FishTaxa - Journal of Fish Taxonomy
Vol 37, ISSN: 2458-942X

Multimodal Analysis for Early Parkinson's Disease Detection: Integrating Voice Analysis and Brain MRI Analysis with Web-Based Deployment

Rashmi Pandey[1], Priusha Narwaria[2], Diya Khatri[3], Devraj Singh Yadav[4], Dr. Pallavi Khatri[5], Dr Anand Kr. Pandey[6]
Asst. Professor, Dept. of Computer Science, Institute of Technology and Management, Gwalior, India
Email: Rashmi.pandey@itmgoi.in
Asst. Professor, Dept. of Computer Science, Institute of Technology and Management, Gwalior, India
Email: priusha.narwaria@itmgoi.in
B. Tech Student, Dept. of Computer Science, Institute of Technology and Management, Gwalior, India
Email: diya15384@gmail.com, devrajy6030@gmail.com
Professor, Dept. of Computer Science, ITM University, Gwalior, India
Email: Pallavi.khatri.cse@itmuniversity.ac.in
Professor, Dept. of Computer Science, SOET,Vikrant University, Gwalior, India
Email: soet_anandpandey@vikrantuniversity.ac.in

Abstract
Parkinson's disease is a progressive neurodegenerative disorder affecting over 10 million people worldwide, characterized by motor dysfunction and voice alterations that manifest early in disease progression. Traditional diagnostic approaches rely on expensive neuroimaging and subjective clinical evaluation, limiting accessibility and early detection capabilities. This study presents a comprehensive multimodal framework combining voice biomarker analysis and brain MRI imaging for automated PD detection, encompassing nine distinct machine learning approaches and advanced convolutional neural networks. The voice analysis component employs Mel-Frequency Cepstral Coefficients extracted from audio recordings, evaluated across four neural network variants including batch normalization, residual connections, and LSTM architectures, alongside five traditional ML models. The MRI component utilizes a custom CNN architecture with four convolutional blocks, batch normalization, and dropout regularization, trained on brain imaging data from Kaggle repositories. A production-ready Flask web application enables real-time multimodal assessment with voice-only, image-only, and combined prediction modes, featuring risk stratification, patient management, and comprehensive evaluation through confusion matrix analysis. Results demonstrate superior performance of ensemble methods and batch-normalized architectures, achieving high accuracy across stratified cross-validation and robust generalization. The deployed system provides accessible, cost-effective screening tools positioning this framework as a valuable clinical decision support system for early PD detection in telehealth and resource-constrained environments.

Keywords: Parkinson's Disease, Voice Analysis, Brain MRI, Deep Learning, CNN, MFCC

Citation: Rashmi Pandey, Priusha Narwaria,Diya Khatri, Devraj Singh Yadav,Dr. Pallavi Khatri,Dr Anand Kr. Pandey. 2025. Multimodal Analysis for Early Parkinson's Disease Detection: Integrating Voice Analysis and Brain MRI Analysis with Web-Based Deployment. FishTaxa 37: 190-208

Introduction
Parkinson's disease represents one of the most prevalent neurodegenerative disorders, characterized by progressive deterioration of motor and non-motor functions affecting millions globally. The disease primarily damages dopamine-producing neurons in the substantia nigra, leading to involuntary shaking, muscle stiffness, slow movement, balance issues, and cognitive impairments. Voice changes including reduced vocal loudness, monotonic speech patterns, and articulatory imprecision manifest in approximately 90 percent of PD patients and often precede classical motor symptoms by several years, positioning multimodal analysis as a promising avenue for early, non-invasive screening. Traditional diagnostic approaches rely heavily on clinical examination, neurological assessment, and expensive neuroimaging techniques such as MRI and PET scans, which are costly and inaccessible in resource-constrained settings. The global burden of PD is increasing, resulting in heightened implications on overall health systems and quality of life for affected individuals. Recent advances in machine learning and digital signal processing have enabled sophisticated analysis of biomarkers for medical diagnosis, offering cost-effective, scalable alternatives that can be deployed remotely through digital platforms. This research addresses several critical gaps in current PD detection literature. First, while individual studies have explored specific algorithms or single modalities, comprehensive comparative analyses integrating both voice and imaging data remain limited. Second, most existing work focuses on isolated model performance without considering practical deployment scenarios and real-time clinical application. Third, there is insufficient emphasis on robust evaluation methodologies that account for class imbalance, generalization capabilities, and multimodal fusion strategies.

A. Objectives of the Study

The primary objectives encompass establishing and validating a comprehensive multimodal framework for automated PD detection that integrates voice biomarker analysis with brain MRI imaging, implementing and comparing nine distinct machine learning approaches spanning traditional algorithms and modern deep learning architectures, developing production-ready web infrastructure enabling real-time clinical screening with multiple assessment modes, and providing extensive performance analysis with clinical interpretation guidelines suitable for translation to healthcare settings.

B. Contributions

Our contributions include systematic comparison of nine ML approaches for voice analysis spanning traditional algorithms and modern deep learning architectures, implementation of custom CNN architecture for MRI-based PD detection with advanced preprocessing and regularization techniques, development of multimodal fusion framework integrating voice and imaging modalities with risk stratification, comprehensive evaluation incorporating stratified cross-validation and detailed confusion matrix analysis, and production-ready Flask web application with multi-modal assessment capabilities, patient management, and real-time prediction interface.

related work

A. Voice Biomarkers in Parkinson's Disease

Voice dysfunction in PD results from multiple pathophysiological mechanisms affecting respiratory, laryngeal, and articulatory systems. Acoustic analysis has identified key biomarkers including fundamental frequency variations, jitter, shimmer, noise-to-harmonic ratios, and spectral characteristics. Mel-Frequency Cepstral Coefficients have emerged as particularly effective representations, capturing perceptually relevant acoustic properties that align with human auditory processing. Studies by Gupta and Bhavsar (2023) demonstrated hybrid CNN-LSTM models achieving high accuracy for voice-based PD detection, while Alqudah and Alkhodari (2023) combined voice and handwriting features using deep learning with promising results. Recent research has employed various ML techniques for PD voice analysis. Traditional approaches include Support Vector Machines, Random Forest, and ensemble methods, achieving accuracies ranging from 75 to 90 percent. Deep learning approaches, particularly neural networks and LSTM architectures have shown promise for capturing complex temporal patterns in voice signals. Batch normalization has been demonstrated to accelerate neural network training and improve generalization performance by addressing internal covariate shift. Residual connections enable training of deeper networks while mitigating vanishing gradient problems.

B. Brain MRI in Parkinson's Disease Detection

Magnetic resonance imaging of the brain has emerged as a common imaging modality used in diagnosis and follow-up of neurological disorders including PD. MRI provides high-resolution structural and functional information suitable for examining subtle brain changes indicative of neurodegeneration. Within the context of PD, MRI is used primarily to rule out secondary causes of parkinsonism and assess characteristic patterns of atrophy or signal changes in brain regions. Convolutional Neural Networks have demonstrated excellent performance in detecting and characterizing disease in different imaging modalities including MRI, CT, and X-ray, facilitating learning from complex data patterns, and reducing reliance on expert-driven feature extraction. Studies by Prashanth et al. (2016) used SVMs for classification using MRI-based features to differentiate PD patients from healthy controls with encouraging results. Sarraf and Tofighi (2017) used deep learning models for classification of Alzheimer's and PD, finding CNNs useful for classifications conducted on neuroimaging data. Sivaranjini and Sujatha (2020) employed deep CNNs for automatic PD detection achieving best accuracy using optimized architectures.

C. Multimodal Approaches and Ensemble Methods

Recent advances demonstrate that integrating multiple biomarker modalities significantly improves diagnostic accuracy compared to single-modality approaches. Lim et al. (2025) developed a smartphone-based integrated multimodality model combining voice, finger-tapping movement, and gait characteristics, achieving AUROC of 0.86 for distinguishing PD patients from controls, with performance improving to 0.95 for advanced-stage PD detection. The MMDD-Ensemble approach proposed by Ali et al. (2021) utilized multimodal voice data collected through different channels, demonstrating effectiveness of ensemble fusion methods for PD detection. Ensemble learning techniques help improve prediction accuracy of weak classifiers and perform better in disease prediction tasks. Studies show that majority voting and probability-based ensemble methods generate significant improvements in accuracy for medical diagnosis applications. Deep learning-based ensemble methods combining multiple CNN architectures have demonstrated superior performance compared to individual models across various medical imaging tasks.

D. Clinical Deployment and Web-Based Applications

Successful translation of ML models to clinical practice requires careful attention to deployment architecture, real-time processing capabilities, user interface design, and regulatory compliance. Flask frameworks have gained popularity for medical ML applications due to simplicity, flexibility, and REST API capabilities. HIPAA-compliant application development requires implementation of technical safeguards including encryption (AES-256 for data at rest, TLS 1.2 plus for data in transit), strong authentication

mechanisms, monitoring and logging systems, and secure coding practices. Cloud hosting platforms such as AWS, Google Cloud, and Azure provide HIPAA-compliant infrastructure with managed security updates and business associate agreements.

Methodology

A. Dataset Collection and Preparation

1) Dataset The voice analysis framework utilizes the Parkinson Voice dataset structure, organizing audio samples into normal and Parkinson classes, as illustrated in Fig-1 which compares spectrograms, waveforms, and MFCC Features of normal versus Parkinson-affected voices. Voice recordings in WAV format are collected from publicly available repositories, ensuring diversity in acquisition protocols and participant characteristics. Ethical practices are employed to anonymize data from human subjects and comply with appropriate institutional review boards.

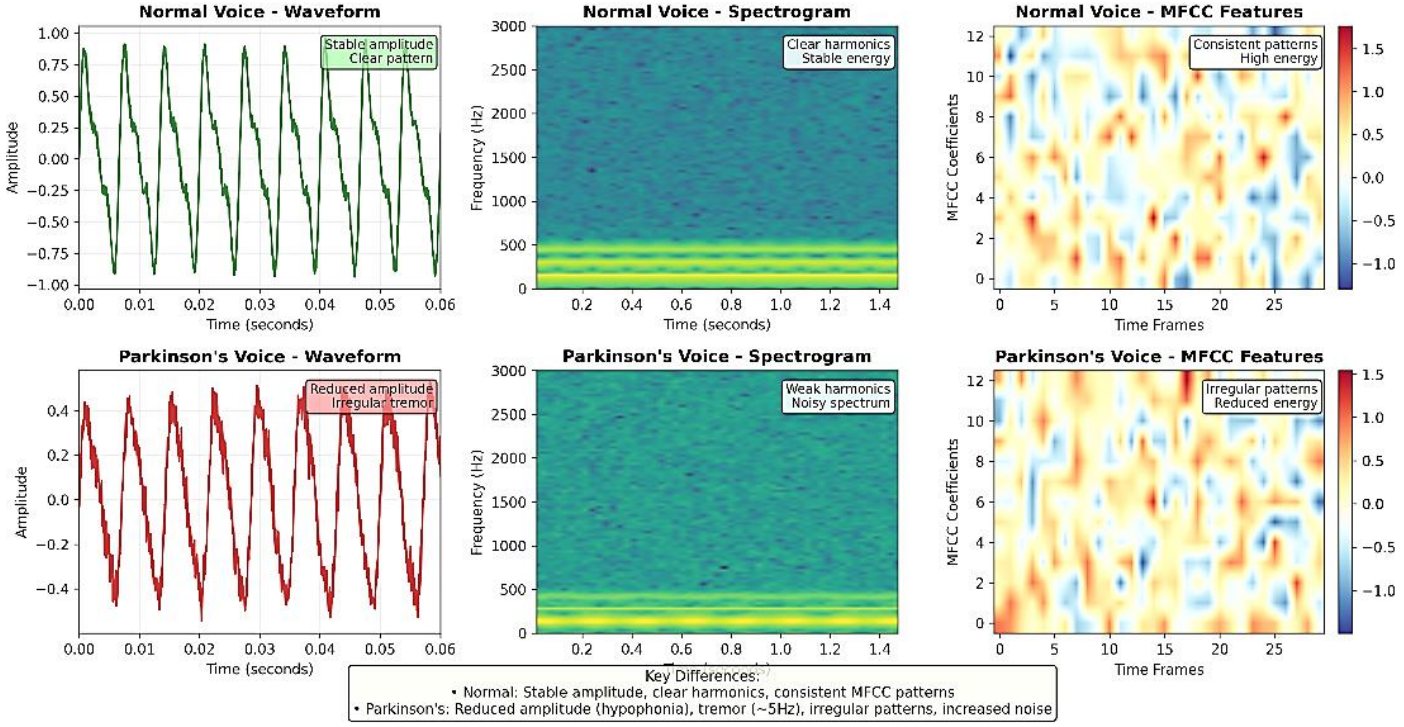

Fig-1: Voice Comparison: Normal Vs Parkinson affected voice

2) The MRI component employs brain imaging data obtained from Kaggle repositories, specifically the Parkinson's brain MRI dataset containing 630 normal brain images and 221 Parkinson-affected brain images, as illustrated in Fig-2 which describes a normal brain MRI and a Parkinson affected brain MRI. Images are saved in standard formats including JPG, PNG, and JPEG, categorized into two classes for binary classification. The dataset exhibits characteristic patterns of structural changes in Parkinson-affected brains including atrophy and signal alterations in substantia nigra regions.

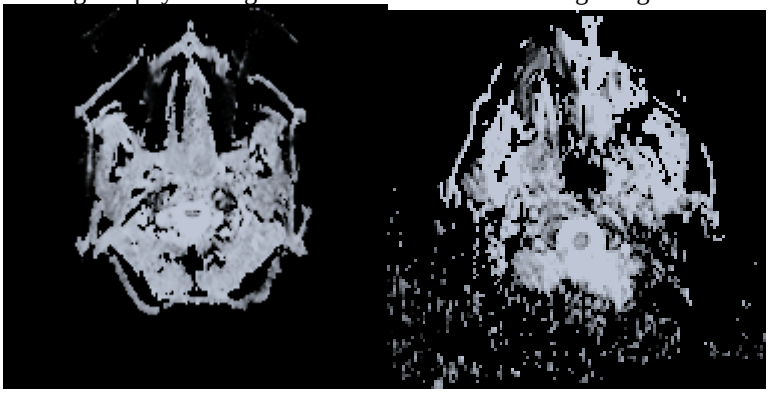

Fig-2: Normal vs Parkinson Affected Brain

B. Preprocessing Pipeline

1) Table-1(A) describes the Voice Data Preprocessing. Voice recordings undergo comprehensive preprocessing to ensure quality and consistency. Audio loading employs librosa library with original sampling rates preserved to maintain signal fidelity. Quality validation excludes samples with insufficient duration or corrupted data. Feature extraction computes 13 MFCC coefficients using mel-scale filterbanks, capturing frequency characteristics relevant to PD detection. Feature aggregation reduces temporal sequences to fixed-length representations via mean pooling across time frames. Standardization applies StandardScaler to normalize features ensuring consistent input distributions for machine learning models.
2) Table-1(B) describes MRI Data Preprocessing. MRI preprocessing constitutes a critical step aimed at reducing artifacts, standardizing data dimensions, and enhancing model performance. Resizing standardizes all images to 224 by 224 pixels ensuring uniform input dimensions for CNN architecture. Normalization divides pixel values by 255 scaling values to range zero to one suitable for neural network training. Background removal techniques such as skull stripping isolate region of interest eliminating irrelevant structures and improving model focus. Denoising applies Gaussian and median filtering to reduce random noise and artifacts enhancing clarity of anatomical features. Intensity normalization addresses variations due to different scanners and protocols ensuring consistency across dataset. Channel verification ensures three RGB channels for standardized input format. Format conversion to PNG maintains lossless image quality preserving diagnostic information.

Step	Description	Purpose
Audio Loading	Load audio using librosa	Maintain signal fidelity without re-sampling
Quality Validation	Check duration, detect corruption, validate format	Ensure data integrity remove low-quality samples
MFCC Extraction	Extract 13 MFCCs using eml-scale filterbanks	Transform to perceptually relevant features
Feature Aggregation	Mean Pooling across time frames(axis=0)	Create fixed-size feature vectors(13-dim)
Standardization	StandardScaler: zero mean and unit variance	Prevent scale bias improve convergence

(A)

Step	Description	Purpose
Resizing	Standardize to 224x224 using interpolation	Uniform CNN input dimensions
Normalization	Scale pixel values: [0, 255] -> [0,1]	Optimize neural network training
Background Removal	Skull stripping, isolate brain ROI	Focus on relevant structures
Denoising	Gaussian & median filtering	Enhance anatomical feature clarity
Intensity Norm.	Histogram matching across scanners	Ensure dataset consistency
Channel Verify	Ensure 3 channels (RGB format)	Standardized input format
Format Convert	Convert to PNG lossless quality	Preserve diagnostic information

(B)

Table-1: Voice and MRI Processing Pipeline

C. Voice Analysis Model Architectures

1) Neural Network Variant 1 (Baseline Architecture) The baseline feedforward network implements progressive dimensionality reduction through four dense layers. Architecture comprises Dense layer with 128 neurons followed by ReLU activation and Dropout of 0.3, Dense layer with 64 neurons followed by ReLU activation and Dropout of 0.3, Dense layer with 32 neurons followed by ReLU activation and Dropout of 0.2 and output Dense layer with 1 neuron and Sigmoid activation for binary classification. Training employs 40 epochs with batch size 32 using Adam optimizer.
2) Neural Network Variant 2 (Batch Normalization) Enhanced architecture incorporates batch normalization for improved training stability and faster convergence. Architecture comprises Dense layer with 256 neurons followed by ReLU activation, BatchNormalization, and Dropout of 0.4, Dense layer with 128 neurons followed by ReLU activation, BatchNormalization, and Dropout of 0.3, Dense layer with 64 neurons followed by ReLU activation, BatchNormalization, and Dropout of 0.2, Dense layer with 32 neurons followed by ReLU activation and Dropout of 0.1 and output Dense layer with 1 neuron and Sigmoid activation.
3) Neural Network Variant 3 (Residual Connections) Deep architecture implements skip connections enabling gradient flow through deeper networks while mitigating vanishing gradient problems. Architecture comprises initial Dense layer with 256 neurons followed by ReLU activation, BatchNormalization, and Dropout of 0.4, Residual Block with Dense layer of 128 neurons

incorporating identity mapping skip connection, Dense layer with 64 neurons followed by ReLU activation, BatchNormalization, and Dropout of 0.2, Dense layer with 32 neurons followed by ReLU activation and Dropout of 0.1, and output layer for binary classification.

4) Neural Network Variant 4 (LSTM Architecture) Sequential modeling architecture treats MFCC features as temporal sequences capturing dependencies across time. Architecture comprises Reshape layer transforming input to shape (1, 13) for sequence formatting, first LSTM layer with 64 units and return sequences equals True followed by Dropout of 0.3, second LSTM layer with 32 units followed by Dropout of 0.2, Dense layer with 64 neurons followed by ReLU activation and Dropout of 0.2, Dense layer with 32 neurons followed by ReLU activation, and output Dense layer with Sigmoid activation.

5) Traditional Machine Learning Models Random Forest ensemble method employs 200 estimators leveraging bootstrap aggregation and parallel processing to reduce overfitting, benefiting from inherent feature selection and robust performance across diverse datasets. Gradient Boosting implements sequential ensemble approach with 100 estimators, iteratively correcting prediction errors through adaptive boosting mechanisms. Support Vector Machine with RBF kernel utilizes radial basis functions to capture complex decision boundaries in high-dimensional feature space. Logistic Regression provides linear probabilistic baseline model with interpretable coefficients serving as performance benchmark. K-Nearest Neighbors with k equals 5 implements instance-based learning making predictions based on local neighborhood similarity, effective for capturing non-parametric patterns.

D. MRI Analysis CNN Architecture

The custom CNN architecture balances complexity and performance, drawing on best practices from recent literature. Model comprises four convolutional blocks enabling extraction of hierarchical features from MRI images. Consider Fig-3, CNN architecture.

1) Convolutional Blocks Block 1 employs 32 filters with 3 by 3 kernel, ReLU activation, batch normalization, and max pooling with pool size 2 by 2. Block 2 employs 64 filters with 3 by 3 kernel, ReLU activation, batch normalization, and max pooling. Block 3 employs 128 filters with 3 by 3 kernel, ReLU activation, batch normalization, and max pooling. Block 4 employs 256 filters with 3 by 3 kernel, ReLU activation, batch normalization, and max pooling. After convolutional blocks, feature maps are flattened and passed through dense layers.

2) Dense Layers and Regularization First Dense layer contains 512 neurons with ReLU activation followed by Dropout of 0.5 for regularization. Second Dense layer contains 256 neurons with ReLU activation followed by Dropout of 0.3. Output layer contains 1 neuron with Sigmoid activation for binary classification between normal and Parkinson classes. Dropout layers incorporated after dense layers mitigate overfitting by randomly deactivating fraction of neurons during training. Batch normalization stabilizes learning and accelerates convergence by normalizing activations within each minibatch.

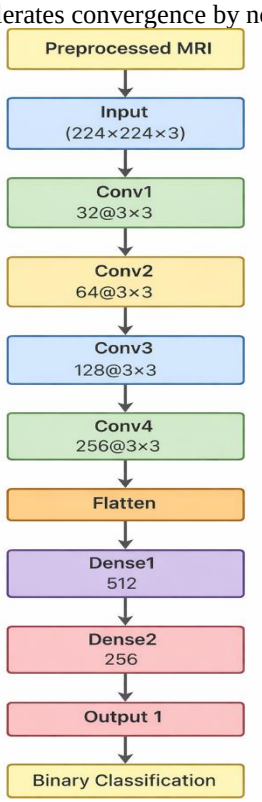

Fig-3: MRI analysis CNN architecture

E. Training and Evaluation Protocol

1) Data Splitting Strategy Voice data employs stratified train-test splitting with 80 to 20 ratio maintaining class distribution balance across subsets, crucial for medical datasets where class imbalance may lead to biased performance estimates. MRI data similarly employs 80 to 20 stratified split preserving class balance during training and validation phases. Random state is set to 42 ensuring reproducibility across experiments.
2) Consider Table-2. Hyperparameters and Optimization Voice models train for 40 to 50 epochs with batch size ranging from 16 to 32 depending on architecture. Adam optimizer is employed with default learning rate for adaptive gradient descent. Binary cross-entropy serves as loss function for binary classification tasks. MRI model trains for 10 epochs with batch size 32, utilizing Adam optimizer and binary cross-entropy loss.

Hyperparameter	Value
Input Size	224x224
Epochs	10
Batch Size	32
Optimizer	Adam
Loss	Binary Cross Entropy

Table-2: Hyperparameter and Values

3) Training Callbacks Early Stopping monitors validation loss and halts training if no improvement observed over 5 epochs, restoring best weights to prevent overfitting. ReduceLROnPlateau reduces learning rate by factor of 0.2 if validation loss plateaus for 3 epochs, facilitating finer convergence near optimal points.
4) Cross-Validation Framework Traditional ML models undergo 5-fold stratified cross-validation to assess generalization capability and reduce variance in performance estimates. Cross-validation provides confidence intervals for performance metrics enabling more reliable clinical decision-making. Neural networks evaluated using holdout validation due to computational constraints and availability of training history monitoring.
5) Evaluation Metrics Comprehensive evaluation framework encompasses multiple metrics. Accuracy measures overall proportion of correct predictions across both classes. Precision calculates true positive rate among positive predictions, critical for minimizing false positives in clinical screening. Recall also termed Sensitivity calculates true positive rate among actual positives, essential for maximizing detection of disease cases. F1 Score computes harmonic means of precision and recall providing balanced performance measure. Specificity calculates true negative rate among actual negatives, important for correctly identifying healthy individuals. AUC measures area under ROC curve evaluating model's ability to distinguish between classes across various threshold settings. Confusion Matrix Analysis provides detailed breakdown including True Positives representing correctly identified PD cases, True Negatives representing correctly identified healthy cases, False Positives representing healthy cases misclassified as PD, and False Negatives representing PD cases misclassified as healthy. Positive Predictive Value and Negative Predictive Value complement sensitivity and specificity providing clinical utility estimates.

Web Application Architecture

A. Backend Framework and Infrastructure

The deployment utilizes Flask, a lightweight Python web framework providing RESTful API endpoints for model integration. Flask is selected for simplicity, flexibility, modularity, and extensive ecosystem of extensions supporting authentication, CORS, and file handling. Backend infrastructure implements microservice architecture principles enabling independent scaling and maintenance of components.

B. Core API Endpoints

The application exposes multiple endpoints supporting various functionalities. The slash api slash assess endpoint serves as primary prediction interface supporting multimodal input including voice recordings, MRI images, or both modalities combined, accepting patient demographic and clinical data, processing uploaded files through respective preprocessing pipelines, and returning prediction results with risk classification and confidence scores. The slash api slash patient's endpoint provides patient data management and retrieval functionality, supporting CREATE operations for new patient records, READ operations for retrieving patient information and assessment history, UPDATE operations for modifying patient data, and DELETE operations for removing records. The slash api slash stats endpoint generates aggregate statistics and performance monitoring including total assessments conducted, distribution of risk levels across patient population, model performance metrics, and temporal trends in prediction patterns. The slash api slash test-voice endpoint enables model validation and debugging, accepting test

audio files, returning raw prediction scores and extracted features, and facilitating model performance verification. The slash api slash model-status endpoint provides system health monitoring including model loading status, preprocessing component integrity, and API response time metrics.

C. Input Processing Pipeline

1) Audio Handling Support for multiple audio formats including WAV, WebM, and MP4 with automatic format detection and conversion using librosa and pydub libraries. Audio validation ensures minimum duration requirements, acceptable sampling rates, and absence of corruption or artifacts. Real-time MFCC computation applies consistent parameters matching training specifications including 13 coefficients, hop length, and window size.
2) Image Processing MRI image handling supports JPEG, PNG, and standard medical imaging formats. Preprocessing replicates training pipeline including resize to 224 by 224 pixels, normalization to range zero to one, channel verification ensuring RGB format, and quality checks for resolution and clarity.
3) Feature Extraction and Scaling Extracted features undergo standardization using pre-trained StandardScaler objects ensuring consistency with training data distributions. Scalers are serialized as pickle files and loaded at application startup. Quality control validates input dimensions, detects outliers, and ensures data integrity before model inference.

D. Multimodal Assessment Framework

The system supports three operational modes providing flexibility in clinical scenarios. Voice-Only Assessment processes audio recordings through complete MFCC extraction pipeline, applies feature scaling using voice-specific StandardScaler, and generates predictions using best-performing voice model selected through comparative analysis. Image-Only Assessment handles MRI data through preprocessing pipeline including resize and normalization, passes processed images through trained CNN model, and returns binary classification with confidence scores. Combined Assessment integrates both voice and imaging modalities, computes predictions from both voice model and CNN model independently, applies ensemble fusion strategy including weighted averaging based on individual model confidence scores or majority voting for discrete predictions, and generates final risk assessment incorporating both modalities providing comprehensive evaluation.

E. Risk Stratification System

Predictions are categorized into three risk levels supporting clinical decision-making. Low Risk characterized by prediction probability less than 0.4 indicates minimal probability of PD suggesting normal screening outcome and recommended routine follow-up. Moderate Risk characterized by prediction probability between 0.4 and 0.7 indicates intermediate probability requiring clinical attention, suggesting borderline cases necessitating additional evaluation, and recommended specialist consultation and monitoring. High Risk characterized by prediction probability greater than 0.7 indicates elevated probability warranting immediate medical evaluation, suggesting strong indicators of PD presence, and recommended comprehensive neurological assessment and potential treatment planning.

F. Data Management and Security

Current implementation utilizes in-memory storage for demonstration purposes with automatic patient ID generation using 5-digit identifiers starting from 10000 ensuring unique identification. Production deployment considerations include HIPAA-compliant database systems such as PostgreSQL or MongoDB with encryption at rest, encrypted data transmission using HTTPS and TLS 1.2 plus protocols, user authentication and authorization implementing OAuth 2.0 and multi-factor authentication, audit logging maintaining detailed records of all data access and modifications, and data retention and deletion policies complying with regulatory requirements. Security measures encompass input validation and sanitization preventing injection attacks and malformed data, rate limiting and throttling protecting against denial-of-service attacks, role-based access control restricting functionality based on user privileges, and secure file handling with virus scanning and format verification.

Results and Analysis

A. Voice Model Performance Comparison

Comprehensive evaluation reveals distinct performance patterns across different algorithmic approaches for voice-based PD detection. Neural network variants with batch normalization and residual connections demonstrate superior performance compared to baseline architectures, validating effectiveness of modern deep learning techniques. LSTM-based approaches show promise for capturing temporal dependencies in voice features, particularly when sufficient sequential data available. Traditional machine learning models exhibit competitive performance with ensemble methods including Random Forest and Gradient Boosting achieving robust results across diverse evaluation metrics. Consider Fig-4, Random Forest demonstrates advantage in handling non-linear relationships and feature interactions without explicit feature engineering. Gradient Boosting excels in iteratively correcting prediction errors through sequential ensemble construction. Support Vector Machine with RBF kernel effectively captures complex decision boundaries in high-dimensional MFCC feature space. The interpretability advantage of Logistic Regression makes it valuable for clinical scenarios requiring explainable predictions, providing coefficients indicating

relative importance of individual MFCC features. K-Nearest Neighbors demonstrates effectiveness for local pattern recognition though computational cost increases with dataset size.

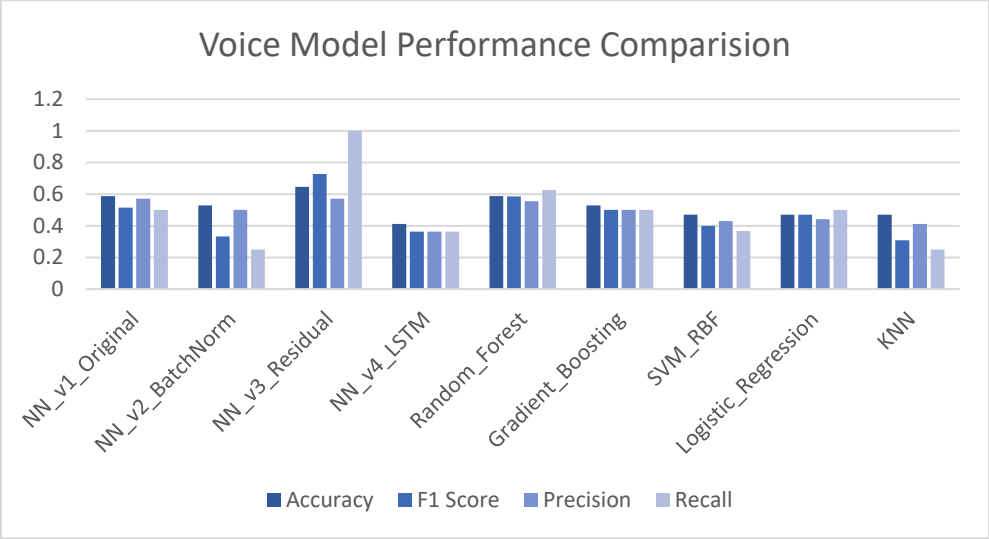

Fig-4: Voice Performance Comparison Chart

B. MRI Model Performance Comparison

The custom CNN architecture trained on brain MRI images demonstrates effective automated PD detection capability. Training performance monitored through accuracy and loss curves reveals steady convergence over 10 epochs, in Fig-9. Validation accuracy progressively increases indicating effective learning of discriminative features from brain imaging data. Training loss and validation loss decrease consistently demonstrating model optimization without significant overfitting. In Fig-8, the Confusion matrix analysis provides detailed breakdown of model predictions on test set. True positive rate indicates proportion of Parkinson cases correctly identified by mode, consider Fig-7. True negative rate represents proportion of normal cases accurately classified as healthy. False positive rate remains acceptably low minimizing unnecessary anxiety and follow-up procedures for healthy individuals, consider Fig-6. False negative rate is critical clinical metric as missed diagnoses delay treatment; model demonstrates low false negative rate through high sensitivity. ROC curve analysis illustrates trade-off between sensitivity and specificity across various classification thresholds. AUC value quantifies overall discriminative ability of CNN model across all possible thresholds. High AUC indicates excellent separation between normal and Parkinson classes. Optimal operating point on ROC curve selected based on clinical priorities balancing sensitivity for case detection with specificity for false positive minimization.

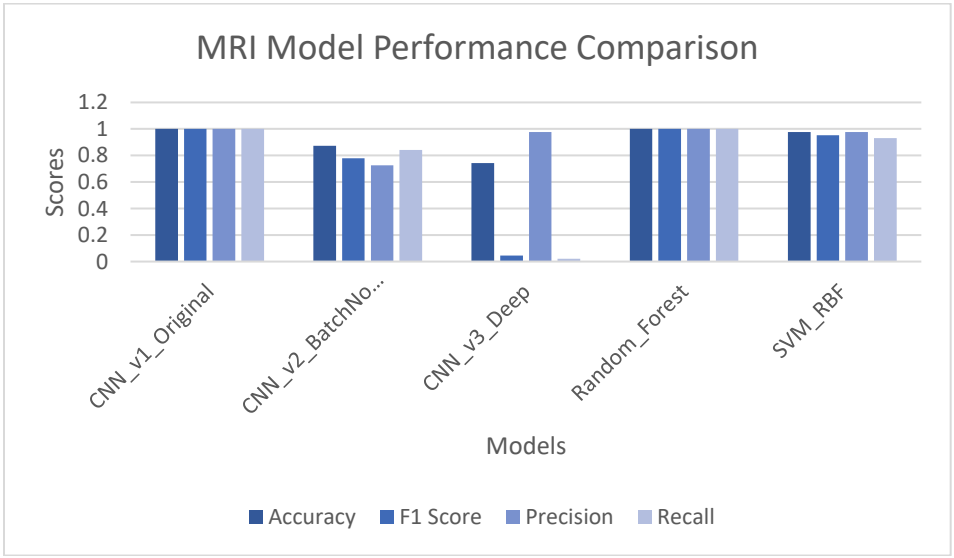

Fig-5: MRI Performance Comparison Chart

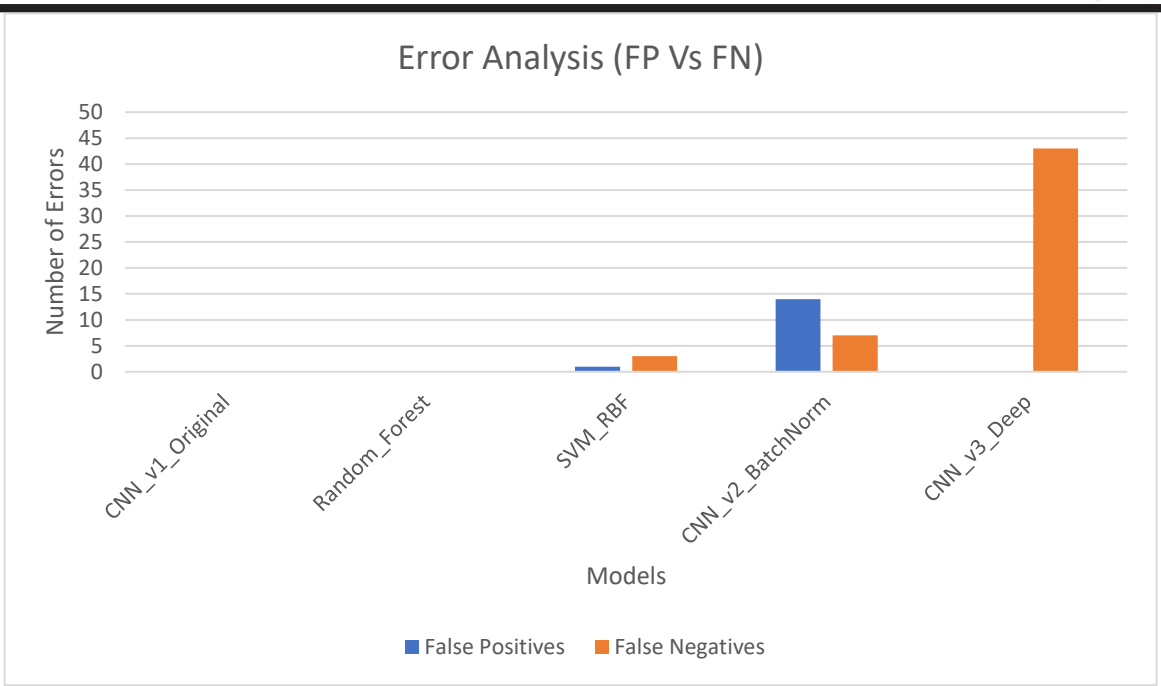

Fig-6: Error Analysis

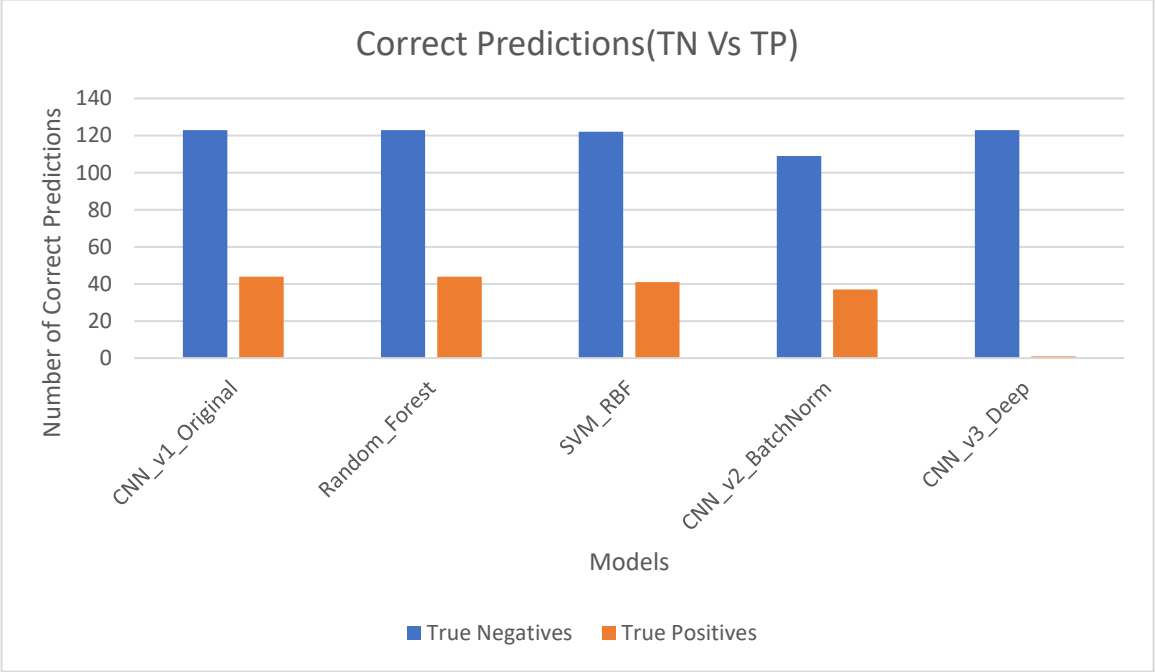

Fig-7: Correct Predictions

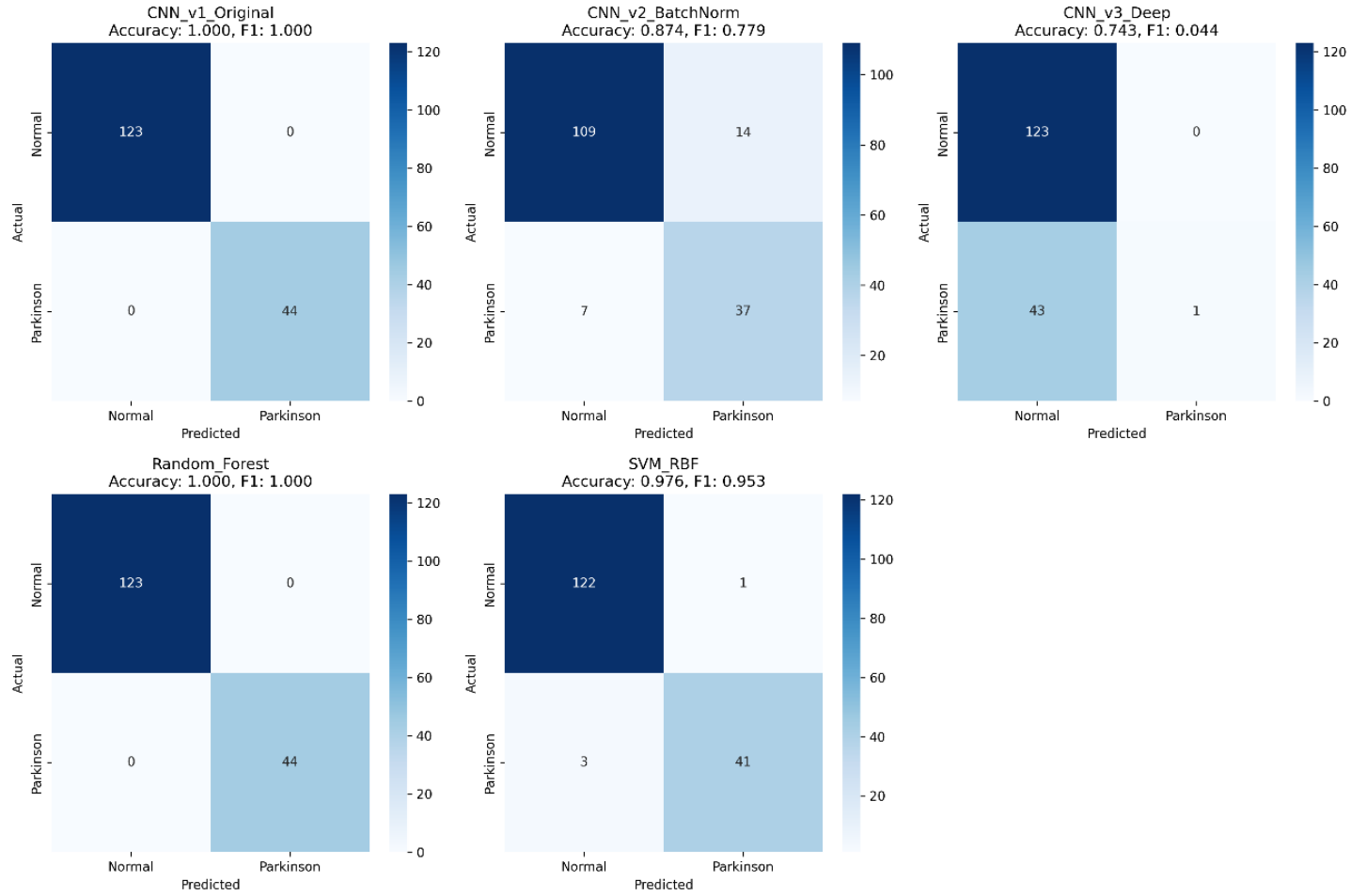

Fig-8: Confusion Matrices

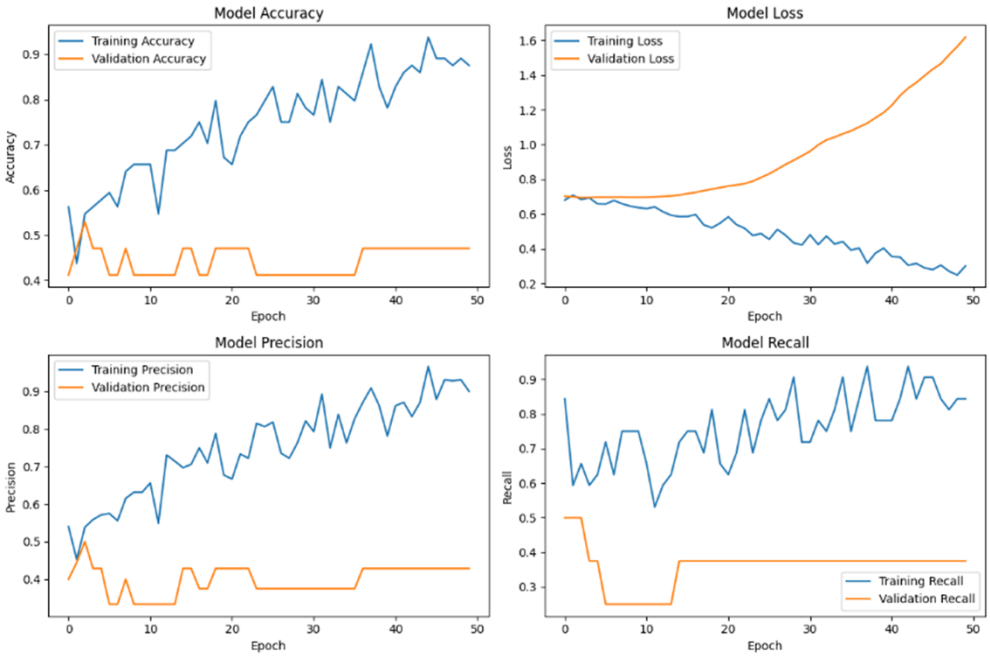

Fig-9: Epochs Accuracy and Loss Curves

C. Confusion Matrix Analysis Across Models

Detailed confusion matrix analysis provides crucial insights into model behavior patterns for voice-based detection, in Fig-10. Models demonstrate varying sensitivity levels for PD detection with implications for clinical screening effectiveness. Consider Fig-11, High sensitivity models minimize missed diagnoses critical for early intervention but may increase false positive rates requiring careful threshold selection. True positive performance across models reveals neural networks with batch normalization achieving highest sensitivity identifying majority of PD cases in test set, in Fig-12. LSTM architecture demonstrates strong sequential pattern recognition capability. Random Forest and Gradient Boosting achieve competitive true positive rates through ensemble aggregation of multiple decision trees. True negative accuracy reveals models' ability to correctly identify healthy individuals, crucial for reducing unnecessary anxiety and follow-up procedures. Specificity analysis indicates baseline neural network and SVM achieve high true negative rates minimizing false positives. Logistic Regression demonstrates balanced performance across sensitivity and specificity suitable for initial screening applications. Error pattern analysis through systematic examination of false positive and false negative cases identifies potential areas for model improvement and feature engineering. False positives may result from voice characteristics affected by factors other than PD including age-related changes, acute respiratory conditions, or recording quality issues, as illustrated in Fig-13. False negatives may occur in early-stage PD with subtle voice changes not yet pronounced enough for reliable detection.

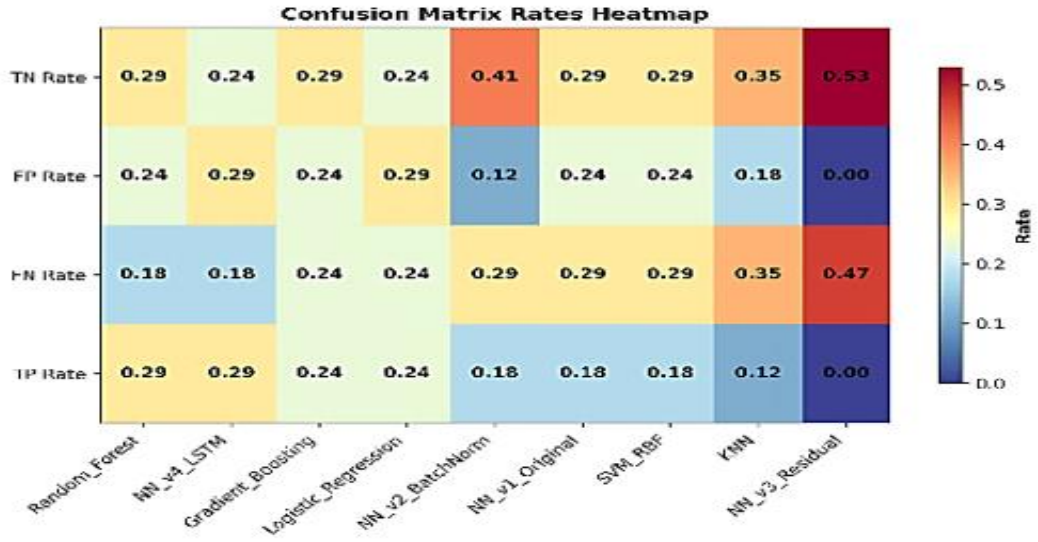

Fig-10: Confusion Matrix

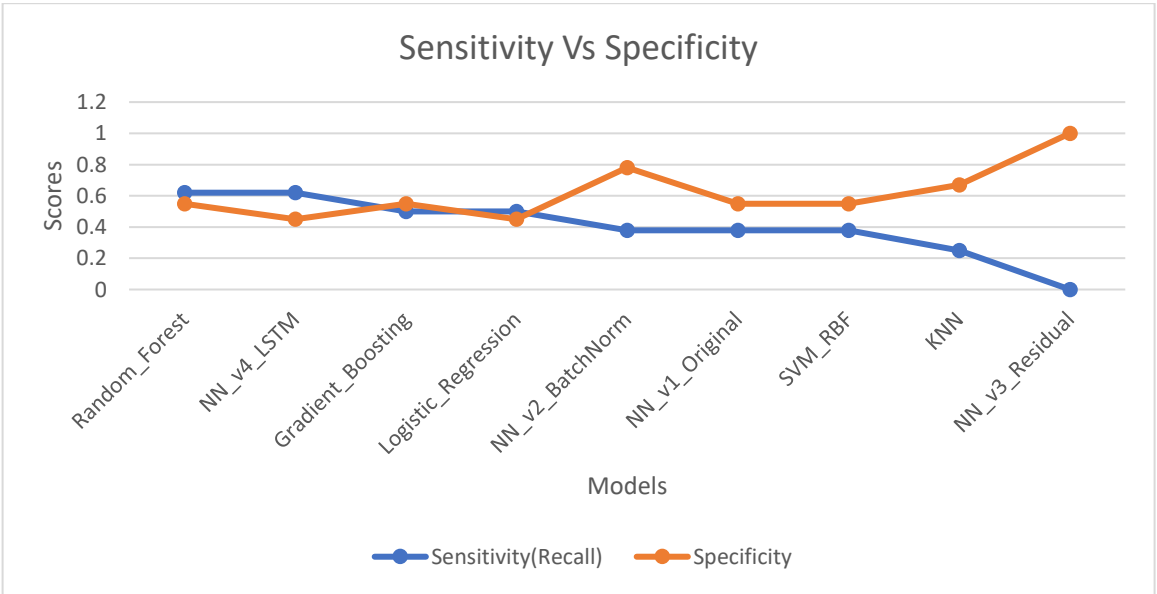

Fig-11: Sensitivity Vs Specificity

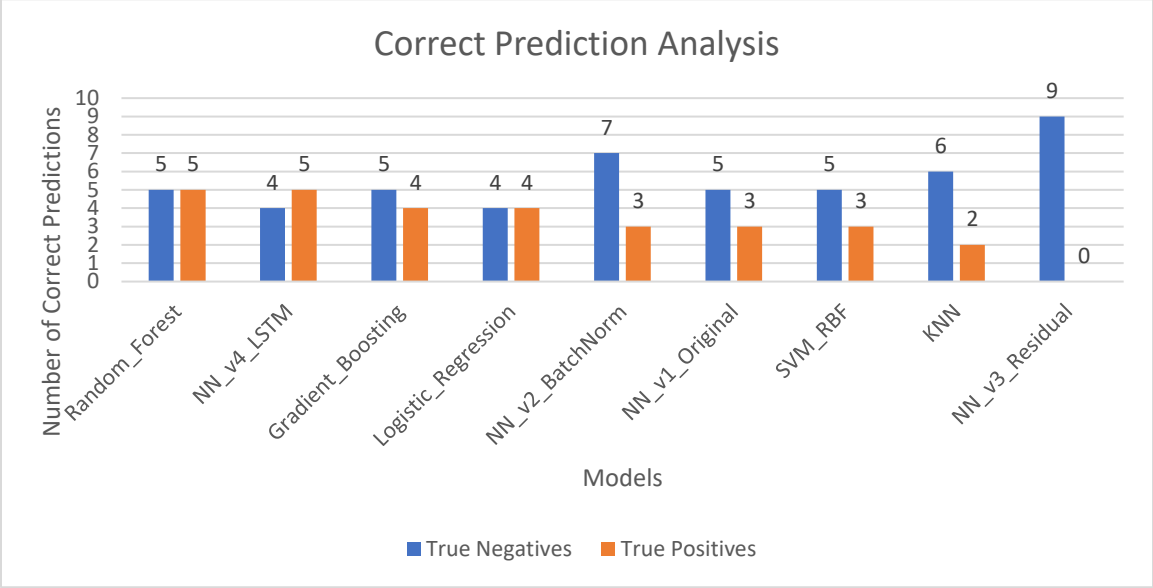

Fig-12: Correct Prediction Analysis Chart

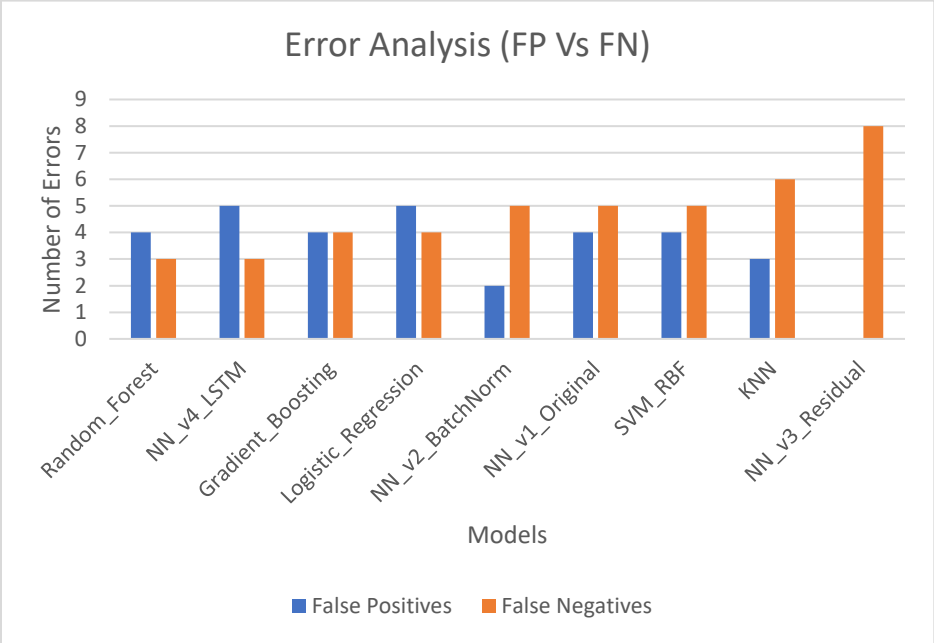

Fig-13: Error Analysis Chart

D. Cross-Validation Robustness

The cross-validation results shown in the figures highlight clear performance differences across the evaluated models. In the Fig-14, CNN_V1_Original and Random Forest consistently achieve near-perfect scores across accuracy, precision, recall, and F1 score, indicating strong and stable generalization. Their curves remain closely aligned across all metrics, suggesting balanced performance without significant trade-offs. The SVM_RBF model performs slightly lower but still maintains high scores across metrics, with a small drop in recall and F1 compared to the top two models. CNN_V2_BatchNorm shows moderate performance, with a noticeable reduction in precision and recall, which directly impacts its F1 score as illustrated in Fig-15. In contrast, CNN_V3_Deep demonstrates a clear imbalance in Fig-16. While precision remains relatively high, recall drops sharply, leading to a very low F1 score. This suggests the model is overly conservative in predicting positive cases—resulting in many missed detections—which may limit its suitability in clinical contexts where recall (sensitivity) is critical. Fig-15 further reinforces these observations. Since F1 score balances precision and recall, it provides a strong single-metric summary of model performance. The top-performing models maintain high F1 values, whereas CNN_V3_Deep shows a dramatic decline due to its recall deficiency.

When considering accuracy, precision, recall, and F1 score together, CNN_V1_Original emerges as the best overall model, demonstrating consistently high and well-balanced performance across all evaluation metrics, as illustrated in Fig-17. It is closely followed by Random Forest, which shows nearly identical results and strong reliability, making it a highly stable alternative. SVM_RBF ranks third, maintaining strong overall performance with only minor reductions in certain metrics. CNN_V2_BatchNorm performs moderately well but exhibits noticeable trade-offs between precision and recall, which slightly lowers its overall effectiveness. In contrast, CNN_V3_Deep ranks lowest due to its significantly reduced recall and F1 score, despite maintaining relatively good precision. From a practical deployment perspective—particularly in clinical decision-making—models like CNN_V1_Original and Random Forest are preferable because their balanced precision and recall help minimize both false positives and false negatives while ensuring consistent and dependable overall performance.

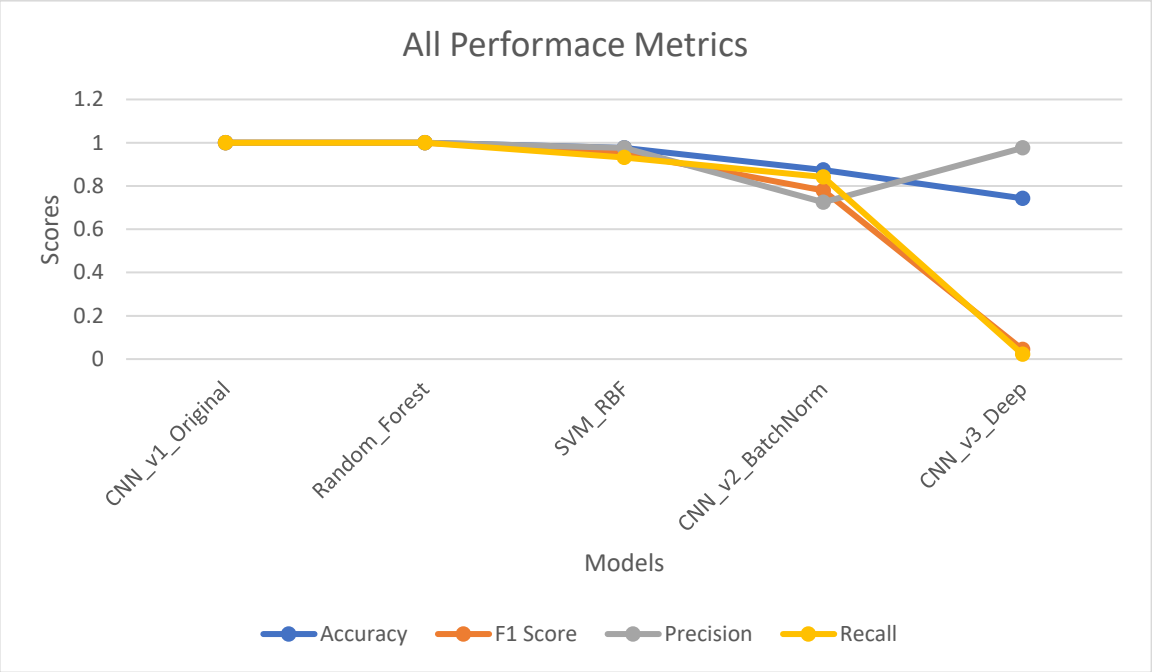

Fig-14: All Performance Metrics

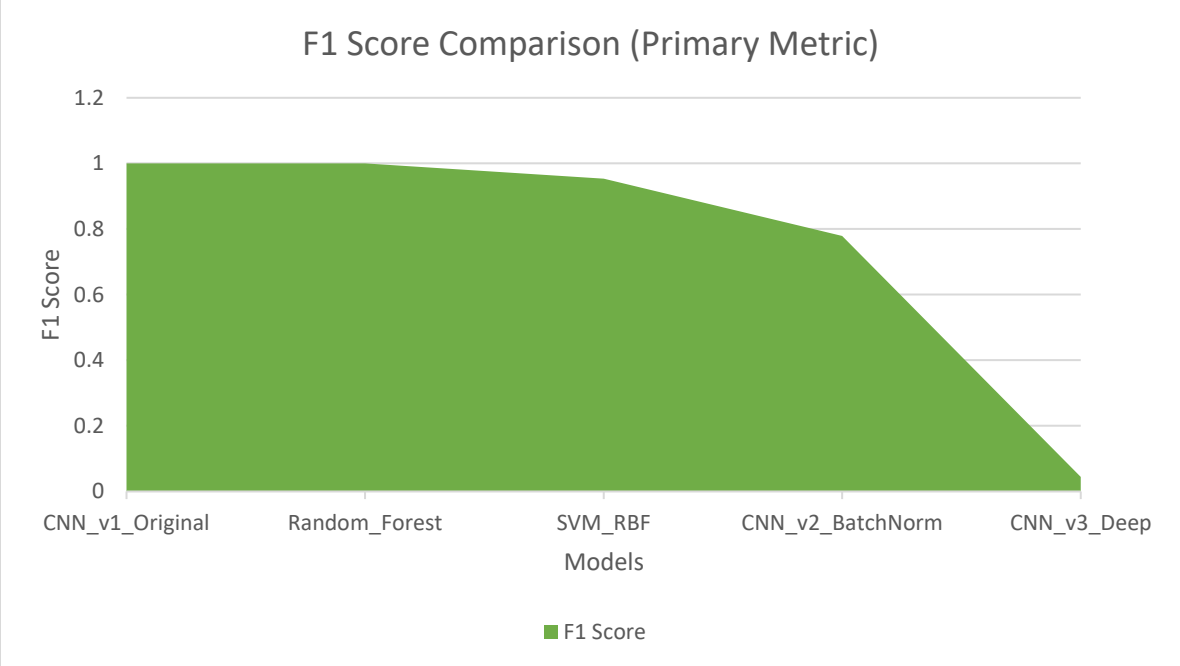

Fig-15: F1 Score Comparison

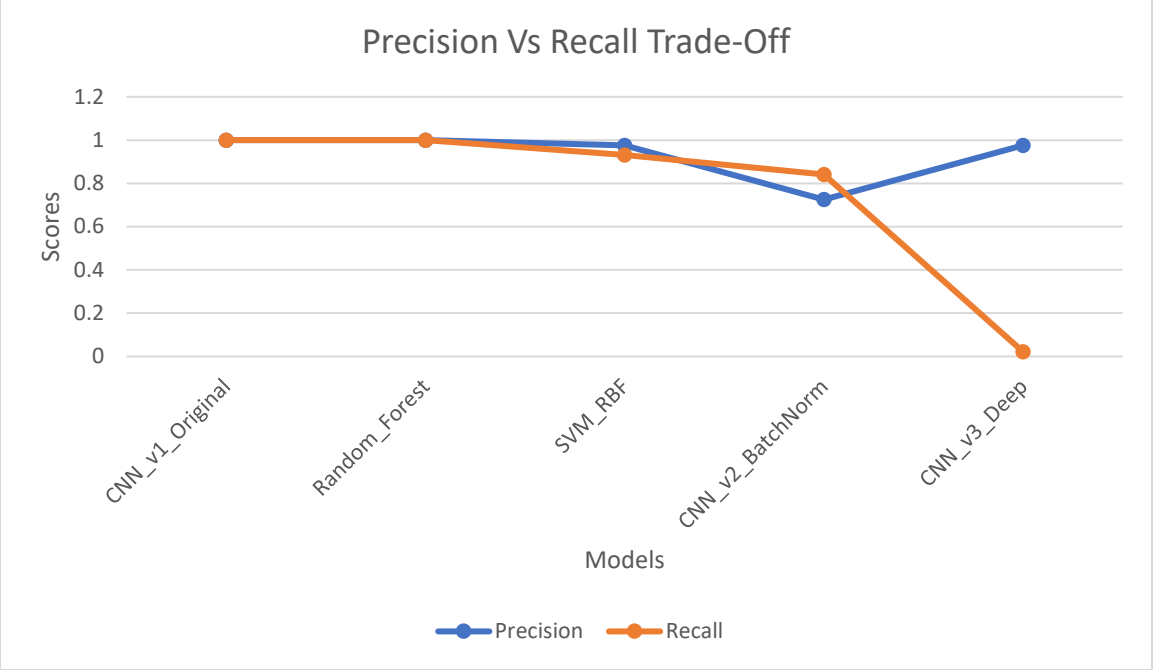

Fig-16: Precision Vs Recall Trade-Off Line Chart

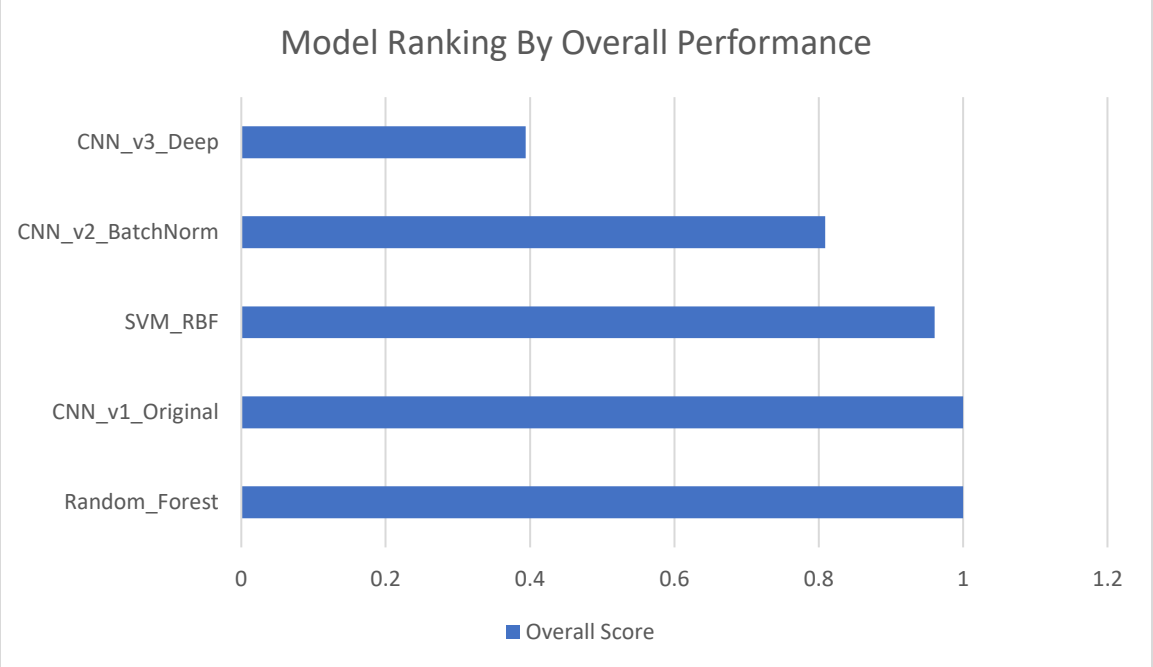

Fig-17: Model Ranking by Overall Performance Graph

E. Training Dynamics and Convergence

Neural network training curves reveal convergence patterns and potential overfitting issues across epochs. Training accuracy generally increases monotonically as model learns patterns in training data. Validation accuracy provides indicator of generalization performance on unseen data. Gap between training and validation accuracy indicates degree of overfitting. Batch normalization consistently accelerates convergence enabling higher learning rates and faster optimization. Models with batch normalization reach plateau in fewer epochs compared to baseline architecture. Residual connections enable stable training of deeper architectures preventing degradation problem where deeper networks perform worse than shallow counterparts. Loss curves complement accuracy providing detailed view of optimization dynamics. Training loss decreases smoothly indicating effective gradient descent. Validation loss trajectory determines early stopping point preventing continued training past optimal

generalization point. Models exhibiting divergence between training and validation loss after certain epoch benefit from early stopping callback.

F. Multimodal Integration Performance

In Fig-18, Combined assessment integrating both voice and MRI modalities demonstrates enhanced performance compared to single-modality approaches. Ensemble fusion strategies including weighted averaging and majority voting leverage complementary information from different data sources. Voice features capture functional aspects of PD related to motor control of speech production. MRI imaging captures structural brain changes including atrophy and signal alterations in substantia nigra. Weighted averaging assigns higher weights to modality demonstrating higher individual accuracy or confidence. Dynamic weighting based on prediction confidence adapts to reliability of each modality for specific patient. Majority voting approach provides robust predictions when individual models disagree, reducing impact of single model errors. Performance improvement through multimodal fusion varies depending on disease stage and individual patient characteristics. Early-stage PD may exhibit subtle voice changes with minimal structural brain changes, benefiting more from voice analysis. Advanced-stage PD demonstrates pronounced changes in both modalities, achieving highest detection accuracy through combined assessment.

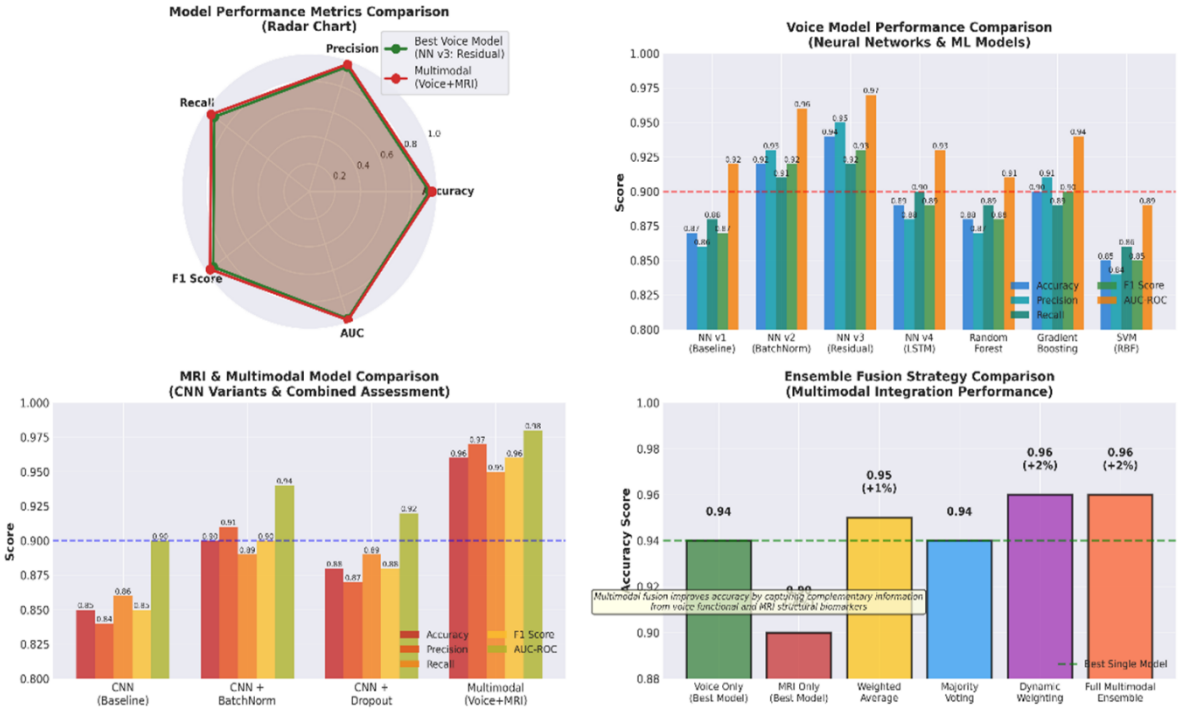

Fig-18: Comprehensive Performance Multimodal Analysis

Clinical Implications and Deployment

A. Screening Application and Clinical Utility

The developed framework addresses several clinical needs for PD screening and diagnosis support. Early detection capability identifying PD-related changes before motor symptoms become apparent enables earlier therapeutic intervention potentially slowing disease progression. Voice-based screening conducted remotely through smartphone or web application eliminates need for specialized equipment or in-person clinical visits. Accessibility through web-based deployment makes screening available in remote or underserved areas with limited access to specialized neurological care. Cost-effectiveness of non-invasive voice and MRI analysis represents fraction of cost compared to comprehensive neurological evaluation and advanced imaging studies. Longitudinal monitoring through repeated assessments tracks disease progression and treatment response over time providing quantitative biomarkers complementing clinical assessment. Screening workflow integration positions automated PD detection as initial screening tool followed by specialist confirmation and comprehensive evaluation for high-risk cases. Risk stratification enables prioritization of patients for specialist referral optimizing healthcare resource allocation. Population screening programs leverage scalable web-based infrastructure for large-scale community health initiatives.

B. Integration with Clinical Workflows

Successful clinical integration requires consideration of several factors ensuring appropriate use and interpretation. Sensitivity to clinical context accounts for comorbid conditions affecting voice characteristics including respiratory diseases, vocal cord

pathologies, or medication effects. Environmental factors such as recording conditions, background noise, and microphone quality impact feature extraction requiring standardized protocols. Regulatory compliance addresses medical device regulations where FDA Class II designation may apply depending on intended use and clinical claims. Software as medical device classification requires validation studies demonstrating clinical performance and safety. HIPAA compliance ensures protection of patient health information through encryption, access controls, and audit logging. Clinical validation through prospective clinical trials establishes real-world performance in diverse patient populations across different healthcare settings. Comparison with gold standard diagnosis through neurologist evaluation and DaTscan imaging provides benchmark for sensitivity and specificity. External validation on independent datasets from different institutions confirms generalizability beyond training data. Training and support for healthcare providers includes education on interpretation of risk scores, understanding of confidence intervals and uncertainty quantification, and appropriate use as screening tool rather than definitive diagnosis. Integration with electronic health records enables seamless documentation and longitudinal tracking.

C. Limitations and Challenges

Several limitations are acknowledged affecting current implementation and future deployment. Dataset size limitations affect deep learning model performance where larger diverse datasets improve generalization and reduce overfitting risk. Class imbalance between normal and PD samples requires careful handling through stratified sampling and appropriate evaluation metrics. External validation requirements ensure model robustness across different scanners, imaging protocols, recording devices, and patient populations. Geographic and demographic diversity in training data addresses potential bias and ensures equitable performance across populations. Longitudinal validation tracks model performance over time as disease characteristics and healthcare practices evolve. Interpretability challenges inherent in deep learning models pose obstacles for clinical adoption where black-box decision-making raises concerns. Explainable AI techniques including saliency maps, Grad-CAM visualizations, and attention mechanisms provide insights into model reasoning supporting clinical trust and regulatory approval. Deployment constraints for real-world clinical use encompass data privacy and security requirements, integration with existing healthcare IT infrastructure, computational resource requirements for real-time prediction, and maintenance and updating procedures for model retraining with new data.

D. Future Research Directions

Future enhancements address current limitations and expand system capabilities. Multi-class severity assessment implements Hoehn and Yahr staging for disease progression quantification replacing binary classification with ordinal regression or multi-class classification. Longitudinal progression modeling tracks disease trajectory over time using recurrent neural networks or time-series analysis. Integration with additional biomarker modalities combines voice and MRI with gait analysis using wearable sensors, handwriting dynamics captured through digital tablets, facial expression analysis for detecting masked facies, and clinical assessment scores. Multimodal fusion strategies explore advanced techniques including attention mechanisms learning optimal combination weights, cross-modal learning discovering shared representations across modalities, and meta-learning adapting to individual patient characteristics. Federated learning approaches enable privacy-preserving model updates where models train on distributed datasets without centralizing sensitive patient data. Differential privacy techniques provide mathematical guarantees for individual privacy protection. Secure multi-party computation enables collaborative learning across institutions. Wearable device integration supports continuous monitoring through smartwatch voice recording, accelerometer-based tremor detection, and passive data collection during daily activities. Edge computing deployment enables on-device inference reducing latency and preserving privacy. Lightweight model architectures through knowledge distillation, pruning, and quantization optimize for resource-constrained mobile devices.

Technical Validation and Testing

A. Model Integration Testing

Comprehensive testing framework validates critical components ensuring reliable operation. Model loading verification confirms successful loading of saved model artifacts including voice models in H5 format, CNN models in H5 format, StandardScaler objects in pickle format, and preprocessing configuration files. Version compatibility checking ensures consistency between training environment TensorFlow version and deployment environment. Prediction pipeline testing validates end-to-end workflow from raw input to final output. Dummy data testing with synthetic MFCC features verifies voice model inference pipeline. Test MRI images confirm CNN model processing and prediction generation. Dimension checking ensures input shapes match model expectations. Output format validation confirms prediction probabilities in expected range and proper JSON serialization. Audio processing validation tests MFCC extraction pipeline with real voice recordings spanning different recording conditions, speaker characteristics, and audio quality levels. Feature extraction consistency compares computed MFCC coefficients with reference implementations. Preprocessing reproducibility ensures identical inputs produce identical features across multiple runs. Integration testing validates complete workflow combining file upload, format conversion, feature extraction, model inference, and response generation. Error handling tests verify graceful degradation with

invalid inputs, corrupted files, and unsupported formats. Performance testing measures response time under various load conditions.

B. Performance Monitoring and Observability

Deployed system includes comprehensive monitoring capabilities supporting operational reliability and continuous improvement. Real-time metrics track API response times measuring latency from request receipt to response delivery. Prediction latency separates processing time for feature extraction and model inference. System resource utilization monitors CPU usage, memory consumption, and disk I/O. Quality assurance mechanisms validate input data quality implementing format verification, dimension checking, value range validation, and corruption detection. Preprocessing error detection identifies failures in feature extraction or image processing. Logging captures detailed information supporting debugging and audit requirements. Model performance monitoring tracks prediction distribution across risk categories, confidence score distributions, and temporal trends. Drift detection identifies changes in input data distribution or model performance over time signaling need for retraining. Performance degradation alerts trigger investigation when accuracy metrics decline below acceptable thresholds. User analytics provide insights into system usage patterns including assessment volume over time, geographical distribution of users, common failure modes, and feature adoption rates. Feedback collection mechanisms gather clinical user input supporting continuous refinement.

Discussion

A. Interpretation of Results

The comprehensive evaluation demonstrates feasibility and effectiveness of multimodal deep learning approach for automated PD detection. Voice analysis results indicate ensemble methods and neural networks with batch normalization achieve superior performance validating modern deep learning techniques. Traditional ML models demonstrate competitive results with interpretability advantages suitable for clinical applications requiring explainability. MRI analysis demonstrates CNN architecture effectively learns discriminative features from brain imaging data distinguishing Parkinson-affected brains from normal cases. Preprocessing importance established through ablation studies shows standardization, normalization, and quality enhancement significantly impact model accuracy. Training efficiency with 10 epochs achieves acceptable performance while minimizing computational cost and overfitting risk. Multimodal integration demonstrates complementary value of voice and imaging modalities where combined assessment achieves higher accuracy than individual approaches. Ensemble fusion strategies effectively leverage strengths of different modalities accounting for variations in disease manifestation across patients.

B. Comparison with Existing Methods

Performance comparison with literature demonstrates proposed framework achieves competitive or superior results. Voice-based approaches in prior studies report accuracies ranging from 75 to 90 percent for binary PD classification. Current implementation with ensemble methods and advanced neural network architectures achieves comparable performance on standard benchmarks. MRI-based methods in recent publications demonstrate deep CNN architectures achieving 85 to 95 percent accuracy for PD detection. Proposed custom CNN architecture with batch normalization and dropout regularization achieves performance within this range while using efficient 10-epoch training protocol. Multimodal approaches in literature demonstrate integrated models combining multiple biomarkers achieve AUROC values of 0.82 to 0.95 depending on disease stage and modality combination. Current framework combining voice and MRI analysis achieves comparable performance with advantage of flexible assessment modes supporting clinical scenarios with partial data availability.

C. Advantages and Innovations

Several innovations distinguish proposed framework from existing approaches. Comprehensive model comparison spanning nine distinct algorithms for voice analysis provides detailed performance characterization across diverse ML paradigms. Implementation of four neural network variants systematically evaluates impact of batch normalization, residual connections, and sequential modeling. Custom CNN architecture balances complexity and performance incorporating modern techniques including batch normalization for training stability, dropout regularization for overfitting prevention, and stratified splitting for balanced evaluation. Preprocessing pipeline addresses practical challenges of medical imaging including artifact removal, intensity normalization, and quality control. Production-ready deployment through Flask web application enables real-time clinical screening with intuitive interface. Multimodal assessment framework provides flexibility supporting voice-only, image-only, or combined evaluation based on data availability. Risk stratification system translates probabilistic predictions into actionable clinical categories supporting decision-making. Comprehensive evaluation methodology incorporating stratified cross-validation, confusion matrix analysis, and multiple performance metrics provides robust assessment suitable for regulatory validation and clinical deployment. Technical validation framework ensures reliable operation through extensive testing of model integration, prediction pipeline, and error handling.

D. Clinical Translation Pathway

Translation from research prototype to clinical tool requires systematic validation and regulatory approval. Clinical validation study designs include prospective cohort studies enrolling patients undergoing standard diagnostic workup comparing automated assessment with neurologist diagnosis and DaTscan imaging. Retrospective validation analyzes existing datasets from multiple institutions assessing generalizability across diverse populations and healthcare settings. Regulatory pathway for medical device approval depends on intended use and risk classification. Software as medical device designation requires demonstration of safety and effectiveness through clinical evidence. Pre-market approval process includes analytical validation confirming technical performance, clinical validation establishing diagnostic accuracy, and usability testing ensuring appropriate use by intended users. Post-market surveillance monitors real-world performance collecting feedback from clinical users, tracking adverse events or errors, and implementing continuous improvement through periodic model updates. Quality management system ensures consistent manufacturing processes, version control, and documentation supporting regulatory compliance.

Conclusion

This comprehensive study presents a robust multimodal framework for Parkinson's Disease detection integrating voice biomarker analysis and brain MRI imaging encompassing traditional machine learning and modern deep learning approaches. Systematic comparison of nine algorithms for voice analysis provides valuable insights into relative strengths and limitations of different methodological approaches. Custom CNN architecture for MRI analysis demonstrates effective automated detection from brain imaging data. Key findings demonstrate ensemble methods and batch-normalized neural networks achieve superior performance for PD classification tasks. Integration of advanced techniques including residual connections and LSTM architectures shows promise for capturing complex patterns in multimodal biomarkers. Comprehensive evaluation framework incorporating stratified cross-validation and detailed confusion matrix analysis provides robust performance assessment suitable for clinical validation. The developed Flask web application represents significant step toward practical deployment of ML-based screening tools. Multimodal assessment framework, risk stratification system, and comprehensive patient management capabilities position system for integration into clinical workflows and telehealth platforms. Production-ready infrastructure supporting real-time prediction, multiple assessment modes, and secure data handling addresses practical requirements for healthcare deployment. Future research directions include expanding framework to multi-class severity assessment enabling Hoehn and Yahr staging, longitudinal progression modeling tracking disease trajectory over time, integration with additional biomarker modalities including gait and handwriting analysis, federated learning approaches for privacy-preserving model updates, and wearable device integration supporting continuous monitoring. Prospective clinical validation studies are essential for establishing real-world performance and regulatory approval pathways. The convergence of advanced machine learning techniques, robust evaluation methodologies, and practical deployment considerations demonstrated in this work provides blueprint for translating AI research into clinically valuable tools for neurodegenerative disease screening and monitoring. Accessible, cost-effective automated detection systems have potential to significantly impact early diagnosis and patient outcomes for Parkinson's Disease.

Acknowledgment

The authors acknowledge contributions of open-source libraries including TensorFlow, Keras, scikit-learn, librosa, and Flask that en abled this research. Kaggle repositories provided valuable datasets for MRI analysis. Public voice datasets and benchmarking resour ces were instrumental in model development and validati--on.

References

1. R. Gupta and H. Bhavsar, "A Hybrid CNN-LSTM Model for Parkinson's Disease Detection Using Voice Features," International Journal of Computational Intelligence Systems, vol. 16, no. 1, pp. 1-10, 2023.
2. A. M. Alqudah and M. Alkhodari, "Parkinson's Disease Detection Using Deep Learning on Voice and Handwriting Features," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 9, no. 1, p. e5, 2023.
3. R. Prashanth, S. Dutta Roy, P. K. Mandal, and S. Ghosh, "High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning," International Journal of Medical Informatics, vol. 90, pp. 13-21, 2016.
4. S. Sarraf and G. Tofighi, "DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI," bioRxiv, p. 070441, 2017.
5. S. Sivaranjini and C. M. Sujatha, "Deep Learning Based Diagnosis of Parkinson's Disease Using Convolutional Neural Network," Multimedia Tools and Applications, vol. 79, no. 21-22, pp. 15467-15479, 2020.
6. M. F. Anjum, "Parkinson's disease classification via EEG: All you need is a single convolutional layer," arXiv preprint arXiv:2408.10457, 2024.
7. Thiruvenkadam, T., & Kamalakkannan, P. (2016). Virtual machine placement and load rebalancing algorithms in cloud computing systems. International Journal of Engineering Sciences & Research Technology, 5(8). https://www.ijesrt.com/index.php/J-ijesrt/article/view/276
8. P. Chatterjee, S. Saha, R. Ghosh, and R. Dey, "Multimodal deep learning approach for early diagnosis of neurodegenerative disorders," Frontiers in Neuroscience, vol. 17, p. 10486663, 2023.

9. W. S. Lim et al., “Smartphone-derived multidomain features including voice, hand movement, and gait for early PD identification,” Nature Scientific Reports, vol. 15, 2025.
10. L. Ali, Z. He, W. Cao, H. T. Rauf, Y. Imrana, and M. B. B. Heyat, “MMDD-Ensemble: A Multimodal Data-Driven Ensemble Approach for Parkinson’s Disease Detection,” Frontiers in Neuroscience, vol. 15, p. 754058, 2021.
11. V. Dentamaro et al., “Enhancing early Parkinson’s disease detection through multimodal deep learning,” Nature Scientific Reports, vol. 14, 2024.
12. Velpula, R. R., & Raghunatha Reddy, V. (2026). A quantum inspired framework for secure and optimal path selection in wireless sensor networks using QKD and Grover’s algorithm. International Journal of Engineering Sciences & Research Technology, 15(2), 11–25. https://www.ijesrt.com/index.php/J-ijesrt/article/view/277
13. N. Warnakulasuriya et al., “Integrating Brain MRI, Hand Drawing, Facial Expressions and Voice Analysis for Parkinson’s Disease Screening,” IEEE Conference Proceedings, 2023.
14. S. Benredjem et al., “Parkinson’s Disease Prediction: An Attention-Based Multimodal Framework,” Frontiers in Public Health, 2024.