*Electronic Letters on Computer Vision and Image Analysis 25(2):55-78, 2025*

# Binary Grey Wolf Optimizer-Based Feature Selection for Parkinson's Disease Detection

Abderrazak Benchabane[1],Raouia Mokni[2,3], Fella Charif [1] and Monji Kherallah[3]

[1] *Department of Electronics and Telecommunications, University of Kasdi Merbah, Ouargla, Algeria*
[2] *Higher Institute of Management of Gabès, Gabès, University of Gabès, Tunisia*
[3] *Advanced Technologies for Environment and Smart Cities (ATES Unit), Sfax University, Sfax, Tunisia*
Received 1st of February 2025; accepted 20th of May 2026

**Abstract**

Parkinson's disease is a progressive neurodegenerative disorder whose early detection remains challenging. One emerging method for diagnosing PD involves analyzing hand-drawing patterns, such as drawing spirals, circles, or specific shapes. Recent advances in computational techniques, particularly machine learning and deep learning algorithms, have been applied to quantify and assess these drawings, allowing for the differentiation between healthy individuals and those with early-stage PD. In this paper, we propose a multiple deep learning architectures; AlexNet, ResNet-50, DenseNet-201, and a Vision Transformer models to extract rich, complementary features from spiral, wave, and meander drawings obtained from the NewHandPD dataset. The feature vectors from these modalities are fused to create a comprehensive representation. To handle the resulting high dimensionality and select the most discriminative features, we propose an empirical evaluation of metaheuristic-based feature selection methods, including Binary Grey Wolf Optimizer (BGWO), Binary Particle Swarm Optimizer (BPSO) and Binary rime-ice algorithm (BRIME) into the pipeline. Experimental results indicate that integrating BGWO with deep feature extractors achieves a peak accuracy of 98.48%. In comparison with BPSO- and BRIME-based feature selection, BGWO consistently yields higher accuracy while selecting fewer features. For the Vision Transformer, however, the accuracy improvement obtained through feature selection is negligible, despite a reduction in feature dimensionality

*Key Words*: Parkinson's disease, Deep learning, Vision Transformer, BGWO, BRIME, wave diagrams, spiral diagrams.

## 1 Introduction

Parkinson's Disease (PD) is a neurological condition that typically impacts older individuals, leading to motor difficulties and mental health issues. It is a progressive disorder with symptoms such as hand tremors and speech difficulties. While there is no cure for PD yet, there are medications and treatments available that can help manage its symptoms, improving the patient's quality of life and morale [1]. Clinical detection of PD

Correspondence to: benchabane.abderrazak@univ-ouargla.dz

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2076
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

often involves a combination of methods, including physical examination and sometimes imaging scans. Detecting PD using scans typically involves various imaging techniques such as MRI, CT scans, or DaTscan [2]. These scans can help visualize changes in the brain's structure or function that are associated with Parkinson's disease. However, these imaging techniques are sophisticated and require specialized equipment and expertise to perform and interpret which makes the PD diagnosis very expensive for the patient [3]. Handwriting and speech can play significant roles in detecting Parkinson's disease. Then, analyzing handwriting and speech patterns can provide valuable insights into the presence and progression of Parkinson's disease [4, 5]. Technologies such as computerized analysis of handwriting or voice recordings can help detect these subtle changes, aiding in early diagnosis and monitoring of the disease. Recently automatic Parkinson's disease detection using machine learning algorithms is widely investigated [6]. This approach often involves analyzing various data sources, including medical imaging, voice recordings, handwriting samples, and other clinical data, to detect patterns associated with PD.

Hand drawing images have been used as datasets to train deep learning models in PD detection. These images are typically collected from patients performing different kinds of hand-drawing tests [7]. There is a set of databases relevant to PD detection. The HandPD database contains spiral and meander images [8], while the Kaggle database includes wave and spiral datasets [9]. Additionally, the NewHandPD database comprises images of spiral, circle, and meander drawings [10]. In [10], Pereira et al have proposed to use the NewHandPD dataset and various machine learning algorithms. They found that the ImageNet architecture achieved an overall accuracy of nearly 93.5 % when using a combined assessment from the available images and signals of the NewHandPD dataset. Using NewHandPD database, Rania Salah El-Sayed have proposed to hybridize the HOG descriptor based CNN with LSTM to improve the feature extraction process and classification accuracy [11]. They have achieved an accuracy of 84.60 % for meander images, 94.02 % for circles and 96.25 for spiral drawings. To improve these accuracies, Aite Zhao [12], has proposed to use the Spatio-temporal Siamese neural network to detect PD individuals. They have achieved an accuracy of 93.39 % for meander images, 90.21 % for circles and 95.86 for spiral drawings. On the other hand, Razzak et al. [13] conducted a comprehensive analysis using both the HandPD and NewHandPD datasets, focusing on spiral, circle, and meander images and using different deep learning models. Fusing features from both spiral and meander images; they achieved an accuracy of 90.41% when using vgg16 model. The highest accuracy of 98.31% was achieved for spiral, meander and circle images together and using the ResNet50 model [5]. Also transformer-based architectures have been considered in handwriting-based Parkinson's disease detection. In [14],N. Wang et al. introduced a Coordinate Attention–Enhanced Swin Transformer (CAS Transformer), which incorporates the Coordinate Attention (CA) mechanism into the Swin Transformer framework. The CAS Transformer achieved an accuracy of 92.68% when using the NewHandPD dataset. In addition, S. ehata et al. [15] proposed MobileViT which combines the architectural advantages of convolutional neural networks (CNNs) and transformers, preserving the efficient characteristics of CNNs while integrating the transformer's self-attention mechanism to capture long-range dependencies and global contextual information. Using NewHandPD dataset, this combination achieved an accuracy of 89.62%

## 2 Contributions

Fusion of features refers to the process of combining multiple sets of features or representations from different sources or modalities to create a single, more informative feature representation. Feature selection, on the other hand, is the process of selecting a subset of the most relevant features from a larger set of features. Feature selection in machine learning is the key to any successful classification model. Including only relevant features can lead to simpler models that generalize better reducing overfitting. Also using fewer features can speed up the training process. In the case of PD, Combining feature fusion and feature selection techniques can lead to enhanced diagnostic models. By fusing features from different types of hand drawings such as spiral and wave diagrams, a more comprehensive representation of the data could be obtained. This will be very helpful to capture different aspects of PD symptoms. Furthermore, by applying feature selection techniques to the fused

features, the model's performance becomes more efficient for PD detection.
In this paper, the spiral, circle and meander drawings from the NewHandPD data set are used for the training process. The AlexNet, DenseNet-201, ResNet-50 and Vision Transformer (ViT) models serve to extract features from the considered drawings. The obtained features are concatenated to form a vector of features representing rich information present in different types of data related to Parkinson's disease. Combining multiple modalities leads to a high-dimensional feature vector. Consequently, a feature selection is suitable. To do this, the Binary Grey Wolf Optimizer (BGWO), the Binary particle Swarm optimizer (BPSO) and the Binary Rime-Ice Optimization Algorithm (BRIME) which are adapted for binary optimization problems, are considered for feature selection tasks.
The main contribution of this work lies in the integration of multimodal deep feature fusion with feature selection based on metaheuristic methods within a unified framework. Unlike approaches based on single architectures or fine-tuned models, the proposed method aims to combine complementary deep representations with optimized feature reduction, enabling high accuracy while improving computational efficiency and robustness. The paper is organized as follows: Section 3 provides an overview of deep learning,vision transformer and meta-heuristic optimizers. Section 4 outlines the proposed methodology. Section 5 presents the results, a discussion, and a comparative analysis. Finally, Section 6 concludes the study, summarizing the key findings.

# 3 Background

## 3.1 Deep learning

AlexNet , DenseNet-201 and ResNet50 are powerful architectures for feature extraction. AlexNet consists of eight primary layers: five convolutional layers, which are responsible for automatic feature extraction, followed by three fully connected layers that perform the final classification [16]. The DenseNet-201 architecture consists of multiple dense blocks, each containing a series of convolutional layers with batch normalization and ReLU activation functions. The output of each dense block is fed into a transition block, which consists of a convolutional layer followed by average pooling to reduce the spatial dimensions of the feature maps. The final output layer is a global average pooling layer followed by a fully connected layer with softmax activation for classification [17]. On the other hand, The ResNet-50 architecture consists of several residual blocks, each containing multiple convolutional layers with batch normalization and ReLU activation functions. The network also includes downsampling layers (stride-2 convolutional layers) to reduce the spatial dimensions of the feature maps [18].

## 3.2 Vision Transformer

ViT is a deep learning architecture originally designed for image recognition tasks, inspired by the Transformer model used in natural language processing [19]. Instead of processing spatial data through convolutional layers, ViT divides an image into fixed-size patches and treats each patch as a token in a sequence. This representation allows the model to capture long-range dependencies between patches using a self-attention mechanism.

Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, it is first divided into $N$ non-overlapping patches $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where each patch is flattened and linearly projected into a $D$-dimensional embedding:

$$\mathbf{z}_i^0 = \mathbf{E}\,\mathbf{x}_i + \mathbf{E}_{pos,i}, \quad i = 1, 2, \ldots, N \tag{1}$$

where $\mathbf{E}$ is a trainable linear projection matrix and $\mathbf{E}_{pos,i}$ represents the positional embedding.

The sequence of patch embeddings $\mathbf{Z}^0 = [\mathbf{z}_1^0, \mathbf{z}_2^0, \ldots, \mathbf{z}_N^0]$ is passed through $L$ Transformer encoder layers. Each layer consists of a Multi-head Self-Attention (MSA) mechanism and a Feed-Forward Network (FFN):

$$\mathbf{Z}^{l'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{Z}^{l-1})) + \mathbf{Z}^{l-1} \tag{2}$$

$$\mathbf{Z}^l = \mathrm{FFN}(\mathrm{LN}(\mathbf{Z}^{l'})) + \mathbf{Z}^{l'} \tag{3}$$

where LN denotes layer normalization, and $l = 1, 2, \ldots, L$.

In the context of feature selection, ViT provides an effective mechanism to measure the importance of features or patches through its self-attention weights. The attention matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ quantifies how strongly each token (feature) attends to others:

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \tag{4}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are the query and key matrices derived from $\mathbf{Z}^{l-1}$, and $d_k$ is the attention dimension. Features associated with higher attention scores are considered more informative and are selected for subsequent learning or classification tasks [20].

### 3.3 Binary grey wolf optimizer

BGWO is one of the meta-heuristic optimization algorithms that are inspired by the hunting and social behavior of grey wolves [21]. It has been adapted for binary optimization problems, making it particularly suitable for feature selection tasks. Within a wolf pack, there are three main groups: alpha, beta, and gamma. Each group has a distinct role to play in the pack's dynamics and survival, and follow a strict social dominant hierarchy. Hunting techniques consist on tracking, encircling the prey until it stops moving and attacking the prey. In GWO, each grey wolf position represents the searched parameters optimizing the considered problem.

The gray wolf population is classified into four groups: $\alpha$, $\beta$, $\delta$ and $\omega$. The $\alpha$ wolves are regarded as the strongest, followed by $\beta$ and $\delta$. The remaining wolves, representing other candidate solutions, fall under the $\omega$ category. Let $X_t = x_1^t, x_2^t, \cdots, x_n^t$ the position of the $n$ wolves at time $t$, where each of $x_i^t = u_1^t, u_2^t, \cdots, u_r^t$ represents a mapping of the problem parameters.

During a hunt, the top three wolves take the lead, while the rest adjust their positions according to [21]:

$$D = |CX_p^t - x_i^t| \tag{5}$$

$$x_i^{t+1} = x_p^t - A \times D \tag{6}$$

where $X_p^t$ is the prey location at time $t$, $x_i^{t+1}$ is the $i$th wolf position at time $t+1$. $D$ is the Euclidean distance between the prey and the wolf $x_i^t$. $A$ and $C$ are coefficients vectors which can be expressed as [21]:

$$A = 2ar_1 - a \quad , \quad C = 2r_2 \tag{7}$$

$$C = 2r_2 \tag{8}$$

where $a = 2 - 2(t/T_{max})$ is decreased from 2 to 0 linearly through the number of iterations, $T_{max}$is total number of iterations. $t$ indicates the current iteration and $r_1$,$r_2$ are randomly distributed over the range 0 and 1. The GWO algorithm is carried out in the following steps:

1. Grey wolves position initialization

    - Evaluate the fitness of each candidate and find the best candidate minimizing the fitness
    - Find the first $\alpha$, the second $\beta$ and the third $\delta$ best fitness solutions,$x_\alpha, x_\beta, x_\delta$
    - Calculate encircling behavior according to equations (1) and (2).

2. Update candidate solution
   The alpha wolf is primarily responsible for directing the hunting pattern of the pack. However, both the beta and delta wolves are also involved in hunting. Therefore, the algorithm selects the first three best solutions as the hunting wolves, with their current positions used to update the positions of all wolves in the pack. The updating formulas for all wolves' positions are as follows [21]:

$$x_i^{t+1} = \frac{x_1 + x_2 + x_3}{3}, \quad \begin{cases} x_1 = x_\alpha - A_1 D_\alpha, & D_\alpha = |C_1 x_\alpha - x_i^t| \\ x_2 = x_\beta - A_2 D_\beta, & D_\beta = |C_2 x_\beta - x_i^t| \\ x_3 = x_\delta - A_3 D_\delta, & D_\delta = |C_3 x_\delta - x_i^t| \end{cases} \tag{9}$$

3. Evaluate fitness and update the first $\alpha$, the second $\beta$ and the third $\delta$ .

While the Grey Wolf Algorithm (GWO) is commonly used for solving continuous optimization problems, a binary version of GWO is essential for feature selection tasks. In this paper, we utilize BGWO2, one of the binary grey wolf optimization algorithms recently introduced by Emary et al. [22], specifically designed for feature selection. It retains the core principles of GWO but includes modifications to adapt it to binary decision spaces. BGWO2 updates the position of each wolf as [22]:

$$x_i^{t+1} = \begin{cases} 1, & S\left(\frac{x_1+x_2+x_3}{3}\right) \geq Th, \\ 0, & \text{Otherwise.} \end{cases} \tag{10}$$

Where Th is a random vector in the range [0, 1], and S represents the sigmoid function, defined as [22]:

$$S(x) = \frac{1}{1 + e^{-10(x-0.5)}} \tag{11}$$

### 3.4 Principle of Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique originally proposed by Kennedy and Eberhart in 1995 [24]. It was inspired by the collective social behavior observed in flocks of birds and schools of fish, where individuals cooperate and share information to locate food sources efficiently. In PSO, each individual in the population, called a particle, represents a potential solution to the optimization problem. The particles move through the search space by updating their positions and velocities based on both their own experience and the experience of neighboring particles.

Let $X_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$ denote the position of the $i$-th particle in a $d$-dimensional search space, and $V_i = [v_{i1}, v_{i2}, \ldots, v_{id}]$ its corresponding velocity. Each particle keeps track of its personal best position $P_i$ and the best position discovered by the entire swarm, denoted by $G$. At each iteration, the velocity and position of each particle are updated according to the following equations:

$$V_i(t+1) = w\, V_i(t) + c_1\, r_1\, (P_i - X_i(t)) + c_2\, r_2\, (G - X_i(t)), \tag{12}$$

$$X_i(t+1) = X_i(t) + V_i(t+1), \tag{13}$$

where $w$ is the inertia weight controlling the impact of the previous velocity, $c_1$ and $c_2$ are the cognitive and social acceleration coefficients, and $r_1$ and $r_2$ are random numbers uniformly distributed in $[0, 1]$. $t$ is the current iteration, and $T$ is the total number of iterations.

During the optimization process, each particle evaluates the objective function at its current position. If the new position yields a better fitness value than its previous best, the personal best $P_i$ is updated. Likewise, if any particle achieves a fitness better than the global best $G$, this new position becomes the updated global best. The process is repeated iteratively until a stopping criterion is met, such as a maximum number of iterations or a satisfactory objective value.

### 3.5 Rime-Ice Optimization Algorithm

The Rime-Ice Optimization Algorithm (RIME) is a recently proposed metaheuristic optimization method inspired by the natural process of rime-ice formation [25]. This physical phenomenon occurs when supercooled water droplets freeze upon contact with solid surfaces under subzero temperatures, gradually building complex rime structures. RIME models this growth process mathematically and adapts it to the search for optimal solutions in a multidimensional space. The algorithm is composed of four main stages: rime cluster initialization, soft-rime search, hard-rime puncture, and positive greedy selection.

In the first stage, known as rime cluster initialization, a population of $n$ agents is randomly distributed within the search space. Each agent represents a potential solution, where the position of the $i$-th agent in the $j$-th dimension is given by

$$x_{ij} = Lb_j + \text{rand}(0,1) \times (Ub_j - Lb_j), \tag{14}$$

with $Lb_j$ and $Ub_j$ denoting the lower and upper bounds of the $j$-th variable, respectively. The initial population is then evaluated using the objective function, and the best agent is recorded as $R_{best}$.

The second stage, the soft-rime search mechanism, simulates the smooth accumulation of soft rime under mild wind conditions. This process enhances the exploratory behavior of the algorithm, allowing agents to search broadly across the solution space. The position of each agent is updated relative to the best-known solution as follows:

$$R_{ij}^{new} = R_{best,j} + r_1 \cdot \cos\left(\frac{t\pi}{10T}\right) \cdot \left(1 - \frac{\text{round}(5t/T)}{5}\right) \cdot \big(h\,(Ub_j - Lb_j) + Lb_j\big), \tag{15}$$

provided that $r_2 < E$, where $E = \sqrt{t/T}$ is the condensation coefficient that increases with the number of iterations. Here, $t$ and $T$ denote the current and maximum iteration numbers, respectively, $r_1$ and $r_2$ are random numbers uniformly distributed in $[0, 1]$, and $h$ is an adhesion coefficient that controls the intensity of the soft-rime growth. At early iterations, $E$ remains small, and the occurrence of soft-rime updates is limited; as the search progresses, $E$ increases, promoting a more dynamic exploration of the solution space.

The third stage, referred to as the hard-rime puncture strategy, reflects the formation of hard rime caused by stronger winds, which produce a denser and more directional ice structure. In this phase, agents refine their search by exchanging information with other individuals in the population, thereby improving local exploitation. The update equation for the hard-rime phase can be expressed as

$$R_{ij}^{new} = R_{ij} + r_3 \cdot (R_{k,j} - R_{ij}), \tag{16}$$

where $R_k$ denotes a randomly selected agent and $r_3$ is a random coefficient within $[0, 1]$. This mechanism enables controlled perturbation of the current solution, helping the population to avoid premature convergence.

Finally, the positive greedy selection strategy ensures that only improved solutions are retained in the population. After computing the fitness of the new candidate $R_i^{new}$, the replacement rule is applied as

$$\text{if } F(R_i^{new}) < F(R_i), \quad R_i \leftarrow R_i^{new}, \tag{17}$$

and

$$\text{if } F(R_i^{new}) < F(R_{best}), \quad R_{best} \leftarrow R_i^{new}. \tag{18}$$

Due to its hybrid structure and natural inspiration, RIME has demonstrated strong robustness and convergence properties in complex optimization and feature selection tasks [26].

## 4 Materials and Methods

The NewHandPD dataset comprises 66 individuals, with 31 diagnosed with PD and 35 considered healthy. It includes data collected during four distinct tasks: drawing circles, spirals, and meanders, as well as recording

signals. Each participant was tasked with providing one circle drawing, four meander and four spiral drawings. In the meander category, there are 140 images from healthy subjects and 124 from PD subjects. Similarly, the spiral images contain 140 samples from healthy subjects and 124 from PD subjects. The circle images have 35 samples from healthy subjects and 31 from those with PD [10].

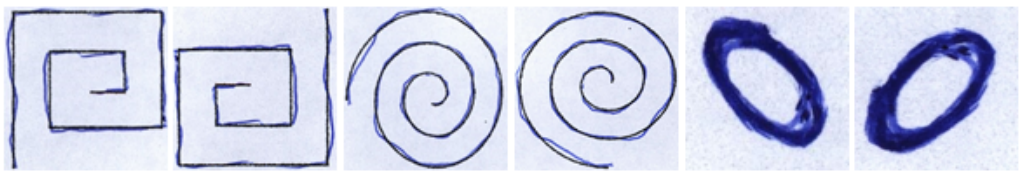
### 4.1 Data augmentation

To increase the diversity and robustness of our training dataset, we applied data augmentation techniques to the hand-drawn spirals, circles, and meanders from both healthy individuals and PD patients. Importantly, in our 5-fold cross-validation setup, the dataset was first split into five disjoint folds. At each iteration, one fold was used as the validation (test) set and the remaining four folds as the training set. Data augmentation was performed only after this fold splitting, and exclusively on the training folds, ensuring that no augmented or near-duplicate samples could appear in the validation folds. This procedure eliminates the risk of data leakage and prevents artificially inflated performance. The augmentation process involved three types of transformations: rotations of 90°, 180°, and 270°; horizontal flips (left and right); and random translations ranging from -10 to 10 pixels along both the horizontal and vertical axes. These augmentations introduced variations in orientation, symmetry, and spatial positioning, helping the model generalize better to unseen samples. After augmentation, the training dataset expanded to a total of 3,168 images. Finally, all images were normalized according to ImageNet parameters to align with the pre-training standards of deep learning models. Notably, no augmentation was applied to the validation (test) images to ensure an unbiased evaluation of the model's performance.

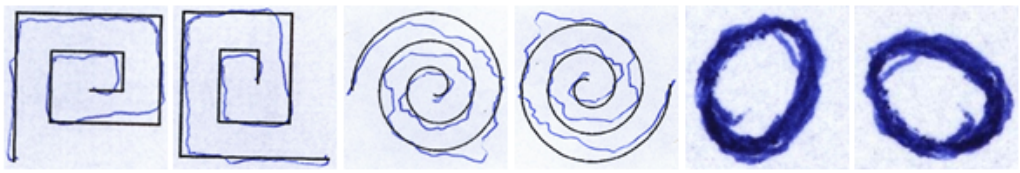
Some samples of the dataset already labelled as healthy or Parkinson are shown in Figure 1.

| **Modality** | **Train** | | **Test** |
|---|---|---|---|
| | **Before data-augmentation** | **After data-augmentation** | |
| Meander | H=105, PD=93 | H=525, PD=465 | H=35, PD=31 |
| Spiral | H=105, PD=93 | H=525, PD=465 | H=35, PD=31 |
| Circle | H=35, PD=31 | H=525, PD=465 | H=35, PD=31 |

Table 1 Summary of modalities before and and after data-augmentation



(a) Healthy



(b) Parkinson

Figure 1 Samples from the augmented NewHandPD dataset.

## 4.2 Methodology

An overview of the proposed method is depicted in the figure 2. From the used modalities; the input images, a deep learning algorithm is used to extract corresponding features. Next, obtained features are concatenated to form a unique feature vector. In the next step, a vector reduction is performed through feature selection while training to obtain the most relevant features. Once the relevant features are obtained, a classifier is used for disease detection. Finally, we apply some criteria to evaluate the approach. In the following, different stages of the method will be presented.

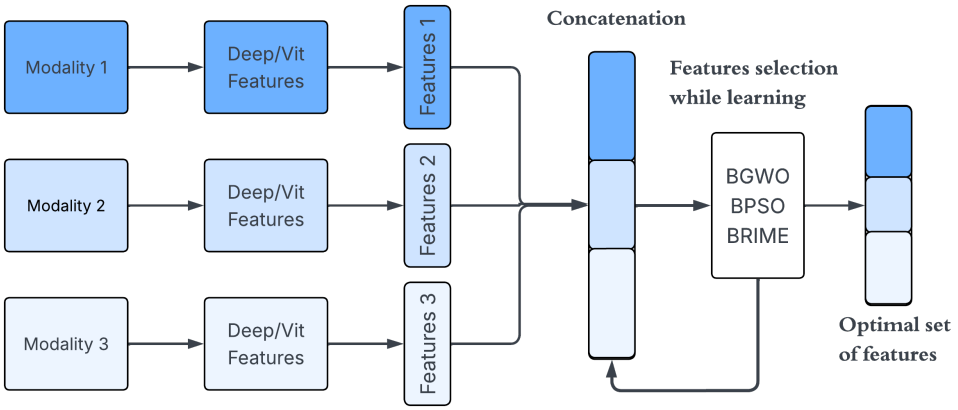



Figure 2 Flowchart of proposed approach for PD detection.

### 4.2.1 Deep and ViT features extraction

This stage involves extracting features from the three types of hand-drawn images (spirals, waves, and meanders) using deep learning techniques, resulting in three distinct feature vectors. We employed the pretrained CNN architectures AlexNet, DenseNet201, and ResNet50 for feature extraction, after removing their final fully connected layers.

Each model was used independently on each modality to capture its specific visual characteristics. Importantly, these CNN models were used solely as fixed feature extractors: their weights were kept frozen and were not updated or fine-tuned during any subsequent steps. AlexNet produces a 4,096-dimensional feature vector for each image, representing high-level abstract features from its penultimate layer (‘fc7‘). ResNet50, known for its residual skip connections and deep feature encoding, provides a 2,048-dimensional feature vector obtained from the global average pooling layer preceding its final classification head. DenseNet201, characterized by dense connections between layers that promote feature reuse, outputs a 1,920-dimensional feature vector from its global average pooling layer. Unlike CNNs that rely on local receptive fields, ViT divides each input image into non-overlapping patches, embeds them into a sequence of tokens, and processes them through multiple self-attention layers. This architecture allows ViT to model long-range spatial relationships across the entire image. The pre-trained ViT-B/16 model was adopted, providing a 768-dimensional feature embedding from its final transformer encoder layer for each modality.

### 4.2.2 Features concatenation

The three feature vectors obtained from the different drawing types are combined into a single feature vector through concatenation. While this integration captures complementary information across modalities, it also produces a high-dimensional representation that requires more computational resources for processing. In the case of the AlexNet model for multimodal feature extraction, each modality-specific feature vector contains

4,096 features, and their concatenation results in a combined feature vector of 12,288 features (3 × 4,096). For ResNet50, the fused feature vector dimension reaches 6,144 features (3 × 2,048) and for DenseNet201, the combined feature vector comprises 5,760 features (3 × 1,920). Finally, with ViT-B/16, the concatenation of the three modality-specific embeddings yields a total of 2,304 features (3 × 768). Figures 3,4 illustrate the feature extraction process using AlexNet and ViT models.

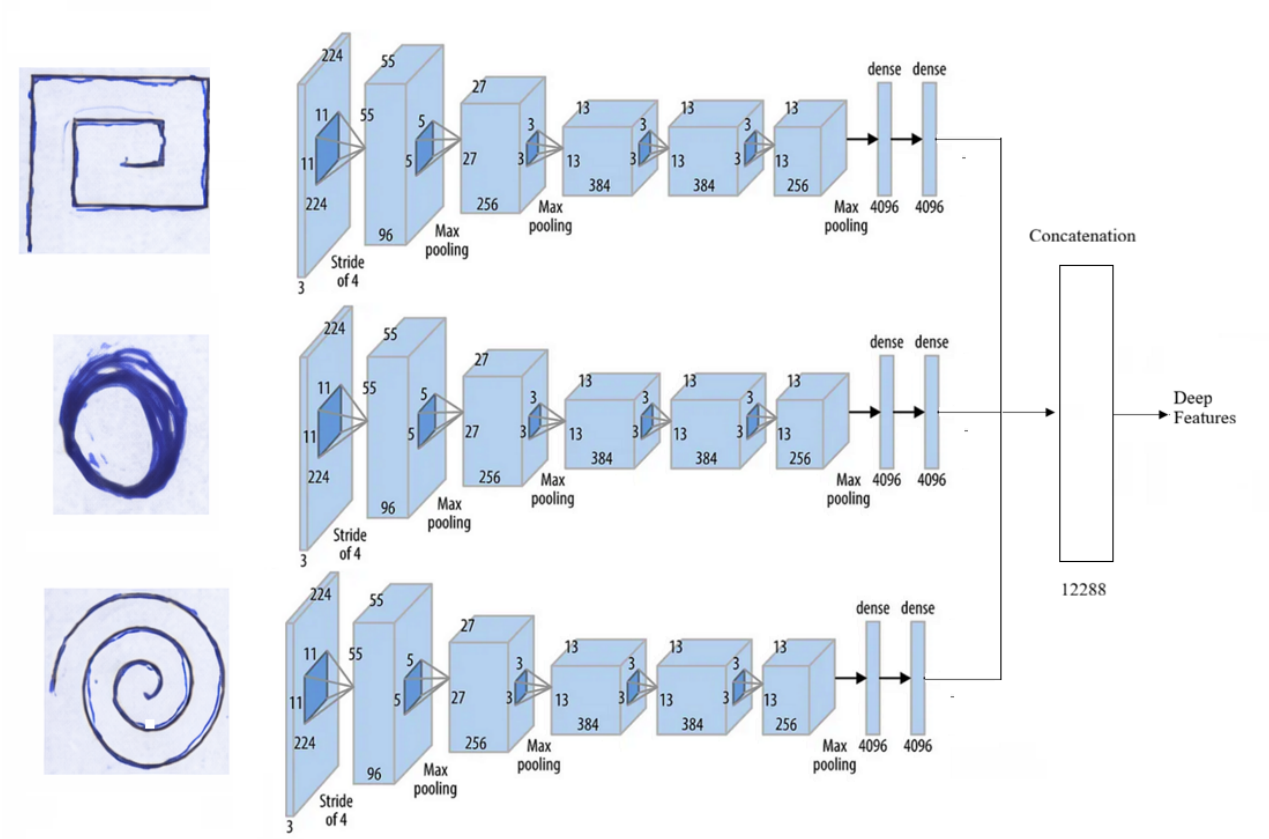



Figure 3 flowchart of feature extraction using AlexNet.

### 4.2.3 Feature selection while training

To reduce dimensionality and retain only the most discriminative information, we applied three meta-heuristic algorithms (BGWO, BPSO and BRIME) algorithms to the concatenated feature vectors. These algorithms search for the optimal subset of features that maximizes classification accuracy, thereby improving model efficiency and generalization. During the optimization process, each candidate feature subset is evaluated using a wrapper-based fitness function that balances classification performance and feature compactness. The fitness is computed as

$$Fitness = \alpha(1 - Accuracy) + \beta\frac{N_s}{N_t} \tag{19}$$

where Accuracy is obtained via internal cross-validation on the training data, $Ns$ is the number of selected features, and $Nt$ is the total number of features. $\alpha$, $\beta$ are weighting coefficients controlling the trade-off between accuracy and feature reduction so that $\alpha + \beta = 1$ .

During the feature selection task, a neural network consisting of a single fully connected layer followed by a softmax activation function is used for classification. The model is trained with the cross-entropy loss function and optimized using gradient-based methods It is important to note that the population-based optimizers operate

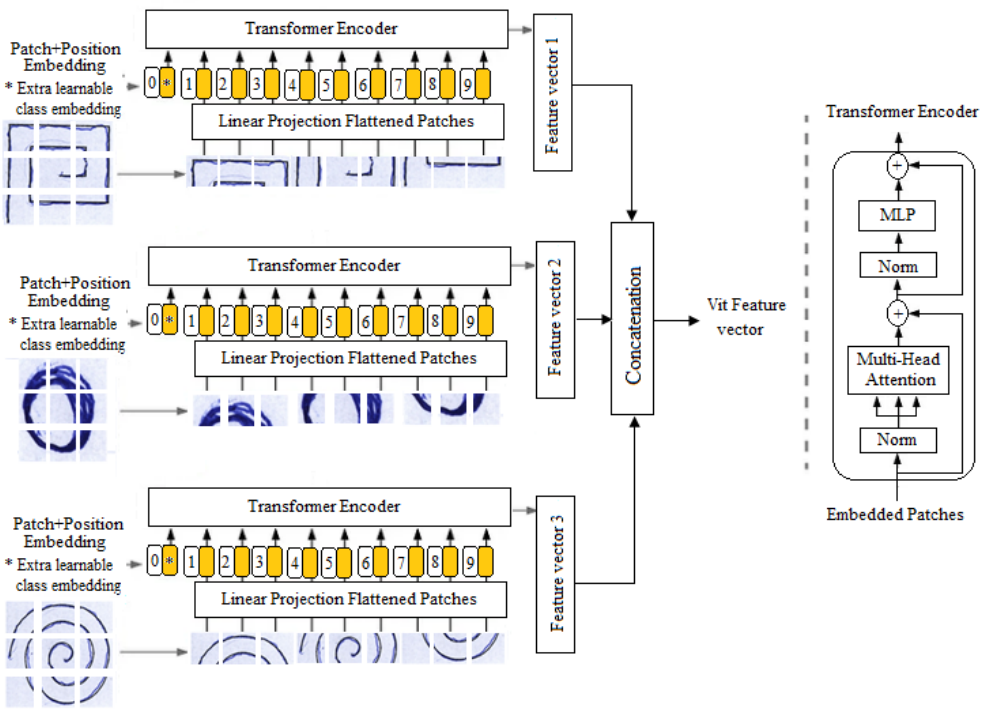



Figure 4 Flowchart of feature extraction using Vision Transformer.

exclusively on the pre-extracted, concatenated feature vectors. The CNN feature extractors remain frozen during this stage, and no weight updates or fine-tuning are performed. This ensures that the feature selection process is completely decoupled from CNN training and focuses solely on selecting the most relevant features for the subsequent classifier.

While fine-tuning pre-trained networks can improve task-specific performance, we used fixed CNN and ViT models to ensure computational efficiency, reproducibility, and a fair evaluation of the proposed feature selection framework. By keeping feature extractors fixed, the evaluation of the efficiency of BGWO, BPSO, and BRIME in selecting discriminating features from multimodal representations becomes consistent, thus avoiding the confounding effects of weight updates.
It should be noted that although the proposed feature selection methodology significantly reduces dimensionality, the selected features represent high-level deep representations that are not directly interpretable in clinical terms. These features capture relevant patterns, such as stroke irregularities and motor impairments associated with Parkinson's disease; however, their relationship to specific clinical markers remains implicit.

The complete PD detection pipeline, including subject-wise data splitting, last-layer deep feature extraction, feature selection using BGWO, BPSO, and BRIME, and classifier evaluation, is summarized in Algorithm 1 for clarity and reproducibility.

**Algorithm 1** PD Detection Pipeline

**Require:** PD dataset , deep models $\{AlexNet, DenseNet, ResNet50, ViT\}$, feature selection methods $\{BGWO, BPSO, BRIME\}$, Number of folds $K$ and number of trials $T$
**Ensure:** Performance metrics
Randomly split subjects into training (80%) and testing (20%)
Apply data augmentation to training data only
**for** each deep model $M$ **do**
  Extract features from the last fully connected / embedding layer for both training and test sets
**end for**
Reorganize training feature matrix
Divide training features into $K$ stratified folds
**for** $k = 1$ to $K$ **do**
  **for** $t = 1$ to $T$ **do**
    Generate candidate feature subset using optimizer (BGWO / BPSO / BRIME)
    Evaluate fitness using equation (19)
    Select optimal feature subset based on minimum fitness
    Train classifier using selected features
    Evaluate on the fixed test set (unseen subjects)
    Store accuracy
  **end for**
**end for**
Aggregate accuracy results across folds and trials
**return** Final performance metrics

## 5 Experimental Result

In the following, we evaluate our proposed multi-modal classification approach using the NewHandPD dataset, which consists of hand drawings from PD patients and healthy individuals. The proposed approach involves three main steps. First, deep and ViT learning models are assigned to the three modalities to extract features. The extracted features are then concatenated and passed through a fully connected layer classifier after undergoing feature selection using BGWO, followed by a single-layer softmax classifier. In the following experiments, k-fold cross-validation is adopted to obtain a robust estimate of the model's performance by averaging results across multiple random data splits. The value of $k$ is set to 5, partitioning the dataset into five subsets, training on four subsets and the fifth as validation set. To ensure a fair comparison, all feature-selection algorithms BGWO, BPSO, and BRIME were implemented under identical experimental conditions. The population size was set to 15, and the maximum number of iterations to 50. The thresholding scheme for binary conversion followed the standard sigmoid transfer function with threshold equal to 0.5.

The proposed method is implemented in MATLAB R2020 on a Windows 11 laptop with 16 GB of RAM and an Intel Core i7-8000U processor. In order to evaluate the performance of each deep learning model and ViT transformer under feature reduction, different evaluation metrics have been applied in this study. These metrics provide a detailed understanding of the model's classification capability by analyzing different prediction outcomes. Their formulas are formulated as: [27]:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{24}$$

where:
True Positive (TP): The model correctly predicts PD disease when the image belongs to a Parkinson's patient.
True Negative (TN): The model correctly predicts a healthy individual when the image belongs to a healthy subject.
False Positive (FP): The model incorrectly predicts PD when the image actually belongs to a healthy individual, leading to a false alarm.
False Negative (FN): The model incorrectly predicts a healthy status when the image actually belongs to a Parkinson's patient, resulting in a missed diagnosis.

### 5.1 Mono-modal case

To highlight the effectiveness of the proposed approach, we initially focus on disease detection using a mono-modal classification strategy. In this context, the three modalities were used separately to train deep learning models: AlexNet, ResNet50, DenseNet201 and ViT. Each model was trained on a specific modality, allowing us to evaluate how well individual features contribute to accurate disease detection. For the four pre-trained networks, the learning rate was set to $0.0001$. The training process was limited to a maximum of 10 epochs with a minimum batch size of 32. Adaptive Moment Estimation (Adam), Stochastic Gradient Descent with Momentum (SGDM), and Root Mean Square Propagation (RMSprop) optimizers were used, along with the cross-entropy loss function, to optimize model performance. As shown in Table 2, Adam consistently achieves the highest mean accuracy in the three deep-based models and modalities, suggesting its effectiveness in optimizing deep learning models. The spiral modality provides superior performance in PD detection compared to meander and circle shapes due to several critical factors related to motor control and neuromuscular function. Firstly, drawing a spiral requires continuous and smooth hand movements with precise motor coordination, leading to noticeable distortions in the spiral patterns of PD patients, which are easier to identify. Secondly, A spiral involves both radial and angular hand movements simultaneously, making it more demanding than simple circular or meander patterns. The Meander modality also performs well, while the Circle modality shows the lowest accuracy since it involves more repetitive movement, which might be easier for PD patients to compensate for, reducing the detection accuracy. ViT achieves mean accuracies between 88.59% and 98.79%, depending on the optimizer and modality. This indicates that ViT effectively captures discriminative features from the hand-drawing tasks, confirming its suitability for handwriting-based PD detection. Compared with CNN-based models such as AlexNet and ResNet50, ViT maintains competitive performance, particularly in the Circle modality.

### 5.2 Multimodal without features selection

In multimodal detection, the three handwriting modalities were utilized simultaneously to enhance the system's ability to extract and integrate diverse features. Table 3 summarizes the performances of the four models using augmented data. As we can see, the ViT model achieves the highest average accuracy (97.23%) among the four models, followed by ResNet50 and DenseNet201 (95.76%,95.45%) and AlexNet (94.85%). The minimum accuracy across the ten runs shows that AlexNet and ResNet50 have the same lowest value (93.94%), whereas DenseNet201 records a slightly lower minimum accuracy (92.42%). Regarding maximum accuracy,

| Model | Modality | ADAM | | | SGDM | | | RMSprop | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| **AlexNet** | Meander | 90.61 | 87.88 | 92.42 | 88.18 | 86.36 | 90.91 | 87.28 | 78.79 | 92.42 |
| | Spiral | 92.12 | 89.39 | 96.97 | 92.12 | 89.39 | 93.94 | 87.27 | 77.27 | 95.45 |
| | Circle | 85.88 | 70.59 | 94.12 | 77.65 | 70.59 | 82.85 | 78.82 | 58.82 | 94.11 |
| **ResNet50** | Meander | 94.55 | 89.39 | 96.97 | 87.27 | 84.85 | 90.91 | 94.55 | 92.42 | 96.97 |
| | Spiral | 96.06 | 93.94 | 96.97 | 93.03 | 89.39 | 96.97 | 96.36 | 95.45 | 96.96 |
| | Circle | 86.76 | 70.58 | 94.12 | 77.65 | 70.59 | 82.35 | 83.53 | 64.70 | 94.12 |
| **DenseNet201** | Meander | 94.24 | 93.94 | 95.45 | 89.39 | 84.85 | 90.91 | 92.42 | 89.39 | 93.94 |
| | Spiral | 95.15 | 92.42 | 96.97 | 87.27 | 84.85 | 90.91 | 95.76 | 92.42 | 96.97 |
| | Circle | 88.24 | 82.35 | 94.12 | 80 | 64.71 | 88.24 | 85.64 | 69.23 | 92.31 |
| **ViT** | Meander | 89.70 | 87.88 | 93.94 | 91.72 | 87.88 | 96.97 | 90.40 | 86.36 | 96.97 |
| | Spiral | 88.59 | 86.36 | 92.42 | 87.47 | 81.82 | 90.91 | 86.57 | 81.82 | 89.39 |
| | Circle | 98.79 | 96.97 | 100 | 99.85 | 98.48 | 100 | 99.39 | 96.97 | 100 |

Table 2 Accuracy (%) of Deep-ViT Learning Models with Different Optimizers

ViT achieves a peak performance of 98.48%, while AlexNet reaches 95.45%. Additionally, ViT model offers better classification performance with fewer extracted features, which could be advantageous for computational efficiency. In addition, the standard deviation (std) indicates how much the classification accuracy of each model varies between different experimental runs. Among the tested models, ViT shows the lowest standard deviation (0.62), which reflects its strong stability and robustness under repeated evaluations.

| Model | Accuracy (%) | | | | Size of Features |
|---|---|---|---|---|---|
| | Average | std | Min | Max | |
| AlexNet | 94.85 | 1.04 | 93.94 | 95.45 | 12288 |
| ResNet50 | 95.76 | 0.76 | 93.94 | 96.97 | 6144 |
| DenseNet201 | 95.45 | 0.79 | 92.42 | 96.97 | 5760 |
| ViT | 97.23 | 0.62 | 95.45 | 98.48 | 2304 |

Table 3 Accuracy of various models without feature selection

## 5.3 Multimodal with features selection

In this experiment, for each model, the feature extraction process was conducted with feature reduction during training. This approach allowed for a comprehensive evaluation of how feature reduction impacts model performance in terms of accuracy and computational efficiency. To ensure the robustness of the results, each strategy was executed 10 times. The average, minimum and maximum values of accuracy were recorded, as well as the length of the reduced feature vector, for each run. The comparison between Tables 3 and 4 highlights the significant impact of feature selection on models performance. Without feature selection, AlexNet, ResNet50, and DenseNet201 require large feature sets to achieve high accuracy. However, after applying feature selection the feature vector lengths are greatly reduced. In the case of BGWO, the features from ResNet50 and DenseNet201 are reduced to $1074.8$ and $1017.4$ as average dimensions respectively, while maintaining nearly the same accuracy. AlexNet, despite achieving the highest accuracy, still requires a larger feature set compared to the other models. Interestingly, ResNet50 and DenseNet201 demonstrate more balanced performance by achieving high accuracy (96.36% and 96.52%, respectively) with significantly reduced feature dimensions, making them more computationally efficient.

The ViT achieved the highest average accuracy of 97.42%, confirming its superior capability in extracting discriminative and compact features from the hand-drawing data. Despite the dimensionality reduction, the

model maintained high classification performance, indicating that BGWO effectively preserved the most relevant features while eliminating redundant ones. In terms of feature vector reduction, ViT produced the lowest feature size (457 features), which is significantly smaller than those of AlexNet, ResNet50, and DenseNet201. This highlights the efficiency of ViT–BGWO integration in generating a lightweight and optimized representation without compromising accuracy. The standard deviation (1.25) remains low, showing consistent behavior under multiple runs. Compared to CNN-based models, ViT provides a better balance between accuracy, stability, and feature compactness. Concerning the BRIME optimizer, the ViT achieved the highest average accuracy (96.52%) with the lowest feature vector length (132.3). AlexNet and ResNet50 also provided competitive accuracies (around 94%), but with much larger optimal feature sizes (974 and 250, respectively), reflecting higher dimensional feature spaces. DenseNet201 yielded slightly lower performance (93.03%) and moderate feature length. Figures 5,6 illustrates the comparative performance of different feature selection algorithms (BPSO, BGWO, and BRIME) and the baseline without feature selection (NoFS) across various deep feature and vision transformer extractors. In general, all selection methods provided accuracy levels comparable to or slightly better than the baseline, confirming their ability to preserve discriminative information while reducing feature dimensionality.

| **Optmizer** | **Model** | **Accuracy (%)** | | | | **Feature Vector Length (%)** | | | | **Optimal** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Mean** | **std** | **Min** | **Max** | **Mean** | **std** | **Min** | **Max** | **size** |
| **BGWO** | AlexNet | 96.97 | 1.23 | 95.45 | 98.48 | 2303.9 | 34.8 | 2242 | 2353 | 2353 |
| | ResNet50 | 96.36 | 1.62 | 93.94 | 98.48 | 1074.8 | 25.7 | 1038 | 1113 | 1089 |
| | DenseNet201 | 96.52 | 1.60 | 93.94 | 98.48 | 1017.4 | 20.4 | 995 | 1054 | 1006 |
| | ViT | 97.42 | 1.2 | 95.45 | 98.48 | 474.5 | 14.37 | 457 | 495 | 457 |
| **BPSO** | AlexNet | 96.97 | 1.01 | 95.45 | 98.48 | 5787.2 | 62.2 | 5694 | 5880 | 5748 |
| | ResNet50 | 96.51 | 1.02 | 95.45 | 98.48 | 2854.2 | 53.2 | 2741 | 2906 | 2833 |
| | DenseNet201 | 96.06 | 1.28 | 93.94 | 98.48 | 2678.8 | 36.9 | 2614 | 2730 | 2647 |
| | ViT | 97.27 | 0.96 | 95.45 | 98.48 | 1346 | 31.4 | 1294 | 1386 | 1332 |
| **BRIME** | AlexNet | 94.84 | 2.39 | 90.91 | 98.48 | 320.2 | 264.69 | 136 | 974 | 974 |
| | ResNet50 | 94.24 | 2.84 | 89.39 | 96.97 | 419.2 | 267.2 | 205 | 1036 | 250 |
| | DenseNet201 | 93.03 | 2.87 | 87.88 | 96.97 | 340.80 | 163.49 | 223 | 706 | 289 |
| | ViT | 96.51 | 1.89 | 92.42 | 98.48 | 132.3 | 78.81 | 70 | 316 | 70 |

Table 4 Accuracy and reduced feature vector length using different optimizers

The combination of the Vit and the BPSO achieves the highest average accuracy (97.27%) with a smallest feature vector (1346 features). The relatively low standard deviation (0.958) indicates stable performance across runs. Compared to CNN-based models, ViT maintains superior accuracy while significantly reducing feature size, demonstrating the effectiveness of BPSO in enhancing model efficiency without sacrificing classification performance.

## 5.4 Statistical Significance Analysis

In the previous section, the reported mean and standard deviation values were computed from 10 independent trials, conducted with and without feature selection (FS). To ensure the robustness and statistical reliability of the observed improvements, the experiments are extended to 50 independent trials, and a formal statistical analysis based on p-values and confidence intervals (CIs) with a significance threshold of $\alpha = 0.05$ is performed.

Table 5 summarizes the obtained confidence intervals for the considered models with and without feature selection, as well as the p-values between them, while Figure 7 illustrates the overlap of the 95% confidence intervals. For DenseNet, both BGWO and BPSO show statistically significant improvements ($p \ll 0.05$), while BRIME yields a smaller yet still significant gain (p = 0.03642). These findings align with the observed upward

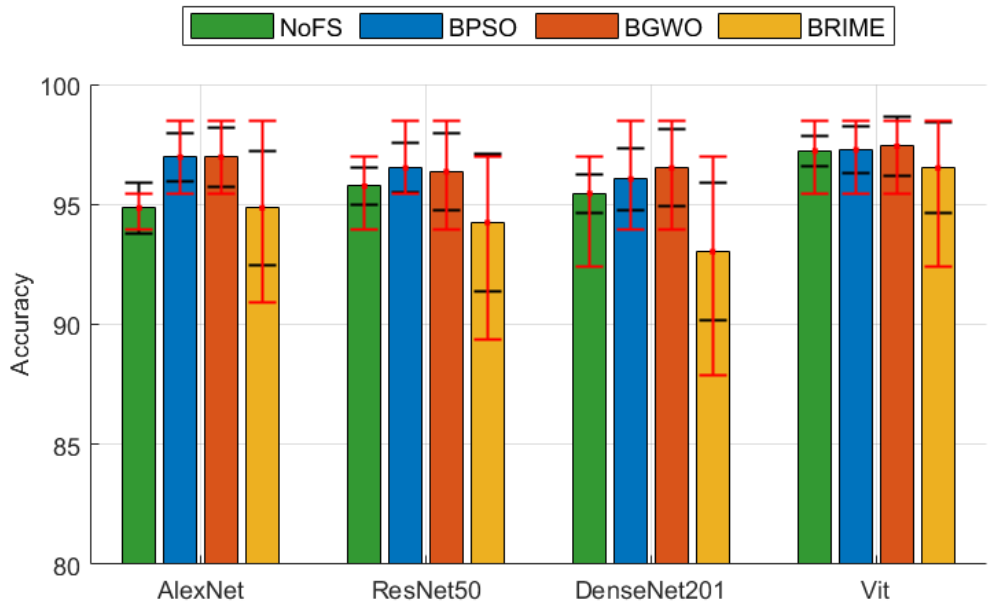



Figure 5 Comparative performance of the selection methods, black dot=$\pm std$, red dot= $Min - Max$

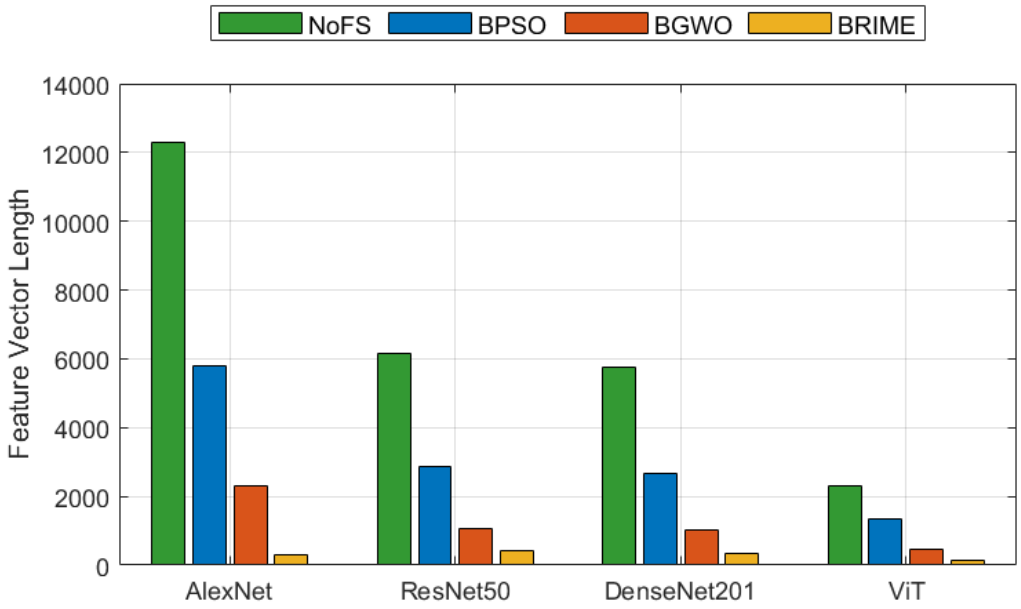



Figure 6 Feature vector length with and without feature selection methods

shift and limited overlap of the corresponding confidence intervals compared to the baseline. For ResNet50, BGWO ($p = 0.000103$) and BPSO ($p = 0.00248$) result in statistically significant accuracy improvements, whereas BRIME does not (p = 0.1187), which is consistent with the noticeable overlap between its confidence interval and that of the NoFS case. For AlexNet, BGWO and BPSO produce very small p-values, indicating clear statistical significance, and BRIME also achieves a significant improvement ($p = 0.003043$). Although some variability is present, the confidence intervals generally remain above the baseline. In contrast, for ViT,

none of the feature selection methods lead to statistically significant improvements ($p > 0.05$), and the confidence intervals largely overlap with the baseline. The limited impact of feature selection on Vision Transformer (ViT) features is due to their inherently compact and informative nature. Unlike CNN features, ViT representations capture global dependencies through self-attention, resulting in low redundancy and well-structured feature spaces. As a result, feature selection provides minimal benefit and may even remove useful information, explaining the negligible performance gains observed. When also considering feature vector length, BGWO appears to offer a more favorable balance between accuracy and dimensionality reduction, achieving statistically significant improvements with fewer selected features for CNN-based architectures. BPSO remains a viable alternative, but tends to retain larger feature sets, making it comparatively less compact and efficient in this context.

Table 5 Statistical comparison using P-values and Confidence Intervals

| Model | NoFS | BGWO | | BPSO | | BRIME | |
|---|---|---|---|---|---|---|---|
| | 95% CI | P-value | 95% CI | P-value | 95% CI | P-value | 95% CI |
| AlexNet | 94.80-94.98 | $< 0.001$ | 96.57-96.85 | $< 0.001$ | 96.54-96.88 | 0.003043 | 94.89-95.50 |
| ResNet50 | 95.64-95.78 | 0.000103 | 95.88-96.16 | 0.00248 | 95.82-96.04 | 0.1187 | 95.28-95.74 |
| DenseNet | 94.81 , 95.23 | $< 0.001$ | 95.88-96.16 | $< 0.001$ | 95.90-96.24 | 0.03642 | 94.40-94.93 |
| ViT | 97.17-97.47 | 0.5262 | 97.24-97.51 | 0.7089 | 97.06-97.24 | 0.1322 | 96.85-97.18 |

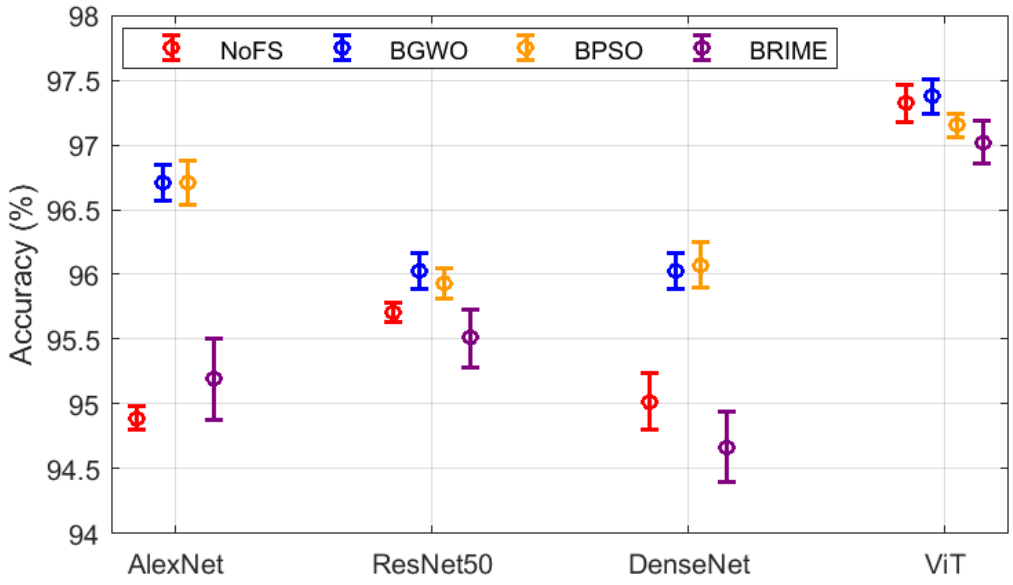



Figure 7 Overlapping 95% Confidence Intervals.

## 5.5 Computational Complexity Analysis

As shown in table 6, the computational cost of the proposed framework is mainly dominated by the feature selection stage, while classifier training remains negligible and varies between 0.025 and 0.04 s. Also, the runtime increases with feature dimensionality, for 12,288 AlexNet features, a higher computation is required, whereas lower-dimensional features such as ViT show reduced runtimes. Among the optimizers, BPSO is consistently the most computationally expensive, while BRIME is the fastest, and BGWO provides a good

trade-off between performance and cost. Feature selection significantly reduces dimensionality, leading to lower memory usage and faster classifier training. Note that the additional training cost incurred during the optimization step does not negatively impact the proposed method because this task is performed offline.

Table 6 Computational Time and Feature Reduction Analysis for Different Models and Optimizers

| Model | Optimizer | Feature Selection Time (s) | Classifier with FS (s) | Classifier without FS (s) | Dimension After Reduction |
|---|---|---|---|---|---|
| AlexNet | BGWO | 42.8191 | 0.0279 | 0.0423 | 3107 |
| | BRIME | 37.3498 | 0.0396 | | 655 |
| | BPSO | 70.1672 | 0.0275 | | 5715 |
| ResNet50 | BGWO | 15.2831 | 0.0265 | 0.0329 | 1487 |
| | BRIME | 14.3703 | 0.0293 | | 610 |
| | BPSO | 24.2258 | 0.0305 | | 1523 |
| DenseNet | BGWO | 16.3186 | 0.0330 | 0.0391 | 1420 |
| | BRIME | 16.0318 | 0.0289 | | 869 |
| | BPSO | 27.3679 | 0.0378 | | 1468 |
| ViT | BGWO | 13.9403 | 0.0298 | 0.0412 | 658 |
| | BRIME | 14.6204 | 0.0255 | | 407 |
| | BPSO | 21.4388 | 0.0260 | | 1348 |

## 5.6 Overall discussion of the proposed approaches

To more point out the effectiveness of the proposed approaches, confusion matrices were conducted for the four models with and without feature reduction. Some models having similar performance were grouped as shown in figure 8. From confusion matrices, performance metrics including accuracy, precision, recall, specificity and F1-score were evaluated.

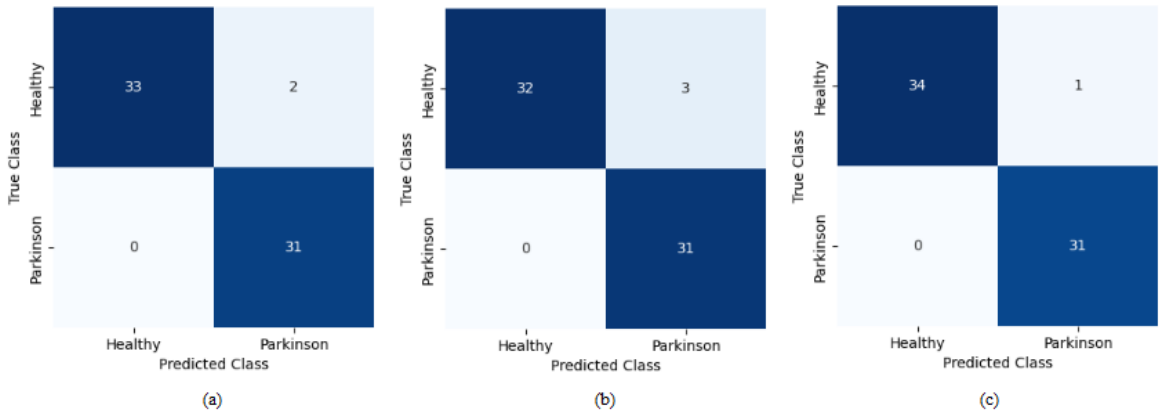



Figure 8 Confusion matrix for ResNet50, DenseNet201 without reduction (a) , AlexNet without reduction (b) , AlexNet, ResNet50, DenseNet201, ViT model with reduction (c) .

As shown in Table 7, ViT without reduction provides a highly effective PD detection rate. The perfect recall of 100% ensures no patients are missed. The slightly lower precision of 96.88% shows that a few healthy subjects were incorrectly diagnosed, but this trade-off is acceptable in a healthcare context where early detection is often prioritized. In the same context, AlexNet without reduction can achieves an accuracy of 95.45% which is lower than the formers but still acceptable. Also, as shown in Figure 8, despite the perfect recall of 100% indicating that all patients with Parkinson's are identified; the three cases of false positives have slightly lower

| Optimizer | Model | Acc (%) | Pr (%) | Re (%) | Sp (%) | F1 Score (%) |
|---|---|---|---|---|---|---|
| NoFS | AlexNet | 95.45 | 91.18 | 100 | 91.43 | 95.38 |
| | ResNet50 | 96.97 | 93.94 | 100 | 94.29 | 96.87 |
| | DenseNet201 | 96.97 | 93.94 | 100 | 94.29 | 96.87 |
| | ViT | 98.48 | 96.88 | 100 | 97.14 | 98.48 |
| With BGWO | AlexNet | 98.48 | 96.88 | 100 | 97.14 | 98.48 |
| | ResNet50 | 98.48 | 96.88 | 100 | 97.14 | 98.48 |
| | DenseNet201 | 98.48 | 96.88 | 100 | 97.14 | 98.48 |
| | ViT | 98.48 | 96.88 | 100 | 97.14 | 98.48 |

Table 7 Best performance obtained from several trials

precision and specificity, indicating that some healthy individuals are being misclassified. Results obtained from the four models with reduction demonstrate a highly effective PD detection models. Achieving an accuracy of 98.488% means that only very few classification errors could be occur.

Also, With a precision of 96.88% and a specificity of 97.14%, the proposed models demonstrate high performance and may serve as supportive tools for PD diagnosis, yielding only one false positive and correctly classifying nearly all healthy individuals.

Additionally, we have used both the Precision-Recall (PR) curve and the Receiver Operating Characteristic (ROC) curve to more evaluate the proposed approach. The PR curve focuses on the trade-off between the proportion of correctly identified Parkinson's cases among all predicted positives and the proportion of actual Parkinson's cases correctly identified. A higher area under the PR curve indicates better classification performance, particularly in distinguishing positive cases. Meanwhile, the ROC curve is a graphical representation of the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) [28]:

$$TPR = \frac{TP}{TP + FN} \tag{25}$$

$$FPR = \frac{FP}{FP + TN} \tag{26}$$

The ROC curve plots TPR vs. FPR at different classification thresholds. Ideally, a perfect classifier would have a curve that reaches the point with coordinates TPR = 1and FPR = 0. The Area Under the ROC Curve (AUC-ROC) serves as a key metric, where a value closer to 1.0 indicates strong discriminatory power. As shown in Figure 9, The Area Under the Curve (AUC) values confirm that BGWO-based models achieve a perfect AUC of 1.0, while the original models have slightly lower AUC values. In the same context, Figure 10 presents Precision-Recall (PR) curves for the four models with and without feature selection. The BGWO-enhanced models exhibit higher Average Precision (AP) scores for the four architectures, achieving AP = 0.9714, while the original models show slightly lower AP values ( AlexNet: 0.9587, ResNet50: 0.9691, DenseNet201: 0.9699). This indicates that using BGWO for feature selection can achieve a nearly perfect classification ability for the three models.

A detailed examination of the misclassified samples revealed that a small number of healthy subjects were incorrectly identified as PD patients. These misclassifications often occurred in drawings that exhibited tremor-like irregularities or inconsistent stroke pressure, which visually resemble early Parkinsonian symptoms. For instance, some healthy participants produced slightly shaky or uneven trajectories in the spiral and meander tasks, leading the model to assign them to the PD class. Figure 11 shows an example of drawings from healthy subjects that was incorrectly classified as PD.

Despite the high performance achieved, the used dataset which consists of only 66 subjects which may limits the generalizability of the proposed approach. Although, data augmentation was applied, it does not increase the intrinsic diversity at the subject level. To overcome data leakage, a strict subject-wise splitting strategy was

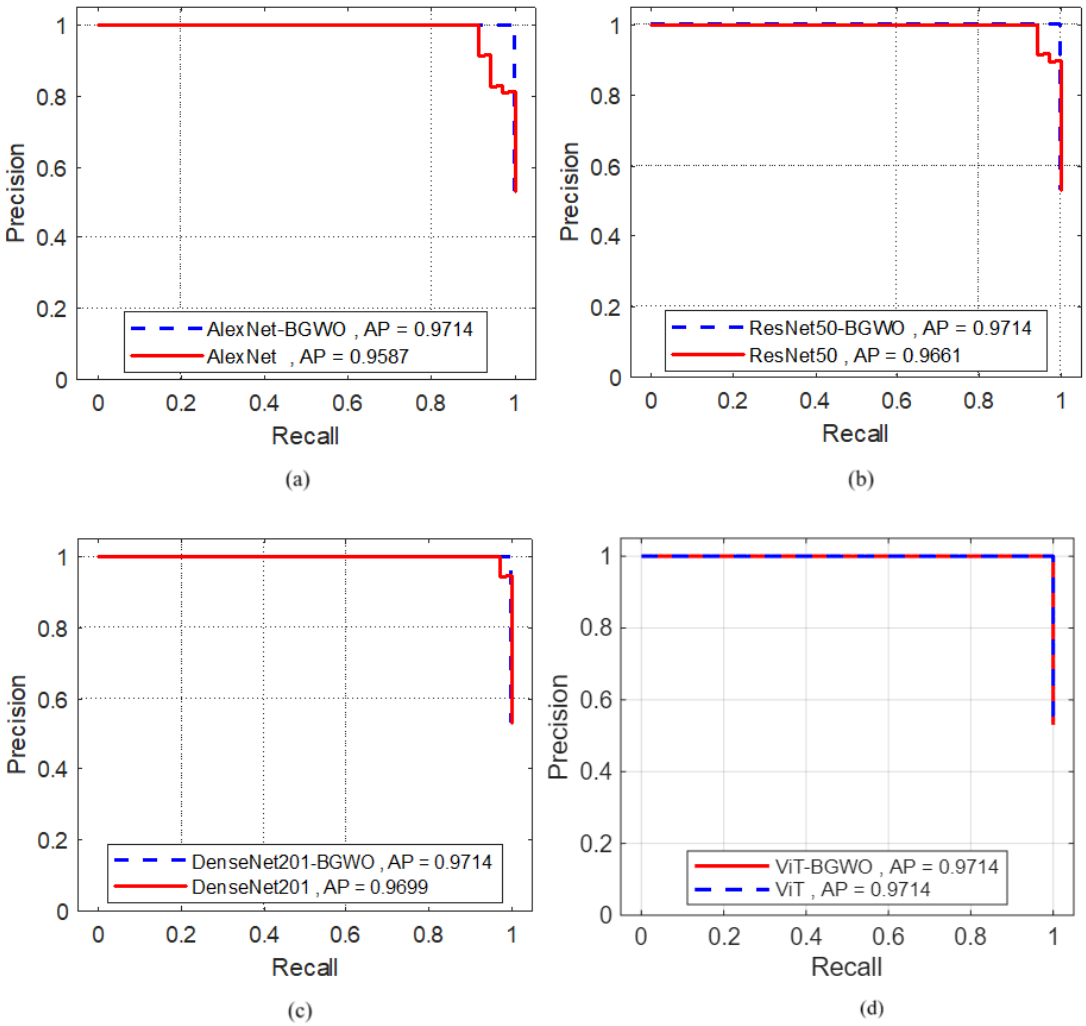



Figure 9 Precision-Recall curves for (a) AlexNet, (b) ResNet50, (c) DenseNet201 and (d) Vit with and without feature selection.

considered and data augmentation was performed only on the training folds after splitting. Nevertheless, the models still capture dataset-specific characteristics and a risk of overfitting may be introdueced and the reported performance should be interpreted with caution.

While the proposed method achieves high classification performance, the observed improvements should be interpreted with caution from a clinical perspective, as small accuracy increases may not translate into meaningful diagnostic benefits when baseline performance is already high. Nevertheless, the approach offers clinically relevant advantages, including high robustness and consistency across folds, high recall ensuring minimal false negatives and significant feature dimensionality reduction. Additionally, the use of frozen feature extractors with a simple classifier enhances model interpretability compared to fully fine-tuned deep models. Despite these strengths, the clinical applicability remains limited without validation on larger datasets.

## 5.7 Cross-Dataset Generalization Study

To further evaluate the generalization capability of the proposed framework, experiments were conducted on an independent dataset, the HandPD dataset [29] . This dataset consists of 92 participants, including 74 pa-

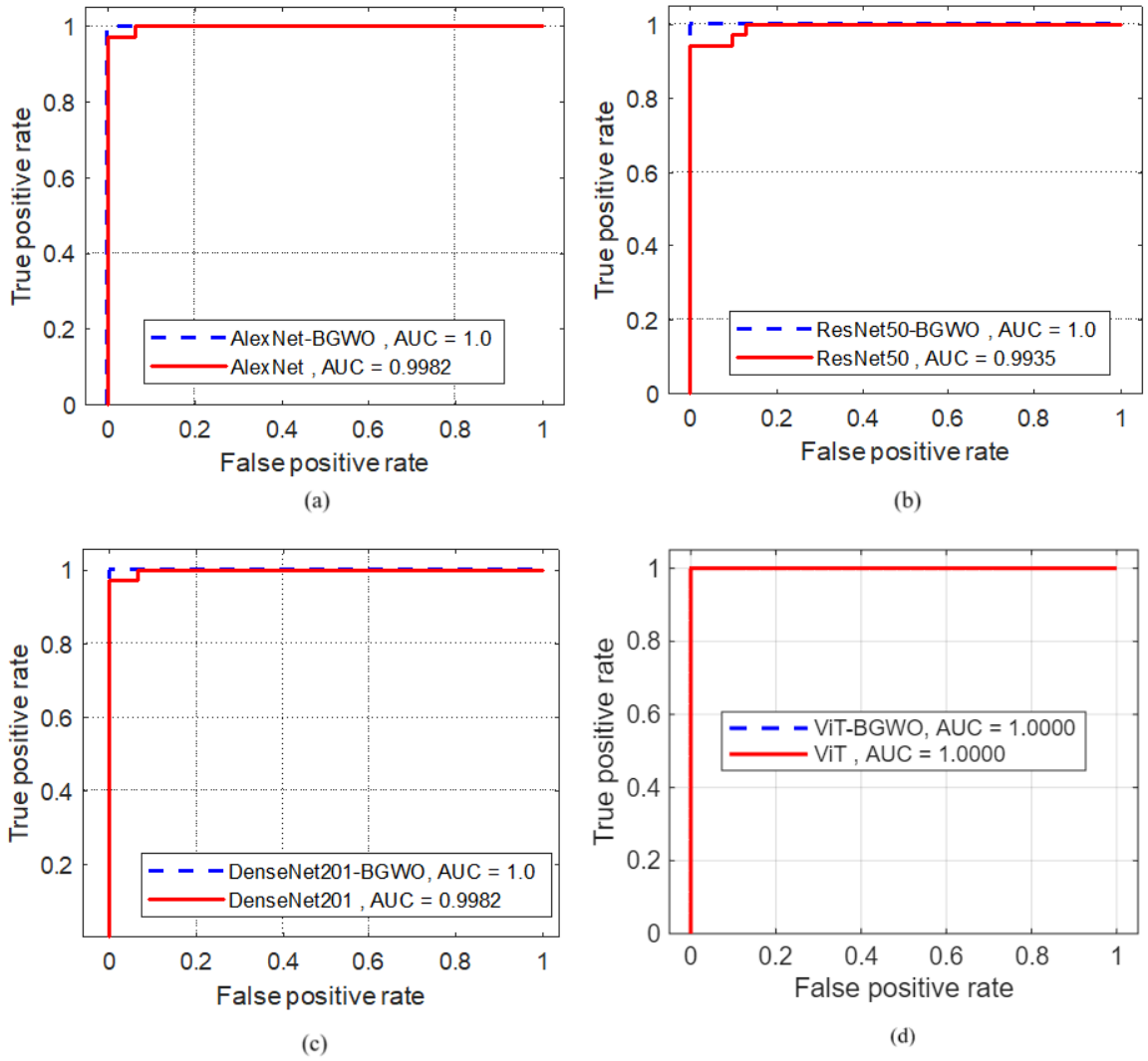



Figure 10 ROC curves for (a) AlexNet, (b) ResNet50, (c) DenseNet201 and and (d) Vit with and without feature selection.

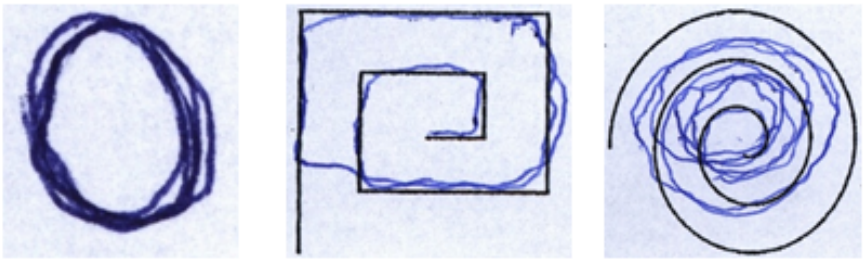
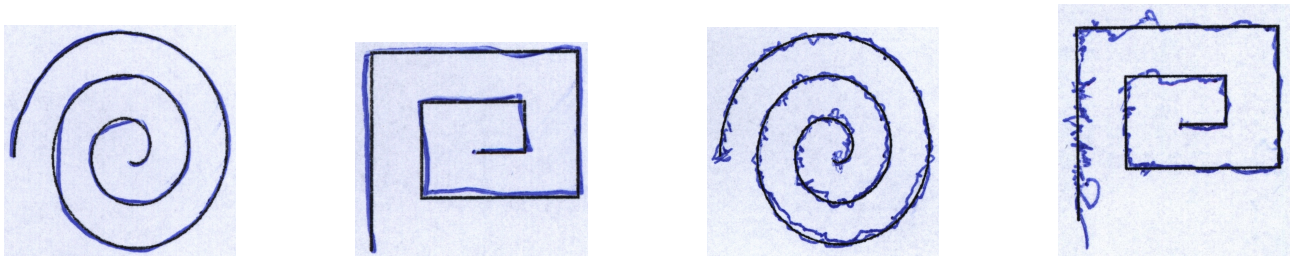


Figure 11 Misclassified samples from healthy subjects incorrectly identified as PD.

tients with Parkinson's disease and 18 healthy control subjects. Each participant performed spiral and meander drawings. For each modality, the dataset includes 72 images from healthy individuals and 296 images from Parkinson's patients. Unlike standard evaluation protocols, the feature selection process was not re-applied

to the external dataset. Instead, the optimal feature subset identified on the NewHandPD dataset was directly reused. Specifically, deep features were extracted from the external dataset using the same pretrained models, and only the features corresponding to the previously selected indices were retained. The classifier trained on the selected features was then applied without modification.



(a) Spiral (H) (b) Meander (H) (c) Spiral (PD) (d) Meander (PD)

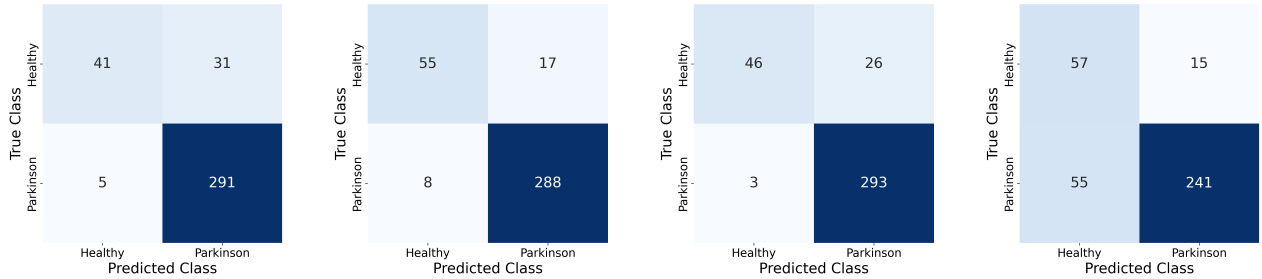
Figure 12 Handwriting samples from the HandPD dataset.

This strategy ensures a strict evaluation of generalization, as no information from the external dataset is used during feature selection. It is important to note that the HandPD dataset does not contain circle drawing samples. To maintain consistency in the multimodal feature representation, the corresponding feature components were set to zero. This allows preserving the same feature structure without introducing additional bias.

| **Model** | **Acc (%)** | **Pr (%)** | **Re (%)** | **Sp (%)** | **F1-score (%)** |
|---|---|---|---|---|---|
| AlexNet-BGWO | 90.22 | 90.37 | 98.31 | 56.94 | 94.17 |
| ResNet50-BGWO | 93.21 | 94.43 | 97.30 | 76.39 | 95.84 |
| DenseNet201-BGWO | 92.12 | 91.85 | 98.99 | 63.89 | 95.28 |
| ViT-BGWO | 80.98 | 94.14 | 81.42 | 79.17 | 87.32 |

Table 8 Cross-Dataset Evaluation Results

The cross-dataset evaluation highlights the effectiveness of BGWO in improving model generalization. As shown in table 8, ResNet50 achieved the best overall performance, offering the highest accuracy and F1-score with a good balance between precision and recall. DenseNet and AlexNet showed very high recall, indicating high detection of Parkinson's cases, but suffered from lower specificity, leading to more false positives. In contrast, ViT exhibited the lowest performance overall, with reduced accuracy and recall. All these observations are further confirmed by the confusion matrices presented in Figure 13




(a) AlexNet (b) ResNet50 (c) DenseNet (d) ViT

Figure 13 Confusion matrices of deep models under cross-dataset evaluation.

### 5.8 Comparative study

Table 9 shows the comparison of diverse Parkinson's disease classification using the NewHandPD dataset and across different models and approaches. In [11], CNN with LSTM approach achieved moderate accuracy of 84.60% for meander images and a higher accuracy of 96.25% when using spiral drawings. The Spatio-temporal Siamese Neural Network [12] improved generalization, achieving up to 95.86% for spirals and 93.39% for meander images. VGG16 [13] achieved a relatively modest accuracy of 90.41%, whereas the fine-tuned ResNet50 model considerably outperformed earlier architectures. Specifically, [13] reports 96.61% accuracy for ResNet50 with frozen weights and 98.31% for the fine-tuned model. In comparison, our proposed approach achieves 98.48% using ResNet50 with frozen weights, demonstrating that the combination of last-layer feature extraction and metaheuristic feature selection can surpass even fine-tuned models while preserving the computational efficiency of frozen networks.

| **Proposal** | **Model** | **Modality** | **Acc (%)** |
|---|---|---|---|
| [11] | CNN with LSTM | Meander images | 84.60 |
| | | Circles drawings | 94.02 |
| | | Spiral drawings | 96.25 |
| [12] | Spatio-temporal Siamese neural network | Meander images | 93.39 |
| | | Circles drawings | 90.21 |
| | | Spiral drawings | 95.86 |
| [13] | VGG16 model | Spiral and meander | 90.41 |
| [13] | ResNet50 with frozen weights | Spiral, meander, and circle images | 96.61 |
| [13] | Fine-tuned ResNet50 model | Spiral, meander, and circle images | 98.31 |
| [14] | CAS Tranformer | Spiral, meander, and circle images | **92.68** |
| **Proposed** | With BGWO | Spiral, meander, and circle images | **98.48** |

Table 9 Comparison of accuracy of proposed method with available approaches in the literature

## 6 Conclusion

The paper presents a comprehensive framework for Parkinson's disease detection based on multimodal deep feature extraction and optimization-based feature selection. Drawings of circles, spirals, and meanders from the NewHandPD dataset are utilized, providing diverse hand-drawn data from both Parkinson's disease patients and healthy individuals. AlexNet, ResNet-50, DenseNet-201, and ViT models are employed as feature extractors, while population-based optimization algorithms; BGWO, BPSO, and BRIME are applied for feature selection to retain only the most relevant features for the final classification stage. The obtained results demonstrate that multimodal analysis significantly outperforms unimodal approaches. The integration of optimization algorithms effectively reduces feature dimensionality while maintaining or improving classification performance, highlighting their potential to enhance computational efficiency without compromising accuracy. Notably, an accuracy of 98.48% is achieved in most cases when combining the three deep learning models and ViT with feature selection. Furthermore, an extensive statistical significance analysis based on 50 independent trials confirms that feature selection significantly improves the performance of CNN-based models, particularly when using BGWO and BPSO, with p-values well below the 0.05 threshold. BRIME exhibits limited or inconsistent gains, especially for ResNet-50. In contrast, none of the feature selection methods yield statistically significant improvements for the Vision Transformer, indicating that ViT already produces highly discriminative features. In terms of feature compactness and performance stability, BGWO proves to be the most effective and consistent feature selection method for CNN-based architectures.

## References

[1] M. Ariful Islam, M. Ziaul Hasan Majumder, M. Alomgeer Hussein, K. Murad Hossain and M.-S. Miah, "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets", *Heliyon* 10(2), 2024. https://doi.org/10.1016/j.heliyon.2024.e25469

[2] E. Adeli F. Shi and C.-Y. Wee, "Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data", *NeuroImage*, 141:206-219,2016. https://doi.org/10.1016/j.neuroimage.2016.05.054

[3] A. Aljohani , "Late feature fusion using neural network with voting classifier for Parkinson's disease detection", *BMC Med Inform Decis Mak*, 269, 2024. https://doi.org/10.1186/s12911-024-02683-0

[4] R. Abumalloh M. Nilashi ,S. Samad, H. Ahmadi, A. Alghamdi, et al., "Parkinson's disease diagnosis using deep learning: A bibliometric analysis and literature review", *Ageing Research Reviews*, 96, 2024. https://doi.org/10.1016/j.arr.2024.102285

[5] I. Kamran, S.Naz, I. Razzak, M. Imran, "Handwriting dynamics assessment using deep neural network for early identification of Parkinson's disease", *Comput. Syst*, 117: 234-244,2021. https://doi.org/10.1016/j.future.2020.11.020

[6] R. Arti, "Imperative Role of Machine Learning Algorithm for Detection of Parkinson's Disease: Review, Challenges and Recommendations", *Diagnostics*, 12(8), 2022. https://doi.org/10.3390/diagnostics12082003

[7] O. Alniemi, H. Mahmood, "Convolutional neural network for the detection of Parkinson disease based on hand-draw spiral images", *Indonesian Journal of Electrical Engineering and Computer Science*, 30, 2023. http://doi.org/10.11591/ijeecs.v30.i1.pp267-275

[8] C.R. Pereira, D.R. Pereira, F.A. Silva, J.P. Masieiro, S. Weber, C. Hook, J. Papa, "A new computer vision-based approach to aid the diagnosis of Parkinson's disease",*Comput. Methods Programs Biomed*, 136:79-88, 2016. https://doi.org/10.1016/j.cmpb.2016.08.005

[9] P. Zham, D.K. Kumar, P. Dabnichki, S.P Arjunan, S. Raghav, "Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral", *Front. Neurology*, 8, 2017. https://doi.org/10.3389/fneur.2017.00435

[10] C. Pereira, S. Weber, C. Hook, G.H. Rosa, J,P. Papa, "Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics",*Proceedings of the SIBGRAPI - Conference on Graphics, Patterns and Images*, 340-346, 2016. https://doi.org/10.1109/SIBGRAPI.2016.054

[11] R. Salah El-Sayed, "A Hybrid CNN-LSTM Deep Learning Model for Classification of the Parkinson Disease", *International Journal of Applied Mathematics*, 53(4):1427-1436,2023.

[12] A. Zhao, H. Wu, M. Chen, N. Wang, "A Spatio-Temporal Siamese Neural Network for Multimodal Handwriting Abnormality Screening of Parkinson's Disease", *International Journal of Intelligent Systems*, 7474, 2023. https://doi.org/10.1155/2023/9921809

[13] I. Razzak, I. Kamran and S. Naz, "Deep Analysis of Handwritten Notes for Early Diagnosis of Neurological Disorders",*2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK*, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207087

[14] N. Wang, X. Niu and Y. Yuan,Y. Sun, R. Li, G. You, A. Zhao, "A Coordinate Attention Enhanced Swin Transformer for Handwriting Recognition of Parkinson's Disease",*IET Image Processing*, 17(9), 2023, pp. 2686-2697. https://doi.org/10.1049/ipr2.12820

[15] S. Mehta, M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer", preprint arXiv:211002178, 2021. https://doi.org/10.48550/arXiv.2110.02178

[16] A. Krizhevsky, I. Sutskever, G. Hinton , "ImageNet Classification with Deep Convolutional Neural Networks",*Advances in Neural Information Processing Systems*, 25, 2012. https://doi.org/10.1145/3065386

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks", *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[18] K. He, X. Zhang, S. Renand J. Sun, "Deep Residual Learning for Image Recognition", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,pp. 770-778,2016. https://doi.org/10.1109/CVPR.2016.90

[19] A. Dosovitskiy, "An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale", *In Proceedings of the 9th International Conference on Learning Representations*, Austria, May 3-7, 2021.

[20] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey", *Technologies*, 13(1), 2025 https://doi.org/10.3390/technologies13010032

[21] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili and H. Alhussian, "Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection",*IEEE Access*, 7:39496-39508,2017. https://doi.org/10.1109/ACCESS.2019.2906757

[22] E. Emary, M. Hossam H. Zawbaa, . Aboul Ella, "Binary grey wolf optimization approaches for feature selection", *Neurocomputing*, 172:371-381, 2016. https://doi.org/10.1016/j.neucom.2015.06.083

[23] H.Pan, S.Chen, H.Xiong,"A high-dimensional feature selection method based on modified Gray Wolf Optimization", *Applied Soft Computing*, 135, 2023. https://doi.org/10.1016/j.asoc.2023.110031

[24] J.Kennedy, R.Eberhart, "Particle swarm optimization," *International Conference on Neural Networks, Perth, WA, Australia*,4:1942-1948 ,1995. https://doi.org/10.1109/ICNN.1995.488968

[25] H.Su, D.Zhao, AA.Heidari, L.Liu, X.Zhang, M.Mafarja, H.Chen, " Rime: a physics-based optimization", *Neurocomputing* 532:183–214.2023. https://doi.org/10.1016/j.neucom.2023.02.010

[26] R.Abu Khurma, M.Braik, A.Alzaqebah et al.,"Advanced RIME architecture for global optimization and feature selection", *J Big Data 11* 89, 2024. https://doi.org/10.1186/s40537-024-00931-8

[27] S. Chakraborty, S. Aich, J Sim, E. Han, J. Park and H. -C. Kim, "Parkinson's Disease Detection from Spiral and Wave Drawings using Convolutional Neural Networks: A Multistage Classifier Approach", *22nd International Conference on Advanced Communication Technology (ICACT), Phoenix Park, Korea (South)*, 298-303, 2020. https://doi.org/10.23919/ICACT48636.2020.9061497

[28] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters* , 27(8):861-874, 2006. https://doi.org/10.1016/j.patrec.2005.10.010

[29] https://wwwp.fc.unesp.br/ papa/pub/datasets/Handpd/