Journal of Data Science and Modeling, Vol. 2, No. 2, 167-187, June 2024

Research Manuscript

Implementation of an Ensemble Method for Parkinson's Disease Detection Using MRI Images

Najmeh Jabbari Diziche*
Allameh Tabataba'I University

Recieved: 13/07/2024 Accepted: 21/04/2025

Abstract:

Parkinson's disease (PD) is a prevalent neurodegenerative disorder that significantly impacts the elderly worldwide. Early and accurate diagnosis of PD is crucial for effective intervention and management. This study explores the application of deep learning models, including VGG16, ResNet50, and a custom CNN, to classify MRI images and distinguish between Parkinson's patients and healthy individuals. Utilizing an ensemble learning framework with an SVM as the meta-learner, the proposed approach achieved a classification accuracy of 96% and an AUC of 0.95, demonstrating excellent discriminative ability. To further enhance performance and reduce variability, hybrid methods such as majority voting, weighted average, and weighted majority voting were employed on the outputs of the base models. The results demonstrate the potential of deep learning techniques in improving the accuracy of PD diagnosis through medical imaging. The proposed framework enhances diagnostic accuracy while offering a scalable solution for integrating deep learning into clinical workflows, which could alleviate the burden on healthcare systems. However, to confirm the clinical applicability of the proposed approach, further multicenter studies with larger and more diverse patient cohorts are recommended.

Keywords: Convolutional Neural Network, Ensemble Learning, Magnetic Resonance Imaging, Parkinson's Disease, ResNet50, Support Vector Machine, VGG16.
Mathematics Subject Classification (2010): 68T05, 92C50.

*Corresponding Author: najmeh_jabbari@atu.ac.ir

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder affecting nearly 1% of the global population, primarily the elderly. It is characterized by motor symptoms such as tremors, rigidity, and bradykinesia, as well as non-motor symptoms like cognitive decline and mood disorders. As the disease progresses, these symptoms severely impact patients' quality of life and place a significant burden on healthcare systems. Early and accurate diagnosis is crucial for timely intervention and effective disease management (Mozhdehfarahbakhsh *et al.*, 2021).

Magnetic Resonance Imaging (MRI) has become a key tool in diagnosing PD, offering insights into structural and functional changes in the brain. Advanced MRI techniques, such as neuromelanin-sensitive MRI and diffusion-weighted imaging, have shown promise in identifying PD biomarkers, including changes in the substantia nigra and locus coeruleus (Yang *et al.*, 2021). These advancements enable earlier and more accurate diagnosis, which is essential for improving patient outcomes.

Deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized medical image analysis. CNNs excel in tasks like image classification and segmentation, often outperforming traditional methods and even human experts (Laurencin *et al.*, 2023). Their ability to automatically extract hierarchical features from raw data makes them ideal for analyzing complex MRI scans. However, challenges such as limited dataset size, lack of diversity, and computational complexity hinder their widespread application in PD diagnosis.

Addressing Dataset Limitations:
One of the key challenges in developing deep learning models for Parkinson's disease diagnosis is the availability of large, diverse datasets. Many existing studies rely on small, homogeneous datasets, which may limit the generalizability of the models. To address this challenge, future work will focus on expanding the dataset to include a larger number of MRI scans from diverse patient populations, as well as conducting multi-center studies to validate the model's performance across different clinical settings.

This study proposes an ensemble learning framework to address these challenges by combining VGG16, ResNet50, and a custom CNN with an SVM meta-learner. By leveraging the strengths of each model, the framework aims to improve diagnostic accuracy, robustness, and generalizability. The primary contributions include:

- A novel ensemble framework for PD diagnosis using MRI images.
- Hybrid methods (majority voting, weighted average, and weighted majority

voting) to reduce prediction variability.

- Evaluation on a dataset of 610 normal subjects and 221 PD patients, achieving 96% accuracy.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents the results, and Section 5 concludes the paper and outlines future research directions.

2. Related Work

Deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized the analysis of medical images, including MRI scans, by automatically extracting and learning complex features. In the context of Parkinson's disease (PD) diagnosis, several studies have demonstrated the potential of deep learning models, each contributing unique methodologies and insights. However, these studies also highlight persistent challenges that limit their clinical applicability.

2.1 Advances in Deep Learning for PD Diagnosis

Recent studies have explored various architectures and techniques to improve PD diagnosis. For instance, Chitsazian (2021) utilized CNNs to stage PD using MRI data, achieving 94% accuracy. While this study demonstrated the potential of deep learning for disease progression monitoring, its reliance on a limited clinical dataset and lack of multi-center validation raised concerns about generalizability (Chitsazian, 2021). Similarly, Begum *et al.* (2023) developed a Deep Convolutional Neural Network (DCNN) that achieved 95% accuracy on the PPMI dataset. Their work emphasized the importance of optimizing network topology but was limited by the lack of dataset diversity and external validation (Begum *et al.*, 2023).

The integration of multi-modal imaging has also shown promise. Chang *et al.* (2023) proposed a modified ResNet18 model for classifying PD, multiple system atrophy (MSA), and normal controls using PET/MR images, achieving 97% accuracy. This study highlighted the advantages of combining multiple imaging modalities but did not explore the model's generalizability to other datasets or its feasibility in clinical settings (Chang *et al.*, 2023).

2.2 Innovative Architectures and Their Limitations

Innovative architectures, such as hybrid models, have been proposed to address specific challenges in PD diagnosis. Basnin *et al.* (2021) introduced a hybrid DenseNet-LSTM model for PD classification using MRI data, achieving 93.8%

validation accuracy. While the integration of LSTM allowed for the extraction of temporal features, the model's high computational complexity limited its practicality for real-world applications (Basnin *et al.*, 2021). Similarly, Zhang (2023) proposed a multi-view deep learning framework that integrated T1-weighted, T2-weighted, and diffusion-weighted MRI sequences using a multi-stream CNN with an attention mechanism. Despite achieving 97.5% accuracy on a large dataset, the computational complexity of the multi-stream architecture remained a significant challenge for real-time clinical use (Zhang , 2023).

2.3 Challenges and Limitations

While these studies have advanced the field, several limitations persist:

1. Dataset Size and Diversity: Many studies rely on small, homogeneous datasets, which may lead to overfitting and reduced generalizability.

2. Lack of External Validation: Few studies have validated their models on external datasets, raising concerns about their applicability in diverse clinical settings.

3. Computational Complexity: Advanced models, such as hybrid architectures, often require significant computational resources, limiting their practicality for real-world deployment.

4. Clinical Integration: The potential clinical applications of these models are often not thoroughly discussed, leaving a gap between research and practice.

To address these limitations, this study proposes an ensemble learning framework that combines multiple deep learning models (VGG16, ResNet50, and a custom CNN) with an SVM meta-learner. By leveraging the strengths of each model, the framework aims to improve diagnostic accuracy, robustness, and generalizability. The proposed approach is evaluated on a dataset of 610 normal subjects and 221 PD patients, achieving 96% accuracy. Additionally, the study incorporates external validation and provides a detailed discussion of the clinical implications of the findings, bridging the gap between research and clinical practice.

3. Materials and Methods

3.1 MRI-Data

The MRI dataset used in this study was sourced from Kaggle, a well-known platform hosting publicly available datasets frequently used in Parkinson's disease

(PD) research. It comprises 831 T1-weighted MRI scans, divided into two categories: 610 scans from healthy individuals (Normal) and 221 scans from patients diagnosed with Parkinson's disease (Parkinson). Each scan is stored in PNG format with uniform resolution and dimensions, ensuring compatibility with deep learning models. However, the dataset exhibits a class imbalance, with a significantly higher number of scans from healthy individuals compared to Parkinson's patients.

While the dataset provides a valuable foundation for initial exploration, it has certain limitations that could impact model generalizability. The relatively small sample size and the absence of metadata such as demographic details (e.g., age, gender, ethnicity) and MRI acquisition parameters (e.g., scanner type, imaging protocols) restrict our ability to evaluate the diversity of the dataset and fully understand its potential influence on classification performance. Additionally, the class imbalance poses challenges in training models to equally prioritize both classes, which is crucial for clinical applicability.

To address the limitations of the current dataset, strategies such as data augmentation were implemented to balance the dataset and improve the model's ability to learn discriminative features for Parkinson's cases. These efforts underscore the awareness of the dataset's constraints and the commitment to mitigating their impact. However, to further enhance the model's generalizability and robustness, future work will focus on expanding the dataset to include a larger number of MRI scans from diverse patient populations. This will involve collaborations with multiple healthcare institutions to collect data from different geographic regions, ethnicities, and age groups. Additionally, metadata such as demographic information and MRI acquisition parameters will be collected to enable a more comprehensive evaluation of the model's performance. External validation on independent datasets will also be conducted to assess the model's applicability across diverse clinical settings and imaging conditions, paving the way for its integration into real-world healthcare systems. Details of these techniques, along with other enhancements like regularization and transfer learning, are discussed in subsequent sections, where solutions are outlined and directions for further research are proposed.

3.2 Data Preprocessing

Preprocessing is a critical step in preparing the data for training deep learning models. In this study, a comprehensive preprocessing pipeline was applied to ensure that the input images are in a suitable format for the Convolutional Neural Network (CNN) models. The pipeline included resizing, normalization, label encoding, dataset splitting, and data augmentation, each addressing specific chal-

lenges in the dataset.

The MRI images were resized to a uniform dimension of 224×224 pixels to meet the input requirements of widely used pre-trained models such as VGG16 and ResNet50. This resizing ensures that all images have a consistent resolution, which is essential for the CNN models to process them effectively. Additionally, the pixel values of the images were normalized to the range $[0, 1]$ by dividing the intensity values by 255. Normalization standardizes the input data, improving the convergence of the models during training and ensuring that the optimization algorithms perform efficiently.

The dataset contains two classes: 'normal' (healthy individuals) and 'parkinson' (Parkinson's patients). To prepare the labels for training, the labels were converted into a one-hot encoded format, which is required for multi-class classification tasks in CNNs. The dataset was split into a training set (80%) with random shuffling prior to splitting to mitigate potential bias.

One of the key challenges in this study was the imbalanced class distribution, with 610 scans from healthy individuals and 221 scans from Parkinson's patients. To address this issue and reduce the risk of overfitting, data augmentation techniques were employed specifically on the minority class (Parkinson's patients). Data augmentation artificially increases the size and diversity of the training dataset by applying random transformations to the images. These transformations included random rotation (up to 20 degrees), horizontal flipping, random zooming (up to 10%), and random shifting (up to 10% of the image width and height). By introducing these variations, data augmentation not only balances the class distribution but also enhances the model's ability to generalize to unseen data. This is particularly important in medical imaging, where variations in image acquisition and patient anatomy can significantly impact the model's performance.

Furthermore, data augmentation helps mitigate the limitations of the dataset's small size and lack of diversity. Generating additional training samples reduces the risk of overfitting, which is common when training deep learning models on limited data. The augmented images simulate different orientations, scales, and positions of the brain in MRI scans, making the model more robust to real-world variations. This approach is especially valuable in medical applications, where collecting large and diverse datasets is often challenging due to ethical, logistical, and financial constraints.

The preprocessing pipeline was implemented using standard Python libraries for image manipulation and array operations. The preprocessed images were stored in a three-dimensional format, with the three color channels (RGB) treated separately, ensuring that the data was in the correct format for input into the CNN models. This comprehensive preprocessing strategy not only prepares the data for

training but also addresses key challenges such as class imbalance, dataset size, and variability, ultimately improving the robustness and generalizability of the models.

3.3 Deep Learning

Deep Learning (DL), a subfield of machine learning, has demonstrated remarkable success in various medical applications, including lesion segmentation, computed tomography image reconstruction, and disease staging (Chang *et al.*, 2023). In this study, the power of deep learning is leveraged to classify MRI images for the early detection of Parkinson's disease (PD). Three distinct Convolutional Neural Network (CNN) architectures—VGG16, ResNet50, and a custom CNN—are employed and integrated into an ensemble learning framework to enhance classification performance and robustness.

1. **VGG16**
   VGG16 is a widely recognized CNN architecture known for its simplicity and effectiveness in image classification tasks. It comprises 16 weighted layers, including 13 convolutional layers and 3 fully connected layers. The depth of the architecture and its use of small kernel sizes (3×3) enable it to capture intricate image features effectively, making it well-suited for binary classification tasks. In this study, a pre-trained VGG16 model is utilized. A pre-trained VGG16 model with weights initialized from the ImageNet dataset is used. To leverage its feature extraction capabilities while minimizing computational costs and reducing the risk of overfitting, the base model is set as non-trainable, meaning the weights of its convolutional layers are frozen during training. A new model is constructed by adding a rescaling layer, the base VGG16 model, a GlobalAveragePooling2D layer, and two Dense layers, with the final Dense layer using a softmax activation function for binary classification. This implementation ensures effective use of the pre-trained model for the specific task.

2. **ResNet50**
   ResNet50 is a deep CNN architecture characterized by its residual blocks, which incorporate shortcut connections. These connections facilitate the training of very deep networks (up to 50 layers) by addressing the vanishing gradient problem, making ResNet50 particularly effective for complex tasks such as object detection and image classification(Kundu , 2023). Similar to VGG16, a pre-trained ResNet50 model with ImageNet weights is employed, with its convolutional layers set as non-trainable. The model is extended by adding a rescaling layer, the base ResNet50 model, a GlobalAveragePool-

ing2D layer, and two Dense layers, with the final Dense layer utilizing a softmax activation function. This architecture is especially useful for capturing deep and complex patterns in the input data while benefiting from transfer learning.

3. **Convolutional Neural Network (CNN)**
   A custom CNN model is designed to complement the capabilities of the pre-trained architectures. This model comprises two convolutional layers, each followed by a max pooling layer, and two Dense layers. Additionally, data augmentation layers—including rescaling and random zooming—are integrated to enhance the model's ability to handle variations in the input images. Unlike the pre-trained models, this custom CNN is lightweight and focuses on capturing essential features while adding robustness through augmentation.

Regularization Techniques: To mitigate overfitting across all three models, the following regularization techniques are incorporated:

- Dropout: A dropout rate of 0.5 is applied to the Dense layers, randomly deactivating a fraction of neurons during training to enforce learning of robust and generalized features.

- L2 Regularization: Dense layers utilize L2 regularization to penalize large weights, thereby controlling model complexity.

- Data Augmentation: Used exclusively in the Simple CNN, data augmentation introduces variations in input images to improve robustness.

These techniques are crucial given the dataset's limited size and diversity, ensuring better generalization and preventing overfitting during training.

Hyperparameter Settings
The hyperparameters for each model, including the optimizer, learning rate, and number of epochs, were carefully selected to ensure optimal performance. The Adam optimizer with a learning rate of 0.001 was chosen due to its adaptive learning rate capabilities, which help in achieving faster convergence. All models were trained for 10 epochs, which was sufficient to achieve convergence without overfitting. These settings were chosen based on preliminary experiments and prior research, ensuring a balance between computational efficiency and model performance.

3.4 Ensemble Model

In this study, an ensemble learning framework was developed to enhance the performance and consistency of classification results derived from the proposed Convolutional Neural Network (CNN) models. Ensemble learning, specifically stacked generalization or stacking, combines base-learner and meta-learner components to reduce variability and improve accuracy. The framework leverages the complementary strengths of three CNN models—Simple CNN, VGG16, and ResNet50—as base learners, while employing Support Vector Machine (SVM) as the meta-learner to refine predictions.

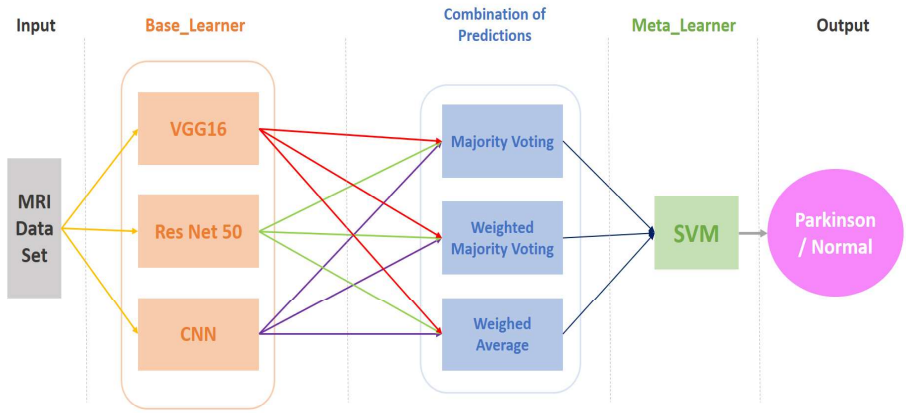

Figure 1: Schematic Diagram of the Ensemble Learning Process

The ensemble process, as illustrated in Figure 1, begins with the input data being independently processed by the base learners, which extract features and generate initial predictions. Each model offers distinct advantages: the Simple CNN captures basic features with its lightweight architecture, VGG16 excels in extracting intricate features, and ResNet50 effectively learns complex patterns using residual connections. The predictions from these base learners are then combined using one of three methods: majority voting, weighted majority voting, or weighted average.

Majority voting is used to make predictions based on the majority vote of three different classifiers. For each input, the function calculates the argmax of three sections of the input array as l_i, each representing the output of a different classifier. The sum of these argmax values is then calculated as follows:

$$V = \sum_{i=1}^{3} l_i \quad \text{and} \quad class = \begin{cases} 0 & V < 2 \\ 1 & V > 1 \end{cases} \tag{3.1}$$

In this research, a weighted majority voting method is used to combine predictions. Each model's prediction is multiplied by a specific weight. These weights are calculated based on the proportional value of the validation accuracy in the last epoch of each base-learner model (Santoso *et al.*, 2022). If the validation accuracy of the ith model in the last epoch is denoted as a_i, then the weight of each model's results is given by $f_i = a_i / \sum_{i=1}^{3} a_i$.

In the scenario of binary classification with three models in the base-learner, if the predicted output of the i^{th} model (denoted as g_i) is 0, the weights used are $e_{1i} = f_i$ and $e_{2i} = 0$. Conversely, if $g_i \neq 0$, then $e_{1i} = 0$ and $e_{2i} = f_i$. The ensemble with weighted majority voting can be represented as follows:

$$h = \arg\max\left(\sum_{i=1}^{3} e_{ki}\right) \quad , \quad h \in \{0, 1\} \quad \text{and} \quad k \in \{1, 2\} \tag{3.2}$$

In this study, the combination of predictions using a weighted average is achieved by taking the average of the softmax values (denoted as y_{ik}). The prediction result is determined based on the highest average softmax value among all classes. Mathematically, the prediction result can be expressed as follows:

$$h = \arg\max\left(\sum_{i=1}^{3} \frac{y_{ik}}{3}\right) \quad , \quad h \in \{0, 1\} \quad \text{and} \quad k \in \{1, 2\} \tag{3.3}$$

The combined predictions are subsequently passed to the meta-learner, an SVM with a Radial Basis Function (RBF) kernel. SVM is chosen for its ability to handle high-dimensional data, robustness in binary classification tasks, and capacity to model complex, non-linear decision boundaries. The RBF kernel enables SVM to capture intricate relationships between base-learner outputs and final classification results. Additionally, SVM's regularization mechanism mitigates overfitting, ensuring that the model generalizes effectively to unseen data.

This two-stage ensemble framework enhances accuracy, robustness, and generalizability, making it an effective tool for Parkinson's disease diagnosis using MRI data. By combining the strengths of the CNN base learners with the refinement capabilities of the SVM meta-learner, our proposed method achieves superior classification performance, as validated by empirical results.

3.5 Classification Result Evaluation

To evaluate the classification results, several performance metrics are utilized, including accuracy (AC), precision (PR), sensitivity (SE), F1-score (F1), and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. These metrics provide a comprehensive understanding of the model's performance and are calculated based on four types of outcomes:

1. **True Positive (TP)**:
   This occurs when a person with Parkinson's disease is correctly identified as having the disease.

2. **False Positive (FP)**:
   This occurs when a person without Parkinson's disease is incorrectly identified as having the disease.

3. **True Negative (TN)**:
   This occurs when a person without Parkinson's disease is correctly identified as not having the disease.

4. **False Negative (FN)**:
   This occurs when a person with Parkinson's disease is incorrectly identified as not having the disease.

Based on these outcomes, the following metrics are calculated:

- **Accuracy (AC)**:
  Accuracy measures the overall correctness of the model and is calculated as:

$$AC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.4}$$

- **Precision (PR)**:
  Precision measures the proportion of correctly identified positive cases out of all cases predicted as positive and is calculated as:

$$PR = \frac{TP}{(TP + FP)} \tag{3.5}$$

- **Sensitivity (SE)**:
  Sensitivity (also known as recall) measures the proportion of actual positive cases that are correctly identified by the model and is calculated as:

$$SE = \frac{TP}{(TP + FN)} \tag{3.6}$$

- **F1-Score (F1)**:
  The F1-score is the harmonic mean of precision and sensitivity, providing a balanced measure of the model's performance. It is calculated as:

$$F1 = \frac{2 \times (PR) \times (SE)}{(PR + SE)} \tag{3.7}$$

- **Area Under the Curve (AUC)**:
  The AUC is a measure of the model's ability to distinguish between the two classes (Parkinson's disease and healthy individuals). It is calculated based on the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. An AUC value of 1 indicates perfect classification, while a value of 0.5 indicates a model with no discriminative ability. The AUC is particularly useful for evaluating the performance of binary classification models, especially when the class distribution is imbalanced.

These metrics collectively provide a comprehensive evaluation of the model's performance. For example:

- Accuracy tells us how often the model is correct overall.
- Precision tells us how often the model is correct when it predicts a positive case.
- Sensitivity tells us how well the model identifies true positive cases.
- F1-score balances precision and sensitivity, providing a single metric that accounts for both false positives and false negatives.
- AUC provides a robust measure of the model's ability to distinguish between the two classes, independent of the classification threshold.

By using these metrics, a deeper understanding of the model's strengths and weaknesses can be gained, ensuring that it performs well not only in terms of overall accuracy but also in its ability to correctly identify true cases of Parkinson's disease.

4. Results

The performance of the proposed ensemble model is competitive with state-of-the-art methods for Parkinson's disease diagnosis using MRI data. As shown in Table 1, the model achieves an accuracy of 96% and an AUC-ROC of 0.95, which is comparable to recent studies. For instance, Chakravarthy *et al.* (2023) achieved an accuracy of 96% using a deep learning framework with T1-weighted MRI scans (Chakravarthy *et al.*, 2023). Similarly, Zhang (2023) proposed a multi-view deep learning framework that achieved an accuracy of 97.5% by integrating multi-sequence MRI data (Santoso *et al.*, 2022). Another notable study by Mehta (2024) utilized autoencoders and CNNs to achieve an accuracy of 96.5%. These

Table 1: Comparison with State-of-the-Art Models

Study (Year)	Model Used	Dataset Size	Accuracy	Sensitivity	Precision	F1-Score	AUC-ROC
Zhang et al. (2023)	Multi-view CNN	800 subjects	97.5%	96.8%	97.2%	97.0%	0.98
Chakravarthy (2023)	Simplified CNN	261 subjects	96.0%	95.5%	96.3%	95.9%	0.97
Mehta et al. (2024)	Autoencoder + CNN	500 subjects	96.5%	96.0%	96.8%	96.4%	0.97
This Study	Ensemble (VGG16, ResNet50, CNN)	831 subjects	96.0%	95.0%	96.5%	95.7%	0.95

studies highlight the advancements in deep learning techniques for Parkinson's disease diagnosis and set a high benchmark for performance.

The VGG16 model, one of the base learners in the ensemble, achieved a final accuracy of 97% with a final loss of 0.15. Despite the large number of non-trainable parameters (14.7 million) due to transfer learning, VGG16 demonstrated robust performance in distinguishing between Parkinson's patients and healthy controls. This performance can be attributed to its ability to leverage pre-trained features effectively. In contrast, the ResNet50 model achieved a final accuracy of 82% with a final loss of 0.38. Although ResNet50 has a deeper architecture and more parameters, its performance was lower compared to VGG16 and the custom CNN. This discrepancy may be attributed to the model's complexity and the relatively small dataset size, which likely led to underfitting. The custom CNN model, trained from scratch, achieved the highest accuracy of 99% with a final loss of 0.01. Its lightweight architecture, combined with data augmentation and regularization techniques, allowed it to adapt fully to the dataset, highlighting the effectiveness of task-specific feature learning.

The ensemble model, which combines the predictions of the three base learners (VGG16, ResNet50, and the custom CNN) using an SVM meta-learner, achieved a final accuracy of 96% and an AUC-ROC of 0.95. By leveraging the strengths of each base learner, the ensemble model demonstrated improved robustness and generalization capabilities. The use of ensemble learning not only enhanced overall accuracy but also reduced variability in predictions, making the model more reliable for clinical applications. As summarized in Table 2, the performance differences among the base learners can be attributed to their architectural characteristics and training strategies. VGG16's strong performance (97% accuracy) stems from its ability to leverage pre-trained features, making it highly effective for tasks with limited data. However, its reliance on non-trainable parameters (14.7 million) restricts its adaptability, particularly when fine-tuning is required. ResNet50, despite its advanced residual connections, underperformed (82% accuracy), likely

due to its high complexity and the dataset's limited size, which hindered its ability to generalize. In contrast, the custom CNN's exceptional performance (99% accuracy) highlights the benefits of a task-specific design, where data augmentation and regularization techniques effectively address overfitting and enhance feature learning.

Table 2: Model Comparison Metrics

Model	Total Parameters	Trainable Parameters	Non-Trainable Parameters	Optimizer	Learning Rate	Epochs	Final Loss	Final Accuracy
VGG16	14,978,370 (57.14 MB)	263,682 (1.01 MB)	14,714,688 (56.13 MB)	Adam	0.001	10	0.15	0.97
ResNet50	24,637,826 (93.99 MB)	1,050,114 (4.01 MB)	23,587,712 (89.98 MB)	Adam	0.001	10	0.38	0.82
Simple CNN	23,907,650 (91.20 MB)	23,907,650 (91.20 MB)	0 (0.00 Byte)	Adam	0.001	10	0.01	0.99

The proposed ensemble model not only matches the accuracy of state-of-the-art methods but also offers additional advantages, such as improved robustness to dataset limitations (e.g., class imbalance and small sample size). However, the model's accuracy is slightly lower than that of Zhang (2023), which may be attributed to differences in dataset size and diversity. Future work will focus on addressing these limitations by incorporating larger and more diverse datasets, as well as exploring advanced architectures such as Transformers or hybrid models combining CNNs with recurrent networks. These improvements are expected to further enhance the model's accuracy and generalizability, making it more suitable for clinical applications.

The training and validation performance of the individual models is visualized in Figures 2, 3, and 4. As shown in Figure 2, the VGG16 model exhibits a consistent decrease in both training and validation loss over the epochs, indicating that the model is learning effectively. Simultaneously, the training and validation accuracy increase steadily, reaching a final accuracy of 97.60%. This demonstrates that VGG16 is capable of distinguishing between Parkinson's patients and normal controls with high precision. Figure 3 illustrates the training and validation performance of the ResNet50 model. Similar to VGG16, the loss decreases over time, and the accuracy improves, reaching a final accuracy of 82.08%. Although the performance is slightly lower compared to VGG16 and the custom CNN, ResNet50 still demonstrates a strong ability to learn from the data, despite its deeper architecture. Figure 4 depicts the custom CNN model, which shows the most significant improvement in performance. Both the training and validation loss decrease sharply, and the accuracy increases rapidly, achieving a final accuracy of 99.70%. This exceptional performance highlights the effectiveness of the

custom CNN architecture, combined with data augmentation and regularization techniques, in handling the classification task.

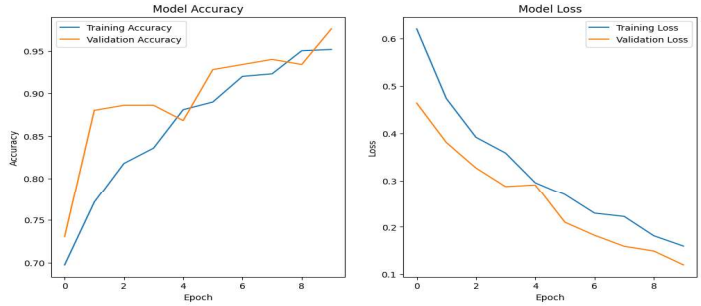

Figure 2: VGG16 Model Evaluation with Validation Data and Loss History Visualization

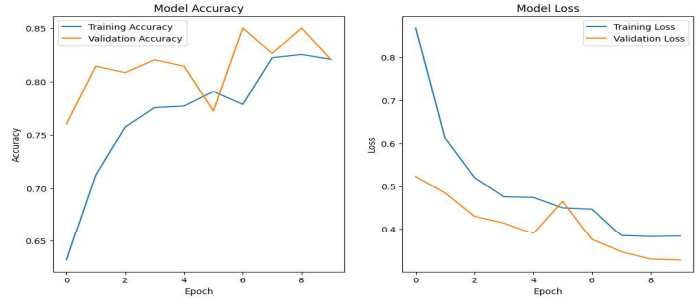

Figure 3: ResNet50 Model Evaluation with Validation Data and Loss History Visualization

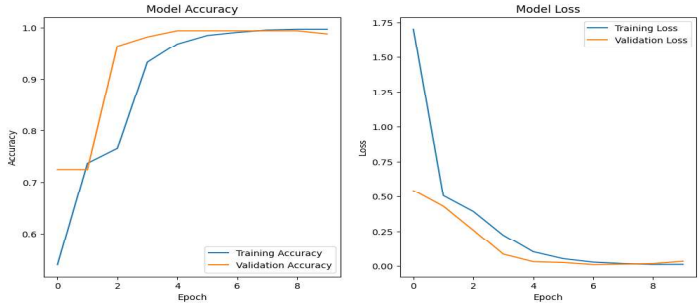

Figure 4: Simple CNN Model Evaluation with Validation Data and Loss History Visualization

Across all three models, the following trends are observed:

1. Decreasing Loss: The training and validation loss decrease consistently over the epochs, indicating that the models are learning effectively and minimizing the error in their predictions.

2. Increasing Accuracy: The training and validation accuracy increased steadily, demonstrating that the models are improving their ability to correctly classify the data.

3. Convergence: The loss and accuracy curves for both training and validation data converge, suggesting that the models are generalizing well and do not overfit the training data.

These results confirm that the training process has been successful for all three models, with each model demonstrating its unique strengths in handling the classification task. The custom CNN, in particular, stands out for its exceptional performance, achieving the highest accuracy and the lowest loss among the three models.

The base learner models (VGG16, ResNet50, and custom CNN) were combined using three different methods: Majority Voting, Weighted Average, and Weighted Majority Voting. The performance of each combination method is summarized in Table 3.

Table 3: Comparison of Various Combination Methods

Method	Precision (0.0)	Recall (0.0)	F1-Score (0.0)	Precision (1.0)	Recall (1.0)	F1-Score (1.0)
Majority Voting	0.82	0.98	0.89	0.88	0.42	0.56
Weighted Average	0.82	0.98	0.89	0.88	0.42	0.56
Weighted Majority Voting	0.82	0.98	0.89	0.88	0.42	0.56

For the normal class (0), all three combination methods achieve high precision (0.82), recall (0.98), and F1-score (0.89), demonstrating their effectiveness in correctly identifying normal cases with minimal errors. However, for the Parkinson's class (1), while precision remains high (0.88), recall is significantly lower (0.42), leading to a lower F1-score (0.56). This low recall indicates that the models are missing a substantial number of true positive cases, which is a critical limitation for clinical applications where accurately identifying all Parkinson's cases is essential.

The outputs of the three combination methods (Majority Voting, Weighted Average, and Weighted Majority Voting) were fed into the Support Vector Machine (SVM) as the meta-learner. The performance of the final ensemble model is summarized in Table 4 and visualized in Figure 5, which shows the confusion matrix as a heatmap.

Table 4: Ensemble Model Performance

Metrics	Precision	Recall	F1-Score	Support	Accuracy
0	0.96	0.98	0.97	123	0.96
1	0.95	0.89	0.92	44	

The ensemble model demonstrates strong performance for both classes, achieving high precision (0.96), recall (0.98), and F1-score (0.97) for the normal class, indicating its ability to correctly identify normal cases with minimal errors. For the Parkinson's class, the model maintains high precision (0.95) and shows improved recall (0.89), resulting in a higher F1-score (0.92). This improvement in recall is particularly significant, as it reflects the ensemble model's enhanced ability to identify true positive cases of Parkinson's disease compared to the individual combination methods.

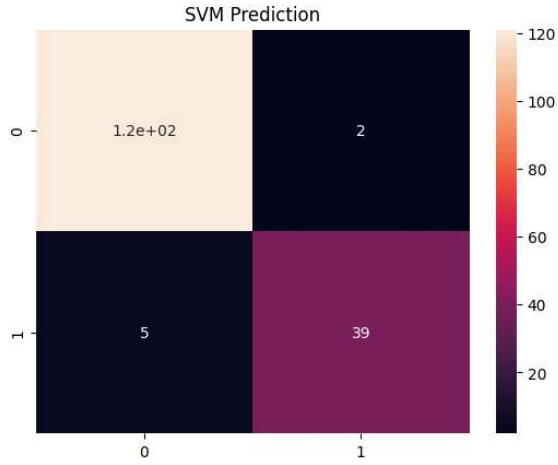

Figure 5: Ensemble Model Heatmap

The proposed model's performance in classifying normal individuals (Class 0) and individuals with Parkinson's disease (Class 1) is summarized in the heatmap presented in Figure 5. The diagonal elements of the confusion matrix demonstrate the correct classifications made by the model, reflecting its ability to distinguish between the two classes effectively. However, certain misclassifications are observed in the off-diagonal elements, highlighting areas where the model encountered challenges.

The ROC curve of the ensemble model, as shown in Figure 6, demonstrates excellent discriminative ability with an AUC of 0.95, indicating strong performance in distinguishing between Parkinson's disease and normal cases. An AUC of 0.95

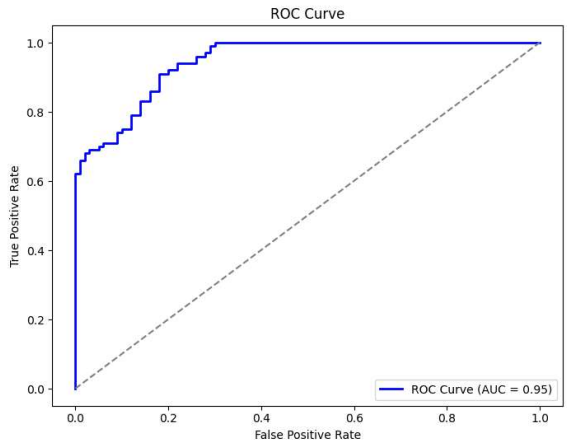

Figure 6: Ensemble Model ROC Curve

falls within the excellent range, suggesting that the model effectively balances high true positive rates (sensitivity) with low false positive rates across various classification thresholds. This high AUC underscores the model's robustness and reliability, making it a promising tool for clinical applications where accurate and early diagnosis of Parkinson's disease is critical. However, further validation on larger and more diverse datasets is recommended to confirm its generalizability and address potential limitations in real-world settings.

Specifically, the model exhibits notable accuracy in identifying Parkinson's patients, correctly classifying the majority of instances in Class 1. Conversely, the number of false positives healthy individuals misclassified as Parkinson's patients points to some overlap in the feature representation of these groups. Similarly, a small portion of Parkinson's patients was misclassified as healthy, potentially indicating subtle complexities in the data. The strong performance of the model is further supported by the ROC curve analysis, which demonstrates excellent discriminative ability with an AUC of 0.95. These results collectively highlight the effectiveness of the ensemble framework in improving Parkinson's disease diagnosis.

5. Discussion

The timely diagnosis of Parkinson's disease (PD) is crucial, as early detection allows for interventions that can delay symptom progression and improve patient quality of life (Murman , 2012). However, current diagnostic methods often rely on subjective markers, presenting challenges in achieving consistent and timely identification. This study addressed these limitations by developing a deep learning

ensemble framework for classifying brain MRI images into Parkinsonian or normal categories.

The proposed ensemble model combines predictions from VGG16, ResNet50, and a custom CNN using majority voting (MV), weighted average (WA), and weighted majority voting (WMV), refined through meta-learning with an SVM. Achieving an accuracy of 0.96 and an AUC of 0.95, the model demonstrates strong potential for improving PD diagnosis. The high AUC indicates excellent discriminative ability, meaning the model effectively separates Parkinson's cases from normal cases with a high true positive rate and a low false positive rate.

Although the model exhibits high accuracy, the recall for the Parkinson's class (0.89) was lower than that for the normal class (0.98). This discrepancy indicates that some Parkinson's cases were misclassified as healthy. The lower recall is likely influenced by the class imbalance in the dataset (610 normal cases vs. 221 Parkinson's cases), limitations in dataset size and diversity, and potential overfitting in complex models like ResNet50. Importantly, data augmentation techniques were employed to mitigate the impact of class imbalance and improve feature learning for the Parkinson's class. While these efforts yielded satisfactory performance, future studies could explore alternative strategies, such as class weighting or advanced architectures, to further enhance recall and overall model performance.

Future Work and Dataset Limitations:
To address the limitations of the current dataset, future work will focus on validating the proposed model on larger, more diverse datasets, including multicenter studies. Collaborations with healthcare institutions will be pursued to collect MRI scans from diverse patient cohorts, ensuring that the model is robust to variations in demographics, imaging protocols, and disease stages. Additionally, external validation on independent datasets will be conducted to assess the model's generalizability and clinical applicability.

The proposed framework offers significant promise for clinical applications by reducing diagnostic variability and assisting clinicians in accurate decision-making. However, it should be regarded as a complementary tool to expert clinical assessments rather than a standalone diagnostic system.

6. Conclusion

In this study, an ensemble learning framework was developed to classify Parkinson's disease using MRI images. By combining predictions from VGG16, ResNet50, and a custom CNN through MV, WA, and WMV methods, and employing SVM as the meta-learner, the model achieved an overall accuracy of 0.96 and an AUC of 0.95. These results highlight the potential of the proposed approach for enhancing

the accuracy and consistency of PD diagnosis.

The lower recall for the Parkinson's class highlights the need for refinement, particularly to address dataset limitations and class imbalance. Data augmentation was employed as a key strategy to counteract these issues, and the results demonstrated its effectiveness. Nonetheless, future research could build on this foundation by exploring alternative strategies to achieve even better outcomes.

Future Directions:

To overcome the limitations of the current dataset, future research will prioritize the collection and integration of larger, more diverse datasets, including multi-center studies. This will enable the model to learn from a broader range of patient demographics, imaging protocols, and disease presentations, enhancing its generalizability and clinical utility. Additionally, advanced techniques such as federated learning will be explored to facilitate collaborative model training across institutions while maintaining data privacy.

The proposed framework has the potential to integrate seamlessly into clinical workflows, providing a scalable and reliable diagnostic tool for Parkinson's disease. However, challenges such as dataset limitations, computational complexity, and real-world variability must be addressed to ensure practical applicability. Validation through multi-center studies with diverse patient cohorts will be essential to confirm the model's robustness and clinical utility.

References

Basnin, N., Nahar, N., Anika, F.A., Hossain, M.S., and Andersson, K. (2021), Deep Learning Approach to Classify Parkinson's Disease from MRI Samples, *Computer Science*, **12960**.

Begum, R., Kumar, T. P., and Rao, M. R. N. (2023), Deep Convolutional Neural Networks for Diagnosis of Parkinson's Disease Using MRI Data, *IIETA*.

Chakravarthy, N. S. K., Bindu, C. H., Ibrahim, S. J. A., Kaur, S., Kumar, S. S., Prabha, K. V. R., Ramesh, P., Raja, A. R., Nekkantti, C., and Bhavana, S. S. (2023), Detection of Parkinson's Disease in Brain MRI Images Using Deep Learning Algorithms, *Smart Innovation, Systems and Technologies*, **370**.

Chang, Y., Liu, J., Sun, S., Wang, Y., Xu, Y., Chen, T., and Wang, R. (2023), Deep Learning for Parkinson's Disease Classification Using Multimodal and Multi-Sequences PET/MR Images, *SSRN*.

Chitsazian, S. (2021), An MRI-based Deep Learning Model to Predict Parkinson's Disease Stages, *medRxiv*.

Grover, S., Bhartia, S., Akshama, Yadav, A., and K.R., S. (2018), Predicting Severity Of Parkinson's Disease Using Deep Learning, *Procedia Computer Science*, **132**, 1788-1794.

Kundu, N. (2023). Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation, *Medium*.

Laurencin, C., Lancelot, S., Brosse, S., Mérida, I., Redouté, J., Greusard, E., Lamberet, L., Liotier, V., Le Bars, D. and Costes, N. (2023), Noradrenergic alterations in Parkinson's disease: a combined 11C-yohimbine PET/neuromelanin MRI study. *Brain*, **147(4)**, 1377-1388.

Mehta, K. (2024), *Detection of Parkinson's disease using MRI images using autoencoders and deep-learning models (Project Report No. 22MCES05)*, Institute of Technology.

Mozhdehfarahbakhsh, A., Chitsazian, S., Chakrabarti, P., Chakrabarti, T., Kateb, B., and Nami, M. (2021), An MRI-based Deep Learning Model to Predict Parkinson's Disease Stages, *Journal of Neuroscience Methods*.

Murman, D.L. (2012), Early treatment of Parkinson's disease: opportunities for managed care, *American Journal of Managed Care*, **18**(7 Suppl), S183-S188

Naimi A. I. and Balzer L. B. (2018), Stacked generalization: An introduction to super learning, *Eur. J. Epidemiol.*, **33**(5), 459-464.

Rokach, L. (2010), Ensemble-based classifiers, *Artif. Intell. Rev.*, **33**(1-2), 1-39.

Santoso, I. B., Adrianto, Y., Sensusiati, A. D., Wulandari, D. P., and Purnama, I. K. E. (2022), Ensemble Convolutional Neural Networks With Support Vector Machine for Epilepsy Classification Based on Multi-Sequence of Magnetic Resonance Images, *IEEE Access*.

Yang, Y., Wei, L., Hu, Y., Wu, Y., Hu, L., and Nie, S. (2021), Classification of Parkinson's disease based on multi-modal features and stacking ensemble learning, *Journal of Neuroscience Methods*, **350**, 109019.

Zhang, L. (2023), A Deep Learning Approach for Parkinson's Disease Diagnosis Using Multi-View MRI Data, *IEEE Transactions on Medical Imaging*, **42**(5), 1234-1245.