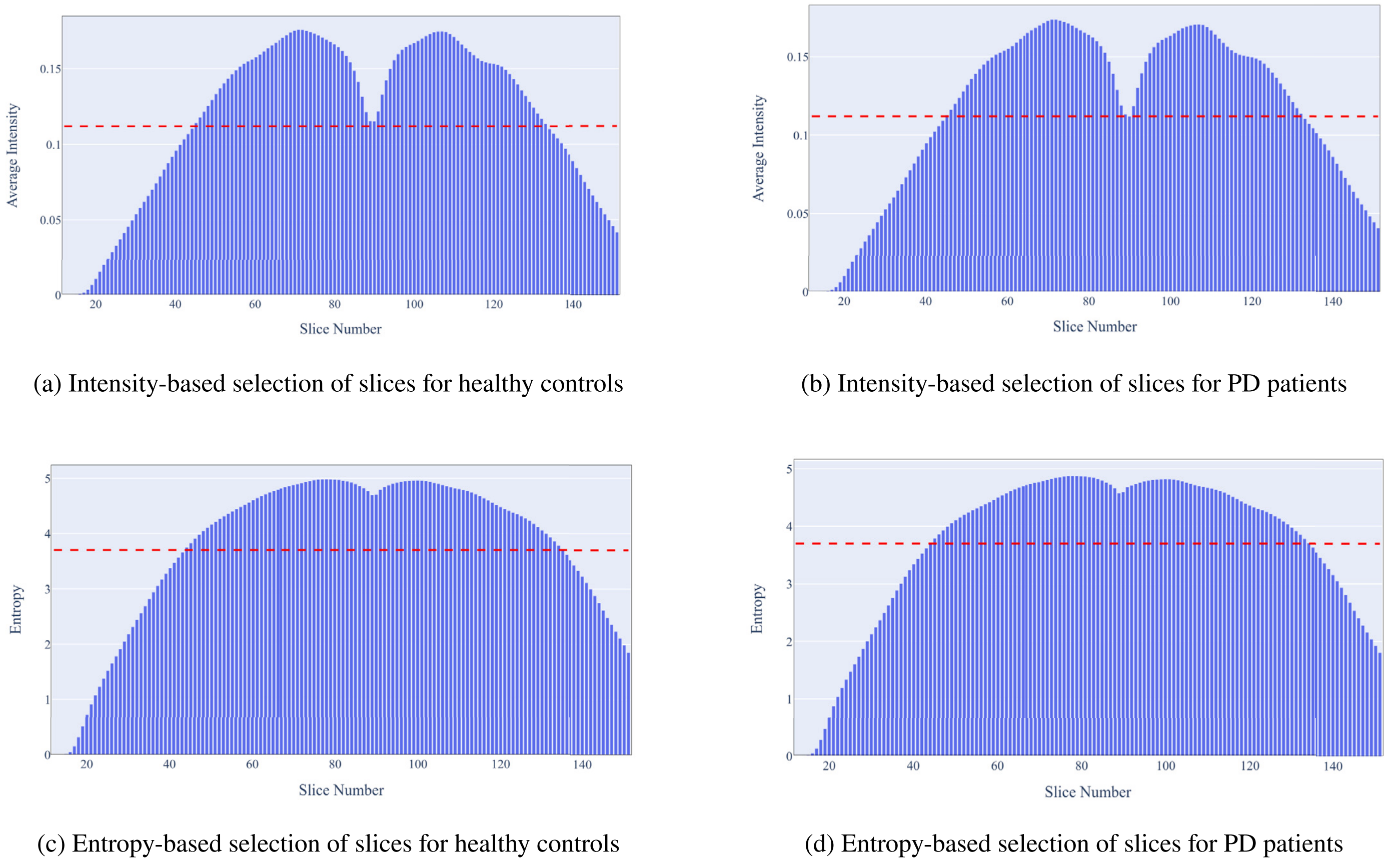
The base network models obtained from the Keras Application Module[3] were pre-trained on the ImageNet database [29]. The weights of the networks in the top layers, learned by training on the ImageNet database was retained by freezing them from further modification. The final 1000 class fully connected layer was removed and the network was modified[4]. The modification to the architecture and the choice of CNN layers is based on experiments conducted in our previous study using T2-weighted MRI [133]. The proposed modification to the CNN is given in Fig. 3. In the figure, the VGG-16 network is used as the example architecture.

On top of the base network pre-trained on the ImageNet database, a flatten layer, a fully connected layer of size 512 with a Rectified Linear Unit (ReLU) activation function, one more fully connected layer of size 128 with ReLu activation function, a dropout layer of value 0.3 and a final softmax decision layer was added. A description of the specific functions of the added layers are given below. The flatten layer is added immediately on top of the base model to convert the two dimensional feature vector of the model into one-dimensional linear vector with sequential features. The flattened feature vector is given as the input to the dense layer or the fully connected layer. The dense layer has connection to all the activations in the previous layer and allows the abstraction from the local information to bigger patterns in the data. The fully connected layer is composed of a weight matrix and an offset bias. The ReLU activation function is a nonlinear threshold function that performs element-wise operation to set negative values to zero. The fully connected layer is then connected to a dropout layer. Given a probability value, the dropout layer randomly sets a portion of the input to zero between iterations. This allows the network (input in the layer) to change in each iteration, thus preventing over-fitting. The dropout layer is connected to another dense layer with two neurons, equal to the number of classes - NC and PD. This final layer has a softmax activation function which returns a probability vector indicating the prediction of the class of the input vector [50,85,74,87,123,91].

The modified CNN architectures are trained on the PPMI dataset after separating out a hold-out test set consisting of twenty subjects from each class. The remaining 386 MR images were used for training. The CNN models were trained for 100 epochs using a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 1e-3 and a momentum of 0.9. After training, the 12 CNN models were then validated through five-fold CV, before testing on the hold-out and external independent test sets.

[3] https://keras.io/api/applications/.

[4] https://keras.io/guides/transfer_learning/.



(a) Intensity-based selection of slices for healthy controls

(b) Intensity-based selection of slices for PD patients

(c) Entropy-based selection of slices for healthy controls

(d) Entropy-based selection of slices for PD patients

Fig. 4. Selection of the Slices of Importance (SOI).

3.4. *Selection of the MRI slices*

Research on the MRI based classification of PD presents a wide range of slice selection strategies, including studies on the segmentation of PD specific brain regions (see section 2). This section discusses the selection of MRI slices based on the methods detailed by Madan et al. [77] and Yagis et al. [143]. This study follows the 2D slice method and treats each slice of the MRI as a separate image. After pre-processing each MRI consisted of 182 slices, from which a visual inspection lead to the removal of 42 slices, which brought down the number of slices to 140. However, even the 140 slices does not always contain meaningful information and the number of MRI slices that are necessary for PD classification could be further optimized. In order to also test validity of the initial decision of choosing only 140 slices, slices from number 15 to 150 were also included in this analysis.

3.4.1. *Intensity-based slice selection*

The first method by Madan et al. [77] used the pixel intensity as a means to measure the importance of the slices. The average pixel intensity of each slice across all the subjects was compared to the overall average pixel intensity across all slices. This is depicted in Figs. 4a and 4b. The average pixel intensity is depicted by the horizontal line in red and the vertical bars represent the value of the average pixel intensity for each slice across the subjects.

3.4.2. *Entropy-based slice selection*

The entropy based method detailed by Yagis et al. [143] utilizes Shannon entropy to measure the information contained in each slice. For each slice, a histogram of the gray levels were calculated. The frequency of the occurrence of the gray level was taken as the probability and used in the entropy calculation as detailed below [55]:

$$E_s = -\sum_{i=1}^{M} p_i \, log \, p_i \tag{1}$$

Here M represents the number of gray levels in the slice, and p_i represents the probability of occurrence of that gray level. The entropy values of the slice and the selection of the slices using the average entropy is depicted in Figs. 4c and 4d.

3.4.3. *Slices of importance (SOI)*

The final range of MRI slices selected using the intensity and entropy for each class as detailed in Fig. 4 are:

- Intensity-based: NC class - 89 slices
- Intensity-based: PD class - 87 slices
- Entropy-based: NC class - 92 slices
- Entropy-based: PD class - 91 slices

Using an overlap of the slice selection criteria matched by both methods, the slice numbers from 45 to 131 have been selected. The SOI study thus reduced the number of important slices per MRI from 140 to 87. An overlay of the Harvard-Oxford (HO) cortical and HO sub-cortical atlases on the 87 SOI is shown in Fig. 5. The image shows that the SOI retain all the anatomical structures relevant to PD and can be used for further analysis. After retaining only the 87 SOI, the top 3 model are tuned for hyper-parameters and further testing.

3.5. *Optimization*

The top 3 models selected from the 12 CNN architectures through 5 fold CV were then tuned for hyper-parameters. For the same, the 87 slices selected through the SOI study was used. During this tuning process, only the 386 scans from the 5 fold CV was used, the 40 scans in the hold-out set were kept separately as the test set. This ensured

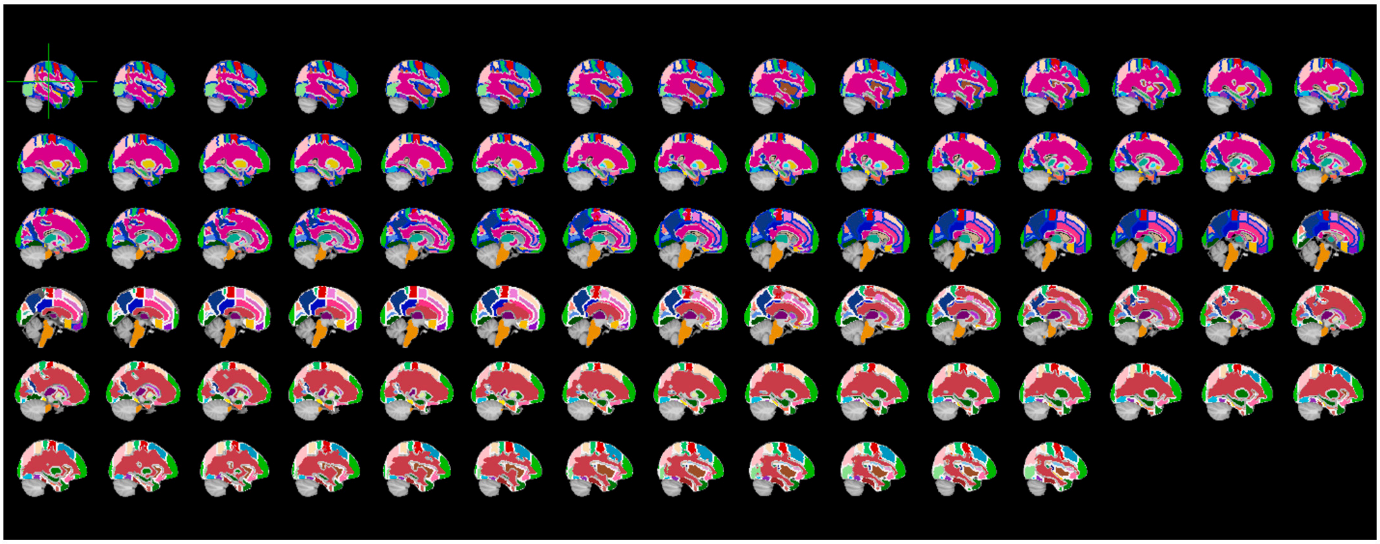


Fig. 5. Harvard-Oxford Cortical and Harvard-Orxford Subcortical atlases overlaid on MRI slices chosen through entropy and intensity based slice selection strategies.

Table 3
Search space of the hyper-parameters for the proposed CNN.

Parameter	Values
Optimizer	Ada, SGD, RMSprop, Adagrad
Learning rate	$1e^{-2}$, $1e^{-3}$, $1e^{-4}$
Size of last dense layer	16, 32, 64, 128
Dropout rate	0.20, 0.25, 0.30

that there was no data leakage and allowed uniformity of the analysis process. The parameter space for the grid search is listed in Table 3. After hyperparameter tuning, the network parameters were fixed. The same parameters were used for retraining the network during the data leakage study as well.

3.6. Data leakage study

Data leakage is a serious problem encountered in some data driven systems due to incorrect data handling. It refers to the process by which the model learns information about the test data as well due to an unintentional leakage of the test data characteristics onto the training stage. Analysis of the previous studies using CNN for the MRI based classification of PD revealed the possibility of data leakage, which is listed in Table 1. The bias due to data leakage results in a higher than actual performance by the CNN and a failure of the model when deployed in real time [140].

3.6.1. Data leakage in MRI based studies

The main causes and the implications of data leakage, specific to MRI based studies is listed below:

1. **Slice-wise data splitting:** An MRI is composed of several slices, each of which can be viewed as a single image. The common practice that has come into medical imaging, as an extension of natural image processing techniques, is to pool all the data (images), randomly shuffle it, and then split the train-test set at a predefined ratio (such as 80-20 or 70-30). This potentially leads to an individual's information to be split across the train and test data, leading to overoptimistic results. This is due to the model having learned the test data characteristics and associated labels during the training process itself [140,77].
2. **Longitudinal data splitting:** The disease progression of a patient is mapped using a timeline of imaging procedures conducted during the initial visit and subsequent follow up checks. Splitting at the MRI level, without considering the longitudinal data can potentially lead to the same person's MRI being present in the training and the test set. This leads to an increased performance of the model because of familiarity with the individual's characteristics [104].
3. **Late split:** Research in data driven systems dealing with small datasets use data augmentation strategies to increase the amount of data available. In such studies, the augmentation has to be carried out on the training data, after separating the validation and test data. Otherwise, the augmented data will contain characteristics learned from the test set as well, leading to an erroneously high performing model which fails when tested on an independent test set [140].
4. **Wrong validation split:** Some MRI based studies have report results using just the cross-validation accuracy scores instead of a hold-out test set. If the same subject's data is split across multiple folds during the cross-validation, it leads to an increase in the performance of the model. Previous studies [80,22,106] showed that the model can learn the characteristics of the individuals instead of the disease in such cases, a process termed identity confounding. Identity confounding also leads to a wrongly inflated accuracy of the model being reported.
5. **Incorrect hyperparameter selection:** Hyperparameter tuning of the deep learning models is an important step that leads to an optimization of the neural network's learning process. Tuning the hyperparameters on the entire dataset can lead to a higher than actual performance because the model has been tuned with a knowledge of the test data's characteristics as well. The correct process would be to tune the hyperparameters using only the training data, after separating out the test data [140].
6. **Intra-domain transfer learning:** This type of data leakage occurs when deep learning models are adapted from one task to another associated task in the same domain. An example would be modifying and adapting a model trained to classify NC vs PD, through transfer learning to predict a three way classification between NC, PD and SWEDD (Scans without evidence of dopaminergic degeneration). If the test set in the second task is made up of subjects who were part of the training data in the first task, then the associated characteristics and labels are already learned by the model. This leads to overoptimistic results due to an inherent bias during the transfer learning process [140].

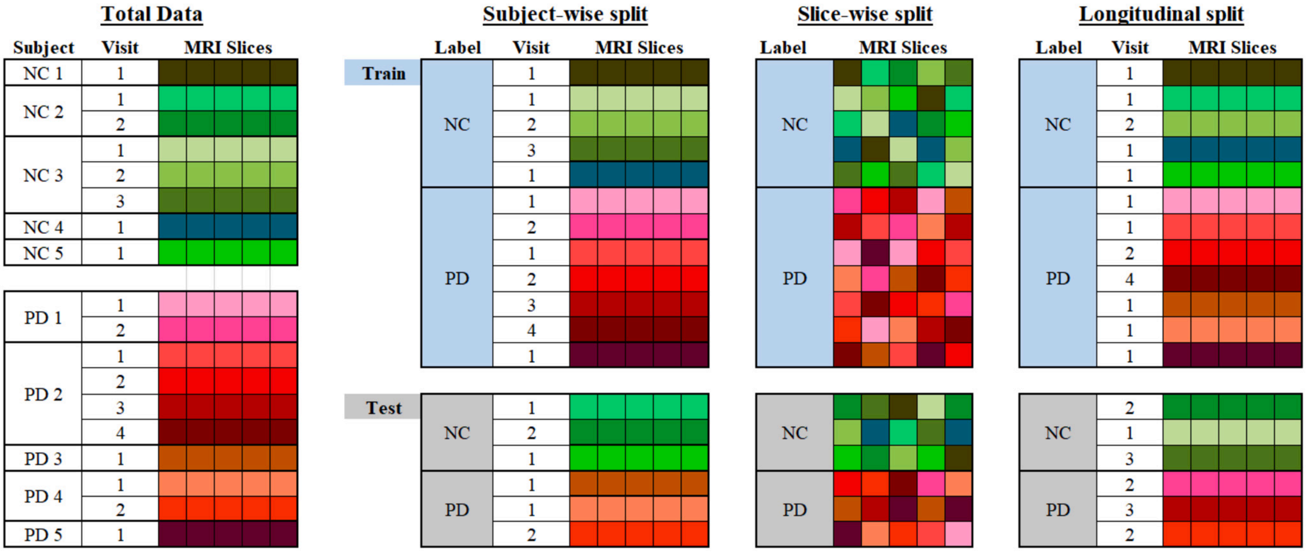


Fig. 6. Representation of the data division strategy to simulate data leakage.

In this study, the effect of slice-wise and longitudinal data leakage in comparison to no data leakage (subject-wise data split) is investigated. The details of the division of the dataset carried out to simulate the effect of data leakage is given below.

3.6.2. Data leakage simulation

The data leakage study was designed to analyze the skewed results in the classification resulting from improper data handling. In this work, the data leakage study was carried out only after the selection of the optimal models and the selection of the MRI slices. Hence, the three top CNNs evaluated in the data leakage study used 87 slices for each MRI. A diagrammatic representation of the subject-wise, slice-wise and longitudinal data leakage considered in this study is shown in Fig. 6. The data splitting to simulate the three data leakage scenarios is detailed below:

1. *No data leakage (or) Subject-wise Split*: The data was split at the subject level and the test set was retained as an independent hold out set (20 subjects in each class). As seen in Fig. 6, the data of each subject from all visits are retained either as part of train data or test data. The results for the PPMI, NEUROCRON and Tao Wu datasets are reported at the subject-level using majority voting [102]. Testing the best models with the independent datasets evaluates the generalizability of the CNN models on realistic, unseen data.
2. *Slice-wise split*: In this case, the MRI slices of the entire dataset are pooled together, shuffled and then randomly split into train, validation and test split at 80%, 10% and 10%, respectively. This erroneous method of splitting the data at the subject level is being simulated in this study to look at how much bias is introduced as a result of slice level splitting. Since NEUROCRON and Tao Wu datasets are independent datasets, the results for the two datasets are still at the subject level. The PPMI dataset on the other hand has only a mixture of slices from various subjects and the performance of the CNN model is reported.
3. *Longitudinal split*: This is a special case where the data is split as individual MRI scans (including all slices) but not at the subject level. Hence, a person who has multiple visits may have their data split between the train and test sets as shown in Fig. 6. This is a sensitive issue in datasets with longitudinal data available [104]. Just like the subject-wise split, the results are reported using majority voting for each MRI in all three datasets.

3.7. Explainable AI framework (XAI)

In order to visualize the regions of importance (ROI) identified by the CNN, a feature map is generated using a method called Gradient-weighted Class Activation Mapping (Grad-CAM) [113]. The method captures the importance of the neurons for the specific target class thorough the gradients of the last convolutional layer of the CNN and uses the same to generate visualizations. Further details of the Grad-CAM can be found in [113]. In this study, the original algorithm proposed by Selvaraju et al. [113] was slightly modified such that the absolute value of the gradients are taken before the application of the ReLU function. This allowed the representation of the negative gradients as well, since the sign is only an indicator of the direction.

The Grad-CAM heatmap generated was overlaid with the pre-processed MRI (registered to MNI152 space) used as input to the CNN, giving a superimposed image. In order to identify the important brain regions, the HO Cortical with 48 brain areas and HO Sub-cortical Atlas with 21 brain areas was used [19]. An overlay of the HO Cortical and Subcortical atlas on the selected 87 MRI slices is shown in Fig. 5. The highlighted portions on the heatmap was mapped using the atlas overlaid on the superimposed image to give the important brain regions identified by the CNN.

4. Results and discussion

4.1. Selection of the optimal CNN architectures through transfer learning

This part of the study investigates the TL based development and evaluation of 12 CNN architectures in order to assess which models are able to capture the characteristics of the MRI data for PD classification better. For this, the 12 CNNs - DenseNet 121, DenseNet 201, Inception V3, Inception ResNet V2, ResNet 50 V2, ResNet 152 V2, MobileNet, VGG 16, VGG 19, NASNet Large, NASNet Mobile, and Xception were used. The data was split at the subject level with 140 slices per MRI. The CNN models were evaluated using a 5-fold CV scheme to evaluate the performance. The architectural details of the 12 CNN models, their performance on the ImageNet dataset, and the accuracy and standard deviation of the models during the five-fold CV when trained on the PPMI dataset is listed in Table 4. The results from Table 4 show that the performance of the CNN models on the PPMI database is not what can be anticipated based on the performance of the models on the ImageNet database. It is seen that VGG16, VGG19 and ResNet50 clock in a better performance for the MRI data compared to the other architectures

Table 4
Results of the five-fold cross validation of the 12 CNN architectures through TL reported in terms of average accuracy and standard deviation of the accuracy across the five folds.

CNN Architecture details			Performance on the ImageNet database		Results of the five-fold CV	
Architecture Name	Depth	No. of Parameters	Top-1	Top-5	Average accuracy	Standard deviation
VGG 16	16	138.4 Million	0.7130	0.9010	0.5229	0.0552
VGG 19	19	143.7 Million	0.7130	0.9000	0.5225	0.0400
ResNet 50 V2	50	25.6 Million	0.7600	0.9300	0.5117	0.0589
Inception ResNet V2	~500	55.8 Million	0.8030	0.9530	0.5094	0.0664
Xception	126	22.9 Million	0.7900	0.9450	0.5048	0.0650
MobileNet	28	4.2 Million	0.7130	0.9010	0.5041	0.0603
NASNet Mobile	~500	5.3 Million	0.7440	0.9190	0.5033	0.0568
ResNet 152 V2	152	60.3 Million	0.7800	0.9420	0.5028	0.0668
DenseNet 121	121	8 Million	0.7500	0.9230	0.5027	0.0725
DenseNet 201	201	20 Million	0.7730	0.9360	0.4988	0.0743
Inception V3	48	23.8 Million	0.7790	0.9370	0.4979	0.0744
NASNet Large	~1000+	88.9 Million	0.8250	0.9600	0.4819	0.0757

Table 5
Final choice of hyperparameters for the top-3 CNN models.

CNN Architecture	Optimizer	Last Dense layer size	Learning rate	Dropout value
VGG 16	RMSprop	128	0.0001	0.2
VGG 19	RMSprop	128	0.001	0.2
ResNet 50	SGD	128	0.01	0.3

considered. These results are in line with the findings of a previous study exploring the performance of transfer learned CNN models on T2-weighted MRI [133]. The VGG models have exhibited better performance than all the other models in both T1-weighted and T2-weighted MRI.

The top 3 architectures - VGG16, VGG19 and ResNet50 models are used for further analysis. The number of slices of MRI was reduced from 140 slices to 87 slices based on the SOI study. Following the SOI selection, the top three CNN models were tuned for the hyperparameters listed in Table 3. Using a grid search scheme, the final hyperparameters fixed are given in Table 5. These parameters were retained across all further analysis, even when the models were retrained to assess the bias due to data leakage.

4.2. Investigation of data leakage and generalizability

4.2.1. Subject-wise split

This part of the study looks at the proper case without data leakage by splitting the data at the subject level. The CNN models trained on the PPMI dataset was tested on the PPMI, NEUROCRON and Tao Wu datasets. Testing on the NEURCRON and Tao Wu datasets enabled the evaluation of the ability of the CNN to generalize to an external patient cohort, acquired with different settings than the dataset the model was trained on. The results of the evaluation are listed in Table 6.

Testing the performance of the CNN on the PPMI dataset shows that ResNet50 performs better than the VGG models with 70% accuracy and F1 score. The VGG models have 58% - 60% accuracy and while the classification of PD is good, the specificity score indicates that the classification of NC are poor, with a high false positive rate. Upon evaluating the generalizability of the CNN architectures, we see that ResNet50 suffers from a loss of accuracy and F1 score when tested on the independent datasets. However, the VGG models showcase a much more stable and consistent performance. From the other metrics, we see that VGG19 remains robust its ability to correctly predict PD throughout the three tests and while ResNet50 has a very good classification of the NC class. The VGG-19 model has a consistent F1 score when tested on all three datasets and is able to generalize well to external cohorts than VGG16 and ResNet50.

4.2.2. Slice-wise split

In this part of the study, the performance of the CNN models on data that had been split randomly at the slice-level allowing a single person's MRI scan to be split between train and test data was evaluated. A representation of the slice level data leakage is shown in Fig. 6. This is an erroneous method that gives a boosted value of accuracy since it does not consider subject-wise accuracy. The results of the study using slice-level data leakage are shown in Table 6. From Table 6, it is clear that for the PPMI database, the accuracy values are now increased from 58% - 70% to 94% - 98%. The difference in accuracy ranges between 34% and 67% as a result of this wrong data split. The data leakage model performed poorly when evaluated on the Tao Wu and NEUROCRON datasets, highlighting the importance of using independent test datasets for testing the generalizability of the developed models. From the results of testing on the external datasets, it is evident that while accuracy scores are a good indicator of the disparity in the results, the F1 scores capture the drastic difference in the performance better.

4.2.3. Longitudinal split

Longitudinal data leakage refers to the case where the data is split at the MRI level, but, a single person's MRI scans from multiple visits could get split between test and train data. The bias in the results due to such an incorrect split was evaluated and the results are presented in Table 6. From the results in Table 6, it is seen that for the PPMI dataset, ResNet50 shows a reduction in the accuracy level from earlier, while VGG19 shows a 30% increase in accuracy value. When deployed to the independent datasets, the ResNet50 model shows better performance, however the VGG models had less than 10% change in accuracy. With the F1 scores, we see that the model's performance is not as bad as the slice level leakage simulation when testing on the external datasets. However, the worst-case detection for PD has been when the models were tested on the Tao Wu dataset. This is the first study to carry out an evaluation of longitudinal data leakage in PD classification using MRI.

Overall, VGG19 had the lowest number of False Negatives among all the three CNN architectures considered, across all the data leakage studies. The model showed consistent performance even when deployed on the additional test datasets. With the specific focus of classifying PD, the VGG19 model would serve as a better choice for the development of a first level screening and in building a hybrid diagnostic decision support tool.

4.3. Visualizing the important regions of the brain using XAI

In order to visualize the regions of the MRI identified by the CNN in the classification between NC and PD, Grad-CAM heat maps were generated. A heat map based visualization of the regions of the brain identified by the VGG-19 model from the slice 58 and slice 108 of the MRI using Grad-CAM is shown in Fig. 7.

Table 6
Summary of the results from the study on the effect of data leakage on the CNN models across three datasets.

Data Leakage	Dataset	CNN	Accuracy	Precision	Recall	Specificity	F1 score	Change in accuracy (%)
None	PPMI	VGG16	0.60	0.57	0.80	0.40	0.67	-
		VGG19	0.58	0.55	0.90	0.25	0.68	-
		ResNet50	0.70	0.68	0.75	0.65	0.71	-
	NEUROCRON	VGG16	0.53	0.61	0.70	0.25	0.66	-
		VGG19	0.56	0.61	0.85	0.06	0.71	-
		ResNet50	0.35	0.00	0.00	0.94	0.00	-
	Tao Wu	VGG16	0.53	1.00	0.06	1.00	0.11	-
		VGG19	0.56	0.56	0.56	0.56	0.56	-
		ResNet50	0.53	1.00	0.06	1.00	0.11	-
Longitudinal	PPMI	VGG16	0.63	0.59	0.85	0.40	0.69	4.17
		VGG19	0.75	0.71	0.85	0.65	0.77	30.43
		ResNet50	0.65	0.62	0.80	0.50	0.70	-7.14
	NEUROCRON	VGG16	0.53	0.59	0.81	0.06	0.69	0.00
		VGG19	0.60	0.62	0.96	0.00	0.75	8.33
		ResNet50	0.42	0.63	0.19	0.81	0.29	20.00
	Tao Wu	VGG16	0.50	0.50	0.06	0.94	0.10	-5.26
		VGG19	0.56	0.75	0.17	0.94	0.27	0.00
		ResNet50	0.58	1.00	0.17	1.00	0.29	10.53
Slice-level	PPMI	VGG16	0.98	0.97	0.99	0.97	0.98	63.79
		VGG19	0.96	0.96	0.96	0.96	0.96	67.58
		ResNet50	0.94	0.94	0.94	0.94	0.94	34.65
	NEUROCRON	VGG16	0.44	0.71	0.19	0.88	0.29	-17.39
		VGG19	0.42	0.58	0.26	0.69	0.36	-25.00
		ResNet50	0.37	0.00	0.00	1.00	0.00	6.67
	Tao Wu	VGG16	0.53	1.00	0.06	1.00	0.11	0.00
		VGG19	0.58	1.00	0.17	1.00	0.29	5.00
		ResNet50	0.50	0.00	0.00	1.00	0.00	-5.26

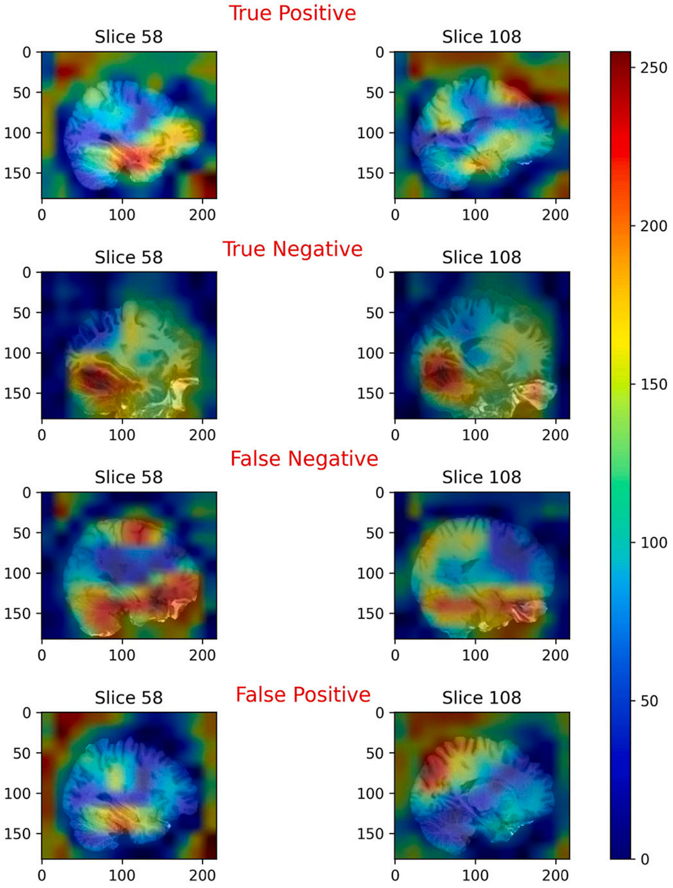


Fig. 7. Grad-CAM heatmaps showing the important regions identified by the CNN.

In the true positive case, the regions highlighted by the heat map included the *Amygdala, Putamen, Pallidum, Brainstem, Cerebral Cortex, Hippocampus, Temporal Pole, Parahippocampal Gyrus, Cerebral White Matter, Insular Cortex, Frontal Pole, Frontal Operculum Cortex, Frontal Orbital Cortex, Temporal Occipital Fusiform Cortex, Lateral Occipital Cortex -Superior Division,and Precuneus Cortex*. We see that a major focus is on the Fronto-Temporo-Pareital region of the brain. This is in line with findings from other studies which indicate progressive atrophy of the Temporal – Hippocampal, Fronto-Parietal areas and the Basal Ganglia [111]. The degeneration of the Amygdala, Caudate and Putamen in PD is well documented across multiple studies [60,111], which were areas highlighted on the heatmap as well. Additionally, the regions highlighted by the heatmap are in line with the Braak staging model [18] where the atrophy due to PD progress from the brainstem upwards towards the Sustantia Nigra, the Mesocortex, Allocortex, and Neocortex.

In contrast to PD, the heatmap of the MRI of a normal subject had the highest concentration on the occipital region, temporal region, and cerebellum, while also focusing on the frontal lobe. The main regions highlighted by the heatmap include *Occipital Fusiform Gyrus, Frontal Orbital Cortex, Temporal Occipital Fusifrom Gyrus, Lingual Gyrus, Temporal Fusiform cortex, Intracalcarine Cortex, Supracalcarine Cortex, Parahippocampal Gyrus, Cerebral White Matter, Cerebral Cortex, the cerebellum, and the Frontal pole*. In NC, it is seen that the network focused on the difference in the tempo-occipital regions of a normal brain compared to that of PD. These regions have been discussed in previous research that has studied the cognitive impairment and decrease in memory due to PD [116].

In the false negative case, the Temporo-Occipital regions were highlighted. The mid-brain regions of pallidum, putamen and thalamus was not focused on by the CNN. The main regions of interest uncovered by the heatmap included the *Temporal Pole, Parahippocampal Gyrus – anterior division, Frontal Orbital Cortex, Temporal Fusiform Cortex – Anterior and Posterior division, Lateral Occipital Cortex – Superior division, Cerebral White Matter, Cerebral Cortex, Precentral Gyrus, Amygdala, Temporal Occipital Fusiform Cortex, and Occipital Fusiform Gyrus*.

Table 7
Important regions of the brain in PD as identified by the Grad-CAM based visualization, normal brain function of the region of interest and how it is affected in PD.

Region of Interest	Function	Effect in PD
Amygdala	Emotional Regulation	Hyposmia, Anxiety/ Depression
Putamen	Learning, Motor control,	Motor symptoms
Pallidum	Motivation, Movement control	Apathy, impulse control disorder, motor symptoms
Brainstem	Autonomic, sensory functions, balance, coordination, reflexes	Dopaminergic degeneration
Cerebral cortex	Voluntary muscle movement, cognition	Tremors, akinesia, cognitive decline
Hippocampus	Memory, Olfaction	Memory deficits
Temporal pole	Complex cognitive functions	Emotion based action regulation, non-motor symptoms
Parahippocampal gyrus	Visuo-spatial, episodic memory	Memory deficits, olfactory disturbance
Cerebral white matter	Information processing	Motor symptoms
Insular Cortex	High level cognitive functions	Non-motor symptoms, cognitive deficits
Frontal Pole	Behaviour, cognition	Executive dysfunction
Frontal Operculum cortex	Linguistic, cognitive, somatosensory functions	Non-motor symptoms, cognitive deficits
Lateral Occipital Cortex	Visual processing, object recognition	Physical frailty, visual dysfunction
Precuneus Cortex	Self-processing function, spatio-visual memory	Cognitive impairment, executive dysfunction, memory deficits

The last case to be analyzed was the false positive case where the Parietal and the Fronto-Temporal regions were focused. The regions covered included the *Temporal Fusiform Cortex – Anterior division, Lateral Occipital Cortex – Superior division, Percuneous Cortex, Cuneal Cortex, Hippocampus, Superior Parietal Lobule, Postcentral Gyrus, Amygdala, Precentral Gyrus, Cerebral Cortex, and Cerebral White Matter*. From the Grad-CAM heat map (Fig. 7), it is clear that the regions concentrated on by the CNN in the True Negative is very different from what is focused on in the False Positive case, leading to wrong classification.

A description of the biological significance and clinical relevance in PD, of the ROIs seen in the true positive case is given in the following subsection.

4.3.1. Biological function and effect due to PD of the ROIs identified in the true positive case

A summary of the ROIs, their function and effect of degeneration of the regions in PD are given in Table 7. Two of the common non-motor precursors in the prodromal stage of PD is hyposmia and anxiety or depression [90,5]. The *amygdala* is an almond shaped brain region that helps regulate emotions and also plays an important role in autonomic and endocrine functions [117]. Amygdala dysfunction in PD has been documented in several studies and is a contributor to the non-motor symptoms of hyposmia and anxiety/ depression [136,59,17]. Further, PD specific lesions is also seen in the amygdala-dependent structures [16,17].

The *putamen* is a part of the basal ganglia region in the brain, and regulates motor functions, learning, and cognitive tasks [46]. The atrophy of the putamen and decreased activity in the region has been noted in a number of studies. Being part of the striatum of the basal ganglia, the degeneration of the putamen plays a significant role in the motor symptoms in PD [31,137,70,115]. The *pallidum* is a subcortical structure that is part of the basal ganglia. The pallidum plays a key role in motivated behavior and motor control [81]. Dopaminergic degeneration in the pallidum due to PD leads to motor symptoms, apathy and impulse control disorders [100,12,90,110].

The *brainstem* is a structure that connects the cerebrum to the cerebellum and the spinal cord. The brainstem is composed of three distinct structures – the midbrain, pons and medulla oblongata; and regulates autonomic and sensory functions, and coordinates balance and bodily reflexes [11]. The midbrain region contains the Substantia Nigra, a primary location of dopaminergic neurons in the brain. The degeneration and depigmentation of the brainstem is thus directly linked to most of PD symptoms. Researchers have also indicated significant changes in the pons and medulla oblongata region in PD, compared to normal subjects [8,112,66].The *cerebral cortex* is the chief nervous system control center in the brain that regulates voluntary muscle movement and cognition, including memory, speech language, emotional regulation, decision making, and learning [62,16]. The degeneration of the cerebral cortex in PD as a result of the progression of the disease from the thalamus has been indicated to contribute towards tremors, akinesia and cognitive decline [16,21,40,95].

The *hippocampus* is a temporal lobe and limbic system structure that plays a key role in the memory processes in the brain. The hippocampus is connected to the olfactory regions and also plays a role in olfaction [6,43,149]. Studies report that changes in the hippocampus cause memory deficits and olfactory disturbances in people with PD [6,76,42]. The *temporal pole* is a region of the brain responsible for many complex, high-level cognitive functions such as semantic information processing, visual recognition, and socio-emotional processing [54]. The temporal pole is a part of the limbic system, along with the hippocampus and the amygdala, and in PD it causes difficulties in emotion-based action regulation and non-motor symptoms such as rapid eye movement sleep behavior disorder and hyposmia [97,52].

The *parahippocampal gyrus* is a brain structure that lies adjacent to the hippocampus and regulated visuo-spatial and episodic memory. The structure also contains the primary olfactory cortex [4,75]. In degeneration of the parahippocampal gyrus in PD has been associated with memory impairments and olfactory disturbances [26,15]. The *cerebral white matter* is composed of nerve bundles that play a crucial role in information processing and cognition [41,103]. Studies have reported a reduction in the white matter volume in PD and the changes due to white matter degeneration has been associated with further deterioration of motor functions in the body due to PD [145,141,89]. The *insular cortex* is a part of the brain that regulates higher cognitive functions including sensori-motor processing, vestibular functions, somatosensory processing, and audio processing, among others [130,73]. The insula is an integrating hub of many brain functions and in PD, it is reported to contribute to the non-motor symptoms and affect cognitive functions as well [28,27,37].

The *frontal pole* is a part of the prefrontal cortex of the brain that regulates cognition, and organized behavior such as planning and executing tasks [14,72]. The impairment of the frontal pole due to PD results in disruption of executive functioning that relies on coordinating and controlling behavior through cognition [34,68]. The *frontal operculum cortex* is a region of the brain linked to linguistic – language, speech, and musical abilities, cognition and somatosensory functions [131]. In PD, the degeneration of the frontal operculum cortex results in non-motor symptoms such as voice impairments, and cognitive disturbances [124,71]. The *Lateral Occipital Cortex – Superior Division* is a region of the brain responsible for object recognition [51,67]. In PD, the lateral operculum cortex's degeneration leads to physical frailty, visual dysfunction – difficulty discerning overlapping images, motion perception, difficulty remembering and recalling complex images, and cognitive functions [138,24,44].

The *precuneus cortex* is a region of the brain that plays a central role in higher cognitive abilities such as self-processing functions, episodic memory and visuo-spatial imagery and memory. The precuneus cortex is a part of the default mode network and plays a significant role in consciousness [20,132]. Degeneration of the precuneus cortex result in cognitive impairment, executive dysfunction and memory deficits [65,45,35].

The Grad-CAM heatmaps show that the regions identified by the CNN in the correctly classified PD class are in line with findings from previous studies, indicating the validity of the model as a decision support system.

4.4. Discussion

This work attempted to answer four research questions identified from previous works (Section 2) on MRI based classification of PD by CNNs:

1. Which architecture trained by transfer learning is most suited for the task of classifying PD using T1 weighted MRI?
2. Methods for MRI slice selection and whether a common range of slices can be selected through multiple methods?
3. How well the developed CNN models trained on a particular dataset does generalizes to an independent test set?
4. What is the effect of slice-level and longitudinal data leakage in hold-out test sets when classifying PD using T1 weighted MRI?

To address the first question, 12 diverse CNN models were evaluated for ability to learn the distinguishing characteristics of T1-weighted MRI for the classification of PD through the TL strategy. All the CNN models were pretrained on the ImageNet dataset. Based on the performance of the selected CNN models on ImageNet, the best performance in the order of highest scores (Table 4 Top-3 and Top-5) was expected to be by NASNet Large, Inception ResNet, Xception and ResNet152. However, when trained on the PPMI dataset and validated through a five-fold CV strategy, VGG16, VGG19 and ResNet-50 were the top performing models. In the accuracy score on the ILSVRC, the VGG models had the lowest scores. The VGG models were also the best models identified from another study using T2 weighted MRI [133], highlighting their suitability in learning the subtle details of the structural changes in the brain seen in the MRI due to PD. The results indicate that the selection of the CNN architecture based on the performance of the model on another task would be sub-optimal. The choice of the CNN architecture has to be made after testing the performance of the models in the target domain. In this case, three models - ResNet50, VGG16 and VGG19 were identified as good contenders to classify PD using MRI. Further investigation of the models by testing them on external datasets were carried out after slice selection.

The slice selection has been shown to improve the performance of the classifier as it removes data that may not be relevant to the problem [77]. Given the diversity in the slice selection in literature, this study presents an objective analysis of two slice selection strategies - entropy-based method and pixel intensity based method as a solution to the second research question. Using the common range of MRI slices from both methods, the number of slices of MRI was reduced from 182 to 87 slices. Using an overlay of the standard HO cortical and subcortical atlas shown in Fig. 5, it was verified that the slices selected contained all the key anatomical structures associated with PD. This number was fixed for further hyper-parameter tuning of the three top performing CNN models identified through CV. The final trained CNN models were tested across various data leakage scenarios and the generalizability of the models were evaluated using two independent datasets.

Despite the proven efficacy of data driven systems, the clinical integration of the models are deterred by questions of generalizability to heterogeneous populations and external patient cohorts [9,134]. In this study, the generalizability of the Top-3 selected architectures are tested on the hold-out data from the PPMI dataset, and the two external datasets - Tao Wu and NEUROCRON. This also forms the first case of the data leakage study - '*No leakage*' or '*Subject-wise split*'. When tested within the same dataset, the ResNet50 model has the highest scores of accuracy and F1 score. However, upon testing the trained models on the Tao Wu and NEUROCRON datasets, the performance of the ResNet50 model dips. The performance of the VGG19 models on the external datasets is on par with the generalization results presented in [19] and [30]. The VGG19 model is more generalizable than the other two architectures considered.

Having established the baseline performance of the developed CNN models within the training dataset and on external datasets, the effect of data leakage, which forms the fourth question, was investigated. Within the same dataset, the slice level and longitudinal data leakage showed an inflation in the accuracy values of above 67% and 30%, respectively. The results highlight how much more optimistic the projected erroneous results could be. Three previous studies have tested the effect of data leakage in MRI based classification of PD. Madan et al. [77] examined the effect of slice level data leakage in T2 weighted MRI and reported up to 30% increase in accuracy. Yagis et al. [144] also evaluated T2 weighted MRI and reported over optimization of up to 26%. Yagis et al. [143] examined the effect of data leakage across the CV folds with slice level data leakage in T1 weighted MRI. The authors report up to 55% inflation of accuracy values on two datasets. The present study presents the results of slice level data leakage using a hold-out test set, as opposed to the CV validation split presented by Yagis et al. [143] and is the first study to address this aspect of data leakage. The results show that while CV based leakage results in identity confounding, at 67%, slice level data leakage using a hold-out partition can be even more dangerous. Further, to the understanding of the authors', this is the first work to evaluate the effect of longitudinal data leakage in PD classification using MRI.

The results from the data leakage study suggest that such erroneous models if deployed in real time will fail and there is extreme necessity for researchers to exercise caution at each stage of data handling. Without the evaluation of the generalizability of the model, the results would not be a projection of the actual performance of the model. When the models with data leakage were tested for generalizability to external datasets, it was seen that the scores reduced drastically with slice-wise split, but relatively less reduction was seen with longitudinal data splitting. This highlights how independent testing of the developed models can be beneficial to catch possible errors overlooked during the training phase. The best approach for handling any such implicit bias would be to split the data at the subject level, after ensuring that longitudinal data of each individual is also taken together at each fold before CV. Data augmentation should be carried out on just the training data after splitting the data at the subject level [140]. Based on the three data leakage tests across the three datasets, the VGG19 model still emerges as the robust architecture for PD classification.

Finally, Grad-CAM based heat maps were used to visualize the ROIs identified by the CNN. The regions identified included the *Amygdala, Putamen, Pallidum, Brainstem, Cerebral Cortex, Hippocampus, Temporal Pole, Parahippocampal Gyrus, Cerebral White Matter, Insular Cortex, Frontal Pole, Frontal Operculum Cortex, Frontal Orbital Cortex, Temporal Occipital Fusiform Cortex, Lateral Occipital Cortex -Superior Division, and Precuneus Cortex*. The ROIs play key roles in characteristic motor and non-motors symptoms seen in PD (see Subsection 4.3.1 and Table 7 for further details). The regions identified are in line with those reported by previous studies on PD [111,60] and they align with the Braak staging of the atrophy due to PD ([18,16]), indicating the validity of the model in classifying PD from T1-weighted MRI.

Overall, based on the results from the experiments carried out, VGG19 emerges as the architecture of choice for the classification of PD using T1 weighted MRI. The data leakage studies highlight the possible bias propagation that can occur in data driven systems. While testing within the dataset showed that selection of transfer learning

CNN architectures based on the source domain is erroneous, testing on independent datasets revealed the architecture that could adapt to external patient cohorts. Testing the generalizability of the models to external datasets showed adverse results in the data leakage cases, further highlighting the need for independent test sets. The results of the CNN were visualized using a Grad-CAM heatmap overlaid on the MRI to enable explainability of the prediction. The results indicated key areas that have been shown to be affected in PD and was in line with expected neurodegeneration patterns in PD.

5. Conclusion

This study focused on the identification of bias propagation in T1 weighted MRI data driven classification of PD by simulating data leakage and testing for generalizability. The study addressed the choice of CNN architectures using transfer learning by evaluating 12 diverse CNN models. Results indicated that choosing based on characteristics other than performance on the target domain is sub-optimal. An empirical analysis of two slice selection strategies was carried out to arrive at a common range of MRI slices and this was used for further testing with the top 3 architectures - VGG16, VGG19 and ResNet50. The top models were tested on PPMI, Tao Wu and NEUROCRON datasets. The results show that testing the generalizability of the models is as important as cross-validation in ensuring that there are no unforeseen errors in the data handling or experimentation stages.

Results show that the VGG19 model had the most consistent performance across all datasets and data leakage cases. Further, it is seen that the over optimistic outputs due to slice level data leakage in hold out test sets could be as high as 67% and this was a potential risk in models, especially if not tested for generalization. The best approach for handling any such implicit bias would be to split the data at the subject level, after ensuring that longitudinal data of each individual is also taken together at each fold before CV. The XAI based analysis using a Grad-CAM visualization of the CNN output was able to identify regions relevant to the disease.

The results of the study reveal that the usage transfer learning to train VGG19 with T1 weighted MRI for the classification of PD has consistent, generalizable performance. The potential implications of bias in data driven systems due to data leakage, the lack of generalizability and non-explainability is well documented. In a step towards building more trustworthy AI models, this study documents the potential sources of data leakage and how to avoid it (splitting the data at the subject level). The analysis of data leakage and testing on independent datasets highlight the necessity of testing for generalizability which could act as a safety net to catch overlooked data handling or experimental errors. Further, the Grad-CAM based XAI interpretation validates the efficacy of the proposed method for use as a diagnostic decision support tool for PD classification using T1-weighted MRI.

Some limitations of the work include that only the T1 weighted MRI was considered. Inclusion of T2 weighted or Flair MRI could potentially improve the accuracy of the model and offer a different view of the data. The present study involved training on the PPMI dataset and testing on PPMI, Tao Wu and NEUROCRON datasets. Inclusion of further multi-center datasets in the training of the CNN models is a future direction. The inclusion of additional datasets with scans under varied conditions would potentially improve the performance of the data driven model. In this study, the modality was limited to MRI and inclusion of other precursory modalities such as gait or speech to build a multi-modal classifier would be beneficial. The study was conducted with the most widely available imaging sequence with possibility of verifying the generalizable nature of the developed CNN. Future directions include the prediction of the disease progression in order to aid the prognosis and treatment of the disease.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Iswarya Kannoth Veetil reports financial support was provided by Council of Scientific and Industrial Research (CSIR) Human Resource Development Group. The funding source did not influence the results reported in this study. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in the preparation of this article was obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data). For up-to-date information on the study, visit ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors.

The Tao Wu and NEUROCRON datasets [10] used in the preparation of this article is available publically and was downloaded from fcon_1000.projects.nitrc.org/indi/retro/parkinsons.html.

Acknowledgment

KVI acknowledges the financial support received from the Human Resource Development Group (HRDG), Council of Scientific and Industrial Research (CSIR), India under the Senior Research Fellowship scheme (09/1034(0007)/2020-EMR-I).

The authors would like to thank Ritwik Murali, Premjith B., Suchithra K. S. and Sajith Variyar V. V. for their help while finalizing the deep learning workflow used in the study.

The authors express their gratitude to the editor and the reviewers for their useful comments and suggestions in improving the quality of this research work.

References

[1] E. Adeli, F. Shi, L. An, C.Y. Wee, G. Wu, T. Wang, D. Shen, Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data, NeuroImage 141 (2016) 206–219.

[2] E. Adeli, G. Wu, B. Saghafi, L. An, F. Shi, D. Shen, Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson's disease, Sci. Rep. 7 (2017) 41069.

[3] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, P. Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020) 1–9.

[4] E.M. Aminoff, K. Kveraga, M. Bar, The role of the parahippocampal cortex in cognition, Trends Cogn. Sci. 17 (2013) 379–390.

[5] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, S. Tangaro, Complex networks reveal early MRI markers of Parkinson's disease, Med. Image Anal. 48 (2018) 12–24.

[6] K.S. Anand, V. Dhikav, Hippocampus in health and disease: an overview, Ann. Indian Acad. Neurol. 15 (2012) 239.

[7] J.L. Andersson, M. Jenkinson, S. Smith, et al., Non-linear registration, aka spatial normalisation, FMRIB technical report TR07JA2, FMRIB Analysis Group of the University of Oxford, 2007, vol. 2, e21.

[8] G. Arribarat, A. De Barros, P. Péran, Modern brainstem MRI techniques for the diagnosis of Parkinson's disease and parkinsonisms, Front. Neurol. 11 (2020) 791.

[9] O. Asan, A.E. Bayrak, A. Choudhury, Artificial intelligence and human trust in healthcare: focus on clinicians, J. Med. Internet Res. 22 (2020) e15154.

[10] L. Badea, M. Onu, T. Wu, A. Roceanu, O. Bajenaru, Exploring the reproducibility of functional connectivity alterations in Parkinson's disease, PLoS ONE 12 (2017) e0188196.

[11] H. Basinger, J.P. Hogg, Neuroanatomy, brainstem, in: StatPearls [Internet], StatPearls Publishing, 2019, pp. 1–3.

[12] M. Béreau, V. Van Waes, M. Servant, E. Magnin, L. Tatu, M. Anheim, Apathy in Parkinson's disease: clinical patterns and neurobiological basis, Cells 12 (2023) 1599.

[13] A. Bhan, S. Kapoor, M. Gulati, Diagnosing Parkinson's disease in early stages using image enhancement, ROI extraction and deep learning algorithms, in: 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), IEEE, 2021, pp. 521–525.

[14] S. Bludau, S.B. Eickhoff, H. Mohlberg, S. Caspers, A.R. Laird, P.T. Fox, A. Schleicher, K. Zilles, K. Amunts, Cytoarchitecture, probability maps and functions of the human frontal pole, NeuroImage 93 (2014) 260–275.

[15] N.I. Bohnen, S. Gedela, P. Herath, G.M. Constantine, R.Y. Moore, Selective hyposmia in Parkinson disease: association with hippocampal dopamine activity, Neurosci. Lett. 447 (2008) 12–16.

[16] H. Braak, E. Braak, D. Yilmazer, R. De Vos, E. Jansen, J. Bohl, Pattern of brain destruction in Parkinson's and Alzheimer's diseases, J. Neural Transm. 103 (1996) 455–490.

[17] H. Braak, E. Braak, D. Yilmazer, R.A. de Vos, E.N. Jansen, J. Bohl, K. Jellinger, Amygdala pathology in Parkinson's disease, Acta Neuropathol. 88 (1994) 493–500.

[18] H. Braak, K. Del Tredici, U. Rüb, R.A. De Vos, E.N.J. Steur, E. Braak, Staging of brain pathology related to sporadic Parkinson's disease, Neurobiol. Aging 24 (2003) 197–211.

[19] M. Camacho, M. Wilms, P. Mouches, H. Almgren, R. Souza, R. Camicioli, Z. Ismail, O. Monchi, N.D. Forkert, Explainable classification of Parkinson's disease using deep learning trained on a large multi-center database of T1-weighted MRI datasets, NeuroImage Clin. 38 (2023) 103405.

[20] A.E. Cavanna, M.R. Trimble, The precuneus: a review of its functional anatomy and behavioural correlates, Brain 129 (2006) 564–583.

[21] D. Cechetto, M. Jog, Parkinson's disease and the cerebral cortex, in: The Cerebral Cortex in Neurodegenerative and Neuropsychiatric Disorders, Elsevier, 2017, pp. 177–193.

[22] E. Chaibub Neto, A. Pratap, T.M. Perumal, M. Tummalacherla, P. Snyder, B.M. Bot, A.D. Trister, S.H. Friend, L. Mangravite, L. Omberg, Detecting the impact of subject characteristics on machine learning-based diagnostic applications, npj Digit. Med. 2 (2019) 99.

[23] S. Chakraborty, S. Aich, H.C. Kim, Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network, Diagnostics 10 (2020) 402.

[24] Y.S. Chen, H.L. Chen, C.H. Lu, M.H. Chen, K.H. Chou, N.W. Tsai, C.C. Yu, P.L. Chiang, W.C. Lin, Reduced lateral occipital gray matter volume is associated with physical frailty and cognitive impairment in Parkinson's disease, Eur. Radiol. 29 (2019) 2659–2668.

[25] F. Chollet, Xception: deep learning with depthwise separable convolutions, arXiv preprint, arXiv:1610.02357, 2016.

[26] L. Christopher, S. Duff-Canning, Y. Koshimori, B. Segura, I. Boileau, R. Chen, A.E. Lang, S. Houle, P. Rusjan, A.P. Strafella, Salience network and parahippocampal dopamine dysfunction in memory-impaired Parkinson disease, Ann. Neurol. 77 (2015) 269–280.

[27] L. Christopher, Y. Koshimori, A.E. Lang, M. Criaud, A.P. Strafella, Uncovering the role of the insula in non-motor symptoms of Parkinson's disease, Brain 137 (2014) 2143–2154.

[28] M. Criaud, L. Christopher, P. Boulinguez, B. Ballanger, A.E. Lang, S.S. Cho, S. Houle, A.P. Strafella, Contribution of insula in Parkinson's disease: a quantitative meta-analysis study, Hum. Brain Mapp. 37 (2016) 1375–1392.

[29] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[30] N.J. Dhinagar, S.I. Thomopoulos, C. Owens-Walton, D. Stripelis, J.L. Ambite, G. Ver Steeg, D. Weintraub, P. Cook, C. McMillan, P.M. Thompson, 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, 2021, pp. 277–286.

[31] D.W. Dickson, Parkinson's disease and parkinsonism: neuropathology, Cold Spring Harb. Perspect. Med. 2 (2012).

[32] E. Dikici, X.V. Nguyen, N. Takacs, L.M. Prevedello, Prediction of model generalizability for unseen data: methodology and case study in brain metastases detection in T1-weighted contrast-enhanced 3D MRI, Comput. Biol. Med. 159 (2023) 106901.

[33] E.R. Dorsey, A. Elbaz, E. Nichols, N. Abbasi, F. Abd-Allah, A. Abdelalim, J.C. Adsuar, M.G. Ansha, C. Brayne, J.Y.J. Choi, et al., Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016, Lancet Neurol. 17 (2018) 939–953.

[34] L.L. Drag, L.A. Bieliauskas, A.W. Kaszniak, N.I. Bohnen, E.L. Glisky, Source memory and frontal functioning in Parkinson's disease, J. Int. Neuropsychol. Soc. 15 (2009) 399–406.

[35] P. Dušek, R. Jech, T. Sieger, J. Vymazal, E. Růžička, J. Wackermann, K. Mueller, Abnormal activity in the precuneus during time perception in Parkinson's disease: an fMRI study, PLoS ONE 7 (2012) e29635.

[36] S. Esmaeilzadeh, Y. Yang, E. Adeli, End-to-end Parkinson disease diagnosis using brain MR-images by 3D-CNN, arXiv preprint, arXiv:1806.05233, 2018.

[37] Y.Y. Fathy, D.H. Hepp, F.J. De Jong, J.J. Geurts, E.M. Foncke, H.W. Berendse, W.D. van de Berg, M.M. Schoonheim, Anterior insular network disconnection and cognitive impairment in Parkinson's disease, NeuroImage Clin. 28 (2020) 102364.

[38] A. Favaro, Y.T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, L. Moro-Velazquez, Interpretable speech features vs. DNN embeddings: what to use in the automatic assessment of Parkinson's disease in multi-lingual scenarios, medRxiv (2023), 2023–05.

[39] S.M. Fereshtehnejad, C. Yao, A. Pelletier, J.Y. Montplaisir, J.F. Gagnon, R.B. Postuma, Evolution of prodromal Parkinson's disease and dementia with Lewy bodies: a prospective study, Brain 142 (2019) 2051–2067.

[40] I. Ferrer, Early involvement of the cerebral cortex in Parkinson's disease: convergence of multiple metabolic defects, Prog. Neurobiol. 88 (2009) 89–103.

[41] C.M. Filley, R.D. Fields, White matter and cognition: making the connection, J. Neurophysiol. 116 (2016) 2093–2104.

[42] H. Foo, E. Mak, R.J. Chander, A. Ng, W.L. Au, Y.Y. Sitoh, L.C. Tan, N. Kandiah, Associations of hippocampal subfields in the progression of cognitive decline related to Parkinson's disease, NeuroImage Clin. 14 (2017) 37–42.

[43] N.J. Fortin, K.L. Agster, H.B. Eichenbaum, Critical role of the hippocampus in memory for sequences of events, Nat. Neurosci. 5 (2002) 458–462.

[44] L.A. Frizon, R. Gopalakrishnan, O. Hogue, D. Floden, S.J. Nagel, K.B. Baker, G.R. Isolan, M.A. Stefani, A.G. Machado, Cortical thickness in visuo-motor areas is related to motor outcomes after STN DBS for Parkinson's disease, Parkinsonism Relat. Disord. 71 (2020) 17–22.

[45] L.l. Gao, T. Wu, The study of brain functional connectivity in Parkinson's disease, Transl. Neurodegener. 5 (2016) 1–7.

[46] M. Ghandili, S. Munakomi, Neuroanatomy, putamen, in: StatPearls [Internet], StatPearls Publishing, 2023, pp. 1–3.

[47] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results, Mov. Disord. Offic. J. Mov. Disord. Soc. 23 (2008) 2129–2170.

[48] H. Greenspan, B. Van Ginneken, R.M. Summers, Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique, IEEE Trans. Med. Imaging 35 (2016) 1153–1159.

[49] D.N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration, NeuroImage 48 (2009) 63–72.

[50] O. Grigas, R. Maskeliūnas, R. Damaševičius, Improving structural MRI preprocessing with hybrid transformer GANs, Life 13 (2023) 1893.

[51] K. Grill-Spector, Z. Kourtzi, N. Kanwisher, The lateral occipital complex and its role in object recognition, Vis. Res. 41 (2001) 1409–1422.

[52] T. Guo, X. Guan, Q. Zeng, M. Xuan, Q. Gu, P. Huang, X. Xu, M. Zhang, Alterations of brain structural network in Parkinson's disease with and without rapid eye movement sleep behavior disorder, Front. Neurol. 9 (2018) 334.

[53] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, corr, arXiv:1512.03385 [abs], 2015.

[54] B. Herlin, V. Navarro, S. Dupont, The temporal pole: from anatomy to function—a literature appraisal, J. Chem. Neuroanatom. 113 (2021) 101925.

[55] M. Hon, N.M. Khan, Towards Alzheimer's disease classification through transfer learning, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2017, pp. 1166–1169.

[56] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv preprint, arXiv:1704.04861, 2017.

[57] C.M. Huang, R. Doole, C.W. Wu, H.W. Huang, Y.P. Chao, Culture-related and individual differences in regional brain volumes: a cross-cultural voxel-based morphometry study, Front. Human Neurosci. 13 (2019) 313.

[58] G. Huang, Z. Liu, K.Q. Weinberger, Densely connected convolutional networks, corr, arXiv:1608.06993 [abs], 2016.

[59] P. Huang, M. Xuan, Q. Gu, X. Yu, X. Xu, W. Luo, M. Zhang, Abnormal amygdala function in Parkinson's disease patients and its relationship to depression, J. Affect. Disord. 183 (2015) 263–268.

[60] N. Ibarretxe-Bilbao, C. Junque, B. Segura, H.C. Baggio, M.J. Marti, F. Valldeoriola, N. Bargallo, E. Tolosa, Progression of cortical thinning in early Parkinson's disease, Mov. Disord. 27 (2012) 1746–1753.

[61] A. Iranzo, Dissecting premotor Parkinson's disease with multimodality neuroimaging, Lancet Neurol. 17 (2018) 574–576.

[62] K.H. Jawabri, S. Sharma, Physiology, cerebral cortex functions, in: StatPearls [Internet], StatPearls Publishing, 2019, pp. 1–3.

[63] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, NeuroImage 17 (2002) 825–841.

[64] M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images, Med. Image Anal. 5 (2001) 143–156.

[65] X. Jia, Y. Li, K. Li, P. Liang, X. Fu, Precuneus dysfunction in Parkinson's disease with mild cognitive impairment, Front. Aging Neurosci. 10 (2019) 427.

[66] T. Jubault, S.M. Brambati, C. Degroot, B. Kullmann, A.P. Strafella, A.L. Lafontaine, S. Chouinard, O. Monchi, Regional brain stem atrophy in idiopathic Parkinson's disease detected by anatomical MRI, PLoS ONE 4 (2009) e8247.

[67] A. Karten, S.P. Pantazatos, D. Khalil, X. Zhang, J. Hirsch, Dynamic coupling between the lateral occipital-cortex, default-mode, and frontoparietal networks during bistable perception, Brain Connect. 3 (2013) 286–293.

[68] A.K. Kendi, S. Lehericy, M. Luciana, K. Ugurbil, P. Tuite, Altered diffusion in the frontal lobe in Parkinson disease, Am. J. Neuroradiol. 29 (2008) 501–505.

[69] H.E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M.E. Maros, T. Ganslandt, Transfer learning for medical image classification: a literature review, BMC Med. Imaging 22 (2022) 69.

[70] K. Kinoshita, T. Kuge, Y. Hara, K. Mekata, Putamen atrophy is a possible clinical evaluation index for Parkinson's disease using human brain magnetic resonance imaging, J. Imag. 8 (2022) 299.

[71] P. Klobusiakova, J. Mekyska, L. Brabenec, Z. Galaz, V. Zvoncak, J. Mucha, S.Z. Rapcsak, I. Rektorova, Articulatory network reorganization in Parkinson's disease as assessed by multimodal MRI and acoustic measures, Parkinsonism Relat. Disord. 84 (2021) 122–128.

[72] E. Koechlin, Frontal pole function: what is specifically human?, Trends Cogn. Sci. 15 (2011) 241.

[73] M.W. Kortz, K.O. Lillehei, Insular cortex, in: StatPearls [Internet], StatPearls Publishing, 2021, pp. 1–3.

[74] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[75] C.H. Lew, K. Semendeferi, Evolutionary specializations of the human limbic system, Evol. Nerv. Syst. 4 (2017) 277–291, https://api.semanticscholar.org/CorpusID:152223771.

[76] L.E. Llewelyn, M. Kornisch, H. Park, T. Ikuta, Hippocampal functional connectivity in Parkinson's disease, Neurodegener. Dis. 22 (2022) 29–33.

[77] Y. Madan, I.K. Veetil, V. Sowmya, E. Gopalakrishnan, K. Soman, Deep learning-based approach for Parkinson's disease detection using region of interest, in: Intelligent Sustainable Systems: Proceedings of ICISS 2021, Springer, 2021, pp. 1–13.

[78] Y. Madan, I.K. Veetil, V. Sowmya, E.A. Gopalakrishnan, Synthetic data augmentation of MRI using generative variational autoencoder for Parkinson's disease detection, in: Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021), Springer, 2022, pp. 171–178.

[79] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al., The Parkinson progression marker initiative (ppmi), Prog. Neurobiol. 95 (2011) 629–635.

[80] R. Maskeliūnas, R. Damaševičius, A. Kulikajevas, E. Padervinskis, K. Pribuišis, V. Uloza, A hybrid U-lossian deep learning network for screening and evaluating Parkinson's disease, Appl. Sci. 12 (2022) 11601.

[81] L.E. Mello, J. Villares, Neuroanatomy of the basal ganglia, Psychiatr. Clin. North Am. 20 (1997) 691–704.

[82] M. Mittermaier, M.M. Raza, J.C. Kvedar, Bias in AI-based models for medical applications: challenges and mitigation strategies, npj Digit. Med. 6 (2023) 113.

[83] D.A. Morales, Y. Vives-Gilabert, B. Gómez-Ansón, E. Bengoetxea, P. Larrañaga, C. Bielza, J. Pagonabarraga, J. Kulisevsky, I. Corcuera-Solano, M. Delfino, Predicting dementia development in Parkinson's disease using Bayesian network classifiers, Psychiatry Res. Neuroimaging 213 (2013) 92–98.

[84] T.A. Mostafa, I. Cheng, Parkinson's disease detection using ensemble architecture from MR images, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020, pp. 987–992.

[85] K.P. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, 2012.

[86] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, M. Wilson, S.M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui, et al., Supervised transfer learning at scale for medical imaging, arXiv preprint, arXiv:2101.05913, 2021.

[87] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.

[88] NHS-England, NHS-Improvement, Diagnostic Imaging Dataset Statistical Release, Department of Health, London, 2023 [Online; accessed 27-December-2023].

[89] C.O. Nyatega, L. Qiang, M.J. Adamu, H.B. Kawuwa, Gray matter, white matter and cerebrospinal fluid abnormalities in Parkinson's disease: a voxel-based morphometry study, Front. Psychiatry 13 (2022) 1027907.

[90] J. Obeso, M. Stamelou, C. Goetz, W. Poewe, A. Lang, D. Weintraub, D. Burn, G.M. Halliday, E. Bezard, S. Przedborski, et al., Past, present, and future of Parkinson's disease: a special essay on the 200th anniversary of the shaking palsy, Mov. Disord. 32 (2017) 1264–1310.

[91] M. Odusami, R. Maskeliūnas, R. Damaševičius, S. Misra, Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of pet and MRI images, J. Med. Biol. Eng. (2023) 1–12.

[92] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2009) 1345–1359.

[93] B. Peng, S. Wang, Z. Zhou, Y. Liu, B. Tong, T. Zhang, Y. Dai, A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease, Neurosci. Lett. 651 (2017) 88–94.

[94] H.R. Pereira, H.A. Ferreira, Classification of patients with Parkinson's disease using medical imaging and artificial intelligence algorithms, in: XV Mediterranean Conference on Medical and Biological Engineering and Computing–MEDICON 2019: Proceedings of MEDICON 2019, September 26-28, 2019, Coimbra, Portugal, Springer, 2020, pp. 2043–2056.

[95] C. Pletcher, K. Dabbs, A. Barzgari, V. Pozorski, M. Haebig, S. Wey, S. Krislov, F. Theisen, O. Okonkwo, P. Cary, et al., Cerebral cortical thickness and cognitive decline in Parkinson's disease, Cereb. Cortex Commun. 4 (2023) tgac044.

[96] W. Poewe, K. Seppi, C.M. Tanner, G.M. Halliday, P. Brundin, J. Volkmann, A.E. Schrag, A.E. Lang, Parkinson disease, Nat. Rev. Dis. Primers 3 (2017) 1–21.

[97] A.R. Potgieser, A. van der Hoorn, A.M. Meppelink, L.K. Teune, J. Koerts, B.M. de Jong, Anterior temporal atrophy and posterior progression in patients with Parkinson's disease, Neurodegener. Dis. 14 (2014) 125–132.

[98] M. Raghu, C. Zhang, J. Kleinberg, S. Bengio, Transfusion: understanding transfer learning for medical imaging, Adv. Neural Inf. Process. Syst. 32 (2019).

[99] K. Rajanbabu, I.K. Veetil, V. Sowmya, E. Gopalakrishnan, K. Soman, Ensemble of deep transfer learning models for Parkinson's disease classification, in: Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, vol. 2, Springer, 2022, pp. 135–143.

[100] A. Rajput, H. Sitte, A. Rajput, M. Fenton, C. Pifl, O. Hornykiewicz, Globus pallidus dopamine and Parkinson motor subtypes: clinical and brain biochemical correlation, Neurology 70 (2008) 1403–1410.

[101] J. Ramya, B.U. Maheswari, M. Rajakumar, R. Sonia, Alzheimer's disease segmentation and classification on MRI brain images using enhanced expectation maximization adaptive histogram (EEM-AH) and machine learning, Inf. Technol. Control 51 (2022) 786–800.

[102] S. Raschka, Python Machine Learning, Packt Publishing Ltd., 2015.

[103] R.E. Roberts, E.J. Anderson, M. Husain, White matter microstructure and cognitive function, Neuroscientist 19 (2013) 8–15.

[104] D.J. Rumala, How you split matters: data leakage and subject characteristics studies in longitudinal brain MRI analysis, in: Workshop on Clinical Image-Based Procedures, Springer, 2023, pp. 235–245.

[105] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[106] S. Saeb, L. Lonini, A. Jayaraman, D.C. Mohr, K.P. Kording, The need to approximate the use-case in clinical machine learning, GigaScience 6 (2017), gix019.

[107] K. Sakai, K. Yamada, Machine learning studies on major brain diseases: 5-year trends of 2014–2018, Jpn. J. Radiol. 37 (2019) 34–72.

[108] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M. Gilardi, A. Quattrone, Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy, J. Neurosci. Methods 222 (2014) 230–237.

[109] S. Sangeetha, K. Baskar, P. Kalaivaani, T. Kumaravel, Deep learning-based early Parkinson's disease detection from brain MRI image, in: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2023, pp. 490–495.

[110] G. Santangelo, P. Barone, L. Trojano, C. Vitale, Pathological gambling in Parkinson's disease. a comprehensive review, Parkinsonism Relat. Disord. 19 (2013) 645–653.

[111] E. Sarasso, F. Agosta, N. Piramide, M. Filippi, Progression of grey and white matter brain damage in Parkinson's disease: a critical review of structural MRI literature, J. Neurol. 268 (2021) 3144–3179.

[112] S.T. Schwarz, Y. Xing, P. Tomar, N. Bajaj, D.P. Auer, In vivo assessment of brainstem depigmentation in Parkinson disease: potential as a severity marker for multicenter studies, Radiology 283 (2017) 789–798.

[113] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[114] P.M. Shah, A. Zeb, U. Shafi, S.F.A. Zaidi, M.A. Shah, Detection of Parkinson disease in brain MRI using convolutional neural network, in: 2018 24th International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1–6.

[115] B. Shen, Y. Pan, X. Jiang, Z. Wu, J. Zhu, J. Dong, W. Zhang, P. Xu, Y. Dai, Y. Gao, et al., Altered putamen and cerebellum connectivity among different subtypes of Parkinson's disease, CNS Neurosci. Ther. 26 (2020) 207–214.

[116] L.C. Silbert, J. Kaye, Neuroimaging and cognition in Parkinson's disease dementia, Brain Pathol. 20 (2010) 646–653.

[117] G. Šimić, M. Tkalčić, V. Vukić, D. Mulc, E. Španić, M. Šagud, F.E. Olucha-Bordonau, M. Vukšić, P.R. Hof, Understanding emotions: origins and roles of the amygdala, Biomolecules 11 (2021) 823.

[118] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556, 2014.

[119] G. Singh, L. Samavedham, Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease, J. Neurosci. Methods 256 (2015) 30–40.

[120] S. Sivaranjini, C. Sujatha, Deep learning based diagnosis of Parkinson's disease using convolutional neural network, Multimed. Tools Appl. (2019) 1–13.

[121] S.M. Smith, Fast robust automated brain extraction, Hum. Brain Mapp. 17 (2002) 143–155.

[122] G. Solana-Lavalle, R. Rosas-Romero, Classification of ppmi MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease, Comput. Methods Programs Biomed. 198 (2021) 105793.

[123] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[124] H. Steurer, E. Schalling, E. Franzén, F. Albrecht, Characterization of mild and moderate dysarthria in Parkinson's disease: behavioral measures and neural correlates, Front. Aging Neurosci. 14 (2022) 870998.

[125] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 1–12.

[126] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[127] Y. Tang, L. Zhao, Y. Lou, Y. Shi, R. Fang, X. Lin, S. Liu, A. Toga, Brain structure differences between Chinese and Caucasian cohorts: a comprehensive morphometry study, Hum. Brain Mapp. 39 (2018) 2147–2155.

[128] L.C. Triarhou, Dopamine and Parkinson's disease, in: Madame Curie Bioscience Database [Internet], Landes Bioscience, 2013, pp. 1–13.

[129] N. Tustison, J. Gee, N4ITK: Nick's N3 ITK implementation for MRI bias field correction, Insight J. 9 (2009).

[130] L.Q. Uddin, J.S. Nomi, B. Hébert-Seropian, J. Ghaziri, O. Boucher, Structure and function of the human insula, J. Clin. Neurophysiol., Offic. Publ. Am. Electroencephalographic Soc. 34 (2017) 300.

[131] N. Unger, M. Haeck, S.B. Eickhoff, J.A. Camilleri, T. Dickscheid, H. Mohlberg, S. Bludau, S. Caspers, K. Amunts, Cytoarchitectonic mapping of the human frontal operculum—new correlates for a variety of brain functions, Front. Human Neurosci. 17 (2023).

[132] A.V. Utevsky, D.V. Smith, S.A. Huettel, Precuneus is a functional core of the default-mode network, J. Neurosci. 34 (2014) 932–940.

[133] I.K. Veetil, E. Gopalakrishnan, V. Sowmya, K. Soman, Parkinson's disease classification from magnetic resonance images (MRI) using deep transfer learned convolutional neural networks, in: 2021 IEEE 18th India Council International Conference (INDICON), IEEE, 2021, pp. 1–6.

[134] I.K. Veetil, V. Sowmya, J.R. Orozco-Arroyave, E. Gopalakrishnan, Robust language independent voice data driven Parkinson's disease detection, Eng. Appl. Artif. Intell. 129 (2024) 107494.

[135] T. Vyas, R. Yadav, C. Solanki, R. Darji, S. Desai, S. Tanwar, Deep learning-based scheme to diagnose Parkinson's disease, Expert Syst. 39 (2022) e12739.

[136] J. Wang, L. Sun, L. Chen, J. Sun, Y. Xie, D. Tian, L. Gao, D. Zhang, M. Xia, T. Wu, Common and distinct roles of amygdala subregional functional connectivity in non-motor symptoms of Parkinson's disease, npj Parkinson's Dis. 9 (2023) 28.

[137] J. Wang, J.R. Zhang, Y.F. Zang, T. Wu, Consistent decreased activity in the putamen in Parkinson's disease: a meta-analysis and an independent validation of resting-state FMRI, GigaScience 7 (2018), giy071.

[138] R.S. Weil, A.E. Schrag, J.D. Warren, S.J. Crutch, A.J. Lees, H.R. Morris, Visual dysfunction in Parkinson's disease, Brain 139 (2016) 2827–2843.

[139] C.P. Weingarten, M.H. Sundman, P. Hickey, N.k. Chen, Neuroimaging of Parkinson's disease: expanding views, Neurosci. Biobehav. Rev. 59 (2015) 16–52.

[140] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, et al., Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation, Med. Image Anal. 63 (2020) 101694.

[141] M.C. Wen, A. Ng, R.J. Chander, W.L. Au, L.C. Tan, N. Kandiah, Longitudinal brain volumetric changes and their predictive effects on cognition among cognitively asymptomatic patients with Parkinson's disease, Parkinsonism Relat. Disord. 21 (2015) 483–488.

[142] C. West, S. Soltaninejad, I. Cheng, Assessing the capability of deep-learning models in Parkinson's disease diagnosis, in: International Conference on Smart Multimedia, Springer, 2019, pp. 237–247.

[143] E. Yagis, S.W. Atnafu, A. García Seco de Herrera, C. Marzi, R. Scheda, M. Giannelli, C. Tessa, L. Citi, S. Diciotti, Effect of data leakage in brain MRI classification using 2D convolutional neural networks, Sci. Rep. 11 (2021) 22544.

[144] E. Yagis, A.G.S. De Herrera, L. Citi, Generalization performance of deep learning models in neurodegenerative disease classification, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 1692–1698.

[145] K. Yang, Z. Wu, J. Long, W. Li, X. Wang, N. Hu, X. Zhao, T. Sun, White matter changes in Parkinson's disease, npj Parkinson's Dis. 9 (2023) 150.

[146] M. Yang, X. Huang, L. Huang, G. Cai, Diagnosis of Parkinson's disease based on 3d resnet: the frontal lobe is crucial, Biomed. Signal Process. Control 85 (2023) 104904.

[147] D. Yin, Y. Zhao, Y. Wang, W. Zhao, X. Hu, Auxiliary diagnosis of heterogeneous data of Parkinson's disease based on improved convolution neural network, Multimed. Tools Appl. 79 (2020) 24199–24224.

[148] X. Zhang, D. Zhai, Y. Yang, Y. Zhang, C. Wang, A novel semi-supervised multiview clustering framework for screening Parkinson's disease, arXiv preprint, arXiv:2003.04760, 2020.

[149] G. Zhou, J.K. Olofsson, M.Z. Koubeissi, G. Menelaou, J. Rosenow, S.U. Schuele, P. Xu, J.L. Voss, G. Lane, C. Zelano, Human hippocampal connectivity is stronger in olfaction than other sensory systems, Prog. Neurobiol. 201 (2021) 102027.

[150] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8697–8710.

Iswarya Kannoth Veetil is currently a Council of Scientific and Industrial Research – Senior Research Fellow pursuing her PhD at Amrita Vishwa Vidyapeetham University, Coimbatore, India. She received her B.E. and M.E. degrees in Electrical and Electronics Engineering; and Power Electronics and Drives from Anna University, Tamil Nadu, India. Her research interests include Complex Systems, Medical Image & Signal Processing, Nonlinear Dynamics, Computational Neuroscience, Complex Networks, and Artificial Intelligence.

Divi Eswar Chowdary is currently enrolled at Amrita Vishwa Vidyapeetham University, pursuing his undergraduate degree. He is majoring in Computer Science with a specialization in Artificial Intelligence. His research interests include Machine learning, Deep Learning, and Generative AI. He is currently involved in research projects related to Efficient Deep Learning and Resource-Scarce Natural Language Processing.

Paleti Nikhil Chowdary is currently with the Amrita Vishwa Vidyapeetham University, pursuing his undergraduate degree. He is majoring in Computer Science with a specialization in Artificial Intelligence. His research interests include Reinforcement learning, machine learning and Generative AI. He is currently engaged in research projects related to controllable reinforcement learning through human interaction.

Sowmya V is working as an Assistant Professor (Selection Grade) in the School of Artificial Intelligence, Coimbatore. She completed PhD in Artificial Intelligence (AI) for Natural Scene Analysis from Amrita Vishwa Vidyapeetham. Her broad research area is AI for Signal and Image Analysis. The sub-areas of her research interests include Biomedical and Agriculture.

Gopalakrishnan E A obtained his Ph.D. from the Indian Institute of Technology, Madras, India, in the year 2016. He is currently the Principal of the Amrita School of Artificial Intelligence, Bengaluru and of the Amrita School of Computing, Bengaluru, India. His research interests are Complex Systems, Data Driven Modelling & Analysis, Artificial Intelligence, Early Warning Systems for Catastrophic Transitions and Time Series Analysis.