Neural Computing and Applications (2022) 34:4567–4593
https://doi.org/10.1007/s00521-021-06612-4

ORIGINAL ARTICLE

An adaptive intelligent diagnostic system to predict early stage of parkinson's disease using two-stage dimension reduction with genetically optimized lightgbm algorithm

Joy Dhar[1]

Received: 4 December 2020 / Accepted: 4 October 2021 / Published online: 21 October 2021

Abstract
Parkinson's disease is one of the most prevalent neurodegenerative sicknesses distinguished by motor function impairment. Parkinson's disease (PD) diagnosis is a complicated job that demands the evaluation of numerous non-motor and motor signs. Throughout the analysis of vocal or speech abnormalities are notable indications that doctors should think. Early diagnosis of PD is essential for preliminary treatment and assisting the doctor to heal and evade the PD's spread in other brain cells and save several lives. So, this study introduces an adaptive expert diagnostic system to predict PD accurately. This suggested system proposes a hybrid methodology: two-stage mutual information and autoencoder-based dimensionality reduction approach with genetically optimized LightGBM (MI-AE-GOLGBM) algorithm, to improve the proposed system's performance and predict the best outcomes. The proposed MI-AE-GOLGBM approach comprises four methodologies: mutual information, autoencoder, genetic algorithm, and LightGBM algorithm, in which mutual information and autoencoder are implemented to form a two-stage dimensionality reduction approach for selecting the informative features from the input dataset and hence producing a reduced dataset with the most significant newly generated features, and genetic algorithm is employed to intelligently optimize the hyperparameters of LightGBM algorithm in which LightGBM algorithm utilizes such newly generated features and the best-optimized hyperparameters provided by the two-stage mutual information and autoencoder-based dimension reduction methods and the genetic algorithm, respectively, to which to classify the PD sufferers and healthy controls and enhance the precision value and reliability of the proposed system. Four different real-world publicly available Parkinson's disease datasets are employed in this proposed research to assess and verify the proposed methodology's performance. This proposed research utilizes different machine learning (ML) algorithms to compare our proposed approach's performance. The outcomes reveal that the proposed methodology can produce the best predictions based on voice data relating to the PD compared to the different ML algorithms.

Keywords Parkinson's disease identification · Autoencoder · Genetic algorithm · LightGBM algorithm

1 Introduction

Parkinson's disease (PD) is a notable neurodegenerative illness around the earth [1]. Based on the American Parkinson's Disease Association (APDA) reports, about 10 million people suffer from PD worldwide. Unfortunately, there is no cure available to heal this disorder due to it does not have any single test to diagnose PD, although doctors technically understand what occurs with the subject [2]. Therefore, it worries about the global prevalence of PD because it increases a lot over time. Moreover, the therapies can only assist in controlling the motor signs [2]. Unlike other illness diagnoses, genetic approaches do not create significant cardinal features of Parkinson's disease [2]. However, there are traditional approaches to diagnose and investigate Parkinson's disease that is invasive approaches, namely, MR, computed tomography scan, X-rays, single photon emission computerized tomography/dopamine transporter scan, PET, and ultrasound, which are expensive and can be helpful when Parkinson's

✉ Joy Dhar
joy.dhar@hatgobindapurschool.co.in

[1] Department of NSQF, Hatgobindapur M. C. High School, East Bardhaman, West Bengal, India

disease is grown over the brain [2]. Therefore, we need a non-invasive and clinical screening analysis for diagnosing Parkinson's disease early and support the doctor in healing and evading the PD's spread in other brain cells and saving several lives. Thus, this study proposes an adaptive intelligent diagnostic system, which comprises a hybrid approach: two-stage mutual information and autoencoder-based dimensionality reduction method with genetically optimized LightGBM (MI-AE-GOLGBM) algorithm detects the early symptoms of the PD and diagnoses the PD most accurately. The proposed MI-AE-GOLGBM approach consists of four methods: mutual information (MI), autoencoder (AE), genetic algorithm (GA), and LightGBM (LGBM) algorithm, in which MI is employed to choose the relevant, informative attributes from the given data and to pass such features to AE for extracting such features and producing a dataset with the most significant newly generated features, and GA is utilized as hyperparameter optimization methodology to which it automatically optimizes the hyperparameters of LGBM algorithm and generates the best optimal hyperparameters, and after then, LGBM algorithm utilizes such optimal hyperparameters and newly generated features to classify the PD sufferers and healthy persons and predicting the most accurate and suitable outcomes.

Several ML algorithms have been suggested to recognize the early signs of PD and related neurodegenerative disorders, and various input mode of the dataset has been employed such as voice, speech patterns data to observe subtly [2]. In this study, based on the above-said input mode, this proposed research employs the proposed MI-AE-GOLGBM methodology on four publicly available real-world Parkinson's disease datasets described in the research methodology section.

Generally, this study investigates the following questions:

1. How is the proposed MI-AE-GOLGBM model performed using a tenfold cross-validation strategy?
2. How does the proposed approach's performance compare to the irrelevant ML algorithms, and which algorithm achieves better performance for datasets 1, 2, and 3?
3. How does the proposed approach's performance compare to the relevant ML algorithms, and which algorithm achieves better performance for datasets 1, 2, and 3?
4. Is the proposed approach enhance the performance of the relevant ML algorithms–explain in detail?
5. How does the proposed approach compared to the relevant and irrelevant ML algorithms, and which algorithm achieves the best performance for the dataset 1, 2, and 3?
6. How does the suggested system verify the performance of the suggested MI-AE-GOLGBM approach [3]?
7. Which model is the best adaptive and reliable model among the suggested methodology and the various ML models for all utilized four datasets in this suggested research? Justify the reasons in brief.

The response to the queries mentioned earlier should be investigated by the proposed MI-AE-GOLGBM method in the experimental results section. However, this proposed research's primary contribution is threefold.

At first, this study developed an adaptive intelligent system to diagnose PD by utilizing a hybrid approach: two-stage mutual information and autoencoder-based dimensionality reduction approach with genetically optimized LightGBM (MI-AE-GOLGBM) methodology to enhance the proposed system's accuracy. Secondly, the MI-AE-GOLGBM method utilizes two different innovative methodologies: two-stage mutual information and autoencoder-based dimensionality reduction and genetically optimized LightGBM (GOLGBM) algorithm. The two-stage dimensionality reduction approach utilizes two different dimension reduction techniques; namely, MI and AE, in which MI is implemented to choose the informative attributes from the input data and AE is implemented to extract the features from the obtained features given by the MI and produces a reduced dataset with the most significant newly generated features. The GA is a hyperparameters optimization methodology to intelligently tune the hyperparameters and generate the best optimal hyperparameters for the LightGBM algorithm. Next, the LightGBM algorithm employs such optimal hyperparameters and the reduced dataset with the newly created features to classify the PD sufferers and healthy persons with the most suitable and reliable prediction. Thirdly, the proposed method's performance assessment is conducted using four real-world, publicly available datasets with several performance assessment metrics.

The rest part of this paper is comprised as follows. Part 2 explains the related studies and their gaps relevant to the proposed adaptive system for diagnosing PD by implementing different ML and deep learning algorithms. Part 3 illustrates the method, which describes the different operations performed by the proposed adaptive intelligent system and describes the proposed MI-AE-GOLGBM approach for diagnosing PD. Segment 4 exhibits the experimental results. A complete investigation of data is conducted to find the solutions that rely on the proposed MI-AE-GOLGBM methodology and various state-of-the-art ML algorithms through different performance assessment metrics and a tenfold cross-validation procedure. In the last part, the conclusion is described.

2 Related studies and research gap

A comprehensive investigation is conducted to review the prevailing research works almost correlated to this proposed research segment. Different types of real-world data are collected from openly accessible datasets repository for developing relevant research works. This paper principally concentrates on the current machine learning and deep learning-based methodologies for Parkinson's diseases' valuable classification. This proposed research includes the result caused by a well-known Parkinson's disease (PD) analysis technique: voice. Several current research works are available based on the above-said PD analysis technique.

In this regard, most researchers have applied various feature selection with optimization or without optimization approaches and feature extraction methodologies to enhance their developed system's performance and diagnose PD accurately. In this respect, Tracy et al. [4] investigated a database of a person with Parkinson's disease and healthy people restraining voice recordings utilized for extracting paralinguistic features that assisted as inputs to ML models for predicting Parkinson's disease severity [4]. Their literature highlighted the potential of voice to be utilized for early identification of Parkinson's disease and designated that voice might assist as a deep phenotype for Parkinson's disease, allowing exactness medication by advancing the precision, speed, availability, and cost of Parkinson's disease management [4]. In contrast, Ali et al. [5] suggested a two-dimensional data selection strategy for sample and feature selection [5]. They employed ranks features by utilizing the chi-square statistical model, searching for an optimal subset of the ranked features, and iteratively choosing examples on Parkinson's speech (PS) dataset [5].

In comparison, Ashour et al. [6] implemented principal component analysis (PCA), the eigenvector enabled centrality attribute selection approaches, and the cubic kernel SVM algorithm to detect PD after employing the PD classification (PDC) dataset [6]. The authors included in [6] generated an accuracy of 94% while detecting PD. In contrast, Karan et al. [7] suggested empirical mode decomposition and extracting features from intrinsic mode function to efficiently describe PS features using ML algorithms: support vector machines (SVM) and random forest (RF) after employing a PS dataset [7]. Whereas, Solana-Lavalle et al. [8] implemented 8 to 20 wrapper-based feature selection approaches along with four classifiers: k-nearest neighbors (KNN), multi-layer perceptron, SVM, and RF to detect vocal-based PD after employing the PDC dataset in which SVM received the best performance in terms of accuracy value of 94.7% while detecting PD [8]. In comparison, Tuncer et al. [9] introduced Minimum average maximum tree and singular value decomposition to elicit prominent features from the voice signals and automatically recognize PD with vowels' help after employing a PDC dataset [9, 10].

However, Haq et al. [11] developed an ML-based prediction system; the SVM was employed as a predictive model to predict PD [11]. Their generated model: L1-norm SVM of feature selection was employed for appropriate and deeply associated attribute selection for precise discriminative classification of Parkinson's disease and non-PD person after employing Oxford PD detection (OPDD) dataset [11]. In contrast, Despotovic et al. [12] introduced the Gaussian processes joined with automatic relevance determination methodologies for effective feature selection after employing the OPDD dataset and Parkinson's telemonitoring (PTM) dataset through which only a tiny subset of deeply related acoustic features is picked for providing more reliable performance and lower complexity [12]. In comparison, Zhang et al. [13] presented energy direction attributes that rely on empirical mode decomposition to detect PD after employing the PDC dataset [13]. Their implemented approach reported an accuracy of 96.54% to detect PD. In comparison, Solana-Lavalle and Rosas-Romero [14] implemented a voice-based detection methodology after applying feature subset selection with four various classifiers to detect PD after employing the PDC dataset [14]. Their developed model reported an accuracy of 95.9% while detecting PD.

On the other hand, Pramanik et al. [15] implemented systematically developed forest and decision forest by penalizing attributes and the RF to detect PD after employing two PD datasets: PDC dataset and acoustic dataset [15]. Their developed models obtained accuracy values from 94.12 to 95%. In contrast, Lysiak and Szmajda [16] presented the results of a comparison between nine selected feature evaluation methods and various sets of classifiers to detect PD after employing the PDC dataset [16]. At the same time, Xiong and Lu [17] utilized adaptive grey wolf optimization (GWO) algorithm and sparse autoencoders to classify PD after employing the PDC dataset [17]. Their generated approach obtained an accuracy of 95% while detecting PD.

In comparison, Pasha & Latha [18] employed two Bio-inspired optimization algorithms: GA and binary particle swarm optimization (PSO), separately on several ML classification models to decide the optimal subset of features of the PD data set giving to the satisfied classification precision [18]. In contrast, Sahu and Mohanty [19] introduced an innovative intelligence model with a combination of a chaos-mapped bat algorithm and an SVM to detect PD after employing two different PD datasets [19]. Their

progressive approach received accuracy values of 98.24% and 99.49% for the two datasets, respectively.

However, some researchers have applied various deep neural networks with optimization methodologies to classify PD accurately. In this matter, Olivares et al. [20] generated an optimized extreme learning machine using the Bat methodology to classify PD after employing the PDC dataset [20]. Their developed approach received an accuracy: 96.74% while classifying PD.

However, in terms of generating hybrid methodology, Gunduz [21] implemented a hybrid dimensionality reduction approach and multi-kernel SVM to classify PD after employing 30 features of the PD dataset [21]. Variational Autoencoders and Fisher score and relief were employed in the hybrid dimensionality reduction approach [21]. Variational autoencoder and relief-based dimension reduction with multi-kernel SVM achieved an accuracy of 0.916 to classify PD [21]. While Cai et al. [22] proposed a hybrid methodology. They employed a relief-based feature selection approach and SVM and bacterial foraging optimization to predict PD accurately after utilizing the UCI-based OPDD dataset [22]. Hoq et al. [23] developed two hybrid models based on an SVM integrating with a PCA and a Sparse Autoencoder (SAE) to detect PD sufferers after employing the PDC dataset [23]. Their developed model: SAE with SVM reported accuracy: 93.5% and F1-score: 95.1% to detect PD [23].

In contrast, some researchers either employed a fuzzy-based methodology with the optimization algorithms or utilized several feature extraction methodologies to predict the PD accurately. In this matter, El-Hasnony et al. [24] proposed a fog-based adaptive neuro-fuzzy inference system with the PSO and the GWO methodology for predicting PD in an IoT environment after utilizing UCI-based PDC dataset for providing accurate outcomes [24]. Chen et al. [25] exhibited an effective and efficient analysis system through a fuzzy KNN to identify PD [25]. PCA was applied for generating the best-classified feature set on which the optimal fuzzy KNN method was built [25].

However, some researchers have applied various feature selection or feature extraction methodologies and hyperparameter optimization methodologies to accurately diagnose PD and enhance its performance. In this matter, Soumaya et al. [26] employed a GA with the SVM approach after employing the PDC dataset [26]. Their progressive approach achieved the performance while classifying PD in terms of accuracy value of 91.18%. In comparison, Ali et al. [27] proposed an intelligent system that employed linear discriminant analysis (LDA) to reduce dimensions and a GA for parameters tuning of the neural network [27]. They employed a PS dataset containing two datasets; each has 20 regular and 25 Parkinson's disease patients to provide suitable accuracy.

In contrast, Lahmiri & Shmuel [28] proposed eight feature selection approaches when joined with a nonlinear SVM to distinguish between Parkinson's disease and healthful people [28]. The hyperparameters of the radial basis function kernel of the SVM classifier were tuned through the Bayesian optimization approach [28]. In contrast, Kaur et al. [29] proposed a framework that relies on a grid search hyperparameter optimization for tuning the multiple hyperparameters of a deep learning model and satisfying accuracy [29].

However, in implementing an imbalanced learning approach, Wang et al. [30] presented an imbalance-XGBoost model to classify PD after employing the PDC dataset [30]. Their generated approach received an accuracy of 93% while classifying PD. While Polat and Nour [31] employed a One aGainst All (OGA)-based data sampling approach that has been employed to partition the PD dataset with acoustic features into five similar segments after employing three various classification models to distinguish these all data partitions [31]. The OGA approach with the weighted KNN approach obtained an accuracy value of 89.46% to classify PD [31].

Apart from the several above-said methodologies based on voice dataset, Maachi et al. [32] introduced a 1D convolutional neural network for developing a Deep Neural Network classifier for predicting Parkinson's disease after employing a public database: Gait in Parkinson's disease collected by Physionet [32]. In contrast, Adams [33] proposed a methodology employed to discriminate the subjects' disease status by unifying several keystroke features analyzed by an ensemble of ML algorithms [33]. When implemented into two separate participant groups, their method could happily distinguish between early-PD patients and healthy people [33]. Tunc et al. [34] applied a wrapper-based Boruta feature selection algorithm with an extreme gradient boosting algorithm to estimate PD severity after employing a PD dataset with speech features [34]. Their applied approach obtained the lowest mean absolute error of 3.87 while estimating PD severity [34]. Karan et al. [35] introduced time–frequency attributes to model discontinuities and sudden alterations in the voice signal because of Parkinson's disease [35]. Their progressive approach comprises feature extraction, classification, time–frequency matrix (TFM) decomposition using non-negative matrix factorization, and TFM design [35]. Their implemented approach was conducted on the PC-GITA dataset and produced the mean accuracy values in vowels and words from 92 to 97%, respectively [35]. De Souza et al. [36] introduced a fuzzy optimum path forest for intelligently PD classification after employing a dataset comprised of attributes obtained from hand-drawn images [36]. Their progressive approach relied on the graph-based structure method and fuzzy logic [36]. Karan and Sahu [37]

developed a merged strategy of Hilbert spectrum analysis and variational mode decomposition to examine the voice tremor of sufferers with Parkinson's disease after employing the PC-GITA dataset [37]. Their generated approach reported an accuracy value of 82% to classify PD. Quan et al. [38] introduced a bidirectional long-short term memory approach to obtain time-series dynamic attributes of a speech signal to detect PD [38]. They employed a mixed-gender database which consists of 45 subjects. Such a database was collected from the GYENNO SCIENCE PD Research Center [38]. Their generated approach received an accuracy of 75.56% while detecting PD.

After analyzing the earlier stated literature, it has been revealed that most of the prior research works involved in generating either feature selection enabled method included in [4–6, 8, 11, 12, 14–16, 18, 26] or optimization-based feature selection approach included in [19, 24] or implementing feature extraction methodologies included in [6–8, 13, 23, 25] or employing hyperparameter optimization methodologies included in [20, 29], or applying imbalanced learning approaches included in [30, 31] or implementing deep learning technique included in [20]. Most of them did not involve developing the hybrid methodology such as any classifier's hyperparameter optimization methodology with the two-stage dimensionality reduction approach, enhancing ML algorithms' performance, diagnosing the PD accurately, and generating the best outcome. Thus, this paper fills this research gap.

In this case, this study introduces a genetic algorithm-based hyperparameter optimization methodology to optimize hyperparameters for LGBM classifier, and adaptive two-stage MI and AE-based dimension reduction approach to enhance machine learning models' performance and diagnose PD and provide the best accurate results than the relevant ML algorithms. Literature included in [17, 21, 22, 27] that only relates to our proposed research. The researchers included in [17, 21, 22, 27] either implementing a variational autoencoder and relief or fisher score-based feature selection approaches with multi-kernel SVM or applying adaptive GWO with sparse autoencoder-based dimension reduction approach or employing relief-based feature selection with bacterial foraging optimization technique to tune hyperparameter for SVM classifier, or applying LDA to reduce dimensions with GA for parameters tuning of the neural network to predict PD and enhance the performance of their suggested systems. However, in this case, those researchers included in [22, 27] utilized either a filter-based feature selection approach: relief or a supervised feature extraction methodology: LDA, rather than utilizing the two-stage dimension reduction approach as our proposed approach, including filter-based feature selection approach: mutual information with the unsupervised feature extraction approach such as autoencoder. In contrast, the researchers included in [17, 21] either employed filter-enabled feature selection approach: fisher score and relief with unsupervised feature extraction approach: variational autoencoder or implemented optimization-based feature selection approach: adaptive GWO with unsupervised feature extraction methodology: sparse autoencoder to predict PD. In this regard, those researchers included in [17, 21] did not involve optimizing their utilized classifiers' hyperparameters to enhance their developed model's performance to predict PD. In this case, this proposed research employs a genetic algorithm-based hyperparameter optimization methodology to tune the hyperparameters of the LGBM algorithm and enhance the capability of our proposed approach. Thus, this proposed research overcomes these drawbacks and utilizes a scalable and robust unsupervised feature extraction methodology: Autoencoder along with the mutual information-based filter enabled feature selection method with genetically optimized LGBM model to form our suggested MI-AE-GOLGBM methodology and provides a solution for diagnosing PD with the most reliable and adaptive precise outcomes. However, our suggested study can be widely employed as its scope is unlimited and unrestricted in any manner. Thus, such above-specified findings differ from the finding of their research works.

3 Preliminaries and research methodology

3.1 Mutual information

Mutual information (MI) can be implemented to assess any random dependence among arbitrary variables. The MI among two arbitrary variables, X and Y, measures the amount of knowledge on Y furnished by X. If X and Y are not dependent, their MI becomes zero. The MI of two arbitrary variables, X and Y, is determined in below

$$\begin{aligned} \mathrm{MI}(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X;Y) \\ &= \int_x \int_x P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} dx\, dy \end{aligned} \tag{1}$$

where $H(X), H(Y), H(X|Y)$, $H(Y|X)$ and $H(X;Y)$ express as the entropy, conditional entropy, and joint entropy of the discrete arbitrary variables X and Y.

Hence, it is solely needed for calculating $P_{X,Y}(x,y)$ in order to measure the mutual information among two arbitrary variables.

3.2 Overview of autoencoder

In autoencoder (AE), the training examples with reduced generated features, $X = [x_1, x_2, \ldots, x_m]^S$, where m = #input features, AE's objectives are to extract the best essential features, F, from the input feature set [39].

However, an autoencoder is utilized to minimize reconstruction error by applying the binary cross-entropy loss explained below.

$$J_{\mathrm{AE}}\left(X, X^D\right) = -\sum_{i}^{N}\left[X_i \log X_i^D + (1 - X_i)\log\left(1 - X_i^D\right)\right] \tag{2}$$

where $J_{\mathrm{AE}}(X, X^D)$, X^D represent the reconstruction error and overall function of the autoencoder, and N = number of samples.

Throughout the autoencoder approach, a weight decay (ψ) technique is added with the reconstruction error of the AE to regenerate the following reconstruction error:

$$J_{\mathrm{AE}}\left(X, X^D\right) = -\sum_{i}^{N}\left[X_i \log X_i^D + (1 - X_i)\log\left(1 - X_i^D\right)\right] + \frac{\aleph}{2}\sum_{i=1}^{2} \parallel W^{(i)} \parallel_2^2 \tag{3}$$

where $\aleph$ signifies the penalty parameter and W signifies a matrix of the weight provided to each layer inputs (either input or hidden layer).

The gradient of the reconstruction error of AE concerning the parameters is estimated using backpropagation, which is employed to update the parameters through a stochastic optimization algorithm: Adam optimization that implements a progressive settlement of several parameters by estimating the gradient first-order moment estimate M_s and second-order moment estimate V_s which are presented in the following Eqs. (4–6) [39].

$$M_s = \alpha_1 M_{s-1} + [(1 - \alpha_1)G_s] \tag{4}$$

$$V_s = \alpha_2 V_{s-1} + \left[(1 - \alpha_2)\left(G_s^2\right)\right] \tag{5}$$

$$\phi_{s+1} = \phi_s - M_s \frac{\delta}{\sqrt{V_s} + \varepsilon} \tag{6}$$

where α_1 and α_2 exhibit the first-order and second-order exponential damping decrement, respectively, G_s denotes a gradient of the parameters at timestep s in the reconstruction error $J_{\mathrm{AE}}(X, X^D)$, δ denotes the update step size, and ε holds a small fixed number for counteracting the denominator from zero [39].

3.3 Overview of light gradient boosting machine algorithm

This segment explains that the LightGBM algorithm applies for classification purposes. Microsoft researchers created the LightGBM algorithm in which is open-source and can be employed by anyone. The LightGBM algorithm is exhibited in Fig. 1.

This suggested system applies the LightGBM algorithm that can bundle mutually exclusive features into a single feature termed the Exclusive feature bundle; the same features histograms are built from the feature bundles by the feature scanning algorithm. This algorithm's time complexity was estimated in the following: $O(n*p)$ where n = number of examples available in the earlier mentioned dataset, and p = number of bundles, where $p <$ #features, where #features imply as the numbers of features in the dataset.

3.4 Proposed intelligent diagnostic system

The proposed expert diagnostic system employs a hybrid methodology: two-stage mutual information and autoencoder-based dimensionality reduction method with genetically optimized LightGBM (MI-AE-GOLGBM) method for generating the most informative newly created features from the given features and also generating the most optimal hyperparameters for the LightGBM algorithm after intelligently optimization process, and thus, to enhance the proposed system's accuracy. The architecture of this suggested system is exhibited in Fig. 2. The above-said proposed methodology is elaborated in the succeeding subsections.

3.4.1 Proposed two-stage mutual information and autoencoder-based dimensionality reduction methodology

The irrelevant occurrence of attributes is one of the main reasons for the overfitting of a classification model. So dimension reduction should be conducted before initiating training for the classifier. The cause for doing the dimension reduction technique is that to improve the classifier's performance, directing to a quicker and more cost-efficient approach. Thus, this proposed system employs a two-stage dimensionality reduction approach, which comprises two different dimension reduction techniques: mutual information (MI) and the autoencoder (AE). MI is employed as a filter-based feature selection approach, and AE is employed as a deep neural-based feature extraction approach. Both methodologies are employed to reduce the input PD dataset's dimension, lessen the overfitting,

LightGBM algorithm:

Input:

(a) Training data as input, T = $\{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)\}$
For each iteration I, $x_i \in X$; $X \subseteq R$; and $y_i \in \{-1, +1\}$;
(b) Iteration: I; Model: M;
(c) Loss functions: $L(y, f(x)) = L(y, y^{HAT})^2$ where $L(y, y^{HAT})^2 \in$ quadratic loss;
(d) Iteration: P;
(e) Large gradient data sampling ratio: a; Small gradient data sampling ratio: b;

1. Many features are mutually exclusive i.e., they never take non zero values simultaneously, therefore Exclusive Feature Bundling (EFB) techniques can be used to produce a single feature and to increase speed for training of this algorithm.
2. To determine which features should be bundled: for each iteration q,
   $Bundling_q$.add($Features_p$); where bundling is the array.
   Set of mutually exclusive features are formed into a single feature namely an exclusive feature bundle.
3. Define $f_0(x) = argmin_d \sum_{j=1}^{n} L(y_j, d)$, where $f_0(x)$ = exclusive feature bundle.
4. For each iteration p from 1 to P:
   For i^{th} examples from 1 to n:
   Prediction = models.predict(T);
   Absolute value of gradient,

$$G(x_i, y_i) = Loss(T, Prediction) = \left|\frac{\partial L(y_i, s)}{\partial s}\right|_{s=f_{p-1}(x_i)} = \Delta L(y_i, s)$$

   Resample dataset using Gradient-based One Side Sampling (GOSS) approach:
   $topN = a \times length(T)$;
   $randN = b \times length(T)$;
   $sorted = GetSortedIndices(abs(G))$;
   $TS = sorted[1:topN]$;
   $RS = RandomPick(sorted[topN:lenght(T)], randN)$;
   $whereTS = Topset\ and\ RS = randomset$;
   $New\_Dataset\ T' = TS + RS$;
   Variance gain $V_J(d)$ is estimated as follows:

$$V_J(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in TS_l} G_i + \left(\frac{1-a}{b}\right)\sum_{x_i \in RS_l} G_i\right)^2}{n_l^J(d)} + \frac{\left(\sum_{x_i \in TS_r} G_i + \left(\frac{1-a}{b}\right)\sum_{x_i \in RS_r} G_i\right)^2}{n_r^J(d)}\right)$$

   where $TS_l = \{x_i \in TS: x_{ij} \le d\}$, $TS_r = \{x_i \in TS: x_{ij} > d\}$, $RS_l = \{x_i \in RS: x_{ij} \le d\}$, $RS_r = \{x_i \in RS: x_{ij} > d\}$;
   Generate a newly formed decision tree F on the dataset T'.
   $F = L(x_i, G_i)$;
   update $f_p = f_p + F$;
   Return F^p

Fig. 1 Light gradient boosting machine (LGBM) algorithm

generate the most significant features that characterize the PD, and improve the proposed system's precision value.

In implementing a filter-based feature selection technique in this suggested system, MI can quickly choose the related, informative features (Z) from the input features (X). The input features maximize the MI between the selected feature subset (Z) and the class variable (Y). Mutual information aims to maximize the pertinence between the input feature set (X) and the target variable (Y) and try to remove unrelated features and diminish the overfitting by selecting the feature subset (Z). Hence, such informative features are passed to the AE for further processing for dimension reduction.

In implementing a feature extraction approach in this suggested system, AE acquires such reduced selected features from MI as an input and gives the output as the newly generated significant features. The training examples with reduced selected features obtained from MI, $Z = \{z_1, z_2, \ldots, z_m\}$, are passed to the AE for feature extraction. The AE aims to generate new significant features from such reduced informative features Z by lessening the unrelated input features, diminishing the overfitting and redundancy from input features Z. Therefore, such newly generated features will improve the performance of the intended method.

A two-stage dimensionality reduction approach is implemented in this proposed system to eliminate the

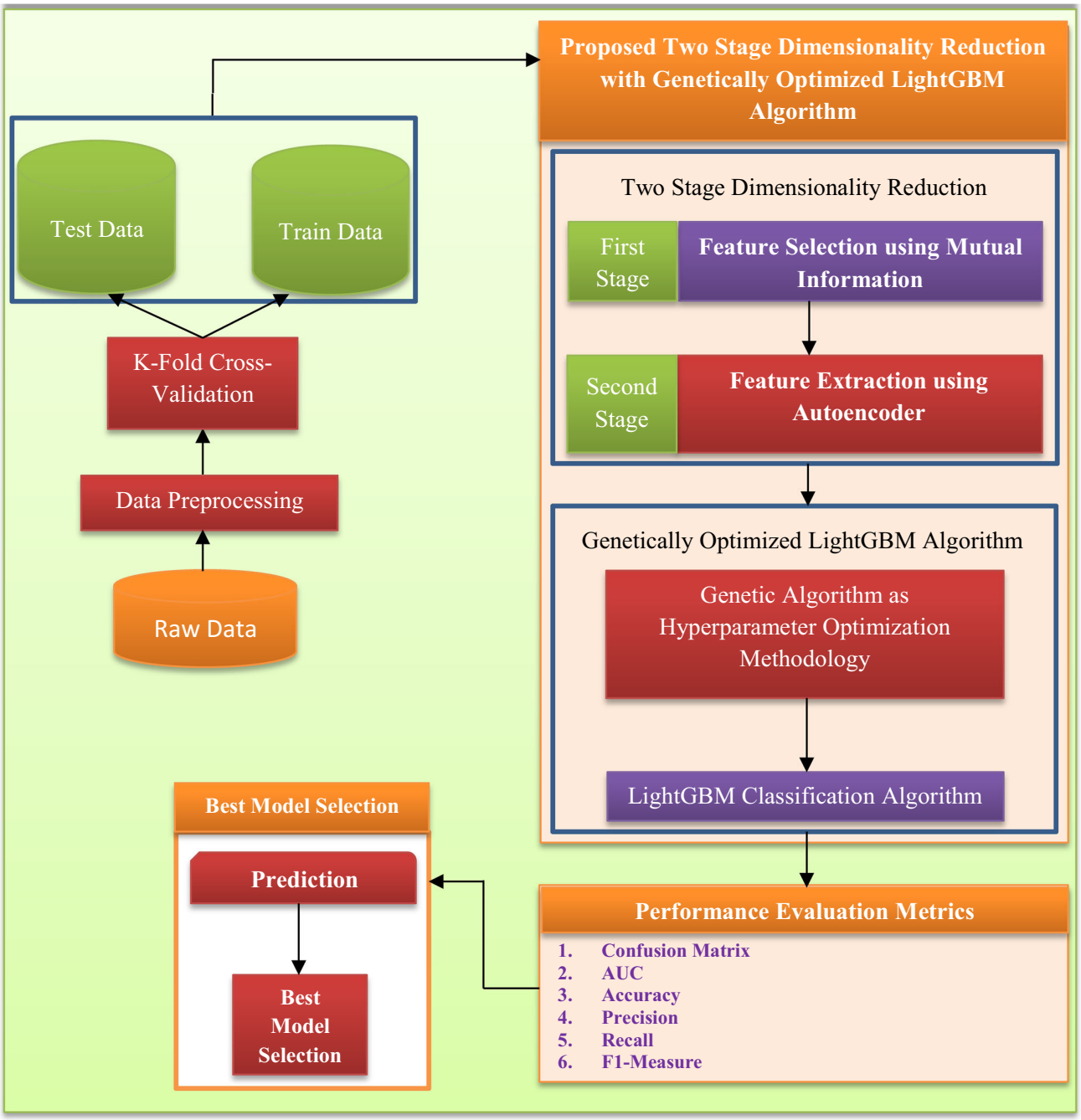

Fig. 2 The architecture of the suggested expert diagnostic system using MI-AE-GOLGBM algorithm

overfitting, irrelevant attributes and diminish the redundancy from the input features X; $X = \{x_1, x_2, \ldots, x_n\}$, where n represents the number of attributes available of the input data and choose the informative attributes subset Z; $Z = \{z_1, z_2, \ldots, z_m\}$, where m exhibits the number of attributes chosen from X by utilizing the MI [40]. Hence, $Z \subset X$. Then such features Z are passed to the AE for feature extraction, and AE produces the most significant newly generated features F; $F = \{f_1, f_2, \ldots, f_p\}$, where p represents the number of newly generated features from Z after utilizing the AE and improves the suggested method's accuracy. Figure 3 exhibits the sequential steps of the two-stage MI and AE-based dimensionality reduction methodology.

This proposed system utilizes a PD dataset supplied as an input to the MI as the first dimensionality reduction

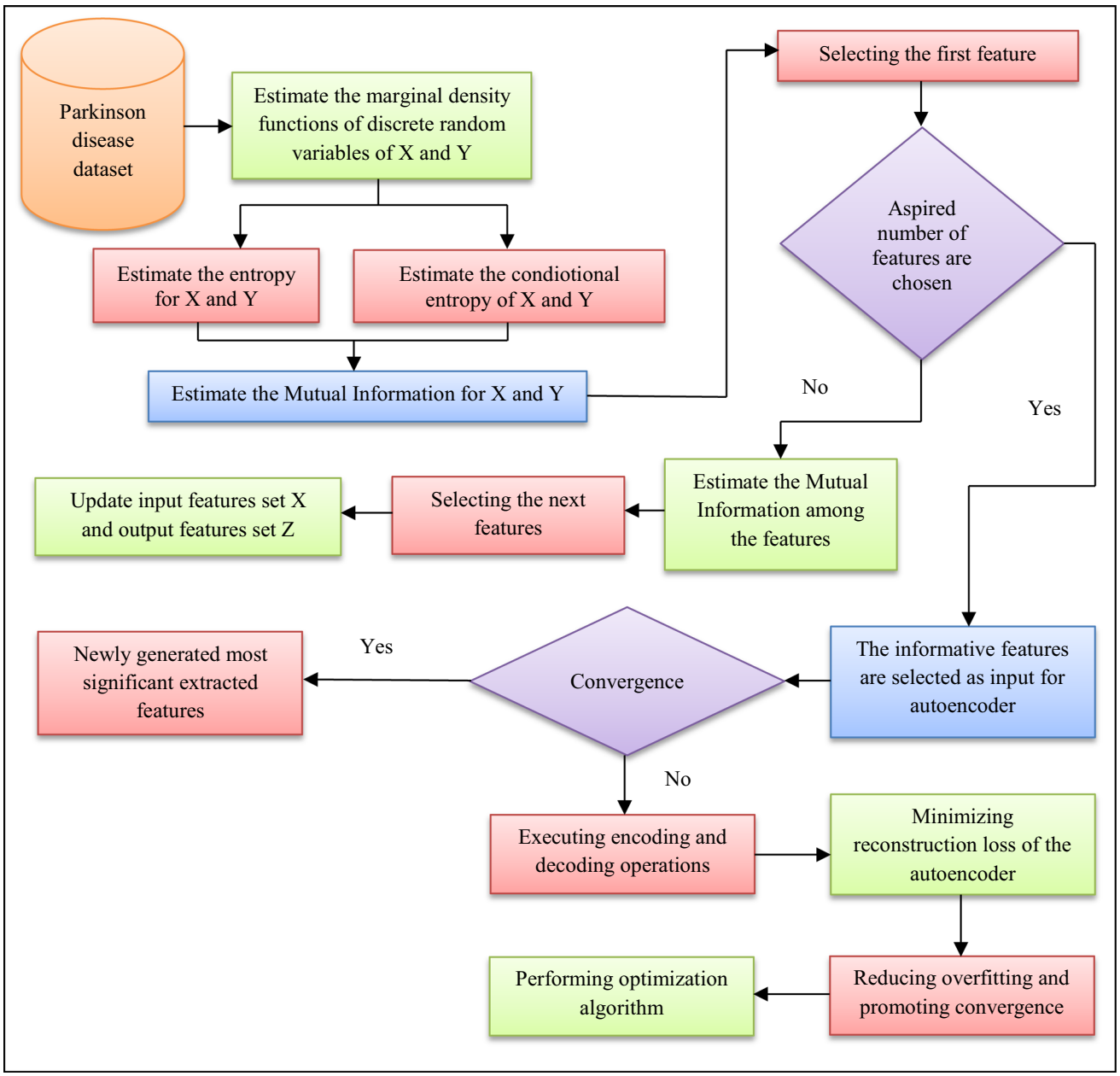

Fig. 3 Sequential steps of two-stage dimensionality reduction approach

stage. Then, MI selects the relevant, informative features from the input features X by diminishing irrelevant features and lessen the overfitting from input features. After then, it passes such selected features Z to the AE to most accurately remove the unrelated features, overfitting, and redundancy from the input features Z. Hence, in the second stage, the AE-based dimensionality reduction approach is implemented in which AE iterations will be performed until the convergence is satisfied. In the end, AE generates the most significant newly generated features F by utilizing such informative features Z and enhancing the performance of the suggested system.

Implementing the MI and AE approaches as a two-stage dimensionality reduction method requires many essential steps, as shown in Fig. 3. In the first stage, this proposed system implements an MI-based feature selection approach, in which it firstly estimates marginal density functions of discrete arbitrary variables X and Y. Then estimate entropy H for the variables X and Y, such as $H(x_i), H(Y)$. After then, estimating conditional entropy $H(Y; x_i)$ and $H(x_i; Y)$, and $MI(Y; x_i)$ where each feature $x_i \in X$. Secondly, selecting the first feature by recognizing the feature x_i that maximizes $MI(Y; x_i)$. Then, assigning the value of X and Z. Thirdly, estimating the MI among the

features. Later, selecting the next feature by determining the feature $x_i \in X$ as the one which maximizes $\mathrm{MI}(Y; x_i) - (\beta) \sum_{x_z} \int_Z \mathrm{MI}(x_z; x_i)$; where β expresses a user-specified parameter. Then, update the value of X and Z. Repeat steps until the aspired number of features are chosen. Finally, assign the selected features to Z as the informative features for further use for the second dimension reduction stage. In the second stage, this automated system implements the AE-based feature extraction method, in which it first assigns the most informative input features Z to X. Then, estimating the encoding and decoding operations based on the AE approach. Secondly, minimizing the reconstruction loss of the AE approach by applying the binary cross-entropy loss reduces overfitting and promotes convergence through weight decay. Thirdly, performing an optimization algorithm relies on tuning the parameter of AE. These operations are continued until the convergence is satisfied. Hence, the AE produces the newly generated most significant features F from the given reduced selected features Z.

To clarify the idea based on Fig. 3, MI and AE are utilized as a two-stage dimensionality reduction approach in this proposed system. Assume that there are seven input features available in PD's dataset that means, $n = 7$; $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$. Such input features X are passed to MI due to reducing the input dataset D's irrelevant features and lessening overfitting. After performing MI on PD's input dataset $\{X, y\}$, the five related informative features are selected, which means $m = 5$. Hence $Z = \{z_1, z_2, z_3, z_4, z_5\} = \{x_1, x_2, x_5, x_6, x_7\}$. Finally, such informative features are then passed to AE to remove unrelated features and overfitting accurately, and additionally to lessen the redundancy of such informative features Z and generate newly generated the most significant extracted features F. Hence, after performing AE on Z, the four new significant features are generated: $p = 4$. Hence $F = \{f_1, f_2, f_3, f_4\}$. Such newly generated extracted features are then passed to the LightGBM algorithm to enhance the suggested system's performance using the proposed methodology. Algorithm 1 describes the two-stage dimension reduction technique briefly in the following.

Algorithm 1: Two-Stage Dimension Reduction

Estimate_MI(Y; X) /* Method for stage 1 dimension reduction */ {
Input: Z = {}; /* empty set*/
1) for i = 1 to n:
Estimating $H(x_i), H(Y), H(Y; x_i)$ and $MI(Y; x_i)$ where each feature $x_i \in X$;
End for
2) Choosing the first feature: Recognize the feature x_i that maximizes $MI(Y; x_i)$;
3) Set $X = X \backslash \{x_i\}$ and $Z = \{x_i\}$;
4) Greedy selection: iterate unto the number of most informative features |Z| = n, are elected:
a) Estimating the MI among the features:
For all pairs of features (x_i, x_z) with $x_i \epsilon X$ and $x_z \epsilon Z$, and x_i, x_z express as the single input feature;
If it is not still estimated: Estimate_$MI(x_i; x_z)$;
b) Following feature selection: determine the feature $x_i \in X$ as the one which maximizes
$MI(Y; x_i) - (\beta) \sum_{x_z \epsilon Z} MI(x_z; x_i)$; where β expresses a user-specified parameter;
Set $X = X \backslash \{x_i\}$ and $Z = Z \cup \{x_i\}$ as chosen features;
return Z; }
Compute_AE(Z) /* Method for stage 2 dimension reduction */ {
Input:
Unlabeled input data, X = Z;
Define the number of hidden layers, h, weight decay regularization term, weight matrix: $W^{(1)}$ and $W^{(2)}$;
Specify encoding and decoding activation function, $\sigma_{\mathrm{ENCODER}}$, $\sigma_{\mathrm{DECODER}}$;
Bias values: b_{hidden}, and b_{out} ;
Procedure:
1) while not converged:
for each i^{th} example in the input dataset X, where i = 1 to n:
a) Compute optimal loss function $J_{AE}(\mathrm{X}, \mathrm{X}^{\mathrm{D}})$ with weight decay of the autoencoder approach using Equation 3;
b) Perform optimization algorithm relies on tuning the parameter of AE elaborated in Equations 4-6;
End for loop
End while loop
2) Generate the most significant newly generated features F from the given Z;
return F; }

3.4.2 Proposed genetically optimized LightGBM algorithm

Hyperparameters of any classifier is one of the main reasons to increase the performance of any machine learning model. So, optimum parameters help improve the correctness of such a machine learning model. Hence, forming the optimum parameters should be conducted before employing the classifier. Therefore, this proposed diagnostic system utilizes a novel method: genetically optimized LightGBM (GOLGBM) algorithm, in which the GOLGBM algorithm comprises two methods: genetic algorithm (GA) and the LightGBM classification algorithm (LGBM). GA is employed as a hyperparameters optimization methodology to automatically tune the hyperparameters of the LGBM algorithm and generate the most optimal parameters to enhance the performance of the LightGBM algorithm.

In contrast, the LGBM algorithm employs such optimal hyperparameters and newly generated most significant extracted features to classify the healthy people and the PD sufferers and provide the most accurate outcomes. Generally, GA is the most notable meta-heuristic methodology based on the evolutionary concept [40]. The individuals with the most desirable survival ability and suitability to the environment are further likely to endure and move on their abilities to forthcoming generations. The succeeding generations will further derive their parents' characteristics and may include more suitable and unsuitable chromosomes. The more suitable chromosomes will be further possibly to endure and have additionally capable offspring. In contrast, the unsuitable chromosomes will slowly vanish. After completing numerous generations, the chromosomes with the fittest suitability will be identified as the global optimum.

A genetic algorithm is a meta-heuristic hyperparameter optimization algorithm in which it chooses the most fitted chromosomes or individuals from the population. It generates the offspring for the next generation after utilizing three operations based on GA: selection, cross-over, and mutation. GA performs its operation as follows. In the initialization phase, it first defines the population, which contains the set of chromosomes. Each chromosome has a set of genes in which a gene is a binary bit.

Regarding this matter, a binary bit 1 denotes the selected gene, and 0 denotes the gene removed from its respective chromosome. Next, evaluating the fitness function of each chromosome from the population and generate its survival point. Then, select the top two most fitted chromosomes from the population by utilizing the "Roulette wheel selection method." Then, in the cross-over phase, selecting a random cross-over point and the tails of both the chromosomes are swapped to produce new offsprings, and then mutation operation is allowed to change in children's genes, making them different parents. Next, the generated offspring are validated using the fitness function, and if it is considered a fitter, it will replace fewer fit chromosomes from the population. These operations are performed until the number of predefined generations has reached its limit. Finally, the most desirable chromosomes of the population are therefore identified and returned to the optimization problem. The time complexity to generate the most desirable chromosomes as the most reliable optimum hyperparameters by the GA is $O(n^2)$. Algorithm 2 of the GOLGBM methodology is described in the following.

The central concept of the GA encrypts the hyperparameters to be tuned into a range of individuals. During the genetic heritage of such individuals, genetic selection happens through the fitness function. In this way, the most suitable individuals are the higher possibility to endure; hence, the whole population's process is achieved after various generations. The reasonably most comprehensive chromosome may be the optimal global answer to the question. The user could specify the GA parameters: estimator, params, scoring, population_size, gene_mutation_prob, gene_crossover_prob, tournament_size, and generations_number. Although, in this proposed system, we have employed the following parameter settings for the GA in the following Table 1. In this proposed research, the GA is adopted to automatically tune the LGBM algorithm's hyperparameters and enhance the LGBM algorithm's performance. Ranges of hyperparameters of the LightGBM algorithm for optimization in this proposed research, exhibited in Table 2.

Algorithm 2: GOLGBM methodology

Compute_Optimal_Hyperparameters_Using_GA (Hyperparameters of LightGBM and their values)

{

Input: Initialize the GA parameters: number of generations, population size, cross-over probability, mutation probability, and size of the tournament, define the estimator algorithm as LightGBM classification algorithm;

Procedure:

1) Assign initial population as hyperparameters of LightGBM classification algorithm.
2) Randomly initialized population, individuals, and genes in which the population expresses itself as an entire search space; individuals signify hyperparameters, and genes exhibit as values of hyperparameters.
3) While termination condition is unsatisfied:
   (a) Assess each chromosome's performance in the current generation by estimating the fitness function as specified as follows: F(GA) = max PEV/100, s.t. PEV ∈ value of any performance evaluation metric (accuracy, precision, recall, f1, AUC). Then chromosomes are ranked based on their fitness values.
   (b) Optimizing the hyperparameters:
   - Genes of each individual are encrypted in the population.
   - Estimate each generation's fitness value of the population.
   - Perform the continuation of the most suitable population.
   - If the performance of the population meets the highest number of genetics:
     Generating the best individuals;
   - Else:
     Performing selection operation based on each chromosome's rank, where a subset of these individuals $R_G \subset$ population; The highest-ranked chromosomes that have a more substantial possibility of regenerating to form the offspring group R'_G, are selected;
   - Performing cross-over and mutation operations structures in R_G and forming offspring group R'_G;
   - Select each individual from R'_G for generating a new population;

   End while loop
4) Max_parameter = Generate the most desirable chromosomes as the most optimum hyperparameters;
   return Max_parameter;
   }

3.4.3 Proposed two-stage mutual information and autoencoder-based dimensionality reduction method with genetically optimized LightGBM algorithm

This proposed diagnostic system introduces a hybrid method: two-stage mutual information and autoencoder-based dimensionality reduction methodology with genetically optimized LightGBM algorithm (MI-AE-GOLOGBM), which is employed to detect PD's early symptoms and predict the most precise outcomes. This suggested method comprises two dimension reduction approaches, mutual information (MI) and autoencoder (AI); this proposed research extracts the most significant features from input features of the dataset and enhances the proposed system's accuracy. Additionally, this suggested system introduces a novel genetically optimized LightGBM (GOLGBM) algorithm that comprises two methodologies: genetic algorithm (GA) and LightGBM algorithm, in which GA is employed to tune the hyperparameters of the LightGBM algorithm intelligently and generates the best optimal hyperparameters for further utilization for the LightGBM classification algorithm. The LightGBM algorithm utilizes such generated significant features and optimal hyperparameters to classify the healthy persons and the PD patients and provides the most accurate and reliable outcomes. The process flow of the suggested MI-AE-GOLGBM method is elaborated in Fig. 4. At the same time, the algorithm of this proposed approach is represented in algorithm 3.

Algorithm 3: Proposed MI-AE-GOLGBM methodology

Input: Input dataset D, where input feature set X = {x_1, x_2, ..., x_n} and target variable, Y = {0, 1}; and define hyperparameters of the LightGBM algorithm and their values.

Procedure:

1. F = Call Two-Stage Dimension Reduction using MI and AE approaches exhibited in Algorithm 1;
2. Optimal_hyperparameter = Call Genetically Optimized LightGBM algorithm described in Algorithm 2.
3. Utilize_LightGBM_Classification_Algorithm (F, **y,** Optimal_hyperparameter)

```
{
Clf = LGBMClassifier(Optimal_hyperparameter).fit(F, y); /* Fitness function of LightGBM classification algorithm */
Pred = Clf.predict(F_test_data); /* LightGBM algorithm to classify the healthy controls and the PD sufferers */
return Pred; /*generate the best prediction */
}
```

Table 1 List of parameters with their value settings of the GA approach

SI no	List of parameter settings for the GA approach	Parameter value
1	Estimator	'LGBMClassifier'
2	Params	Each hyperparameter values of the LGBM model
3	Population_size	1000
4	Gene_mutation_prob	0.05
5	Gene_crossover_prob	0.85
6	Tournament_size	3
7	Generations_number	200
8	Cross-validation	10

Table 2 List of hyperparameters of the LightGBM algorithm with a description and their values utilizes in the proposed research

Index	List of parameters	Description	Range of value settings for tuning the hyperparameters
1	Max_depth	Maximum depth for a tree model	(− 1, 300)
2	Num_leaves	The maximum number of leaves in one tree, where num_leaves = 2^ max_depth	(5 to 600)
3	Min_data_in_leaf	Minimum number of data in one leaf to prevent overfitting	(10 to 500)
4	n_estimators	Number of estimators	(100 to 1000)
5	Bagging_fraction	Bagging fraction utilized for faster speed	(0.1 to 1.0)
6	Feature_fraction	feature sub-sampling utilized for quicker speed	(0.1 to 1.0)
7	Max_bin	Maximum number of bins	(200 to 300)
8	Learning_rate	Learning rate	(0.01 to 0.99)
9	Min_child_weight	Minimum child weight	(0.00001 to 1.0)
10	Reg_alpha	L1 regularization	(1.0 to 3.0)
11	Reg_lambda	L2 regularization	(1.0 to 2.0)
12	Colsample_bytree	Column sample by tree	(0.8 to 1.0)
13	Feature_fraction_seed	Random seed for feature sub-sampling	(1200 to 1500)
14	Boosting_type	Boosting type	(gbdt, dart, and goss)

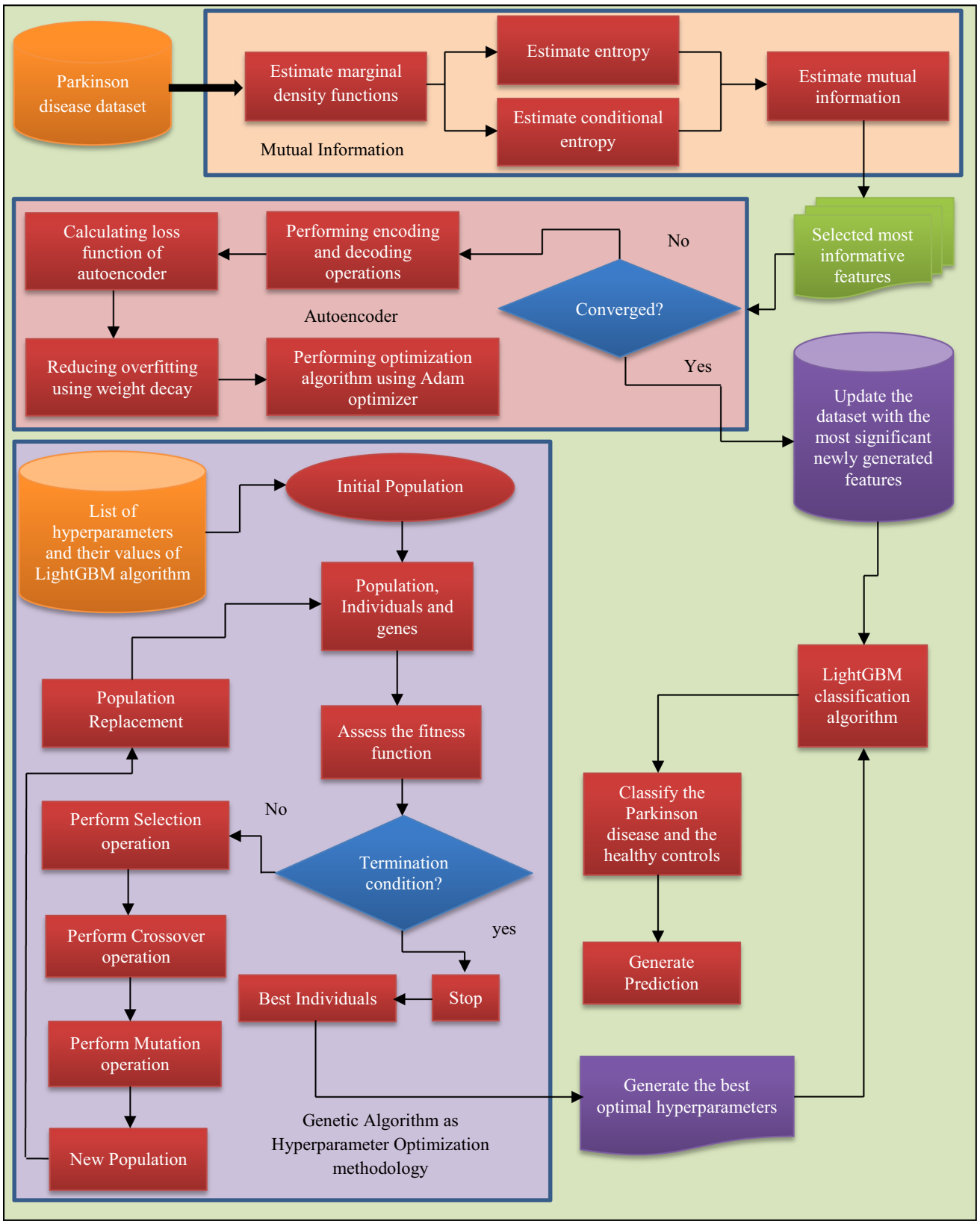

Fig. 4 Process flow of the proposed MI-AE-GOLGBM methodology

Table 3 Summary of Examined Datasets

List of Datasets	Dataset name	Number of Total Examples	Number of Available Features	PD Diagnosis Techniques	References
Dataset 1	PDC dataset	756	753	Voice	[41]
Dataset 2	Parkinson's Data set	197	22	Voice	[44]
Dataset 3	Parkinson Dataset with replicated acoustic features	240	45	Voice	[45]
Dataset 4	PS dataset	1040	25	Voice	[46]

3.5 Dataset and data preprocessing techniques

This proposed research employs the proposed MI-AE-GOLGBM methodology on publicly available real-world datasets based on Parkinson's disease (PD). There are four different real-world datasets considered for the experiments of this proposed research through which the achievement of this proposed methodology is appropriately assessed and verified. In this proposed research, datasets 1, 2, and 3 are utilized to assess the performance, and the remaining data set 4 is utilized for verifying the performance of the suggested methodology.

The suggested research employs the proposed methodology on the PDC dataset [41], which comprises 756 numbers instances and 754 attributes, out of which 188 instances belong to patients with PD with ages varying from 33 to 87, in which 107 individuals are men, 81 individuals are women [41]. The remaining 580 numbers of instances belong to healthy persons. The dataset was gathered from 188 sufferers at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University [10, 42, 43]. The control group comprises 64 healthful people, out of which 23 individuals are men, and the remaining 41 are women with ages ranging between 41 and 82 [41]. This dataset 1 comprises 753 attributes, which are utilized as features. Experiments with the PD classification dataset have focused on generally striving to identify PD existence value: 1 from nonexistence value: 0.

This proposed research employs another dataset: the Parkinsons dataset [44], which comprises 197 instances and 23 attributes designed through Max Little [44]. The data were gathered from the University of Oxford, collaborating with the National Centre for Voice and Speech, Denver, Colorado [10, 42–44]. Dataset 2 is contained a series of biomedical voice determinations from 31 individuals, out of which 23 individuals belong to Parkinson's disease [42, 44]. Each attribute in the table is a specific voice measure, and each row communicates to one of 195 voice records from these individuals [42]. The data's principal objective is to distinguish non-PD people from individuals with PD, in which the target variable is set to 1

Table 4 The performance achieved by the suggested MI-AE-GOLGBM method for datasets 1, 2, and 3 after utilizing a tenfold cross-validation procedure

Dataset	Fold number	AUC	PR	REC
Dataset 1	Fold 1	0.9679	0.9138	0.9464
	Fold 2	0.9589	0.8730	0.9821
	Fold 3	0.9252	0.8485	0.9825
	Fold 4	0.9012	0.8485	0.9825
	Fold 5	0.9695	0.9048	1.0000
	Fold 6	0.9086	0.8871	0.9649
	Fold 7	0.8769	0.8281	0.9464
	Fold 8	0.9530	0.8594	0.9821
	Fold 9	0.9474	0.9153	0.9643
	Fold 10	0.9568	0.8871	0.9821
	Mean	0.9363	0.8765	0.9733
Dataset 2	Fold 1	0.9867	0.9333	0.9333
	Fold 2	0.9867	0.9375	1.0000
	Fold 3	0.9763	0.9375	1.0000
	Fold 4	1.0000	0.8824	1.0000
	Fold 5	1.0000	1.0000	1.0000
	Fold 6	1.0000	1.0000	1.0000
	Fold 7	0.9000	0.9286	0.9286
	Fold 8	0.9143	0.8667	0.9286
	Fold 9	0.9000	0.9375	1.0000
	Fold 10	0.9000	0.9333	0.9333
	Mean	0.9565	0.9357	0.9724
Dataset 3	Fold 1	0.6840	1.0000	0.3333
	Fold 2	0.9062	1.0000	0.6667
	Fold 3	0.8194	0.6471	0.9167
	Fold 4	0.9062	0.7857	0.9167
	Fold 5	0.8333	1.0000	0.7500
	Fold 6	0.8750	0.9000	0.7500
	Fold 7	0.9688	0.9000	0.7500
	Fold 8	1.0000	1.0000	1.0000
	Fold 9	0.9792	0.8571	1.0000
	Fold 10	0.9826	1.0000	0.8333
	Mean	0.8952	0.9090	0.7917

Table 5 Comparing the performances among the suggested methodology and the irrelevant ML algorithms for the datasets 1, 2 and 3

Dataset	Model	AUC	ACC	PR	REC	F1
Dataset 1	RF	0.8510	0.8399	0.8579	0.9432	0.8978
	KNN	0.6275	0.7261	0.7777	0.8864	0.8282
	SVM	0.6282	0.7223	0.7451	0.9539	0.8362
	LR	0.6914	0.7592	0.7717	0.9645	0.8568
	DT	0.6739	0.7285	0.8443	0.7869	0.8020
	Proposed MI-AE-GOLGBM	0.9363	0.8769	0.8765	0.9733	0.9221
Dataset 2	RF	0.9020	0.8342	0.8756	0.9262	0.8956
	KNN	0.7480	0.7618	0.8267	0.8781	0.8475
	SVM	0.6929	0.8147	0.8130	0.9933	0.8916
	LR	0.8367	0.8184	0.8700	0.9133	0.8840
	DT	0.7314	0.7776	0.8534	0.8562	0.8519
	Proposed MI-AE-GOLGBM	0.9565	0.9279	0.9357	0.9724	0.9533
Dataset 3	RF	0.8471	0.7708	0.8196	0.7250	0.7467
	KNN	0.7850	0.7208	0.7294	0.7083	0.7099
	SVM	0.8252	0.7167	0.7409	0.6500	0.6795
	LR	0.7693	0.7250	0.7545	0.6750	0.6907
	DT	0.6750	0.7042	0.7187	0.6833	0.6740
	Proposed MI-AE-GOLGBM	0.8952	0.8417	0.9090	0.7917	0.8230

Table 6 Comparing the performances among the suggested methodology and the relevant ML algorithms for datasets 1, 2, and 3

Dataset	Model	AUC	ACC	PR	REC	F1
Dataset 1	LGBM	0.8958	0.8544	0.8693	0.9465	0.9048
	XGB	0.8588	0.8532	0.8651	0.9537	0.9062
	CB	0.8873	0.8518	0.8630	0.9536	0.9038
	Proposed MI-AE-GOLGBM	0.9363	0.8769	0.8765	0.9733	0.9221
Dataset 2	LGBM	0.9234	0.8703	0.8903	0.9524	0.9186
	XGB	0.9121	0.8763	0.8957	0.9529	0.9217
	CB	0.9189	0.8447	0.8724	0.9462	0.9041
	Proposed MI-AE-GOLGBM	0.9565	0.9279	0.9357	0.9724	0.9533
Dataset 3	LGBM	0.8397	0.7458	0.7795	0.7083	0.7250
	XGB	0.8224	0.7250	0.7785	0.6667	0.6965
	CB	0.8564	0.7792	0.8313	0.7500	0.7593
	Proposed MI-AE-GOLGBM	0.8952	0.8417	0.9090	0.7917	0.8230

for PD-affected individuals and 0 for non-PD individuals. Dataset 2 contains 23 columns, out of which 22 attributes are utilized as features.

While this proposed research employs dataset 3: Parkinson dataset with replicated acoustic features [45]. The dataset contains acoustic features extracted from 3 voice recording replications of the supported /a/ phonation for each of the 80 individuals, in which 40 of them with PD [10, 45]. This dataset comprises 240 instances and 46 attributes, created by Carlos J. Perez of the University of Extremadura, Cáceres, Spain [45]. This dataset comprises 120 instances are belonging to patients with PD, out of which 78 individuals are men, and 42 individuals are women. The remaining 120 numbers of instances belong to healthy persons. This dataset 3 contains 45 attributes that are utilized as features. Experiments with the Parkinson dataset with replicated acoustic features have focused on identifying PD existence value: 1 from nonexistence value: 0.

This proposed research employs a PS dataset with multiple sound recordings [46] for verification purposes so that this dataset will verify the performance of the suggested MI-AE-GOLGBM method. This dataset 4 comprises 1040 instances and 26 attributes [46]. The PD dataset consists of a training dataset and a test dataset [46]. The training data refers to 20 sufferers, out of which six are female, the remaining are male, and 20 healthy individuals, out of whom ten are female, and the remaining are male. The data are gathered at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University

Table 7 The performance achieved by the suggested MI-AE-GOLGBM method for dataset 4, after applying a tenfold cross-validation procedure

Dataset	Fold number	AUC	PR	REC
Dataset 4	Fold 1	1.0000	0.9851	0.9851
	Fold 2	1.0000	0.9844	0.9403
	Fold 3	1.0000	0.9848	0.9701
	Fold 4	1.0000	1.0000	0.9403
	Fold 5	1.0000	0.9565	0.9851
	Fold 6	1.0000	0.9710	1.0000
	Fold 7	1.0000	0.9697	0.9552
	Fold 8	1.0000	1.0000	1.0000
	Fold 9	1.0000	1.0000	0.9851
	Fold 10	1.0000	0.9851	1.0000
	Mean	1.0000	0.9837	0.9761

[10, 42, 43, 46]. Many sound recordings, out of which 26 are voice samples, including supported vowels, numbers, words, and compact sentences, are obtained from all subjects [10, 42, 43, 46]. Throughout collecting this dataset, 28 PD sufferers are urged to answer just the supported vowels 'a' and 'o' three times, respectively, that produce 168 recordings [42, 43, 46]. Similarly, 26 features are elicited from voice specimens of this dataset 4 [42, 43, 46]. This dataset can be employed as an independent test dataset to validating the training dataset effects [42, 43, 46]. Experiments with the Parkinson's speech dataset have focused on generally striving to identify PD existence value: 1 from nonexistence value: 0 [46].

The distribution of the examples regarding these above-said four datasets, which belong to two classes: the number of patients due to PD and healthy patients, is exhibited in Table 3. The above-said datasets are abstracted into Table 3, which exhibits the number of total instances available in the datasets, the number of available features,

Table 8 The performance verification of the suggested methodology after utilizing dataset 4

Dataset	Model	AUC	ACC	PR	REC	F1
Dataset 4	LGBM	0.8603	0.9072	0.9420	0.9273	0.9076
	RF	0.9846	0.9165	0.9426	0.9273	0.9104
	LR	0.9922	0.9098	0.9354	0.9126	0.9101
	KNN	0.7819	0.7083	0.7311	0.7746	0.7430
	DT	0.9001	0.9039	0.9366	0.9273	0.9048
	XGB	0.8993	0.9047	0.9380	0.9273	0.9055
	CB	0.9849	0.9064	0.9273	0.9407	0.9069
	SVM	0.9196	0.8134	0.8580	0.8418	0.8361
	ML-AE-GOLGBM	1.0000	0.9773	0.9837	0.9761	0.9797

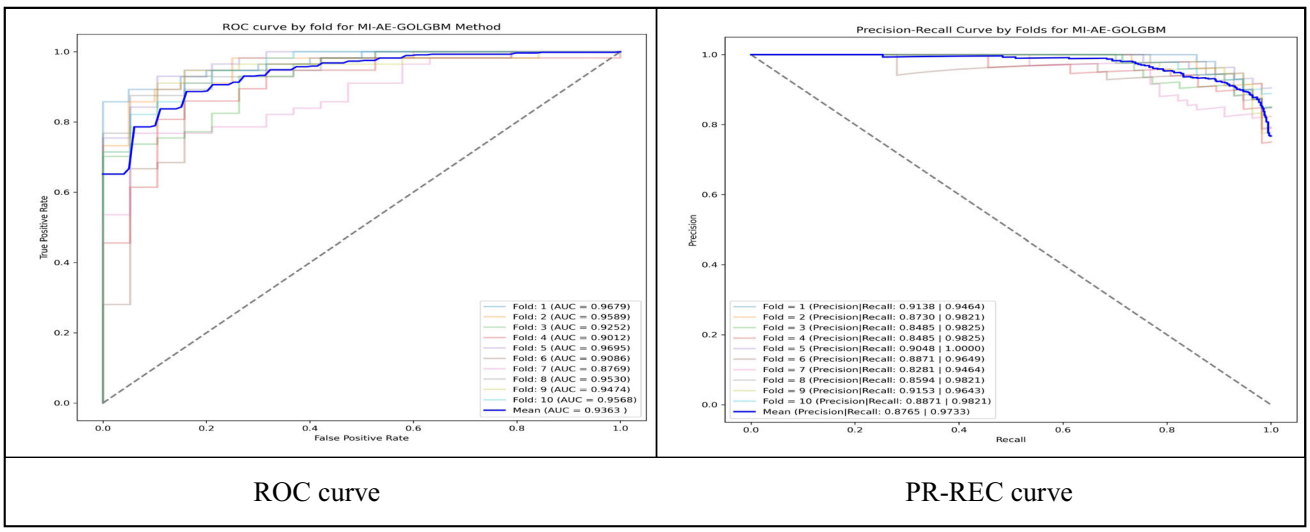

Fig. 5 ROC curve and PR-REC curve of the suggested MI-AE-GOLGBM method for dataset 1 after applying a tenfold CV procedure

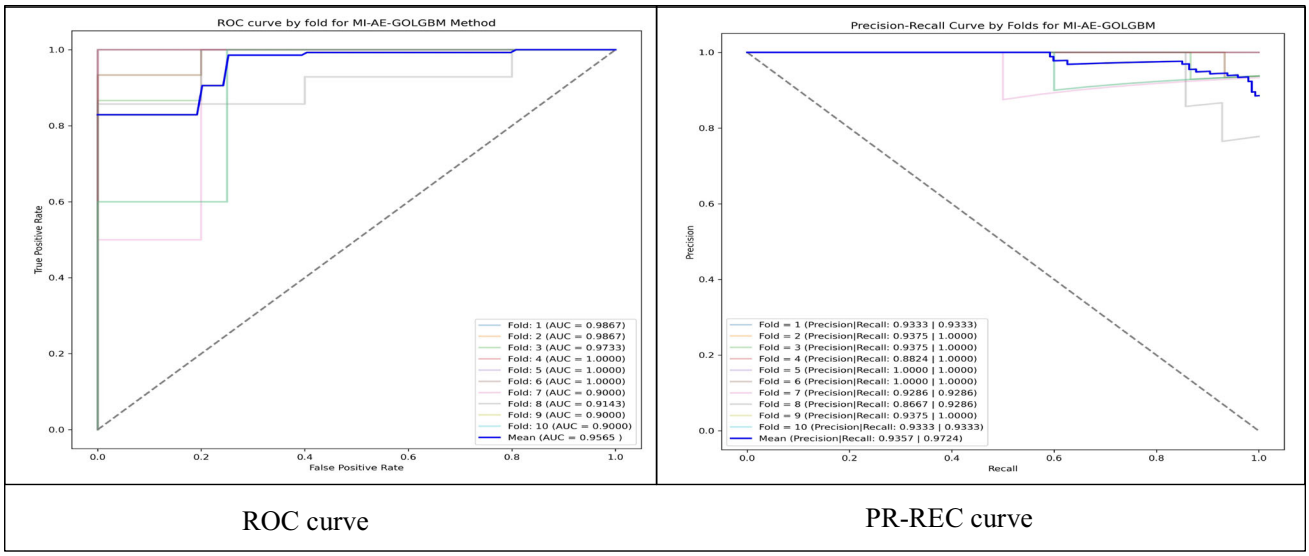

Fig. 6 ROC curve and PR-REC curve of the suggested MI-AE-GOLGBM method for dataset 2 after applying a tenfold CV procedure

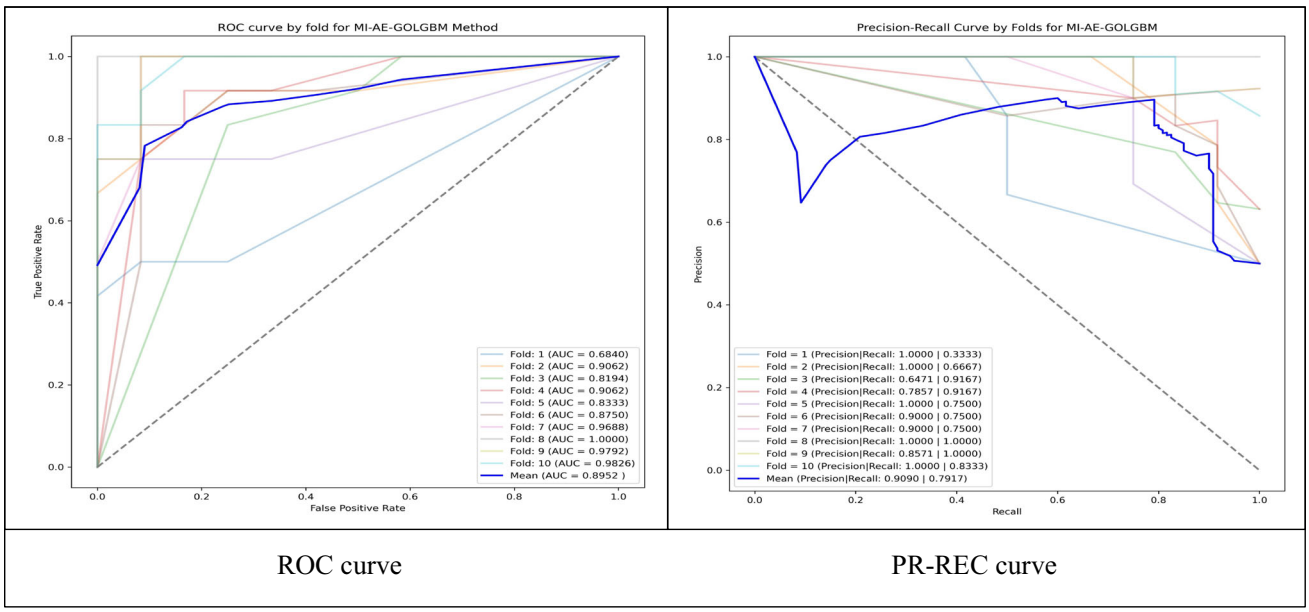

Fig. 7 ROC curve and PR-REC curve of the suggested MI-AE-GOLGBM method for dataset 3 after appying a tenfold CV procedure

the name of the dataset, and references for analyzing and downloading all the used datasets for assessing the performance of the proposed methodologies.

In this proposed system, the proposed MI-AE-GOLGBM approach is performed on these above-said datasets to evaluate the proposed methodology's performance. Moreover, the tenfold cross-validation (CV) strategy is applied to enhance the proposed method's accuracy; next, the performance assessment metrics are applied to assess the performance of our proposed method for each data set [40, 47].

3.6 Performance assessment metrics and overall best model selection

This suggested system applies the suggested MI-AE-GOLGBM methodology on each publicly available real-world dataset to predict the PD accurately. Various state-of-the-art ML algorithms: RF, logistic regression (LR), LGBM, KNN, XGBoost (XGB), CatBoost (CB), SVM, and decision tree (DT), are employed to compare the performance of the suggested MI-AE-GOLGBM methodology with the help of various performance evaluation metrics and tenfold CV technique [47]. This proposed research

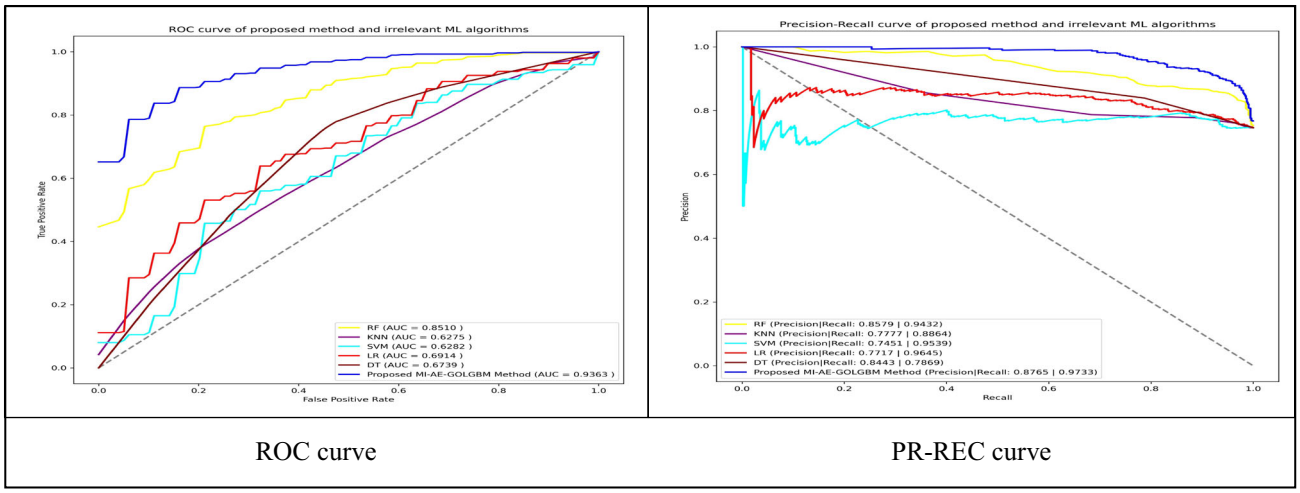

Fig. 8 Comparing the performances among the suggested MI-AE-GOLGBM method and the irrelevant ML algorithms with the help of ROC curve and PR-REC curve for dataset 1

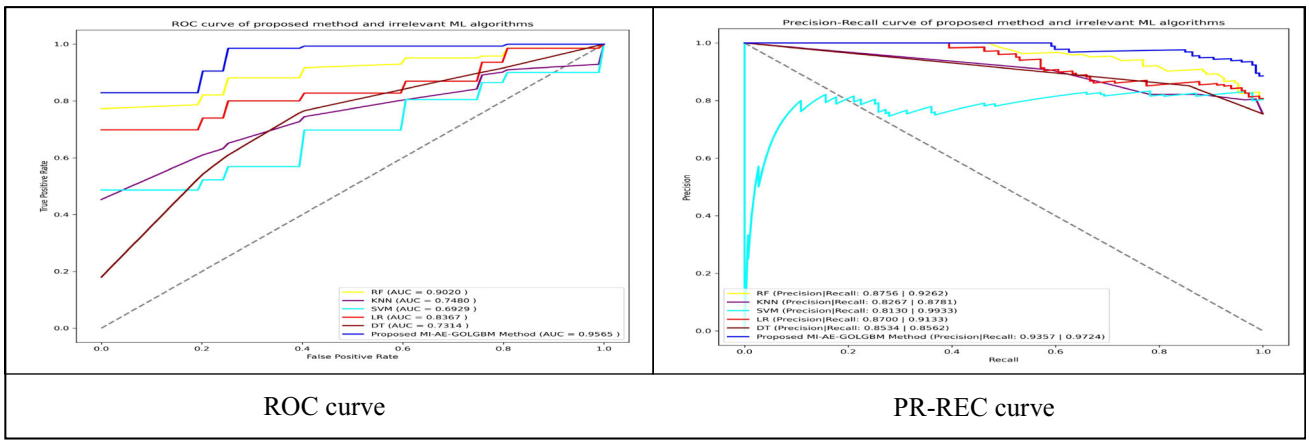

Fig. 9 Comparing the performances among the suggested MI-AE-GOLGBM method and the irrelevant ML algorithms with the help of ROC curve and PR-REC curve for dataset 2

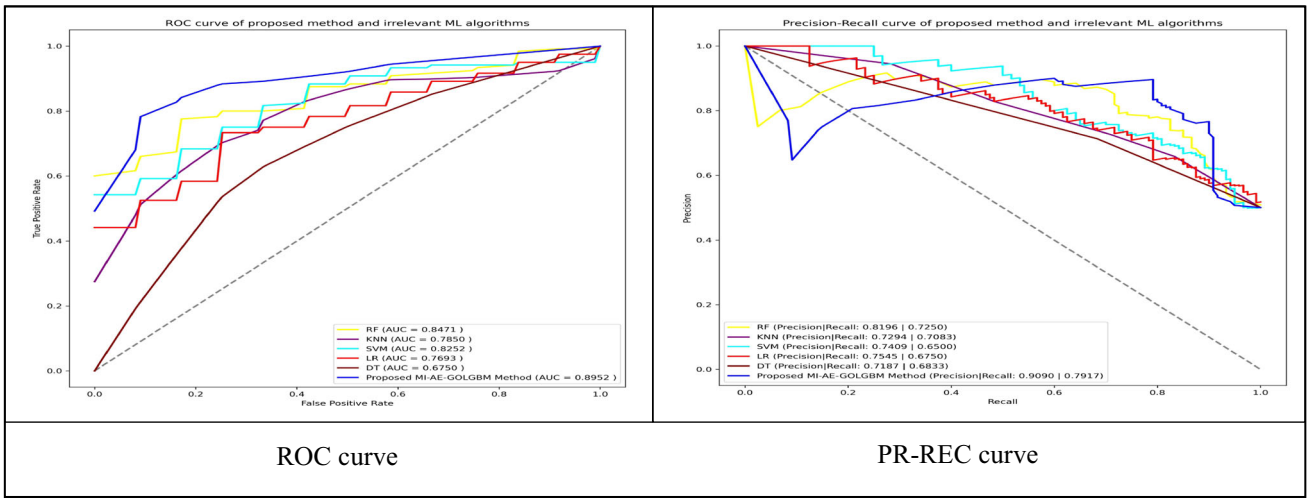

Fig. 10 Comparing the performances among the suggested MI-AE-GOLGBM method and the irrelevant ML algorithms with the help of ROC curve and PR-REC curve for dataset 3

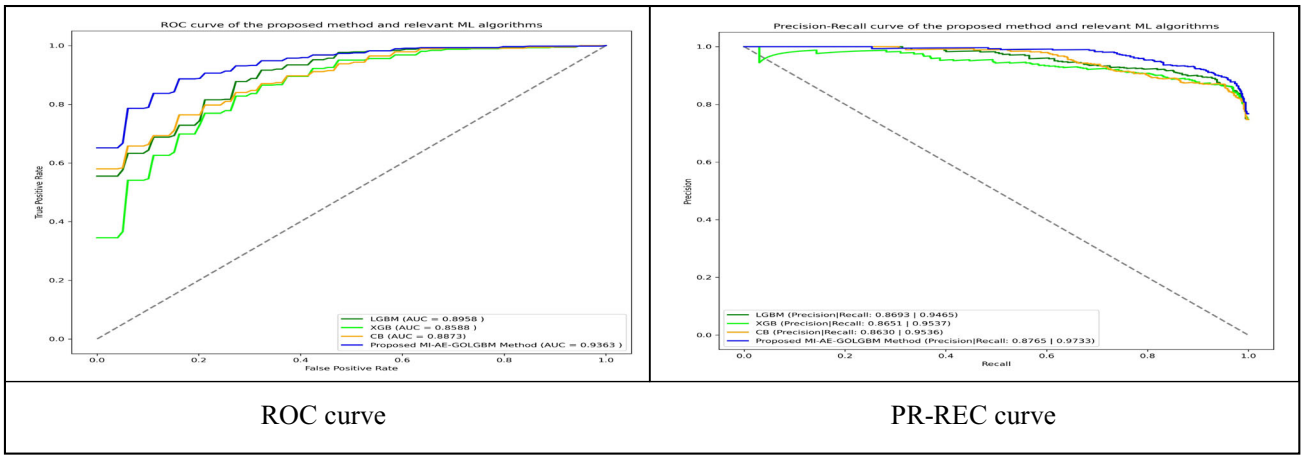

Fig. 11 Comparing the performances among the suggested MI-AE-GOLGBM method and the relevant ML algorithms with the help of ROC and PR-REC curves for dataset 1

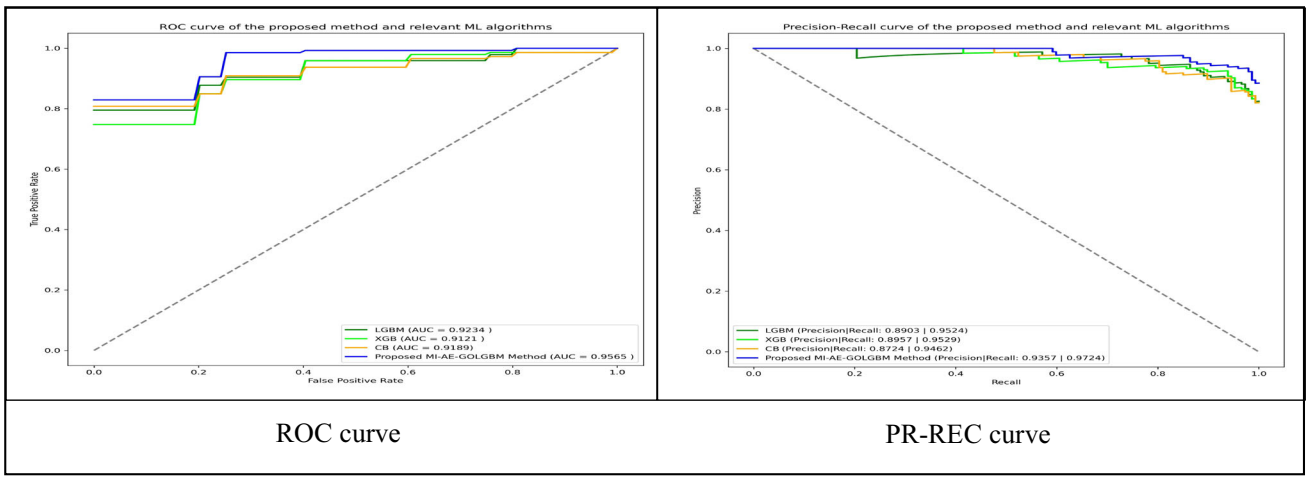

Fig. 12 Comparing the performances among the suggested MI-AE-GOLGBM method and the relevant ML algorithms with the help of ROC and PR-REC curves for dataset 2

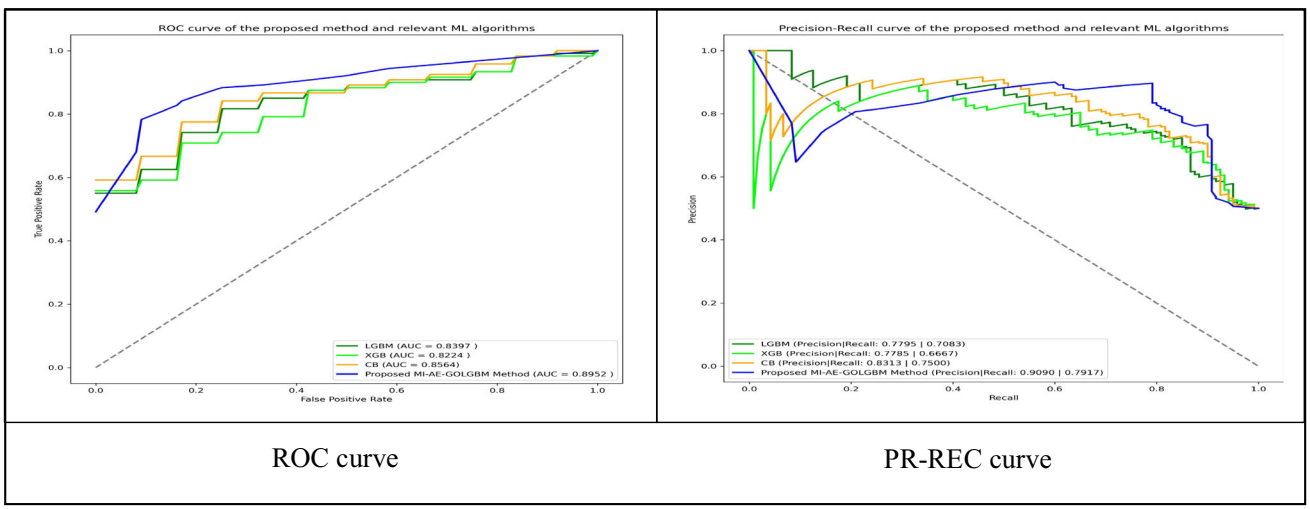

Fig. 13 Comparison of performances between the suggested MI-AE-GOLGBM method and the relevant ML algorithms with the help of ROC and PR-REC curves for dataset 3

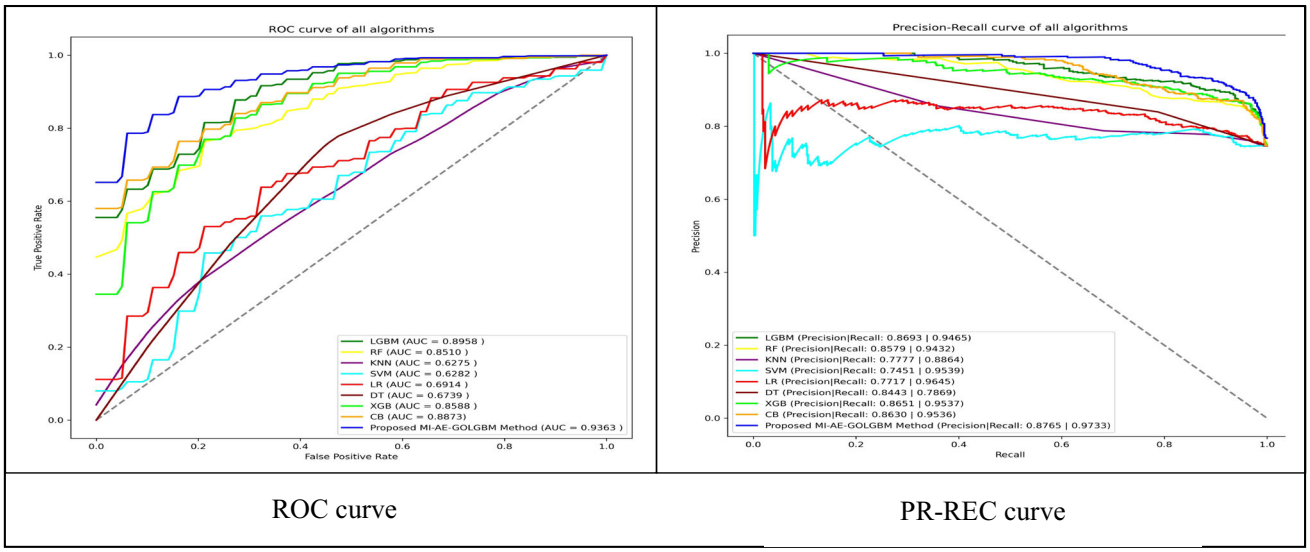

Fig. 14 ROC curve and PR-REC curve of the proposed methodology along with the relevant or irrelevant ML algorithms for the dataset 1

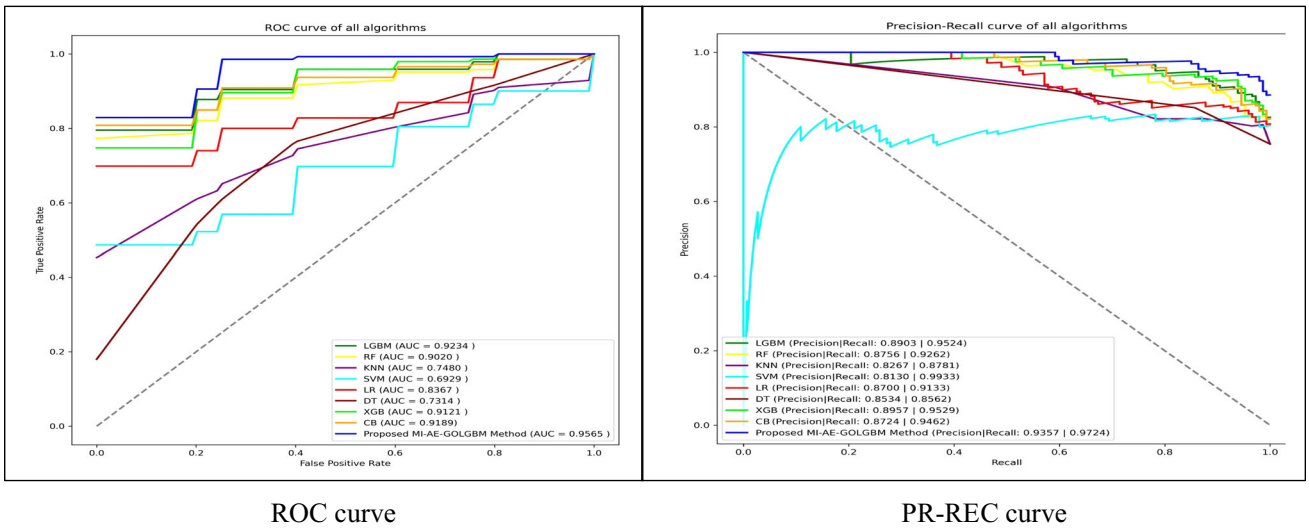

Fig. 15 ROC curve and PR-REC curve of the proposed methodology along with the relevant or irrelevant ML algorithms for dataset 2

assesses the proposed methodology's performance by using various performance evaluation metrics such as confusion matrix, AUC, accuracy (ACC), precision (PR), recall (REC), f1-measure (F1) value.

This tenfold CV strategy is employed to train and validate the performance P_r, where $P_r = \{P_1, P_2, \ldots, P_r\}$, of the set of r models M_r, where $M_r = \{M_1, M_2, \ldots, M_r\}$, for each dataset [47]. The above-stated performance evaluation metrics are applied to the set of r models regarding each model's performance [47].

In this phase, performance P_r for each model M_r is implemented to pick the most desirable model M_{BM} from r model sets [3, 47]. The proposed research employs the formula specified below for determining the overall best model M_{BM} [47].

$$M_{\mathrm{BM}} = M_r, \text{ where } P_{\mathrm{BM}} = \arg\max \sum_{1}^{r} P_r \tag{7}$$

where M_{BM} = The overall most desirable model chosen from the set of r models regarding best performance P_{BM} for each dataset, M_r = set of r models, and P_{BM} = The overall best performance chosen in terms of comparing the performance P_r for the set of r models [3, 47].

4 Results

The resulting performance achieved by the suggested hybrid methodology: MI-AE-GOLGBM, with the help of a tenfold CV procedure for the earlier mentioned datasets 1,

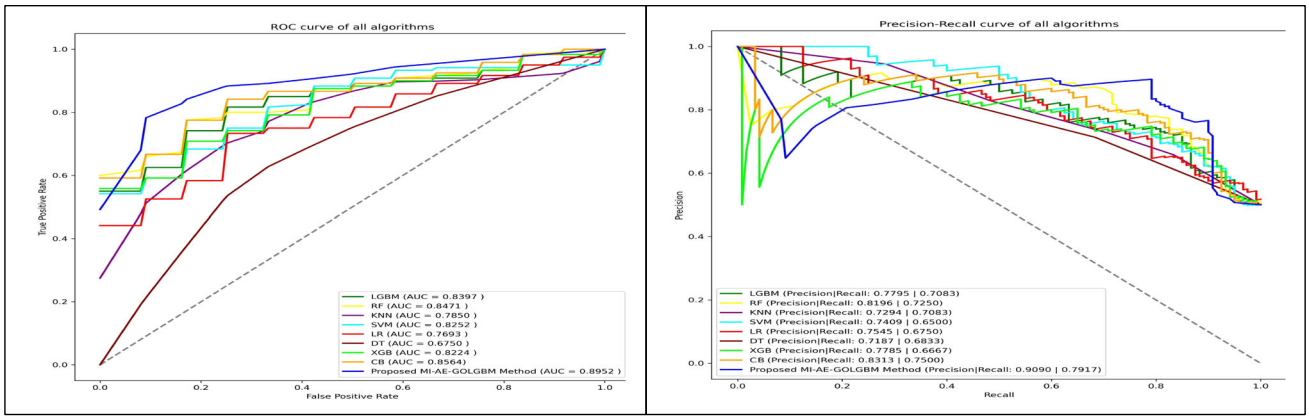

Fig. 16 ROC curve and PR-REC curve of the proposed methodology along with the relevant or irrelevant machine learning algorithms for dataset 3

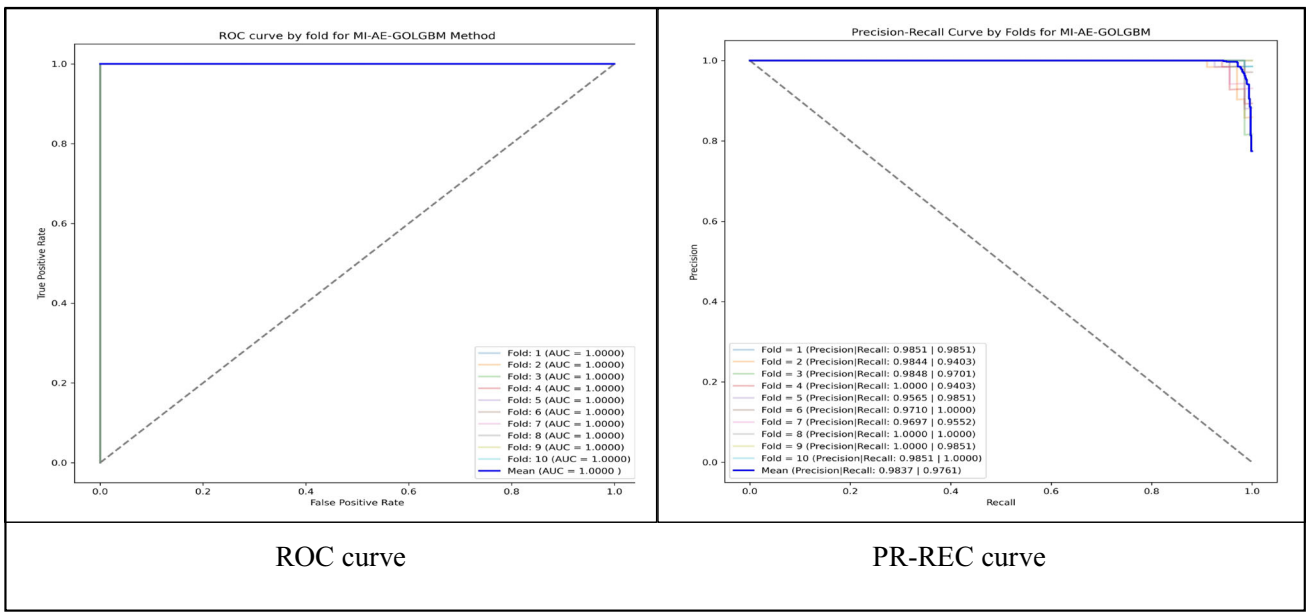

Fig. 17 ROC curve and PR-REC curve of the suggested MI-AE-GOLGBM method for dataset 4 after utilizing a tenfold CV procedure

2, and 3, are exhibited in Table 4 [40]. Comparing the performance between the suggested method and the irrelevant machine learning models: RF, LR, KNN, DT, and SVM [43, 46], for the datasets 1, 2, and 3, are exhibited in Table 5 and identify a model which determines the better performance among them. Furthermore, the performance comparison between the proposed approach and the relevant ML algorithms: LGBM, XGB, CB [46] for the above-said three datasets are also exhibited in Table 6 and determine the suitable model which achieves the more reliable performance among them. Such utilized relevant ML algorithms are entirely based on gradient boosting algorithms. Therefore, Table 6 also describes briefly that the proposed methodology enhances the relevant ML algorithms' performance. Additionally, Tables 5 and 6 determine the best model between the proposed approach and the above-said irrelevant or relevant ML algorithms.

Before verifying the proposed methodology's performance, the MI-AE-GOLGBM model exhibits its performance on data set 4 after employing a tenfold cross-validation strategy displayed in Table 7 [40]. After then, Table 8 exhibits the performance of the recommended methodology after utilizing dataset 4, in which this dataset is employed to verify the performance of the suggested method [40]. In the end, Tables 5, 6, and 8 are exhibited to help and determine the best adaptive model between the proposed methodology and the earlier mentioned relevant

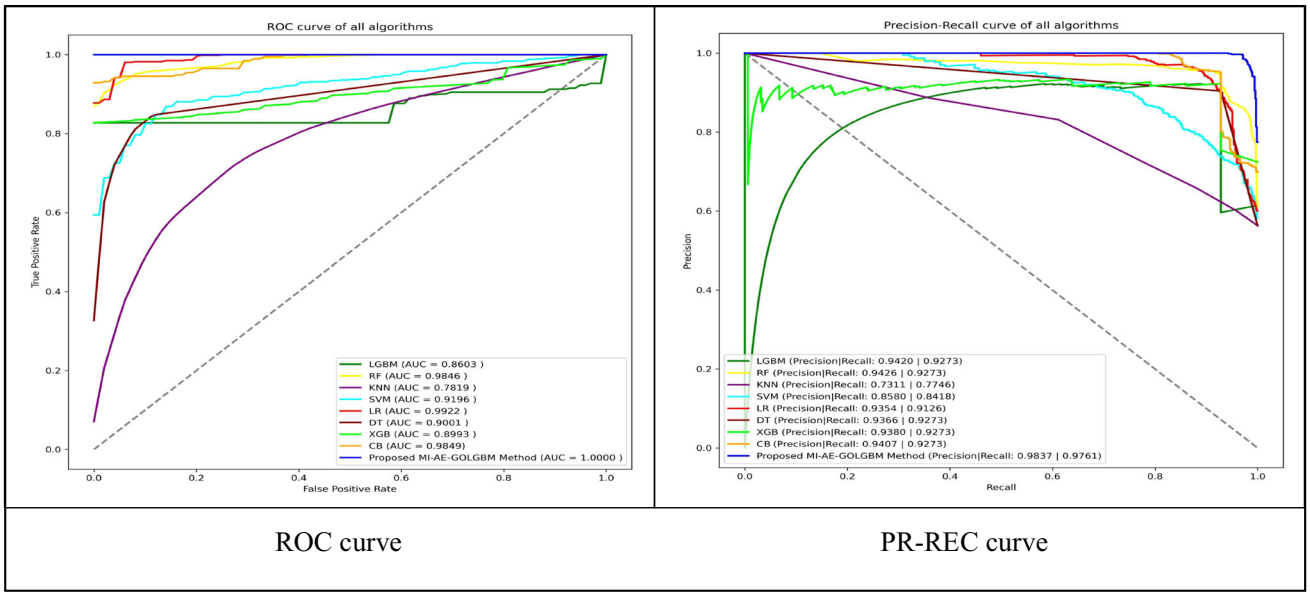

Fig. 18 Comparing the performances among the suggested MI-AE-GOLGBM method and the relevant and irrelevant ML algorithms with the help of ROC and PR-REC curves for dataset 4

or irrelevant ML algorithms after utilizing the four datasets, as mentioned above.

The above-said tables show the performances of the suggested approach with the help of the tenfold CV strategy and the different performance evaluation (PE) metrics mentioned earlier. The acquired outcomes from Tables 4, 5, 6, 7, and 8 can recognize the most reliable elaboration for the above-specified questions represented in the introduction section.

Regarding first question, the solution represents the data displayed in Table 4, in which the proposed MI-AE-GOLGBM approach received the mean AUC values of 0.9363, 0.9565, and 0.8952, accuracy of 0.8769, 0.9279, and 0.8417, PR values of 0.8765, 0.9357, and 0.9090, REC values of 0.9733, 0.9724, and 0.7917, and f1-measure values: 0.9221, 0.9533, and 0.8230 for the dataset 1, 2, and 3, respectively, after utilizing a tenfold CV procedure [40]. ROC and PR-REC curves provide the outcomes after employing the tenfold CV procedure exhibited from Figs. 5, 6 and 7 [40]. Such figures represent the ROC and PR-REC curves for each fold achieved by the suggested approach after utilizing a tenfold CV procedure.

Regarding the second question, the answer represents the data displayed in Table 5, which compares performances between the proposed methodology: MI-AE-GOLGBM and the above-said irrelevant ML algorithms. As exhibited in Table 5, the proposed MI-AE-GOLGBM approach gained the better AUC of 0.9363, 0.9565, and 0.8952, accuracy of 0.8769, 0.9279, and 0.8417, PR values of 0.8765, 0.9357, and 0.9090, and f1-measure: 0.9221, 0.9533 and, 0.8230 for the datasets 1, 2, and 3, respectively, than the irrelevant ML algorithms. While the proposed approach obtained better REC values of 0.9733 and 0.7917 for datasets 1 and 3, respectively, compared to the above-said unrelated ML algorithms. Hence, the proposed methodology's performance: MI-AE-GOLGBM, provides better outcomes, which describe that the proposed approach is a better model that supplies better performance than irrelevant ML algorithms for diagnosing PD accurately. Figures 8, 9, and 10 represent the ROC and the PR-REC curves of our recommended method and the irrelevant ML algorithms for datasets 1, 2, and 3. Such figures are employed to visualize and verify the outcomes provided in Table 5.

Regarding the third and fourth questions, the answers represent the data, which is reflected in Table 6, in which the proposed MI-AE-GOLGBM algorithm provides better performance in terms of AUC values of 0.9363, 0.9565, and 0.8952, the accuracy of 0.8769, 0.9279, and 0.8417, PR values of 0.8765, 0.9357, and 0.9090, REC values of 0.9733, 0.9724, and 0.7917, and f1-measure: 0.9221, 0.9533, and 0.8230 for the dataset 1, 2, and 3, respectively. Whereas in respect of relevant ML algorithms, the CB algorithm obtained the lowest ACC values of 0.8518 and 0.8447, PR values of 0.8630 and 0.8724, and f1-measure values: 0.9038 and 0.9041 for datasets 1 and 2, respectively. Moreover, the CB algorithm obtained the lowest REC value of 0.9462 for dataset 2. In comparison, the XGB algorithm obtained the lowest AUC values of 0.8588, 0.9121, and 0.8224 for datasets 1, 2, and 3, respectively. Additionally, the XGB algorithm obtained the lowest ACC value of 0.7250, PR value of 0.7785, REC value of 0.6667,

and the f1-measure value of 0.6965 for dataset 3. In contrast, the LGBM algorithm obtained the lowest REC value of 0.9465 for dataset 1. Thus, the MI-AE-GOLGBM method's performance provides better outcomes; therefore, it declares that the proposed approach determines as a better model, which provides more reliable performance compared to relevant ML algorithms for diagnosing PD precisely. Figures 11, 12, and 13 represent the ROC and the PR-REC curves of our recommended method and the relevant ML algorithms for datasets 1, 2, and 3. Such figures are employed to visualize and verify the outcomes provided in Table 6.

Thus, the proposed methodology: MI-AE-GOLGBM, provides significant impact and beats the powerful and relevant ML algorithms. Thus, the proposed system exhibits that the proposed method enhances the relevant ML algorithms' performance through the reliable and the best-suited two-stage dimensionality reduction approach with the genetic algorithm-based hyperparameter optimization methodology. In this matter, the proposed MI-AE-GOLGBM algorithm enhances the performance compared to the LGBM algorithm in terms of the AUC values of 4.05%, 3.31%, and 5.55%, ACC values of 2.25%, 5.76%, and 9.32%, PR values of 0.72%, 4.54%, and 12.95%, REC values of 2.68%, 2%, and 8.34%, and the f1-measure values of 1.73%, 3.47%, and 9.8% for the datasets 1, 2, and 3, respectively. Simultaneously, the proposed MI-AE-GOLGBM algorithm enhances the performance compared to the XGB algorithm in terms of the AUC values of 7.75%, 4.44%, and 7.28%, ACC values of 2.37%, 5.16%, and 11.67%, PR values of 1.14%, 4%, and 13.05%, REC values of 1.96%, 1.95%, and 12.5%, and the f1-measure values of 1.59%, 3.16%, and 12.65% for the datasets 1, 2, and 3, respectively. On the other hand, the proposed MI-AE-GOLGBM algorithm enhances the performance compared to the CB algorithm in terms of the AUC values: 4.9%, 3.76%, and 3.88%, ACC values: 2.51%, 8.32%, and 6.25%, PR values: 1.35%, 6.33%, and 7.77%, REC values: 1.97%, 2.62%, and 4.17%, and the f1-measure values: 1.83%, 4.92%, and 6.37% for the datasets 1, 2, and 3, respectively. Hence, such outcomes exhibit that the suggested MI-AE-GOLGBM methodology determines a better model than the relevant, robust ML algorithms in terms of AUC, ACC, PR, REC, and f1-measure values for the earlier mentioned three datasets.

Regarding the fifth question, the answer represents the data displayed in Tables 5 and 6, which exhibit the comparison of performances between the proposed methodology: MI-AE-GOLGBM, and the above-said relevant and irrelevant ML algorithms. Both Tables show that the proposed approach's performance beats the above-said ML algorithms in terms of various performance evaluation metrics values mentioned above. Figures 14, 15, and 16 represent the ROC curves and the PR-REC curves of the suggested method, and the irrelevant and relevant ML algorithms for datasets 1, 2, and 3. Such figures are employed to verify the outcomes, which are already given in Tables 5 and 6. Hence, such figures elaborate that the proposed MI-AE-GOLGBM approach is exhibited as the best model in terms of the AUC values of 0.9363, 0.9565, and 0.8952, ACC values of 0.8769, 0.9279, and 0.8417, PR values of 0.8765, 0.9357, and 0.9090, and f1-measure values of 0.9221, 0.9533, and 0.8230 for datasets 1, 2, and 3, respectively, and REC values: 0.9733 and 0.7917 for datasets 1 and 3, respectively.

Regarding the sixth question, the answer represents the data concerning such question, illustrated in Table 8. Dataset 4 is employed in this proposed research to verify the proposed methodology's performance by comparing it with the relevant or irrelevant ML algorithms. In this regard, the proposed MI-AE-GOLGBM method gained the best AUC value: 1.0, ACC of 0.9773, PR value of 0.9837, REC value of 0.9761, and f1-measure value: 0.9797 for dataset 4. Whereas in respect of the above-said irrelevant and relevant ML algorithms, the KNN algorithm obtained the lowest AUC value: 0.7819, ACC value of 0.7083, PR value of 0.7311, REC value of 0.7746, and f1-measure value of 0.7430 for dataset 4. Verifying the proposed method's performance on data set 4 is achieved with a tenfold CV procedure exhibited in Table 7.

Figure 17 represents the ROC curve and the PR-REC curve by each fold of the suggested method for dataset 4, and Fig. 18 exhibits the respective ROC curve and PR-REC curve of the proposed method and the irrelevant and relevant ML algorithms for dataset 4. Such Figs. 17 and 18 are employed to verify the performance and the outcomes generated by the proposed method. Thus, such figures elaborate that the proposed MI-AE-GOLGBM approach was determined as the best model in terms of the AUC value of 1.0, ACC of 0.9773, PR value of 0.9837, REC value of 0.9761, and f1-measure value of 0.9797 for dataset 4. Hence, the proposed methodology's performance: MI-AE-GOLGBM, is verified and provided the best outcomes after utilizing dataset 4.

After performing the above-said verification process, regarding the seventh question, the solution represents the data which is illustrated on Tables 5, 6, and 8, represents the best-effective approach for all utilized four datasets, in which this proposed research discovers that the proposed MI-AE-GOLGBM approach gained the best AUC values of 0.9363, 0.9565, 0.8952, and 1.0, ACC of 0.8769, 0.9279, 0.8417, and 0.9773, PR values of 0.8765, 0.9357, 0.9090, and 0.9837, and f1-measure values: 0.9221, 0.9533, 0.8230, and 0.9797 for datasets 1, 2, 3, and 4, respectively, and REC values of 0.9733, 0.7917, and 0.9761 for datasets 1, 3, and 4, respectively. Hence, such outcomes determine

that the proposed approach: MI-AE-GOLGBM is the best method for the best adaptive performances for all the above-said Parkinson's disease datasets. Therefore, the proposed hybrid model: a two-stage MI and AE enabled dimensionality reduction method with a genetically optimized LightGBM (MI-AE-GOLGBM) algorithm, supplies the best significant impact on the health sector, and detecting PD accurately, and providing the most scalable and reliable outcomes.

5 Conclusion

This proposed research presents an adaptive intelligent diagnostic system that employs a hybrid methodology: MI-AE-GOLGBM, to predict the PD and supports doctors in providing precise and up-to-date medicine and, hence, protecting many individual lives. A two-stage dimensionality reduction with genetically optimized LightGBM methodologies is employed in this study to generate the proposed model. The first stage of the two-stage dimension reduction approach implements an MI-based feature selection technique, which reduces irrelevant features, lessens the overfitting from the input features, and selects informative features. While in the second stage of the dimension reduction approach, it utilizes an unsupervised feature extraction strategy: AE, which is to lessen the features, and overfitting and, more importantly, reduces the redundancy of the input dataset given by MI and generates the most significant new features. In contrast, a genetically optimized LGBM algorithm implements a GA to tune the LGBM algorithm's hyperparameters automatically. After then, the LGBM algorithm employs such optimized hyperparameters and newly generated most significant features to classify the PD-affected patients and healthy controls most accurately. The conducted examinations completely confirm the effectiveness of the suggested approach after employing four PD-relevant voice-based datasets. The primary outcomes of this proposed research are concluded in the following.

1) The proposed approach is precisely tailored to identify PD and achieving the best performances regarding the AUC values of 0.9363, 0.9565, 0.8952, and 1.0, ACC of 0.8769, 0.9279, 0.8417, and 0.9773, PR values of 0.8765, 0.9357, 0.9090, and 0.9837, and f1-measure values: 0.9221, 0.9533, 0.8230, and 0.9797 for datasets 1, 2, 3, and 4, respectively, and REC values of 0.9733, 0.7917, and 0.9761 for datasets 1, 3, and 4, respectively.
2) The proposed approach beats the various ML classifiers, delivers the most outstanding performances, and validates itself as the most desirable approach to predict PD.

However, this suggested model can be improved after employing the latest dimension reduction approaches or employing the most recent hyperparameter optimization approaches like Harris Hawk optimization strategy and ORCA optimization algorithm or improving the current proposed model after utilizing a supervised or unsupervised deep neural network methodology to identify PD. In this case, the suggested method's future task is to attach the methodologies mentioned earlier and improve the performances to classify PD and upgrade our suggested approach's performance.

Declarations

Conflict of interest Thus the authors declared that they have no conflict of interest.

References

1. Alemami Y, Almazaydeh L (2017) Pathological voice signal analysis using machine learning based approaches. Comput Inform Sci 11(1):8. https://doi.org/10.5539/cis.v11n1p8
2. Naseer A, Rani M, Naz S, Razzak MI, Imran M, Xu G (2019) Refining Parkinson's neurological disorder identification through deep transfer learning. Neural Comput Appl 32(3):839–854. https://doi.org/10.1007/s00521-019-04069-0
3. An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms. (n.d.). IIETA | Advancing the World of Information and Engineering. https://www.iieta.org/journals/isi/paper/https://doi.org/10.18280/isi.250502
4. Tracy JM, Özkanca Y, Atkins DC, Ghomi RH (2020) Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 104:103362. https://doi.org/10.1016/j.jbi.2019.103362
5. Ali L, Zhu C, Zhou M, Liu Y (2019) Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. Expert Syst Appl 137:22–28. https://doi.org/10.1016/j.eswa.2019.06.052
6. Ashour AS, Nour MKA, Polat K, Guo Y, Alsaggaf W, El-Attar A (2020) A novel framework of two successive feature selection levels using weight-based procedure for voice-loss detection in parkinson's disease. IEEE Access 8:76193–76203. https://doi.org/10.1109/access.2020.2989032
7. Karan B, Sahu SS, Mahto K (2020) Parkinson disease prediction using intrinsic mode function based features from speech signal. Biocybern Biomed Eng 40(1):249–264. https://doi.org/10.1016/j.bbe.2019.05.005
8. Solana-Lavalle G, Galán-Hernández J-C, Rosas-Romero R (2020) Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybern Biomed Eng 40(1):505–516. https://doi.org/10.1016/j.bbe.2020.01.003
9. Tuncer T, Dogan S, Acharya UR (2020) Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybern

Biomed Eng 40(1):211–220. https://doi.org/10.1016/j.bbe.2019.05.006
10. Sharma SR, Singh B, Kaur M (2021) Classification of Parkinson disease using binary Rao optimization algorithms. Expert Syst. https://doi.org/10.1111/exsy.12674
11. Haq AU, Li JP, Memon MH, Khan J, Malik A, Ahmad T, Shahid M (2019) Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings. IEEE Access 7:37718–37734. https://doi.org/10.1109/access.2019.2906350
12. Despotovic V, Skovranek T, Schommer C (2020) Speech based estimation of Parkinson's disease using gaussian processes and automatic relevance determination. Neurocomputing 401:173–181. https://doi.org/10.1016/j.neucom.2020.03.058
13. Zhang T, Zhang Y, Sun H, Shan H (2021) Parkinson disease detection using energy direction features based on EMD from voice signal. Biocybern Biomed Eng 41(1):127–141. https://doi.org/10.1016/j.bbe.2020.12.009
14. Solana-Lavalle G, Rosas-Romero R (2021) Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. Biomed Signal Process Control 66:102415. https://doi.org/10.1016/j.bspc.2021.102415
15. Pramanik M, Pradhan R, Nandy P, Bhoi AK, Barsocchi P (2021) Machine learning methods with decision forests for Parkinson's detection. Appl Sci 11(2):581. https://doi.org/10.3390/app11020581
16. Lysiak A, Szmajda M (2021) Empirical comparison of the feature evaluation methods based on statistical measures. IEEE Access 9:27868–27883. https://doi.org/10.1109/access.2021.3058428
17. Xiong Y, Lu Y (2020) Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. IEEE Access 8:27821–27830. https://doi.org/10.1109/access.2020.2968177
18. Pasha A, Latha PH (2020) Bio-inspired dimensionality reduction for Parkinson's disease (PD) classification. Health Inform Sci Syst. https://doi.org/10.1007/s13755-020-00104-w
19. Sahu B, Mohanty SN (2021) CMBA-SVM: a clinical approach for Parkinson disease diagnosis. Int J Inf Technol 13(2):647–655. https://doi.org/10.1007/s41870-020-00569-8
20. Olivares R, Munoz R, Soto R, Crawford B, Cárdenas D, Ponce A, Taramasco C (2020) An optimized brain-based algorithm for classifying Parkinson's disease. Appl Sci 10(5):1827. https://doi.org/10.3390/app10051827
21. Gunduz H (2021) An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. Biomed Signal Process Control 66:102452. https://doi.org/10.1016/j.bspc.2021.102452
22. Cai Z, Gu J, Chen H-L (2017) A new hybrid intelligent framework for predicting Parkinson's disease. IEEE Access 5:17188–17200. https://doi.org/10.1109/access.2017.2741521
23. Hoq M, Uddin MN, Park S (2021) Vocal feature extraction-based artificial intelligent model for Parkinson's disease detection. Diagnostics 11(6):1076. https://doi.org/10.3390/diagnostics11061076
24. El-Hasnony IM, Barakat SI, Mostafa RR (2020) Optimized ANFIS model using hybrid metaheuristic algorithms for Parkinson's disease prediction in IoT environment. IEEE Access 8:119252–119270. https://doi.org/10.1109/access.2020.3005614
25. Chen H-L, Huang C-C, Yu X-G, Xu X, Sun X, Wang G, Wang S-J (2013) An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst Appl 40(1):263–271. https://doi.org/10.1016/j.eswa.2012.07.014
26. Soumaya Z, Drissi Taoufiq B, Benayad N, Yunus K, Abdelkrim A (2021) The detection of Parkinson disease using the genetic algorithm and SVM classifier. Appl Acoust 171:107528. https://doi.org/10.1016/j.apacoust.2020.107528
27. Ali L, Zhu C, Zhang Z, Liu Y (2019) Automated detection of parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. IEEE J Trans Eng Health Med 7:1–10. https://doi.org/10.1109/jtehm.2019.2940900
28. Lahmiri S, Shmuel A (2019) Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. Biomed Signal Process Control 49:427–433. https://doi.org/10.1016/j.bspc.2018.08.029
29. Kaur S, Aggarwal H, Rani R (2020) Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease. Machine Vision Appl. https://doi.org/10.1007/s00138-020-01078-1
30. Wang C, Deng C, Wang S (2020) Imbalance-xgboost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. Pattern Recogn Lett 136:190–197. https://doi.org/10.1016/j.patrec.2020.05.035
31. Polat K, Nour M (2020) Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals. Med Hypotheses 140:109678. https://doi.org/10.1016/j.mehy.2020.109678
32. Maachi IE, Bilodeau G-A, Bouachir W (2020) Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait. Expert Syst Appl 143:113075. https://doi.org/10.1016/j.eswa.2019.113075
33. Adams WR (2017) High-accuracy detection of early Parkinson's disease using multiple characteristics of finger movement while typing. Plos One. https://doi.org/10.1371/journal.pone.0188226
34. Tunc HC, Sakar CO, Apaydin H, Serbes G, Gunduz A, Tutuncu M, Gurgen F (2020) Estimation of Parkinson's disease severity using speech features and extreme gradient boosting. Med Biol Eng Compu 58(11):2757–2773. https://doi.org/10.1007/s11517-020-02250-5
35. Karan B, Sahu SS, Orozco-Arroyave JR, Mahto K (2021) Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction. Comput Speech Lang 69:101216. https://doi.org/10.1016/j.csl.2021.101216
36. De Souza RW, Silva DS, Passos LA, Roder M, Santana MC, Pinheiro PR, De Albuquerque VH (2021) Computer-assisted Parkinson's disease diagnosis using fuzzy optimum- path forest and restricted Boltzmann machines. Comput Biol Med 131:104260. https://doi.org/10.1016/j.compbiomed.2021.104260
37. Karan B, Sekhar Sahu S (2021) An improved framework for Parkinson's disease prediction using variational mode Decomposition-Hilbert spectrum of speech signal. Biocybern Biomed Eng 41(2):717–732. https://doi.org/10.1016/j.bbe.2021.04.014
38. Quan C, Ren K, Luo Z (2021) A deep learning based method for Parkinson's disease detection using dynamic features of speech. IEEE Access 9:10239–10252. https://doi.org/10.1109/access.2021.3051432
39. Yan B, Han G (2018) Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system. IEEE Access 6:41238–41248. https://doi.org/10.1109/access.2018.2858277
40. Dhar J (2021) Multistage ensemble learning model with weighted voting and genetic algorithm optimization strategy for detecting chronic obstructive pulmonary disease. IEEE Access 9:48640–48657. https://doi.org/10.1109/access.2021.3067949
41. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Apaydin H (2019) A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. Appl Soft Comput 74:255–263. https://doi.org/10.1016/j.asoc.2018.10.022

42. UCI machine learning repository: Parkinson's disease classification data set. (2017, November 5). https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification
43. Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation. (2017, September 18). Publishing Open Access Research Journals & Papers | Hindawi. https://www.hindawi.com/journals/pd/2017/6209703/
44. Little MA, Mcsharry PE, Roberts SJ, Costello DA, Moroz IM (2007) Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 6(1):23. https://doi.org/10.1186/1475-925x-6-23
45. Naranjo L, Pérez CJ, Campos-Roca Y, Martín J (2016) Addressing voice recording replications for Parkinson's disease detection. Expert Syst Appl 46:286–292. https://doi.org/10.1016/j.eswa.2015.10.034
46. Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, Kursun O (2013) Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE J Biomed Health Inform 17(4):828–834. https://doi.org/10.1109/jbhi.2013.2245674
47. Dhar J, Jodder AK (2020) An effective recommendation system to forecast the best educational program using machine learning classification algorithms. Ingénierie des systèmes d Inform 25(5):559–568. https://doi.org/10.18280/isi.250502