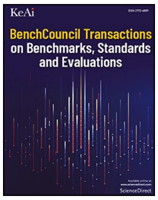
Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/


KeAi
BenchCouncil Transactions on Benchmarks, Standards and Evaluations

Research Article

Comparative study of deep learning models for Parkinson's disease detection☆

Abdulaziz Salihu Aliero *, Neha Malhotra

School of Computer Application, Lovely Professional University Phagwara, Punjab, India

ARTICLE INFO

Keywords:
Deep learning
Multilayer perceptron
Early
Detection
Parkinson's
Disease

ABSTRACT

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects movement and cognition, impacting millions of people worldwide. The diagnosis of PD primarily relies on clinical tests, which can often result in delayed identification of the disease. Recent advancements in data-driven methods using deep learning have demonstrated potential for improving early diagnosis by utilizing clinical and vocal inputs. This study conducted a comparative analysis of five deep learning models: Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Autoencoder, and Generative Adversarial Network (GAN), specifically for the detection of PD using vocal biomarkers. Among these models, the MLP achieved the highest predictive accuracy at 97.4 %. The RNN, GRU, and Autoencoder models attained a similar accuracy rate of 87.2 %. In contrast, the GAN model yielded an accuracy of only 76.9 %. The UCI vocal dataset from Kaggle was utilized in this research, along with extensive data preprocessing techniques to address missing values. Performance evaluation was conducted using multiple metrics. The results indicate that deep learning models can effectively diagnose PD using voice data, suggesting their potential to enhance diagnostic accuracy and support clinical decision-making. Furthermore, these models are feasible for large-scale integration into clinical workflows.

1. Introduction

Currently after Alzheimer's disease, Parkinson's disease (PD) is considered the second most prevalent chronic and rapidly progressing neurodegenerative disorder, affecting millions of people worldwide. Despite receiving a diagnosis, many individuals succumb to the disease. Over time, the number of new cases has continued to rise, as research predicts by 2040, over 17 million individuals will be impacted by (PD) [1].

Besides, the disease causes profound damage to the brain cells that produce dopamine, a very critical neurotransmitter that is in charge of movement and coordination [2]. The neuronal damage is followed by a loss in dopamine, ultimately leading to motor and non-motor symptoms [3].

The motor symptoms typically become evident many years after the onset of the disease after experiencing a prodromal period with non-motor symptoms [4]. The disease targets particular regions of the brain, the substantia nigra and the superior colliculus in the midbrain [5]. (See Fig. 1). The superior colliculus, is another region of the midbrain that handles visual information and eye movement, this reduces the quality of life considerably. Interestingly enough, early identification is crucial for the management of the disease and provision of more treatments that would slow down the progression [6]. With this, having more advanced diagnostic tools for medical practitioners in order to detect the disease earlier before progressing to advanced stages are absolutely essential. In the same way, detection in the prodromal stage rather than the postmotor stage would allow effective neuroprotection with the objective of delaying advanced motor symptoms. It is important to note that reducing the disease burden is eventually capable of increasing the quality of life in the patient for a more extended period by enabling early detection [7].

Conventional techniques particularly Machine Learning (ML) for PD detection heavily depends on clinical evaluation with limited data and motor symptoms. These symptoms are effective for later stages of PD but not as sensitive for detection of the disease in early stages. This occurs as motor symptoms appear following significant damage in the dopaminergic cells of the brain with approximately 50–70 % of dopaminergic cells in the substantia nigra being destroyed [8]. Early diagnosis is thus

☆ Peer review under the responsibility of The International Open Benchmark Council.

* Corresponding author.

E-mail address: abdulaziz.12306122@lpu.in (A. Salihu Aliero).

https://doi.org/10.1016/j.tbench.2025.100219
Received 12 April 2025; Received in revised form 1 June 2025; Accepted 12 June 2025
Available online 13 June 2025

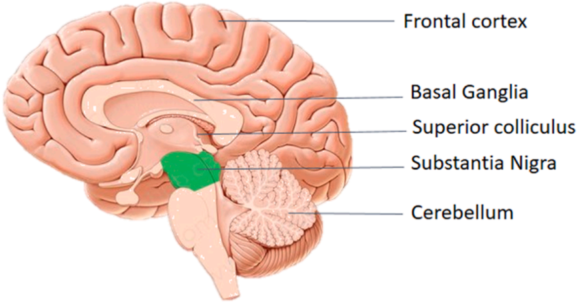


Fig. 1. Substantia nigra in the midbrain.

challenging with ML, prompting diagnostic tools that can diagnose the disease at preclinical or early stages by using deep learning (DL) modalities [9]. This is a highly specific technique in terms of equipment. Moreover while, DL provide some diagnostic assistance in detecting PD, early diagnosis of this disease remains a problem. Because non-motor symptoms become apparent after a long period, most patients receive a diagnosis after a considerable portion of the dopaminergic brain system is affected.

However, the highly subjective nature of current clinical observations and the insensitivity in imaging the identification of early neuronal loss indicate a crucial necessity [10]. With these cutting-edge techniques of using Artificial intelligence (AI), specifically DL algorithms, as a means of obtaining more and better diagnostic tools, AI is among the novel methods that can help fill gaps in early diagnosis of PD using clinical data analysis. DL revolutionized the ML field in recent times this is no longer a secret, as it has made it possible for anyone to be able to design models that can process large and complex sets of information, also it has been an innovative technology in medical diagnosis with enhanced precision in PD identification in all data modalities [11]. Deep learning models, which can analyze progressively higher amounts of information, primarily imaging and clinical parameters, can do so in a way that associates them with PD illnesses with a remarkably high level of precision in clinical diagnosis. In the previous few years, a class of deep learning architectures, e.g., convolutional neural networks (CNNs), are brilliant in learning how to automatically extract features from raw points instead of engineering them. In PD diagnosis, this is especially critical as human experts argue that the detection of subtle patterns in the imaging or clinical data should give rise to significant improvements in early detection[12]. Although traditional ML models have demonstrated promising results in PD detection, but DL have exhibited superior performance, particularly in handling large and complex datasets. Deep learning models have the advantage of automatic feature extraction, reducing the reliance on manual feature engineering and has the capacity and ability to handle highly complex architectures, often comprising hundreds or even thousands of layers.

The main goal of this study is to analyse deep learning models for PD detection, even before the onset of motor symptoms. It aims to create diagnostic models that match or surpass existing methods in accuracy. Additionally, the study analyse five deep learning models for PD detection using voice data, in other to identify the most accurate and robust model. Furthermore, it focuses on building a scalable model that can be effectively deployed across diverse healthcare settings while maintaining high accuracy. Various aspects of model performance evaluation were considered by ensuring that the model is robust enough to function well with any patient dataset [13]. Another way to frame this work is ensuring that the model is robust enough to behave well on any configuration of patient dataset. The goal is to ensure that the model works equally well on different patient types, providing a level of confidence one can have in diagnosis regardless of patient population beyond the one in which it trained.

The structure of the study is elaborated as: Section II. Describes related work, Section III describes methods used and dataset processing. Results and discussion in Section IV, and lastly Section V, encloses conclusion and future directions.

2. RELATED WORK

Recently research efforts has been put on detecting Parkinson's disease (PD) by using clinical data, voice data and MRI image data based on machine learning (ML) and deep learning (DL) techniques. This is because, these non-traditional methods hold promise for spotting stable warning signs which often escapes detection via traditional diagnostic means. In this regard, the diagnosis of PD has traditionally relied on clinical evaluations involving motor function assessments, speech and gait analysis, and neuroimaging techniques. However, these conventional methods often lack sensitivity, particularly in detecting early-stage PD when timely intervention is most beneficial [14]. In recent years, ML and DL techniques have shown promise in enhancing diagnostic accuracy by leveraging diverse clinical datasets. Early applications of ML for PD detection primarily focused on voice data, as vocal impairments are common among PD patients. Various studies have demonstrated the effectiveness of ML models when combined with feature selection techniques. For instance, Saeed et al. [15] uses feature selection techniques based on filters and wrappers to process voice recordings, achieving an accuracy of 88.33 % with k-nearest neighbors (KNN). Another study [16] illustrate using DL techniques on clinical data, specifically voice signals for the early diagnosis of PD and improved diagnostic accuracy. Similarly, Singh et al. [17] applied decision trees, random forests, and logistic regression to voice-based features, reinforcing the potential of ML-based techniques as cost-effective and non-invasive tools for PD detection. Beyond voice analysis, recent research has expanded the scope of clinical data used in PD detection. Lin et al. [18] employed motion data collected via inertial measurement units (IMUs) to develop neural network models capable of distinguishing healthy individuals from PD patients. Their study achieved detection rates of 92.72 % for advanced-stage PD and 99.67 % for early-stage cases, highlighting the growing role of gait analysis in early diagnosis.

Chintalapudi et al. [19] compared three DL architectures RNN, MLP, and LSTM on voice features of PD patients. Their findings revealed that the LSTM model achieving an accuracy of 99 %. This study underscores the ability of DL models, particularly LSTM, to effectively process nonlinear and complex data such as speech recordings. In contrast, traditional ML models require extensive preprocessing and manual feature selection, which can be labor-intensive and less scalable. Similarly Kurmi et al. [20] implemented an ensemble of CNN models, including VGG16, ResNet50, Inception-V3, and Xception, to analyze DaTscan images, achieving a classification accuracy of 98.45 %. Their study highlights the potential of CNN-based architectures in improving diagnostic precision through multi-model integration with DL that has been effectively applied to medical imaging for PD detection. Despite the advantages of DL, traditional ML approaches remain relevant, particularly when computational resources are limited or when datasets are small. For instance, Govindua et al. [21] compared random forests, SVM, and logistic regression for PD detection using voice data, achieving an accuracy of 91.83 %. While DL excels in large-scale datasets, ML techniques can still deliver competitive results when optimized with feature selection techniques. Another study by [22] explores ML and DL models for classifying PD using speech data. The results indicate that DL models outperform traditional ML approaches, with ML achieving an accuracy of 92.18 %, while the most effective DL model attains the highest accuracy of 95.41 %.

Neural networks (NN), particularly CNNs and LSTMs, have been extensively used in medical diagnostics due to their ability to model complex patterns in diverse clinical datasets, including images, time-series data, and auditory signals. Several studies have leveraged LSTM and CNNs [23,24] to analyse different data types such as gait, speech, and handwriting for PD detection. Similarly NN has also been used in multimodal data, where different types of clinical measurements were integrated into providing a more comprehensive diagnostic model. Taleb et al. [25] used CNN-BLSM architectures on motion HandPD_MultiMC_data collected via wearable sensors, achieving 97.62 % for PD stage classification. These findings emphasize the efficacy of ensemble models in enhancing diagnostic accuracy by capturing both static and dynamic characteristics of PD symptoms. Furthermore, LSTMs on the other hand, have been widely adopted for sequential data modeling, making them particularly suitable for voice and motion analysis in PD detection. Neural networks have also been integrated into multimodal frameworks, where multiple clinical data sources are combined for more comprehensive diagnostic models. CNNs have been applied to EEG data analysis for PD classification. Sugden and Diamandis [26] developed a channel-wise CNN model that achieved an accuracy of 80.4 %, this study highlights CNN's has the ability to extract spatial patterns from EEG data, which could be extended to other clinical modalities. Furthermore Majhi et al. [27] explored Hybrid DL models including Grey Wolf Optimization (GWO) optimization, which applied to two images datasets, these metaheuristic algorithm achieved 99.94 % accuracy. Another study by Islam et al. [28] conduct Extensive review for PD detection, their review concluded that voice and handwriting dataset integration significantly enhances diagnostic accuracy, especially in early-stage PD detection rather than just using images data. A study by [29] employed lightweight, pre-trained DL models with two-fold training and merged ideal features for hand-drawn Parkinson's disease screening. Similarly, Keles et al. [30] used Part-Aware Residual Network (PARNet), retrained from a COVID-19 model, and applied to SPECT images for PD detection. Al-Tam et al. [31] used Ensemble learning with stacking and bagging applied to two benchmark PD datasets to Enhance PD diagnosis through stacking ensemble-based ML approach, the technique seeks to improve generalisation performance, rectify dataset class imbalances, and raise the overall accuracy of PD detection. Furthermore, Ismail and Osman [32] performed a Classification of scalograms using AlexNet, GoogleNet, and ResNet50, followed by a hybrid system based on majority voting. Also, DenseNet and NasNet are used for different classifications. Hongyi et al. [33] Used brain images and radiomics-based automated hybrid approach targeting midbrain for early PD detection. Table 1. Presents a comprehensive assessment of research studies on PD detection using ML and DL techniques. Building on existing studies, our focus is to demonstrate that voice data can effectively identify PD in its early stages.

Numerous research studies, including those cited in references [14–16] and [19–20], have explored the application of machine learning (ML) and deep learning (DL) techniques for detecting Parkinson's disease (PD) through voice data. These studies often utilize models such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Recurrent Neural Networks (RNN), and Long Short-Term Memory networks (LSTM), typically examining either individual architectures or making pairwise comparisons. However, there is a significant gap in the literature regarding systematic comparative analyses of multiple deep learning models assessed under consistent experimental conditions. This study aims to fill that gap by comparing five distinct deep learning models within a unified evaluation framework, assessing their performance on a standardized voice dataset.

Our interpretations focused on both the classification accuracy itself and the potential operationalization of the systems, which bring implications for future clinical integration. Therefore, this study conducts a comparative analysis of five deep learning models applied to voice data for Parkinson's disease detection.

3. METHODOLOGY

The proposed method was implemented on the Google Colab platform using Python programming and incorporates five models to predict whether the patient has Parkinson's disease or not. To ensure validation of the model on unseen data and avoid fitting the training data too well, the dataset was divided into training sets, validation sets, and test sets.

3.1. Dataset and preprocessing

This study employs a publicly available voice dataset derived from the UCI Repository, which is also available on Kaggle, comprising the collected acoustic characteristics that are utilised to identify PD. The

Table 1
A summary of research studies focused on Parkinson's Diseases.

Ref.	Study Focus	Dataset Used	Model(s) Used	Best Performance
Al-Nefaie et al. [14]	Highlights the effectiveness of ML classifiers with vice data	UCI Voice dataset	KNN, SVM, RF and LR	SVM and RF achieve 95.00 %
Faisal et al. [15]	Voice data for PD detection	Voice recordings	KNN with feature selection	88.33 %
Awais et al. [16]	DL on voice signals for PD	Clinical voice signals	Deep Learning models	86.00 %
Shikha et al. [17]	ML models on voice-based features	Voice recordings	Decision Tree, RF, Logistic Regression	92.00 %
Lin et al. [18]	Motion data analysis using IMUs	Motion sensor data	Neural Networks	99.67 %
Chintalapudi et al. [19]	Comparison of DL models on voice	Voice features	RNN, MLP, LSTM	99.00 %
Kurmi et al. [20]	DaTscan image analysis	DaTscan images	CNN Ensemble (VGG16, ResNet50, etc.)	98.45 %
Alshammri et al. [21]	ML model comparison using voice	Voice data	RF, SVM, Logistic Regression	91.83 %
Rahman et al. [22]	ML vs DL on speech data	Speech recordings	ML and DL models	95.41 %
Govindu et al. [23]	Use of ML in Telemedicine	MDVP Audio data	SVM, RF, KNN and LR	91.83 %
Rehman et al. [24]	Multimodal data (speech, gait, handwriting)	Various clinical datasets	Hybrid LSTM-GRU	98.00 %
Taleb et al. [25]	CNN-BLSM on Handwriten data	HandPD_MultiMC_data	CNN-BLSM	97.62 %
Sugden et al. [26]	EEG analysis with CNN	EEG recordings	Channel-wise CNN	80.04 %
Majhi et al. [27]	PD Diagnosis using hybrid deep learning and metaheuristic optimization	(T1, T2-weighted) MRI & SPECT DaTscan	DenseNet, InceptionV3, LSTM, Grey Wolf Optimization (GWO), VGG16,	99.94 %
Islam et al. [28]	ML/DL models for PD detection using handwriting and voice data	Handwriting & Voice	Various ML & DL models	95.41 %
Rajinikanth et al. [29]	Hand-Sketch-based PD screening using pre-trained DL models	Hand-Sketch Wave Data	MobileNet, KNN	100 %
Keles et al. [30]	PD detection using a retrained COVID-19 model on SPECT images	SPECT Imaging	PARNet, ResNet	95.43 %
Al-Tam et al. [31]	Stacking ensemble ML models for PD detection	PD Benchmark Datasets	Random Forest, SVM, Gradient Boosting, Logistic Regression	96.18 %
Ismail and Osman [32]	Scalogram-based PD detection with deep learning	Speech Signals	AlexNet, GoogleNet, ResNet50, DenseNet, NasNet	95.00 %
Hongyi et al. [33]	Radiomics-based deep learning for early PD detection	MRI (Midbrain & Substantia Nigra)	YOLO v5, LeNet	96.03 %

dataset contains 195 voice recordings from PD positive and negative participants. It comprises 21 attributes extracted from voice records; the features are describe in (Table 2) [22], they provide full information about vocal patterns to allow effective classification. For example, Srinivasan et al. [34] have used the UCI dataset and voice samples of matched controls and PD patients to classify subjects based on voice feature changes using supervised machine learning methodologies.

The dataset was separated into target labels (y) and input features (x). Using train_test_split, a stratified split was carried out, with 20 % going to testing and 80 % going to training. The target variable indicates whether the subject is PD-positive (1) or healthy (0) [22]. (Fig. 2) illustrates the typical process that are followed, it explains how the dataset was splits into training and test data, train five models on the dataset, and validate the results using test data.

Min-Max Scaling was used to normalise the features to a [0, 1] range prior to model training. In order to express time-steps for recurrent models (LSTM and GRU), the input was transformed into a three-dimensional format.

During pre-processing, particularly when working with clinical data it is essential to address the Outliers and values that are missing. If these issues are not properly addressed, it can result in a shortage of accurate models and exacerbate future challenges. A study by [36], applied different approaches to solve this problem enabling a complete dataset to train the model. Multiple imputation techniques were applied to handle the missing clinical assessment data, and subsequently, the duplicate value was also addressed. Another study by [37] identified the requirement for the identification of outliers in speech data and their exclusion as they play a major role in the estimation of a model with undesirable or abnormal values; data was normalized with the z-score and visually inspected for outliers and marked accordingly. After the exclusion of outliers, highly correlated features were identified with the correlation coefficient validating their appropriateness. (Fig. 3) below illustrates the correlation of the dataset features when outliers are excluded [35].

Table 2
Dataset Features [35].

Feature Name	Description
Average Fundamental Frequency (Fo)	Mean vocal fundamental frequency measured in Hertz (Hz).
Maximum Fundamental Frequency (Fhi)	Highest recorded vocal fundamental frequency (Hz).
Minimum Fundamental Frequency (Flo)	Lowest recorded vocal fundamental frequency (Hz).
Jitter (%)	Percentage-based measure of frequency variation.
Jitter (Abs)	Absolute measurement of frequency instability.
Relative Average Perturbation (RAP)	Short-term variation in fundamental frequency.
Pitch Period Perturbation Quotient (PPQ)	Measure of long-term frequency variations.
Jitter: DDP	Three-point average of absolute jitter values.
Shimmer	Quantifies amplitude variation in the voice signal.
Shimmer (dB)	Measurement of amplitude fluctuation in decibels (dB).
Amplitude Perturbation Quotient (APQ3)	Average of amplitude variations over three cycles.
Amplitude Perturbation Quotient (APQ5)	Average of amplitude variations over five cycles.
MDVP: APQ	Measures overall amplitude perturbation across the signal.
Shimmer: DDA	Three-point averaged shimmer calculation.
Noise-to-Harmonics Ratio (NHR)	Ratio of non-harmonic noise to harmonic components.
Harmonics-to-Noise Ratio (HNR)	Ratio measuring harmonic signal strength relative to noise.
Correlation Dimension (D2)	Nonlinear measure of signal complexity.
Fractal Scaling Exponents	Quantifies the self-similarity and complexity of the signal.

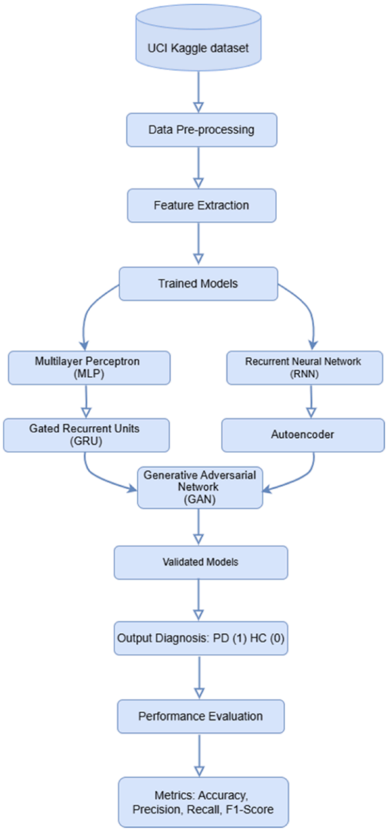

UCI Kaggle dataset
Data Pre-processing
Feature Extraction
Trained Models
Multilayer Perceptron (MLP)
Recurrent Neural Network (RNN)
Gated Recurrent Units (GRU)
Autoencoder
Generative Adversarial Network (GAN)
Validated Models
Output Diagnosis: PD (1) HC (0)
Performance Evaluation
Metrics: Accuracy, Precision, Recall, F1-Score

Fig. 2. Proposed flow of our approaches.

3.2. *Multilayer perceptron (MLP)*

The MLP model was implemented with Scikit-learn's MLPClassifier with two hidden layers, with (100, and 50) neurons respectively, used ReLU as an activation function and Adam as the optimizer, this constitutes a deep neural network by definition, we recognize that it is relatively shallow compared to modern deep learning architectures such as CNNs or LSTMs. This configuration was selected to evaluate how well a simple deep model could perform on the voice dataset for Parkinson's Disease detection. The model was trained up to 500 epochs. The evaluation metrics like accuracy, precision, recall and F1-score were computed based on predicted values on the test set [38].

For visualization, confusion matrices and heatmaps were used. A heatmap of the classification report was also generated to give a more visual summary of the performance metrics across the two classes.

3.3. *RNN-LSTM model*

The LSTM based RNN is built with TensorFlow's Keras API. The architecture consisted of one LSTM layer (with 50 units) followed by Dense output layer with sigmoidal activation for binary classification. Binary cross-entropy was used as the loss function [39], and the Adam optimizer was used to compile the model and it trained for 50 epochs with a batch size of 32.

The reshaping of inputs was performed in a way suitable for the LSTM model as 3D tensors. Accuracy, classification reports and confusion matrices were used to evaluate performance.

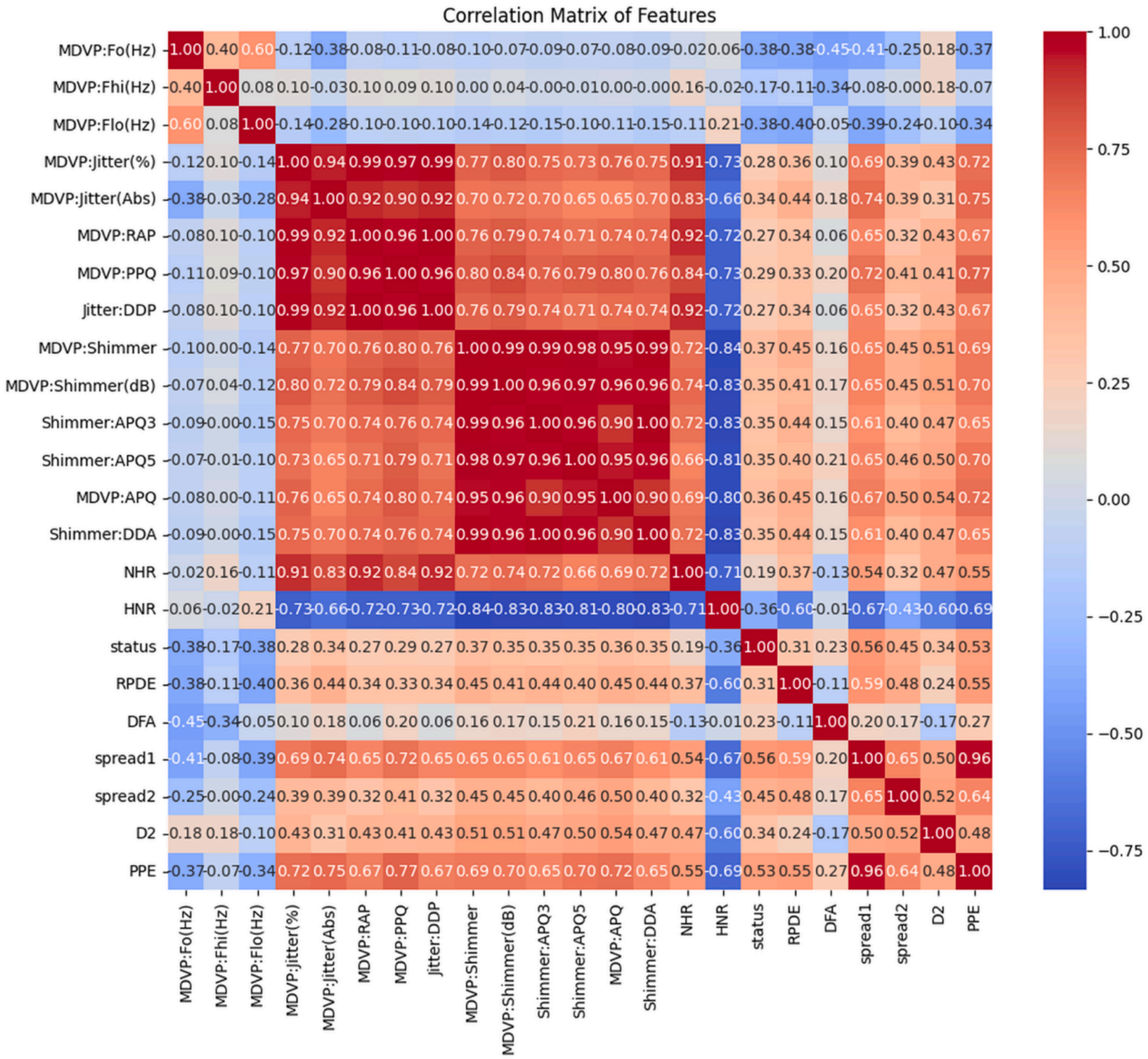


Fig. 3. The Correlation matrix.

3.4. Gated recurrent unit (GRU) network

The GRU-based model had a similar structure as LSTM architecture, but used a GRU layer instead of LSTM layer. It only had a GRU layer with 50 hidden units followed by a dense layer with sigmoid as an activation function. All recurrent models were trained with the same configuration (same optimizer, loss function and evaluation metrics) to maintain consistency.

3.5. Autoencoder for feature extraction

To learn compressed representations of the input voice features, an unsupervised autoencoder was built. An input layer, a 14-neuron encoding layer (dimensionality reduction), and a decoding layer that tried to reconstruct the input were all included in the Autoencoder model. The encoded representations were extracted following the Autoencoder's training on the training set with Mean Squared Error (MSE) loss.

The encoded test data was used to train a Logistic Regression classifier to accomplish Parkinson's classification. After that, performance was assessed using common classification metrics.

3.6. Generative adversarial network (GAN)

In this work we employ a GAN that is composed by two neural networks: a Generator and a Discriminator, which are trained against each other. The Generator takes as input a 100-dimensional random noise vector, takes it through two fully connected layers (with 64 and 128 neurons with ReLU activation) and finally reshapes the output to the original voice feature vector. The output adopts the tanh activation function to produce synthetic feature vectors with normalized entries.

The Discriminator is a binary classifier that takes a 21-dimensional input vector and processes it through two dense layers, 128 and 64 neurons, both with ReLU activation, and a final output neuron with sigmoid activation function to differentiate between real and generated samples.

The models are trained adversarially in a loop for 10 epochs with a batch size equal to 32. In each iteration:

The Generator takes as input noise and generate fake samples.

The Discriminator is trained over a mixed batch of real and fake data, using binary cross-entropy loss with label noise (eg, a small amount of random noise is added to the labels for better generalisation).

The Generator is in turn updated so that it is effective at convincing the Discriminator that the generated samples came from the data.

Both networks are optimized using Adam with a learning rate of 1e-4. Once trained, the discriminator is used here as a classifier in its own right to classify the samples in the test set between Parkinson's and healthy. Its performance is measured by accuracy, precision, recall and confusion matrix representation.

3.7. Experimental environment

The experiments were run in Python using TensorFlow, Keras, and Scikit-learn. Matplotlib and Seaborn were used for data visualisation. Experiments were run on a GPU-accelerated machine to minimize training time and increase computing efficiency.

4. RESULTS AND discussion

The classification performances were visualized in the form of heatmaps for confusion matrices and rich metric overviews. These visualizations facilitated identification of misclassification trends, and each model's robustness. The performance of the proposed study is compared and analysed as shown in (Fig. 4), the results for each model clearly reflect the efficiency, with Multilayer Perceptron (MLP) attaining the highest accuracy (97.4 %). In the below (Fig. 5): The MLP model's classification performance yielded 32 true positives, 6 true negatives, 1 false positive, and 0 false negatives [40]. This demonstrates the strong prediction of the model with minimum misclassification errors. The performance of the Recurrent Neural (RNN-LSTM) model is given in (Fig. 6), where the model achieved 30 true positives and 4 true negatives. But it had 3 false positives and 2 false negatives as well, less accurate than the MLP model. Likewise, (Fig. 7) shows the classification results of the Gated Recurrent Unit (GRU) model that achieved 30 true positives and 4 true negatives. It had 3 false positives and 2 false negatives, showing performance similar to the RNN-LSTM model. The classification results of the Autoencoder model, as presented in the (Fig. 8), are composed of 31 true positives, 3 true negatives, 4 false positives and 1 false negative. However, as a large number of false-positive predictions are present, it indicates that it is probably incorrect for predicting certain instances. Lastly, (Fig. 9) shows the accuracy of the GAN model, which had 30 true positive cases but no true negatives. It had 7 False Positives and 2 False Negatives, showing that the model detects all true positive cases, but has lower specificity due to no true negative classifications.

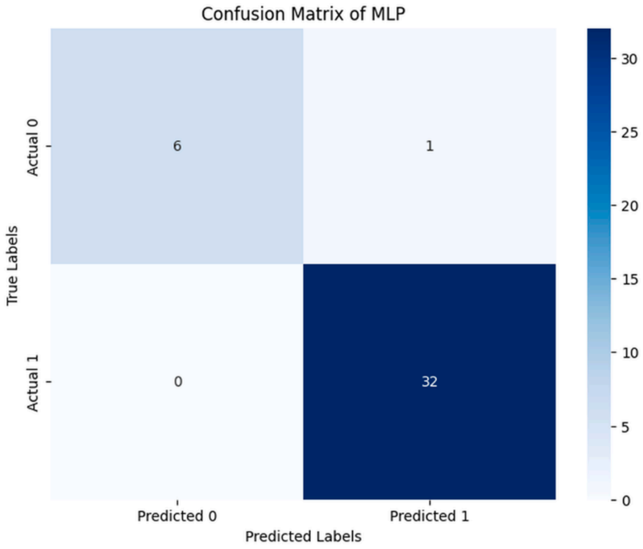


Fig. 5. MLP Confusion Matrix.

Though MLP turned out to be a promising performer, however, the performances of RNN, GRU, Autoencoder, as well as GAN models, were simplistic or less in classification. This could be because the dataset is just not very friendly to architectures with strong temporal or generative structure like they prod and therefore is not taking advantages of some of those features.

Several factors may contribute to better results of the Multilayer Perceptron (MLP) model. First of all, in UCI voice dataset, it is tabularized (structured, non-sequential) features coming from vocal signal processing (jitter, shimmer and fundalmental frequency, etc.), which are better modeled by fully connected layers instead of RNN structures. MLPs are indeed quite effective at extracting patterns from such feature-rich inputs that are not fundamentally based on temporality. On the other hand, models like RNN/LSTM/GRU are specifically for sequential time series data, and they may not perform better when the data has

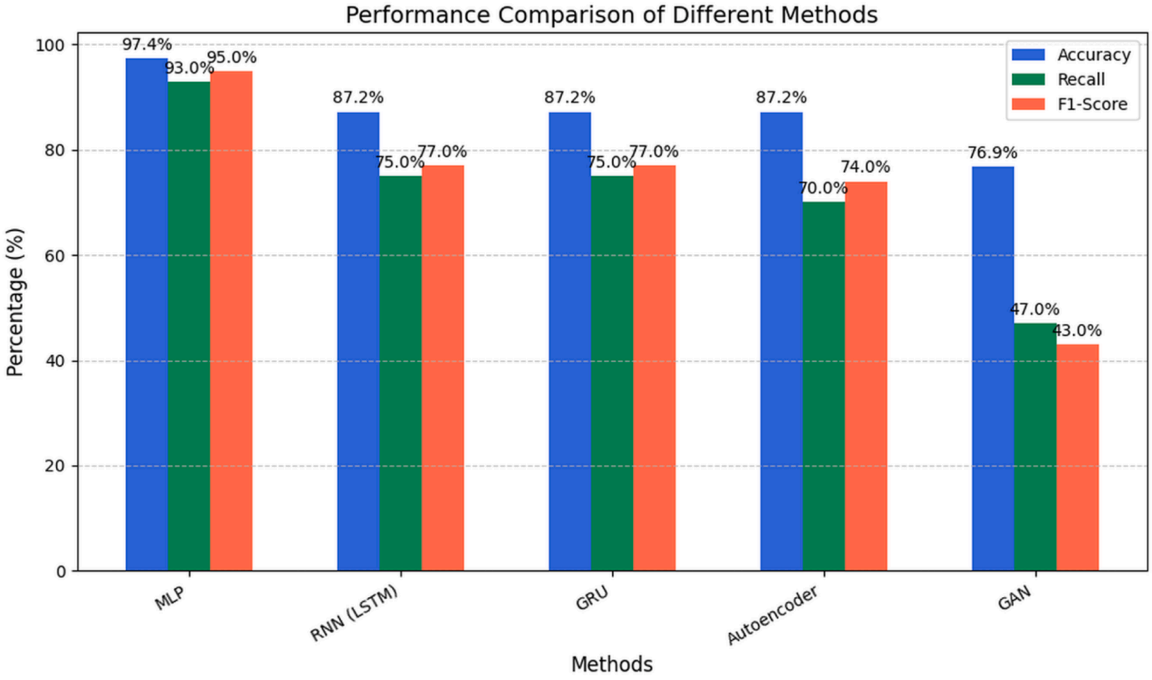


Fig. 4. Performance comparison of different methods.

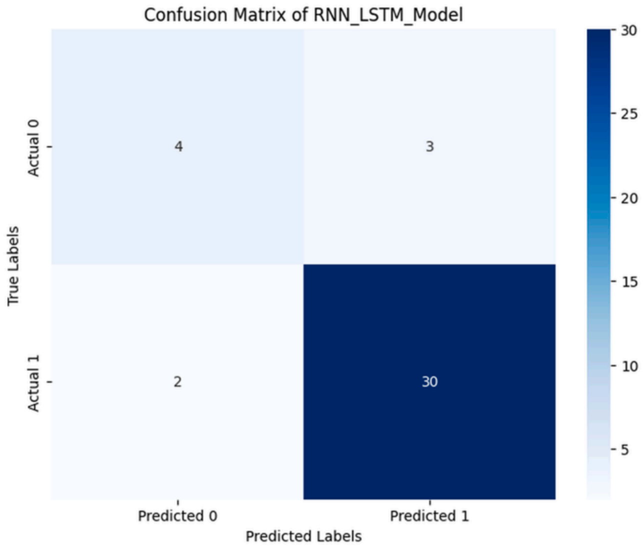


Fig. 6. RNN_LSTM Confusion Matrix.

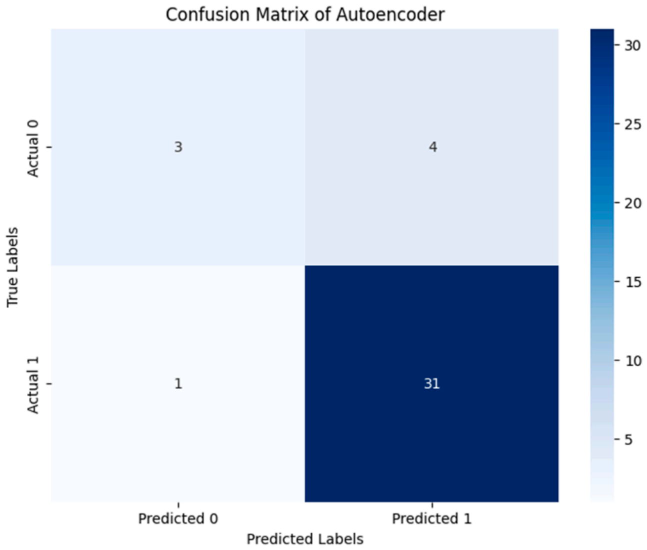


Fig. 8. Autoencoder Confusion Matrix.

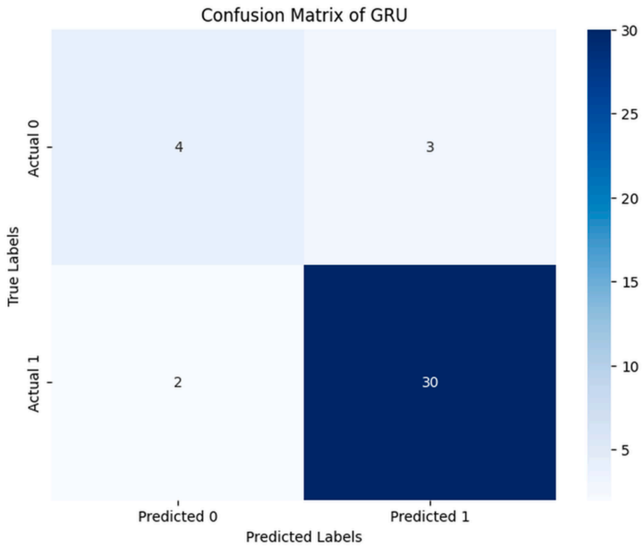


Fig. 7. GRU Confusion Matrix.

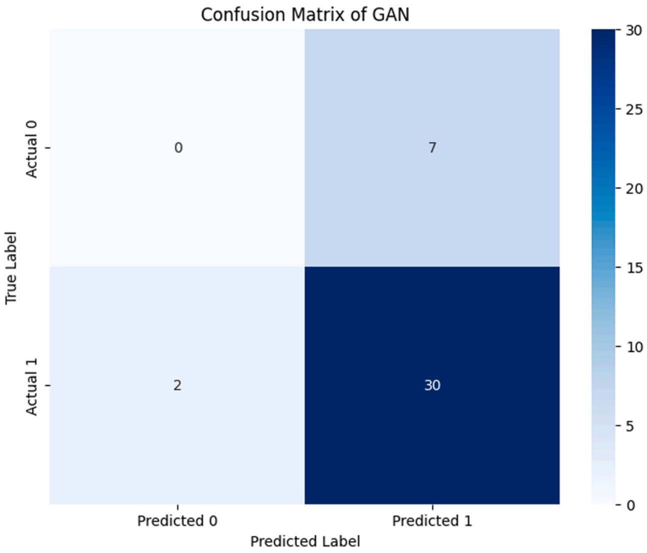


Fig. 9. GAN Confusion Matrix.

mostly independent observations. Also, MLP's simple structure is less prone to overfitting, especially on smaller data sets, and can achieve faster convergence during training. On the other hand, deeper models such as GANs and Autoencoders may require larger datasets or regularization techniques to achieve similar generalization.

While the voice features employed in this study are not structured as sequential time frames, we have also included RNN, and GRU based models to evaluate the possibility of them capturing any intrinsic or hidden temporal dependencies that may be present in the speech patterns. Such architectures are usually beneficial for modeling time-series or sequential data, however trying them here provided us with an opportunity to evaluate their flexibility and capability in non-conventional domains. From the results; although they did not outperform MLP, they show that their incorporation gives a wider perception on how the different model types perform on voice datasets in a feature vector form as opposed to a sequence. This guides future researchers in making a choice of models according to the data structure and the modality.

4.1. Computational efficiency and resource usage

Besides classification accuracy, we also compared the computation time of different models, which are crucial for real clinical-like setting. All experiments were run using the Google Colab on an NVIDIA Tesla T4 and 12 GB of RAM. The MLP model needed around 28 s of training, which was surprising, since the architecture is simple. The Autoencoder took 45 s, better in addition of dimensionality.

On the other hand, sequential models, such as RNN-LSTM and GRU, were slow and each took about 2–3 min due to temporal voice data processing. The GAN model takes the most time to train (approximately 5 min) because of its dual-network structure and the adversarial training loop.

These findings illustrate that although complex models (e.g., GANs) could offer new insights, simpler architectures such as MLP offer significantly faster training times and may be more suitable for time-sensitive clinical environments where resources are limited.

Below (Table 3) gives an overview of the estimated training times and hardware used and the compute requirements of all the deep learning models we employed in this study. These observations assist in

Table 3
Training Time and Computational Resource Comparison for All Deep Learning Models used in the study.

Model	Training Time	Hardware	Remarks
MLP	28 s	GPU (Tesla T4)	Fastest; clinically feasible
Autoencoder	45 s	GPU (Tesla T4)	Moderate time; simple design
RNN-LSTM	2 min	GPU (Tesla T4)	Requires sequence processing
GRU	2 min	GPU (Tesla T4)	Similar to LSTM
GAN	5 min	GPU (Tesla T4)	Longest; adversarial setup

assessing their applicability for clinical use.

4.2. Comparative analysis

In general, the MLP model outperformed other deep learning models attaining the best accuracy with the fewest classification errors. (Fig. 4) compares each of the five models. The importance of temporal modelling in speech-based PD detection is highlighted by the consistent superior performance of RNN_LSTM and GRU over the other three. Despite not being the greatest performers, autoencoder and GAN provided valuable insights into generative and feature extraction techniques.

4.3. Limitations and future work

There are some limitations that should be considered with this work although the results of this study show that deep learning models are effective in diagnosing PD based voice data. First, the study is restricted to the voice data of a unique modality, which may not represent all aspects of PD. Second, there are only 195 samples and no demographic diversity in the dataset that can limit the global generalization of our model. Third, some models (especially LSTM and GAN) demand more computational resources as well as training time, which probably restricts their use in the low-resource clinical setting. We will further investigate multimodal fusion, including handwriting, gait, imaging etc., and the scalability of the proposed models with other datasets of different scales and modalities in our future work.

Although RNN and GRU models are designed for time-series data, in this study, they were applied to non-sequential input vectors containing 21 voice features per sample. As such, their temporal modelling strengths were not fully utilized. This is recognized as a limitation, and future work will explore the use of time-dependent feature sequences to better align with the nature of RNN and GRU architectures.

The RNN and GRU architectures used in this study are generally designed for sequential or time-series data, yet the voice dataset in this study comprises statistical feature vectors rather than raw temporal sequences. Similarly, GANs are a leading example of a model that does well at generating images, or raw audio, but was only able to be applied here for generative tasks involving structured features. Our goal, however, was to evaluate all models under the same conditions to establish a benchmark comparison. In the future, it would be interesting to look into finding more compelling ways to manipulate the data so that each model's natural advantages might be exploited, as providing RNNs with raw audio features or using GANs for data augmentation.

5. Conclusion

This work, which uses five deep learning techniques, demonstrates encouraging advancements in the early identification and treatment of Parkinson's disease using voice data. Under this condition, the results indicate the performance varies among the deep learning archtictures, the MLP model has the best result out of the models that were tried, with 97.4 % accuracy. These results underscore the potential for the use of deep learning models in clinical applications to facilitate early PD diagnosis. Future research should explore the transformer-based models dedicated to audio processing, such as Audio Spectrogram Transformer (AST), SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to imbue faith into AI-based diagnostic tools. They will aid to identify which vocal features contribute the most to classification, rendering the models more interpretable and more clinically relevant. Also to achieve better classification performance. This extend have the opportunity to evaluate temporal and contextual properties in more details and to compare their performance with the models discussed.

CRediT authorship contribution statement

Abdulaziz Salihu Aliero: Writing – original draft. **Neha Malhotra:** Supervision.

Declaration of competing interest

We declare that there are no conflicts of interest regarding the publication of this manuscript.

All authors have reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

References

[1] K.R. Chaudhuri, et al., Economic burden of Parkinson's Disease: a multinational, real-world, cost-of-illness study, Drugs Real World Outcomes 11 (1) (2024) 1–11, https://doi.org/10.1007/s40801-023-00410-1.
[2] F. Latifoğlu, S. Penekli, F. Orhanbulucu, M.E.H. Chowdhury, A novel approach for Parkinson's disease detection using Vold-Kalman order filtering and machine learning algorithms, Neural Comput. Appl. 36 (16) (2024) 9297–9311, https://doi.org/10.1007/s00521-024-09569-2.
[3] S. Gaba and H. Kaur, "Machine Learning Techniques for Parkinson's Disease Prediction and Progression: A Comprehensive Review," *Proc. Int. Conf. Commun. Comput. Sci. Eng. IC3SE 2024*, pp. 430–436, 2024, https://doi.org/10.1109/IC3SE62002.2024.10593626.
[4] S. Roy, T. Pal, S. Debbarma, A comparative analysis of advanced machine learning algorithms to diagnose Parkinson's disease, Procedia Comput. Sci. 235 (2023) (2024) 122–131, https://doi.org/10.1016/j.procs.2024.04.015.
[5] V. Reddy, "PPINtonus : Early Detection of Parkinson ' s Disease Using Deep-Learning Tonal Analysis," 2022.
[6] M. Martinez-Eguiluz, et al., Diagnostic classification of Parkinson's disease based on non-motor manifestations and machine learning strategies, Neural Comput. Appl. 35 (8) (2023) 5603–5617, https://doi.org/10.1007/s00521-022-07256-8.
[7] J. Zhang, Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease, npj Park. Dis. 8 (1) (2022), https://doi.org/10.1038/s41531-021-00266-8.
[8] A.S. Aliero, N. Malhotra, A survey on detection of Parkinson's disease through clinical data using deep learning approach, Proc. - 2024 IEEE 16th Int. Conf. Commun. Syst. Netw. Technol. CICN 2024 (2024) 173–177, https://doi.org/10.1109/CICN63059.2024.10847535.
[9] L. Li, F. Dai, S. He, H. Yu, H. Liu, Automatic diagnosis of parkinson's disease based on deep learning models and multimodal data, Deep Learn. Approaches Early Diagnosis Neurodegener. Dis. (2024) 179–200, https://doi.org/10.4018/979-8-3693-1281-0.ch009.
[10] M. Jyotiyana, N. Kesswani, M. Kumar, A deep learning approach for classification and diagnosis of Parkinson's disease, Soft Comput 26 (18) (2022) 9155–9165, https://doi.org/10.1007/s00500-022-07275-6.
[11] S.N.H. Bukhari, K.A. Ogudo, Ensemble machine learning approach for Parkinson's disease detection using speech signals, Mathematics 12 (10) (2024), https://doi.org/10.3390/math12101575.
[12] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, M. Tsiknakis, Multi-modal deep learning diagnosis of Parkinson's Disease - A systematic review, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 2399–2423, https://doi.org/10.1109/TNSRE.2023.3277749.
[13] T. Aşuroğlu, H. Oğul, A deep learning approach for parkinson's disease severity assessment, Health Technol. (Berl). 12 (5) (2022) 943–953, https://doi.org/10.1007/s12553-022-00698-z.
[14] A.H. Al-Nefaie, T.H.H. Aldhyani, D. Koundal, Developing system-based voice features for detecting Parkinson's disease using machine learning algorithms, J. Disabil. Res. 3 (1) (2024) 1–10, https://doi.org/10.57197/jdr-2024-0001.
[15] F. Saeed, et al., Enhancing Parkinson's Disease prediction using machine learning and feature selection methods, Comput. Mater. Contin. 71 (2) (2022) 5639–5657, https://doi.org/10.32604/cmc.2022.023124.
[16] A. Mahmood, M. Mehroz Khan, M. Imran, O. Alhajlah, H. Dhahri, T. Karamat, End-to-End deep learning method for detection of invasive Parkinson's disease, Diagnostics 13 (6) (2023), https://doi.org/10.3390/diagnostics13061088.
[17] S. Singh, P. Sarote, N. Shingade, D. Yelale, N. Ranjan, Detection of Parkinson's disease using Machine learning algorithm, Int. J. Comput. Appl. 184 (6) (2022) 24–29, https://doi.org/10.5120/ijca2022922016.

[18] C.H. Lin, F.C. Wang, T.Y. Kuo, P.W. Huang, S.F. Chen, L.C. Fu, Early detection of Parkinson's disease by neural network models, IEEE Access 10 (2022) 19033–19044, https://doi.org/10.1109/ACCESS.2022.3150774.

[19] N. Chintalapudi, G. Battineni, M.A. Hossain, F. Amenta, Cascaded deep learning frameworks in contribution to the detection of Parkinson's disease, Bioengineering 9 (3) (2022), https://doi.org/10.3390/bioengineering9030116.

[20] A. Kurmi, S. Biswas, S. Sen, A. Sinitca, D. Kaplun, R. Sarkar, An ensemble of CNN models for Parkinson's Disease detection using DaTscan images, Diagnostics 12 (5) (2022) 1–18, https://doi.org/10.3390/diagnostics12051173.

[21] R. Alshammri, G. Alharbi, E. Alharbi, I. Almubark, Machine learning approaches to identify Parkinson's disease using voice signal features, Front. Artif. Intell. 6 (2023), https://doi.org/10.3389/frai.2023.1084001.

[22] S. Rahman, M. Hasan, A.K. Sarkar, F. Khan, Classification of Parkinson's disease using speech signal with Machine Learning and Deep learning approaches, Eur. J. Electr. Eng. Comput. Sci. 7 (2) (2023) 20–27, https://doi.org/10.24018/ejece.2023.7.2.488.

[23] A. Govindu, S. Palwe, Early detection of Parkinson's disease using machine learning, Procedia Comput. Sci. 218 (2022) (2022) 249–261, https://doi.org/10.1016/j.procs.2023.01.007.

[24] A. Rehman, T. Saba, M. Mujahid, F.S. Alamri, N. ElHakim, Parkinson's disease detection using hybrid LSTM-GRU deep learning model, Electron. 12 (13) (2023) 1–21, https://doi.org/10.3390/electronics12132856.

[25] C. Taleb, L. Likforman-Sulem, C. Mokbel, M. Khachab, Detection of Parkinson's disease from handwriting using deep learning: a comparative study, Evol. Intell. 16 (6) (2023) 1813–1824, https://doi.org/10.1007/s12065-020-00470-0.

[26] R.J. Sugden, P. Diamandis, Generalizable electroencephalographic classification of Parkinson's disease using deep learning, Informatics Med. Unlocked 42 (2023) 101352, https://doi.org/10.1016/j.imu.2023.101352 no. July.

[27] B. Majhi, et al., An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm, BMC Med. Imaging 24 (1) (2024) 1–20, https://doi.org/10.1186/s12880-024-01335-z.

[28] M.A. Islam, M.Z. Hasan Majumder, M.A. Hussein, K.M. Hossain, M.S. Miah, A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets, Heliyon 10 (3) (2024) e25469, https://doi.org/10.1016/j.heliyon.2024.e25469.

[29] V. Rajinikanth, S. Yassine, S.A. Bukhari, Hand-sketchs based Parkinson's disease screening using Lightweight Deep-learning with two-fold training and fused optimal features, Int. J. Math. Stat. Comput. Sci. 2 (2023) 9–18, https://doi.org/10.59543/ijmscs.v2i.7821, no. Ml.

[30] A. Keles, A. Keles, M.B. Keles, A. Okatan, PARNet: deep neural network for the diagnosis of parkinson's disease, Multimed. Tools Appl. 83 (12) (2024) 35781–35793, https://doi.org/10.1007/s11042-023-16940-3.

[31] R.M. Al-Tam, F.A. Hashim, S. Maqsood, L. Abualigah, R.M. Alwhaibi, Enhancing Parkinson's Disease diagnosis through stacking ensemble-based machine learning approach, IEEE Access 12 (2024) 79549–79567, https://doi.org/10.1109/ACCESS.2024.3408680, no. June.

[32] İ. Cantürk, O. Günay, Investigation of scalograms with a deep feature fusion approach for detection of Parkinson's disease, Cognit. Comput. 16 (3) (2024) 1198–1209, https://doi.org/10.1007/s12559-024-10254-8.

[33] H. Chen, et al., An automated hybrid approach via deep learning and radiomics focused on the midbrain and substantia nigra to detect early-stage Parkinson's disease, Front. Aging Neurosci. 16 (2024), https://doi.org/10.3389/fnagi.2024.1397896 no. May.

[34] S. Srinivasan, P. Ramadass, S.K. Mathivanan, K. Panneer Selvam, B.D. Shivahare, M.A. Shah, Detection of Parkinson disease using multiclass machine learning approach, Sci. Rep. 14 (1) (2024) 1–17, https://doi.org/10.1038/s41598-024-64004-9.

[35] Md Abu Sayed, et al., Parkinson's disease detection through vocal biomarkers and advanced machine learning algorithms, J. Comput. Sci. Technol. Stud. 5 (4) (2023) 142–149, https://doi.org/10.32996/jcsts.2023.5.4.14.

[36] N. Islam, M.S.A. Turza, S.I. Fahim, R.M. Rahman, Advanced Parkinson's Disease Detection: a comprehensive artificial intelligence approach utilizing clinical assessment and neuroimaging samples, Int. J. Cogn. Comput. Eng. 5 (2024) 199–220, https://doi.org/10.1016/j.ijcce.2024.05.001, no. November 2023.

[37] L. Ali, A. Javeed, A. Noor, H.T. Rauf, S. Kadry, A.H. Gandomi, Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network, Sci. Rep. 14 (1) (2024) 1–14, https://doi.org/10.1038/s41598-024-51600-y.

[38] M. Ianculescu, C. Petean, V. Sandulescu, A. Alexandru, A.M. Vasilevschi, Early detection of Parkinson's disease using AI techniques and image analysis, Diagnostics 14 (23) (2024), https://doi.org/10.3390/diagnostics14232615.

[39] A. Alotaibi, Ensemble deep learning approaches in health care: a review, Comput. Mater. Contin. 82 (3) (2025) 3741–3771, https://doi.org/10.32604/cmc.2025.061998.

[40] Y. Wang, et al., An automatic interpretable deep learning pipeline for accurate Parkinson's disease diagnosis using quantitative susceptibility mapping and T1-weighted images, Hum. Brain Mapp. 44 (12) (2023) 4426–4438, https://doi.org/10.1002/hbm.26399.