Predicting Parkinson's Disease Related Genes Based on PyFeat and Gradient Boosted Decision Tree (GBDT)

Marwa Helmy
Mansoura University
Eman Eldaydamony
Mansoura University
Nagham Mekky
Mansoura University
Mohammed Elmogy (melmogy@mans.edu.eg)
Mansoura University
Hassan Soliman
Mansoura University

Research Article

Keywords: Parkinson's Disease, PyFeat, GBDT, DNA FASTA

Posted Date: January 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1223538/v1

Predicting Parkinson's Disease Related Genes Based on PyFeat and Gradient Boosted Decision Tree (GBDT)

Marwa Helmy[1], **Eman Eldaydamony**[1], **Nagham Mekky**[1], **Mohammed Elmogy**[1,*], **and Hassan Soliman**[1]

[1]Information Technology Dept., Faculty of Computers and Information, Mansoura University, Mansoura, P.O.35516, Egypt
[*]melmogy@mans.edu.eg

ABSTRACT

Identifying genes related to Parkinson's disease (PD) is an active and effective research topic in biomedical analysis, which plays a critical role in diagnosis and treatment. In recent years, many studies have proposed different techniques for predicting disease-related genes. However, a few of these techniques are designed or developed for PD gene prediction. Most of these PD techniques are developed to identify only protein genes and discard long non-coding (lncRNA) genes, which play an essential role in biological processes and the Transformation and development of diseases. This paper proposes a novel prediction system to identify protein and lncRNA genes related to PD that can aid in an early diagnosis. First, we preprocessed the genes into DNA FASTA sequences from the UCSC genome browser and removed the redundancies. Second, we extracted some significant features of DNA FASTA sequences using five numerical mapping techniques with Fourier transform and PyFeat method with Adaboost technique as feature selection. Finally, the features were fed to the gradient boosted decision tree (GBDT) to diagnose different tested cases. Seven performance metrics are used to evaluate the performance of the proposed system. The proposed system achieved an average accuracy (ACC) equals 78.1%, the area under the curve (AUC) equals 84.9%, the area under precision-recall (AUPR) equals 85.0%, F1-score equals 78.2%, Matthews correlation coefficient (MCC) equals 0.564, Sensitivity (SEN) equals 79.1%, and specificity (SPC) equals 77.1%. The experiments demonstrate promising results compared with other systems. The predicted top-rank protein and lncRNA genes are verified based on a literature review.

Introduction

Parkinson's disease (PD) is a common neurodegenerative disease that is characterized by the loss of dopaminergic neurons in an area of the brain known as the substantia nigra (SN)[1]. This loss in dopaminergic neurons causes unexplained nerve dysfunction, which leads to motor and non-motor disturbances[2]. PD affects an estimated 7-13 million people worldwide[3]. PD is determined rare before the age of 50, but it becomes more common as people get older. It's affecting more than 1% of the people above the age of 60 and $\sim 4\%$ above 80 years. Therefore, PD is considered the most common movement disorder and the second most common neurodegenerative disease after Alzheimer's disease (AD)[4]. There are four essential signs related to PD: tremor, rigidity, bradykinesia, and post instability[5]. However, the cause of PD is still unclear. Furthermore, the disease progresses at a different pace in different people. So that, the disease course varies depending on the patient's age, and the rate of progression differs across the population[2,6]. PD's progression and the degree of symptoms create several socio-economic challenges, which affect not only PD patients and the healthcare system but also their families and caregivers[3,4].

Because of the complexities of PD, there is no single suitable gold standard test to diagnose PD, track its progression, predict risk factors, or assess the PD severity. As a result, there has been an ongoing search for suitable PD biomarkers over the last decade[2,7]. The biomarker is characterized as a noticeable feature that is capable of detecting unusual biological processes[8]. So that, the discovery and validation of PD biomarkers are critical for enhanced clinical evaluation and treatment of the disease.

There are four biomarkers to identify PD: clinical, imaging, biochemical, and genetic markers. Clinical biomarkers have been identified as the most commonly utilized diagnostic measures, which experts use for assessing and diagnosing PD and determining the progression and severity of PD. Observing motor symptoms, such as tremor, rigidity, bradykinesia, and post instability, are considered the primary assessments using the UPDRS scale. However, distinguishing PD from other Parkinsonism and movement diseases, such as progressive supranuclear palsy (PSP) and essential tremor (ET), can be difficult with such markers[2].

In the neuroimaging biomarkers, PD is characterized by the loss and degradation of the dopaminergic neuron. As a result,

neuroimaging techniques for the dopamine system may be good candidates for diagnosis and treatment analysis[8]. Single-photon emission computed tomography (SPECT) and dopamine transporter (DAT) imaging modalities have been used widely for diagnosing PD and other neurodegenerative disorders. Other imaging techniques, such as transcranial sonography (TCS) and magnetic resonance imaging (MRI), are also used to track and monitor brain changes that can be utilized to identify the PD's risk[9].

Biochemical biomarkers benefit over other types of biomarkers. This is because it can be discovered in the body fluids, such as saliva, serum, cerebrospinal fluid (CSF), blood, and biopsies, making them less expensive to extract. As a result, the process includes a non-invasive analysis for the molecules and proteins present in the body fluids[2]. On the other hand, many PD genes are known for genetic biomarkers according to the national center for biotechnology information (NCBI) website[4]. However, there are still many PD genes that have not been identified. In addition, PD has various signs, which appear in the latter stages of the disease. Therefore, we work on the genetic markers for identifying genes to make an early PD diagnosis.

Identifying genes related to diseases is determined as the most challenging task in biological analysis. Nevertheless, it provides significant contributions to the understanding of disease parthenogenesis, medical diagnosis, and drug development[10,11]. As a result, identifying genes related to PD enhances the experience and understanding of this disease and helps in the diagnosis and treatment of the PD[12]. The most existing methods are designed for predicting disease-related genes. However, a few of these methods are used for PD gene prediction. Furthermore, most of these PD's methods are designed to identify genes that can code as protein and discard non-coded elements like Long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) in PD-genes prediction[13].

Most studies in the biological field show that lncRNAs play a critical and essential role in the transformation and development of various diseases. The lncRNA is a transcript consisting of longer than 200 nucleotides that cannot be translated into proteins. Therefore, identifying the lncRNAs associated with diseases is vital for improving the diagnosis and treatment of the diseases. So that, identifying protein as well as lncRNA genes related to PD enhances the diagnosis and treatment of PD[13,14].

Our proposed prediction system used the lncRNA genes as another data source besides the protein genes. The use of lncRNAs overcomes the limitations that only protein genes are expressed as the original data. So that we can identify all genes associated with PD, which can aid in an early diagnosis and treatment. We represent all genes into DNA FASTA sequences that contain the most significant information about the genes. It plays an important role in the extraction of essential and distinguishing features of the genes[15]. The main contributions of our proposed prediction system can be summarized in the following points:

- The novel framework is proposed for predicting genes related to PD based on protein as well as lncRNA genes, which have a critical role in the development of PD.
- All protein and lncRNA genes are presented as DNA FASTA sequences to obtain local and global significant genes.
- The FASTA sequences are fed to multiple feature extraction methods for extracting the most distinguishing and important features. Five numerical mapping techniques with Fourier and PyFeat methods are utilized to achieve this goal.
- To decrease the complexity and the computational time, the Adaboost (AB) technique is used to reduce the dimensionality of PyFeat features generation.
- The most distinguishing features are fed to the gradient-boosted decision tree (GBDT) technique to diagnosis different test cases.
- Various performance metrics are used to evaluate the proposed system. In addition, we validated our proposed system by comparing it to some current systems.
- We verify the predicted top-rank protein and lncRNA genes based on the most recent studies from the literature.

For reader convenience, the used abbreviations in this paper are listed in Table 1. The rest of this paper is divided into five sections. The related work, the current weakness, and how we overcame these limitations in our proposed system are discussed in Section 2. The materials and methods are introduced in Section 3. The datasets, hardware specifications, evaluation metrics, and results are introduced in Section 4. Our experimental results are discussed in Section 5. Finally, Section 6 represents a conclusion and summary of our future work plans.

Related Work

Predicting genes related to a disease is considered an active search topic in the biological field. Many researchers have identified and predicted genes related to diseases, but some of these studies specialized in PD. Table 2 shows the summary of the current studies. For example, Radivojac et al.[16] presented an approach to predict the disease-related genes based on the protein-protein

PD	Parkinson's disease	ACC	Accuracy
lncRNA	Long non coding RNA	PPV	Positive Predictive Value
DFT	Discrete Fourier Transform	FFT	Fast Fourier Transform
MM	monoMonoKGap	MD	monoDiKGap
MT	monoTriKGap	DM	diMonoKGap
DD	diDiKGap	DT	diTriKGap
TM	triMonoKGap	TD	triDiKGap
A	Adenine	C	Cytosine
G	Guanine	T	Thymine
DT	Decision Tree	NB	Naive Bayes
TP	True positive	RF	Random Forest
FP	False positive	AB	Adaboost
LR	Logistic Regression	GBDT	gradient boosting decision tree
SVM	Support Vector Machine	LDA	Linear Discriminant Analysis
AUPR	Area under precision-recall	AUC	Area Under the Curve
FN	False negative	TN	True negative
SE	sensitivity	SPC	specificity
TPR	True positive rate	FPR	False negative rate

Table 1. The used abbreviations.

interaction (PPI) network. First, they used three ways to represent feature vectors, such as disease-protein relationship, protein sequence, and protein function information. Second, they reduced the feature vector dimension to overcome overfitting and computation cost by using information gain to rank features. Finally, in the classification step, they applied the support vector machine (SVM) classifier as a supervised technique with two layers for predicting genes related to the disease.

Yang et al.[17] proposed a novel ensemble-based PU learning method (EPU) to identify genes related to the disease. They used multiple data sources and ensemble machine learning classifiers. First, they built three networks: the PPI, go similarity, and gene expression similarity networks. Then, they applied weighted K-nearest neighbor (KNN), weighted naïve Bayes (NB), and multiple level SVM classier based on the ensemble weighted gene. Based on ensemble-weighted classifiers, they built the EPU learning to predict disease-related genes. Hwang[18] proposed stepwise random forests (SRF) method to select the biological features for identifying genes related to the disease. First, they integrated multiple biological features from the gene characteristics, such as protein domains, gene ontology, and human protein interactions. Then, they conducted phenotype-gene association and preliminary feature selection. Their proposed SRF method consisted of two steps. First, the most important features were selected by using filter-based methods according to one-dimensional random forest regression. Second, the selected biological features were fed to random forest classification to identify genes related to the disease.

Ding et al.[14] proposed a prediction model for identifying the lncRNA-disease relationship via tripartite graph lncRNA-disease-gene (TPGLDA). Their model consists of four steps. First, they built gene-disease and lncRNA-disease adjacency matrix by combining gene-disease and lncRNA-disease interactions. Second, they estimated the relationship profile for each node, combined this vector into the adjacency matrix in order to allocate resources, and built a tripartite graph with lncRNA-disease-gene. Then, they used the resource allocation process according to a tripartite graph to build the relationship between lncRNA and disease. Finally, for each disease-lncRNA relationship, they calculated the resource score consequently.

Xuan et al.[13] proposed a method for identifying lncRNA genes related to the disease. They represented a convolutional neural network (CNN) to predict the lncRNA-disease associations referred to as CNNLDA. First, their system determined the similarities and relationships such as lncRNAs-diseases, lncRNAs-miRNAs, and miRNA-disease relationships. Then, they combined these similarities and relationships to build the matrix of features based on the biological principles about diseases, lncRNAs, and miRNAs. Thus, their framework was designed to extract both the attention and the global feature representations of disease-lncRNA relationships. The first part of their framework was specialized for feature extraction from the similarities and associations of diseases and lncRNAs. In the second part of their framework, the various weights are assigned to each feature and its types by performing their proposed system to predict lncRNAs related to the disease.

Bonidia et al.[19] proposed method to diagnose different lncRNAs cases. They extracted features based on a Fourier transform. They utilized different representations in order to classify the lncRNAs. Four classification techniques were employed, such as SVM, random forest (RF), AB, and NB, to build their system. Zhang et al.[20] performed frequent gene co-expression analysis for identifying genes associated with PD. They employed six known genes related to PD as known genes. They used Pearson correlation coefficients (PCC) between any couple of genes inside each dataset to find genes that frequently co-express with

these known genes. A set of PD genes were identified. This set of genes was analyzed and showed great importance with neurodegenerative diseases and metabolism.

Peng et al.[21] built an integrated network that contained different types of nodes and edges. It represented various biomedical data, such as diseases, genes, ontology terms, and their associations. They developed a simplified laplacian normalization supervised random walk (SLNSRW) algorithm, which consisted of three steps. First, they used multiple datasets and ontologies to build an integrated network. Second, they built a weighted integrated network by using a laplacian normalization. Finally, they applied a supervised random walk (RWR) method to predict disease-related genes based on a weighted integrated network. Peng et al.[12] identified genes related to PD based on node2vec autoencoder-support vector machine (N2A-SVM) method. They aimed to identify the protein genes that are related to PD. Their method consisted of three steps. First, they represented each gene by using the PPI network. Second, they used node2vec for extracting the useful features of these representations. Then, for dimension reduction of features vector, they used the auto-encoder method. Finally, they used the SVM classifier to build their training model.

Lei et al.[10] identified genes related to common diseases including PD. They combined protein genes, lncRNAs, and diseases with building a heterogeneous network. They proposed a network propagation algorithm, which applied to these heterogeneous networks. They employed the information loss model to improve these networks for identifying genes related to the disease. They determined the weights of the similarity networks based on information loss to select the most important relationships by using 3-sigma. They used a network propagation algorithm for scoring the genes. The disease-genes association probabilities are represented based on the final score of these genes. Yang et al.[22] predicted the disease-genes using a novel deep neural network model (PDGNet). They combined multiple views of phenotypes and genotypes features. They enhanced the deep neural network parameters and extracted an accurate features vector for each gene and disease with feedback information from training samples. These vectors were used as input layers in their non-linear network for learning multiple features of genes and disease. The appropriate scores between genes and disease were calculated by determining the similarity among their vectors. Finally, they used the cross entropy between the relevant scores and the true labels of disease-gene relations to optimize their model as the feedback results.

Bi et al.[23] designed a realistic multimodal analysis model by using data from functional magnetic resonance imaging (fMRI) and single nucleotide polymorphisms (SNPs). Their model consisted of three parts. First, they used correlation analysis to build the fusion feature of subjects. Second, they used their neural network as a clustering evolutionary random neural network ensemble (CERNNE) for analyzing the fusion feature. Finally, their method combined neural networks that were randomly constructed and used the clustering technique for optimizing the ensemble learner. The CERNNE was used to create a multi-task research system, identify PD patients, and predict PD-related genes and brain regions.

As mentioned above, The current studies have several limitations, which can be summarized in the following points. First, most studies have developed methods to predict the genes related to disease, but a few of these methods were designed for PD-genes prediction[13,14,16–18]. Second, some of these PD methods identified only protein genes related to PD and ignored lncRNA genes, whether lncRNAs is critical for improving our understanding and diagnosing different diseases[12,20,22,23]. Third, the evaluation measures values for identifying disease-related genes are still challenging[10,12,20,21]. Finally, in some studies using deep learning, their models have some limitations that they are prone to severe overfitting issues, as well as the training takes more time and requires a large memory[12,13,22].

To overcome and solve these limitations, we designed the proposed prediction system that primarily identifies genes related to PD based on the protein and lncRNA genes to benefit from the biological importance of lncRNAs besides the proteins. The proposed system represents all genes as DNA FASTA sequences to get all essential and distinguishing information. We extract the most significant features of these FASTA sequences using five numerical Fourier transform[19,24] and PyFeat method with AB as a feature selection technique[15]. The selected features are fed to the GBDT technique to aid in the diagnosis of different test cases. Finally, for evaluation, seven different performance metrics are applied to validate our proposed system.

Materials and Methods

The main contribution of our system is the identification of PD-related genes: protein and lncRNA, which can aid in the diagnosis and treatment of the disease. The proposed prediction system represents PD genes as DNA FASTA sequences by using the UCSC genome browser. We extracted the most significant features by utilized different feature extraction methods. These extracted features contain the most important and distinguishing information, which represent the DNA sequences, which play an essential role in PD-related gene prediction. These selected features are fed to the GBDT technique to diagnosis different test cases in our proposed system. As a result, the proposed system can analyze two separate datasets: proteins and lncRNAs. To validate our system, we used a variety of performance metrics.

Fig. 1 shows a novel framework of the proposed prediction system that consists of four steps. First, the preprocessing step for removing duplication of genes is followed by representing genes as DNA FASTA sequences and removing redundancies of sequences. Second, from the DNA sequences, the most significant features are extracted using different feature extraction

Study	Year	Analysis	Methodology	Dataset	Performance
Radivojac et al.[16]	2008	Identifying genes related to disease based on PPI network	PPI, SVM	HPRD, Swiss-Prot	AUC 73.1%
Zhang et al.[20]	2011	Predicting genes related to Parkinson's disease based on gene expression	PCC, TOPPGene	NCBI GEO	
Yang et al.[17]	2014	Predicting disease-genes based on PPI, GO, and gene expression similarity	EPUI	HPRD, OPHID	F1-score 78.6%
Peng et al.[21]	2017	Predicting disease-related genes based on genes, diseases, and ontology	SLN-SRW	Clinvar, GO, DO, STRING, OMIM	AUC 0.79
Hwang[18]	2017	Identifying genes related to disease based on random forests	SRF	OMIM, HPRD, OPHID, GO	recall 82.76% F1-score 83.48%
Ding et al.[14]	2018	Predicting lncRNAs genes related to diseases	TPGLDA	LncRNADisease, DisGeNET	AUC 93.9%, SEN 53.5%, SPC=99.0%
Peng et al.[12]	2019	Parkinson's disease genes prediction based on proteins genes	N2A-SVM	ClinVar	AUC 0.7289
Lei et al.[10]	2019	Predicting disease-related genes based on protein , lncRNAs, and disease	InLPCH	LncRNADisease, HPRD, OMIM	AUC 0.7863, recall > 0.5
Xuan et al.[13]	2019	Predicting disease related to lncRNA genes	CNNLDA	LncRNADisease, Lnc2Cancer,GeneRIF, starBase,DincRNA	AUC 0.952, AUPR 0.251
Yang et al.[22]	2020	Predicting disease-related genes based on disease -gene gene-GO, and disease-phenotype	PDGNet	DisGeNet, HPO, OrphaNet, STRING, HPRD, IntAct, PINA,	F1-score 0.227, recall 0.208
Bonidia et al.[24]	2020	Diagnosing between diferent cases lncRNAs	DFT, Entropy, Complex Network	RefSeq, GreeNC Ensembl (v87, v32)	
Bi et al.[23]	2021	Predicting PD-related genes brain regions	CERNNE	PPMI	ACC = 88.57%

Table 2. A comparison of some recent studies.

methods: five numerical representations with Fourier transform and the PyFeat method with AB technique as a feature selection. Then, the feature vectors are fed to the GBDT technique to diagnosis different test cases. Finally, we evaluate the proposed system results through seven performance measures, which show promising results compared with other systems. The proposed prediction system is detailed in the following subsections.

Prepossessing

In the preprocessing step, we prepare and enhance the original data to feed it to the feature extraction methods. First, we clear the original data and remove the protein and lncRNA gene redundancies. Then, we represent the protein and lncRNA genes as DNA FASTA sequences from the UCSC genome browser. These sequences contain the most significant local and global information about the genes, which aid in extracting the most important feature by using feature extraction techniques. Finally, we remove the FASTA sequences redundancies to enhance our proposed system and get accurate results.

Feature Extraction

We tried to extract the most significant features from the DNA FASTA sequences that contain the most valuable and distinguishing information. This information aids in predicting the structure, function, relationship, and expression of the genes. These extracted features help us correctly identify protein and lncRNA genes related to PD. This step is considered the most critical step in our proposed prediction system. We apply different feature extraction methods, which are described in this section. Frist, we apply Fourier transform with Five numerical mapping representation, which are Binary, Integer, Real, Z-curve, and EIIP representations. Second, we utilize PyFeat that used 13 biological methods for features generation, and AB as a feature selection technique. It's important to keep in mind that a biological sequence is defined as $S = (s[0], s[1], \cdots, s[L-1])$ in order for $s \in \{A, C, G, T\}$.

Fourier Transform and Numerical Mappings

For extracting features, the discrete Fourier transform (DFT) is applied. It is commonly utilized in digital image and signal processing fields. DFT can reveal hidden periodicities after translating from the time domain to frequency domain[24]. The DFT for a signal at frequency f, with length $L, x[l] (l = 1, 2, \cdots, L-1)$, is defined by equation (1).

$$X[f] = \sum_{l=0}^{L-1} x[l] e^{-j\frac{2\pi}{L} fl}, \qquad f = 1, 2, \cdots, L-1. \tag{1}$$

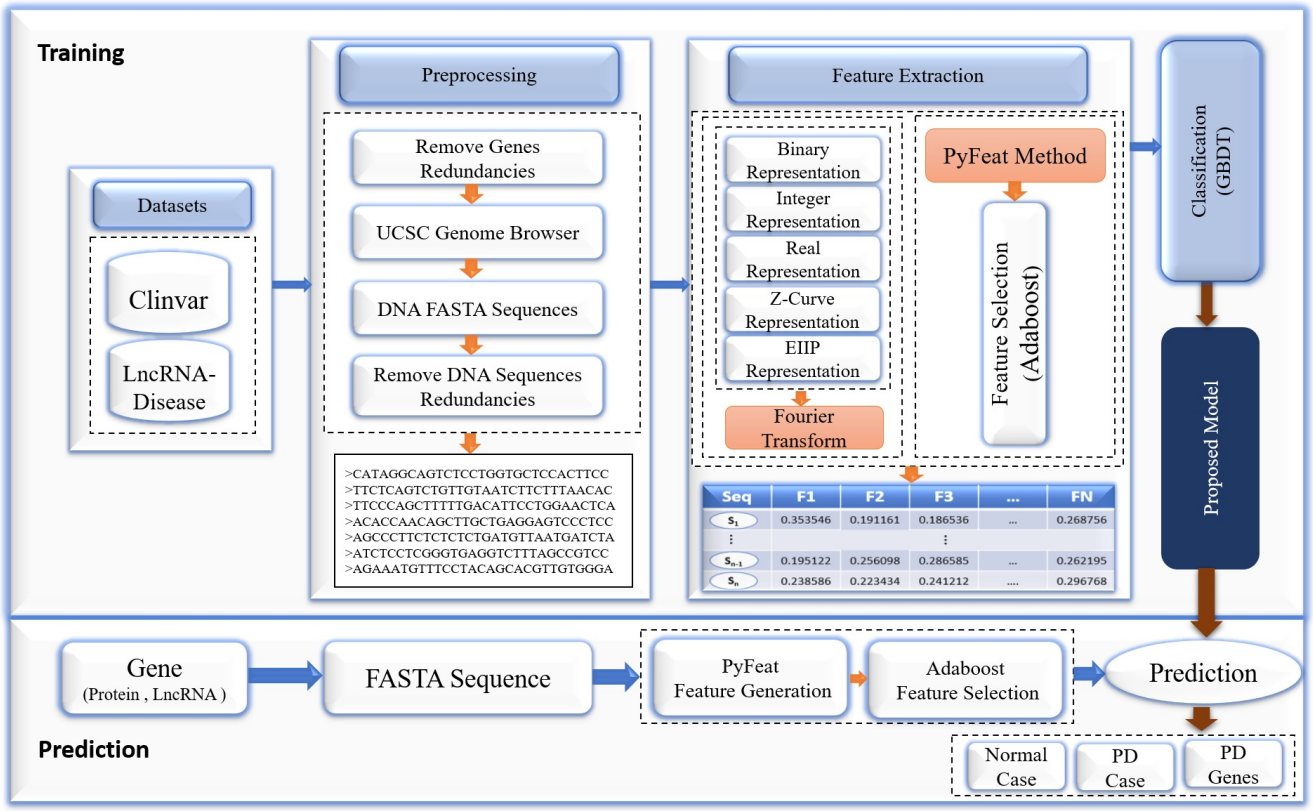

Figure 1. The proposed prediction system for identifying protein and lncRNA genes associated with PD.

This approach has been extensively investigated in bioinformatics, primarily for the study of recurring elements and periodicities in DNA sequences. To compute DFT for a sequence, we employ the fast Fourier transform (FFT), which is a very effective method to calculate the DFT. As a result, we use five numerical mapping representations, which are mentioned below: Binary, Integer, Real, Z-curve, and EIIP representations.

Binary Representation: This representation can use single or multidimensional vectors. Essentially, this method converts a sequence $s \in \{A,C,G,T\}^L$ into a matrix $B \in \{0,1\}^{4L}$ such that $B = [b1,b2,b3,b4]^T$, where T is the transpose operator. equation (2) is used to build each array.

$$b_i[l] = \begin{cases} 1, S[l] = \alpha[i] \\ 0, S[l] \neq \alpha[i] \end{cases}, \quad \text{where} \quad \alpha = (A,C,G,T), \quad l = 0,1,\cdots,L-1. \tag{2}$$

Therefore, in matrix B, each row could be an array that denotes the presence of base A in the first row, base C in the second row, base G in the third row as well as base T in the fourth row. For example, sequence $S = (T,G,A,C,C,G,A,G,A,G,A)$ is represented using binary form, where $b_1 = (0,0,1,0,0,0,1,0,1,0,1)$ stands for A-bases, $b_2 = (0,0,0,1,1,0,0,0,0,0,0)$ stands for C-bases, $b_3 = (0,1,0,0,0,1,0,1,0,1,0)$ stands for the G-bases, and $b_4 = (1,0,0,0,0,0,0,0,0,0,0)$ stands for T-bases. After that, the DFT is used for these sequences to obtain equation (3). Also, we obtain the power spectrum of this biological sequence by using equation (4).

$$B[f] = \sum_{n=0}^{L-1} b[l] e^{-j\frac{2\pi}{L}fl}, \quad \forall i \in [1,4], \quad f = 1,2,\cdots,L-1. \tag{3}$$

$$P_B[f] = \sum_{i=1}^{4} |B_i[f]|^2, \quad f = 1,2,\cdots,L-1. \tag{4}$$

Integer Representation: It is a one dimensional representation. We convert the four nucleotides of a biological sequence (T,C,A,G) to integers $(0,1,2,3)$. For instance, sequence $S=(T,G,A,C,C,G,A,G,A,G,A)$ is represented as $I=(0,3,2,1,1,3,2,3,2,3,2)$, which is defined by equation (5). Then, the DFT and power spectrum of the integer representation are defined by equation (6).

$$i[l]=\begin{cases} 0, & S[l]=T \\ 1, & S[l]=C \\ 2, & S[l]=A \\ 3, & S[l]=G \end{cases}, \qquad l=1,2,\cdots,L-1. \tag{5}$$

$$I[f]=\sum_{l=0}^{L-1} i[l]e^{-j\frac{2\pi}{L}fl}, \qquad P_I[f]=|I[f]|^2, \qquad f=1,2,\cdots,L-1. \tag{6}$$

Real Representation: This representation use the complement property of the complex mapping for real number representation[25]. The real representation is -1.5 for A, -0.5 for G, 0.5 for C, and 1.5 for T, as represented by equation (7). For example, sequence $S=(T,G,A,C,C,G,A,G,A,G,A)$ is represented as $r=(1.5,-0.5,-1.5,0.5,0.5,-0.5,-1.5,-0.5,-1.5,-0.5,-1.5)$. The DFT and power spectrum of the real representation are defined by equation (8).

$$r[l]=\begin{cases} -1.5, & S[l]=A \\ -0.5, & S[l]=G \\ 0.5, & S[l]=C \\ 1.5, & S[l]=T \end{cases}, \qquad l=1,2,\cdots,L-1. \tag{7}$$

$$R[f]=\sum_{l=0}^{L-1} r[l]e^{-j\frac{2\pi}{L}fl}, \qquad P_R[f]=|R[f]|^2, \qquad f=1,2,\cdots,L-1. \tag{8}$$

Z-curve Representation: It is a three-dimensional curve, which is used to describe DNA sequences with biological meaning. We can basically check sequence $S[l]$ with length L, considering the l-th element of the sequence $(l=1,2,\cdots,L)$. After that, we use the aggregate appearance numbers for each base A_l,C_l,G_l, and T_l, which represent the frequency of presence a base from $S[1]$ to $S[l]$. By using this method, we reduce the number of indications for sequences from four to three for all four elements in a symmetrical way[26].

$$A_l+C_l+G_l+T_l=l \tag{9}$$

This Z-curve is made from a set of nodes $P_1,P_2,\cdots,P_L$, which the coordination $x[l],y[l]$, and $z[l]$, where $(l=1,2,\cdots,L)$ are defined exclusively based on the Z-transform, as shown in equation (10).

$$P[l]=\begin{cases} x[l]=(A_l+G_l)-(C_l+T_l) \\ y[l]=(A_l+C_l)-(G_l+T_l) \\ z[l]=(A_l+T_l)-(C_l+G_l) \end{cases}, \qquad x[l],y[l],z[l]\in[-l,l], \qquad l=1,2,\cdots,L. \tag{10}$$

A sequence is fully described by the distributions $x[l],y[l]$, and $z[l]$. As a result, three biologically significant distributions will be available: (1) $x[l]$= purine/pyrimidine, (2) $y[l]$= amino/keto, (3) $z[l]$= weak hydrogen bonds/strong hydrogen bonds. For example, sequence $S=(T,G,A,C,C,G,A,G,A,G,A)$, will be represented with three distributions: $x=(4,5,6,5,4,1,2,3,4,5,5)$, $y=(-2,-3,-2,-1,0,-1,0,-1,0,-1,1)$, $z=(0,-1,0,-1,-2,-1,0,-1,0,-1,-1)$. After that, the DFT and the power spectrum are defined by equations (11) and (12).

$$X[f]=\sum_{l=0}^{L-1} x[l]e^{-j\frac{2\pi}{L}fl}, \qquad Y[f]=\sum_{l=0}^{L-1} y[l]e^{-j\frac{2\pi}{L}fl}, \qquad Z[f]=\sum_{l=0}^{L-1} z[l]e^{-j\frac{2\pi}{L}fl} \tag{11}$$

$$P_P[f]=|X[f]|^2+|Y[f]|^2+|Z[f]|^2, \qquad f=1,2,\cdots,L-1. \tag{12}$$

EIIP Representation: EIIP values of nucleotides for representing DNA sequences and for locating exons were proposed in[27]. According to this study, the EIIP representation is 0.0806 for G, 0.1260 for A, 0.1335 for T, and 0.1340 for C, as shown in equation (13). For example, $S = (T,G,A,C,C,G,A,G,A,G,A)$ is represented as $d = (0.1335, 0.0806, 0.1260, 0.1240, 0.1240, 0.0806, 0.1260, 0.0806, 0.1260, 0.0806, 0.1260)$. The DFT and power spectrum of this representation are defined by equation (14).

$$d[l] = \begin{cases} 0.0806, & S[l] = G \\ 0.1260, & S[l] = A \\ 0.1335, & S[l] = T \\ 0.1340, & S[l] = C \end{cases}, \qquad l = 1,2,\cdots,L-1. \tag{13}$$

$$D[f] = \sum_{l=0}^{L-1} d[l] d^{-j\frac{2\pi}{L}fl}, \qquad P_D[f] = |D[f]|^2, \qquad f = 1,2,\cdots,L-1. \tag{14}$$

Features: For each representation, we employ the feature extraction depending on the peak to average power ratio (PAPR), signal to noise ratio (SNR), minimum, maximum, median, population standard deviation, sample standard deviation, percentile (15/25/50/75), variance, coefficient of variation, amplitude, semi-interquartile range, interquartile range, skewness, and kurto-sis[19,24].

PyFeat

Extracting crucial features is an essential step in representing biological DNA sequences and identifying genes related to diseases. PyFeat is used for the creation of different numeric feature representations for biological sequences. In addition, it can be used to describe the fusion of essential features from broad neighboring residues. It focuses on extracting features that collect information about the relationships of neighboring residues so that more local and global features can be provided. This method can also choose the best and most essential features from a set of features that are created primarily by the gap. For biological DNA sequences, we have selected a similar group of features from different methods, such as Z-curve, gcContent, cumulative skew, Chou's pseudo composition, monoMonoKGap, monoDiKGap, monoTriKGap, diMonoKGap, diDiKGap, diTriKGap, triMonoKGap, and triDiKGap[15]. After the features generation, the AB technique is employed in order to select features with the most discriminatory information possible to reduce the dimensionality, complexity, and computational time. So that, the number of extracted features can be reduced significantly and employ PyFeat to represent the combination of essential features from large neighboring residues.

Features Generation: It intends on catching the frequency distributions of different permutations of the base nucleotides acids in biological DNA sequences. It is used to describe the sequences in the model training phase based on the kGap, efficiently. For DNA sequences, when the value of kGap is small, the number of generated features is also small, and the occurrence frequency of the generated features keeps local or short-range sequence-order information. While the value of kGap is moderately large, the generated features maintain global or long-range sequence-order information. According to the previous analysis, we consider the features where kGap values are equal to five to extract features that include local and global information. Table 3 shows the most significant features that are extracted from these different methods.

Z-curve: It is often used in genomic sequence analysis. It has got three components in three axes. They are defined by equation (13), in which three features are generated based on the Z-curve method.

GCcontent: This measure shows the proportion of G and C elements out of four elements (A, C, G, and T) in a sequence. It is defined by equation (15).

$$GC = \frac{\sum G + \sum C}{\sum A + \sum C + \sum G + \sum T} \times 100\% \tag{15}$$

ATGC ratio: It represents the summation ratio of the A and T elements to the summation of the G and C elements in a DNA sequence. It is defined by equation (16).

$$ATGCRatio = \frac{\sum A + \sum T}{\sum G + \sum C} \tag{16}$$

Cumulative Skew: It considers two measures as the GC skew and AT skew. The GC skew is determined as the normalized excess of G and C in a sequence. Similarly, AT skew is determined as the normalized excess of A and T in a sequence, as defined by equation (17).

$$GCskew = \frac{\sum G - \sum C}{\sum C - \sum G};\ ATskew = \frac{\sum A - \sum T}{\sum T - \sum A} \tag{17}$$

Pseudo Composition: This measure determines the frequencies of sub-sequences, where n is the sub-sequences length. The number of generated features from a sequence is defined by equation (18), where $n = 3$, then only eighty-four (84) features exist. These features are determined by the frequencies of sub-sequences: A, C, G, T, AA, ..., TT, AAA, ..., and TTT in the whole DNA sequence.

$$num(PC) = \sum_{i=1}^{n}(4^i), \qquad i = 1,2,\cdots,I-1. \tag{18}$$

monoMonoKGap: The generated features are determined based on the frequencies of sub-sequences with single nucleotides at the beginning and end as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (19). $kGap = 5$ and only 80 features exist. These features are determined by the frequencies of sub-sequences: A-A, ..., T-T, A- -A, ..., T- -T, A- - -A,..., T- - -T, A- - - -A, ..., T- - - -T, A- - - - -A, ..., and T- - - - -T in the whole DNA sequence.

$$num(MM) = 4 \times 4 \times n \tag{19}$$

monoDiKGap: The generated features are extracted based on the frequencies of sub-sequences with single nucleotide at the beginning and two nucleotides at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (20), where $kGap = 5$, then 320 features exist. These features are determined by the frequencies of sub-sequences: A-AA, ..., T-TT, A- -AA, ..., T- -TT, A- - -AA, ..., T- - -TT, A- - - -AA, ..., T- - - -TT, A- - - - -AA, ..., and T- - - - -TT in the whole DNA sequence.

$$num(MD) = (4) \times (4 \times 4) \times n \tag{20}$$

monoTriKGap: The generated features are extracted based on the frequencies of sub-sequences with single nucleotide at the beginning and three nucleotides at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (21), where $kGap = 5$, then 1280 features exist. These features are determined by the frequencies of sub-sequences: A-AAA, ..., T-TTT, A- -AAA, ..., T- -TTT, A- - -AAA, ..., T- - -TTT, A- - - -AAA, ..., T- - - -TTT, A- - - - -AAA, ..., and T- - - - -TTT in the whole DNA sequence.

$$num(MT) = (4) \times (4 \times 4 \times 4) \times n \tag{21}$$

diMonoKGap: The generated features are extracted based on the frequencies of sub-sequences with two nucleotides at the beginning and single nucleotide at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (22), where $kGap = 5$, then 320 features exist. These features are determined by the frequencies of sub-sequences: AA-A, ..., TT-T, AA- -A, ..., TT- -T, AA- - -A, ..., TT- - -T, AA- - - -A, ..., TT- - - -T, AA- - - - -A, ..., and TT- - - - -T in the whole DNA sequence.

$$num(DM) = (4 \times 4) \times (4) \times n \tag{22}$$

diDiKGap: The generated features are extracted based on the frequencies of sub-sequences with two nucleotides at the beginning and two nucleotides at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (23), where $kGap = 5$, then 1280 features exist. These features are determined by the frequencies of sub-sequences: AA-AA, ..., TT-TT, AA- -AA, ..., TT- -TT, AA- - -AA, ..., TT- - -TT, AA- - - -AA, ..., TT- - - -TT, AA- - - - -AA, ..., and TT- - - - -TT in the whole DNA sequence.

$$num(DD) = (4 \times 4) \times (4 \times 4) \times n \tag{23}$$

diTriKGap: The generated features are extracted based on the frequencies of sub-sequences with two nucleotides at the beginning and three nucleotides at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features

for DNA sequence is defined by equation (24), where $kGap = 5$, then 5120 features are existed, these features are determined by the frequencies of sub-sequences: AA -AAA, ..., TT-TTT, AA- -AAA, ..., TT- -TTT, AA- - -AAA, ..., TT- - -TTT, AA- - - -AAA, ..., TT- - - -TTT, AA- - - - -AAA, ..., and TT- - - - -TTT in the whole DNA sequence.

$$num(DT) = (4 \times 4) \times (4 \times 4 \times 4) \times n \tag{24}$$

triMonoKGap: The generated features are extracted based on the frequencies of sub-sequences with three nucleotides at the beginning and single nucleotide at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (25), where $kGap = 5$, then 1280 features are existed, these features are determined by the frequencies of sub-sequences: AAA-A, ..., TTT-T, AAA- -A, ..., TTT- -T, AAA- - -A, ..., TTT- - -T, AAA- - - -A, ..., TTT- - - -T, AAA- - - - -A, ..., and TTT- - - - -T in the whole DNA sequence.

$$num(TR) = (4 \times 4 \times 4) \times (4) \times n \tag{25}$$

triDiKGap: The generated features are extracted based on the frequencies of sub-sequences with three nucleotides at the beginning and two nucleotides at the ends as well as $kGap$ between them where $kGap = n$. The number of generated features for DNA sequence is defined by equation (26), where $kGap = 5$, then 5120 features exist. These features are determined by the frequencies of sub-sequences: AAA-AA, ..., TTT-TT, AAA- -AA, ..., TTT- -TT, AAA- - -AA, ..., TTT- - -TT, AAA- - - -AA, ..., TTT- - - -TT, AAA- - - - -AA, ..., and TTT- - - - -TT in the whole DNA sequence. Table 3 shows the overall methods utilized by PyFeat and the number of features for each method.

$$num(TD) = (4 \times 4 \times 4) \times (4 \times 4) \times n \tag{26}$$

Feature Selection: Based on PyFeat features generation, different methods are used, which are Z-curve, gcContent, ACGT ratio, Cumulative Skew, Chou's Pseudo composition, monoMonoKGap, monoDiKGap, monoTriKGap, diMonoKGap, diDiKGap, diTriKGap, triMonoKGap, and triDiKGap. As a result, we obtained a large number of features for each biological sequence, as shown in Table 3.

Method	Number Of Features
Z-curve	3
GCcontent	1
ATGC ratio	1
Cumulative Skew	2
Pseudo composition	84
monoMonoKGap	80
monoDiKGap	320
monoTriKGap	1280
diMonoKGap	320
diDiKGap	1280
diTriKGap	5120
triMonoKGap	1280
triDiKGap	5120
# of features	14,891

Table 3. PyFeat feature generation and their numbers.

To reduce the complexity and computational time for the classifier, the AB technique is used to reduce the feature vector dimension obtained by the PyFeat method and concurrently keep informative features. AB technique achieves an average impurity-curtailment, according to dividing each feature on all of the trees trained based on various weight distributions. So that, the features with the maximum score in the trained model are selected by using the SAMME.R AB technique. According to these composite features, we use the SAMME.R technique as feature selection for selecting n features with the maximum score in the trained model[15]. After applying the SAMME.R, we obtain 213 features as average for each biological sequences instead of 14,891 features generated by PyFeat, as shown in Table 3. We represent the algorithm for the proposed preprocessing and feature extraction technique using PyFeat with AB technique as feature selection, as shown in Algorithm 1.

Algorithm 1: The proposed preprocessing and feature extraction

Data: list of genes$L_0 1, 2, \cdots, N$
Result: The matrix of the most significant features F
Remove the redundancies in L_0 and update to L_1;
foreach *gene in* L_1 **do**
  Get the DNA FASTA sequence from the UCSC genome browser,
end
Remove duplication of sequences in L_1;
Get the clear FASTA file L with S sequences;
Initialize matrix of features m;
foreach *sequence(s) in file L* **do**
  apply PyFeat with 13 methods;
  for $th = 1 \rightarrow 13$ **do**
    Select features and store in the matrix m with row s;
  end
  Update the matrix m;
end
Get the final matrix of features$M$;
Apply the AB technique for feature selection, Select the features with the high score in matrix F;

Classification

The features of the DNA sequence are fed to the GBDT technique. This technique is used to diagnosis different test cases and predict the protein and lncRNA genes related to PD. The GBDT is considered one of the best efficient machine-learning technique, which is used for classification as well as regression problems. It is established from several decision trees, which the final result is achieved according to the summation of all trees' consequences. Via numerous iteration rounds, weak classifiers are generated in each iteration by GBDT, and each classifier is trained based on the gradient of classifiers in the previous round. The final classifier is founded based on the summation of weights for the weak classifiers, which are resulted in each round of training[28–30]. The model training is shown in the next steps:

1. The model is initialized as shown in equation (27).

$$h_0(x) = 0.5 * log\left(\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} 1 - y_i}\right) \tag{27}$$

   where N is the number of samples in training set, y_i is the real label of each sample.

2. The loss function is defined by equation (28).

$$F(y, h_{m-1}(x_i)) = log(1 + exp(-y h_{m-1}(x_i))) \tag{28}$$

   where y refers to the actual class, and $m(x)$ refers to the weak model in the m_{th} round.

3. For each round where $h = 1, 2, \cdots, H$

   (a) For i_{th} sample in m_{th} round, determine the negative gradient of the loss function as defined by equation (29).

$$r_{h,i} = \frac{y_i}{1 + (exp\ \ (y_i)\ \ h(x_i))}, i = 1, 2, \cdots, N. \tag{29}$$

   (b) Build the m_{th} decision tree, after that find the corresponding leaf node area $R_{m,j}$, where $j = 1, 2, \cdots, J$, which J is the number of leaf nodes in the tree.

   (c) For each leaf node's samples, th optimal outcome result of fitting the leaf node $(v_{m,j})$ is calculated by equation (30).

$$v_{m,j} = arg\,min_{v} \sum_{x \in R_{h,j}} log(1 + exp(-y_i h(x_i) + v)) \tag{30}$$

(d) Update h_{th} weak model by equation (31).

$$h_m = h_{m-1}(x) + r * \sum_{j=1}^{J} v_{m,j} I(x \in R_{m,j}) \tag{31}$$

where r is the learning rate with $0 < r \leq 1$, and $I(x \in R_{m,j})$ means that if x falls on the leaf node according to $R_{m,j}$, so that this corresponding term is equal 1.

(e) See whether M is lower than m. If M is more than m, then go to step (4) to finish the training. Otherwise, go to the step (1) for the next iteration.

4. The end of training with model H_m.

$$H_m(x) = h_0(x) + r * \sum_{m=1}^{M} \sum_{j=1}^{J} v_{m,j} I(x \in R_{m,j}) \tag{32}$$

We represent the algorithm for the proposed classification based on the GBDT technique as shown in Algorithm 2.

Algorithm 2: The proposed classification with GBDT.

Data: $D_{Train} = (x_1, y_1), (x_2, y_2), \cdots, (x_N, x_N)$ $D_{Test} = (x_1, y_1), (x_2, y_2), \cdots, (x_N, x_N)$
Result: Training model, and Prediction results
Use D_{Train} to train the GBDT;
Initialize the model as $h_0(x)$;
for $m = 1 \rightarrow M$ **do**
  for $i = 1 \rightarrow N$ **do**
    Compute the loss function: $F(y, h_{m-1}(x_i))$;
    Compute the residuals: $r_{h,i}$;
  end
  Build the m_{th} decision tree;
  Get the corresponding leaf node area $R_{m,j},\ j = 1, 2, \cdots, J.$;
  for $j = 1 \rightarrow J$ **do**
    get the $v_{m,j}$;
  end
  Update the weak classifier $h_m(x)$;
end
Get the final model $H_M(x)$;
Use D_{Test} to evaluate the prediction model;
for $s = 1 \rightarrow N$ **do**
  Process GBDT prediction model;
  Get the predicted label;
end
Calculate the evaluation metric based on the real label and the predicted label;

Experimental Results

This section represents the datasets description, hardware and software specifications, evaluation metrics, results, and discussion. In the results subsection, we divide the datasets into training and testing sets. Then, we use the training set and show the results of different feature extraction methods: five numerical representations with Fourier transform and the PyFeat method with AB feature selection technique. Then, we provide the overall proposed system based on proteins and lncRNAs datasets. After that, we represent some tables and figures supporting a target idea by employing seven performance metrics. Finally, we present an objective comparison of the proposed system with some literature studies in the discussion subsection. Also, we provide the strengths and weaknesses of the proposed system. Furthermore, a literature study can be used to verify the top-ranked predicted protein and lncRNA genes.

Datasets Description

This subsection describes the two utilized datasets, which are proteins and lncRNAs datasets.

- Proteins Dataset[12,31]: From the ClinVar, we downloaded proteins genes associated with PD. After removing the redundancies, we got 182 genes associated with PD as a positive case. Also, we got random 182 genes not associated with PD as negative case, as shown in Table 4.
- lncRNAs Dataset[32]: We downloaded lncRNAs genes associated with PD from the LncRNADisease v2.0. We got 137 genes associated with PD as a positive case. Also, we got random 137 genes not associated with PD as negative case as shown in Table 4.

Datasets	Site	Positive	Negative
Proteins	ClinVar	182	182
lncRNAs	LncRNADisease v2.0	137	137

Table 4. Datasets Description.

Hardware and Software Specifications

This subsection describes the specifications of the used software/hardware in our research. We developed this work using Python 3.7.6 and PyCharm 2019.3.3 with pandas, itertools, numpy, sklearn, and matplotlib libraries. We ran our system on a machine of core i7/4.5. It has 16 GB RAM and an NVIDIA GeForce GTX with 4 GB VRAM.

Evaluation Metrics

We used seven metrics for measuring the performance of our proposed system, which are accuracy (ACC), area under the curve (AUC), area under precision-recall curve (AUPR), F1-Score, Matthews correlation coefficient (MCC), sensitivity (SEN), and specificity (SPC)[33,34], which are defined by equations (33)$\cdots$ (40).

$$ACC = \frac{TN+TP}{TN+FP+TP+FN} \tag{33}$$

$$AUC = \frac{TP+TN}{TP+FP+FN+TN} \tag{34}$$

$$Precision = \frac{TP}{FP+TP} \tag{35}$$

$$Recall = SEN = \frac{TP}{FN+TP} \tag{36}$$

$$AUPR \approx 0.5(Precision+Recall) \tag{37}$$

$$F1-score = \frac{TP}{TP+0.5(FN+FP)} \tag{38}$$

$$MCC = \frac{TP.TN-FP.FN}{\sqrt{(FP+TP).(FN+TP).(FP+TN).(FN+TN)}} \tag{39}$$

$$SPC = \frac{TN}{FP+TN} \tag{40}$$

It is essential to clarify that true positive (TP) is the rate of the genes that are correctly predicted as PD-genes. True negative (TN) is the rate of the genes that are correctly predicted as not PD-genes. False positive (FP) is the rate of the genes that are incorrectly predicted as PD genes. Moreover, false negative (FN) is the rate of the genes that are incorrectly predicted as not PD-genes. ACC is the rate of the correct result over the total results based on TP and TN. It determines the accuracy of the proposed system. AUC is the average of the TP and FP rates at different classification thresholds, which is used as a summary of the ROC curve. The higher value of AUC gives the best performance when distinguishing between positive and negative genes of PD.

The precision is the rate of the correct predicted results over the amount of correct and incorrect prediction results, which the term "results" refers to the positive genes. The SEN or recall is the rate of the correct predicted results over the all correct predicted results, which the term "results" refers to the negative genes. Similar to ROC AUC, to define PR AUC, AUPR measures the average precision and recall under the ROC curve. The MCC is considered a contingency matrix method to calculate the Pearson product-moment correlation coefficient between actual and predicted values. SPC is the rate of the correct predicted results over the all correct predicted results, which the term "results" refers to the negative genes.

Results

To evaluate our proposed system, we applied 4-fold and 10-fold cross-validation techniques to validate the proteins and lncRNAs datasets and overcome the overfitting limitations. First, the most significant features are extracted by different feature extraction methods, such as binary, integer, real, Z-curve, and EIIP representations with Fourier transform. Besides, we applied the PyFeat method with AB technique for feature selection. Then, the selected features are fed to nine different classifiers for diagnosis, which are the GBDT, logistic regression (LR), decision tree (DT), Naive Bayes (NB), bagging, RF, AB, SVM, and linear discriminant analysis (LDA). Finally, we evaluated the results by using seven different performance measures.

For the proteins dataset, the most significant features are fed to nine classifiers. The GBDT classifier obtained the best results (6) followed by LR (four best results), RF (one best result), LDA (one best result), and DT, NB, Bagging, AB, and SVM (no better results), as showed in Table 5. Figs. 2 and 3 represent the classifiers' accuracy results using 4-fold and 10-fold cross-validation techniques with 5 Fourier representations and the PyFeat based on the proteins dataset, respectively. Similarly, for the lncRNAs dataset, the selected features are fed to nine classifiers. Again, the GBDT classifier obtained the best results (5) followed by SVM (two best results), LR (two best results), AB (one best results), RF (one best result), LDA (one best result) and DT, NB, Bagging (no better results), as showed in Table 6. Figs. 4 and 5 represent the classifiers' accuracy results using 4-fold and 10-fold cross-validation techniques with 5 Fourier representations and the PyFeat based on the proteins dataset, respectively.

Method	K-Fold	LR (%)	DT (%)	NB (%)	Bagging (%)	RF (%)	AB (%)	SVM (%)	LDA (%)	GBDT (%)
Binary	4	57.1	54.4	50.0	58.7	54.7	55.2	57.4	**58.8**	47.5
	10	**58.2**	53.3	49.4	56.0	57.4	55.2	58.0	58.0	55.8
Integer	4	60.4	54.4	50.9	54.4	61.2	60.4	61.2	61.5	**61.8**
	10	60.9	50.0	49.4	61.2	61.7	59.3	63.6	61.8	**62.0**
Real	4	**61.2**	54.9	52.7	60.1	57.7	56.0	56.3	60.4	57.1
	10	**60.7**	54.4	56.3	53.6	57.4	57.3	59.6	60.4	59.0
Z-curve	4	56.9	53.0	51.7	55.5	53.6	56.9	58.8	57.7	**59.0**
	10	56.0	54.1	51.4	55.3	50.3	52.5	57.6	56.9	**59.6**
EIIP	4	59.6	53.5	52.2	56.3	**61.8**	58.0	60.4	60.6	61.5
	10	**61.5**	58.5	51.9	59.8	56.5	56.6	59.7	58.4	56.6
PyFeat	4	67.1	60.8	51.0	66.6	72.6	71.0	69.3	55.6	**77.8**
	10	66.0	59.2	49.6	71.0	75.6	73.7	68.0	58.1	**76.2**

Table 5. The classifiers' accuracy using 4-fold and 10-fold cross-validation with 5 Fourier representation and the PyFeat based on proteins dataset.

As shown on Table 5 and Table 6, the GBDT technique gives the best results compared to other techniques. First, the most distinguishing and essential features are extracted using five numerical representations (binary, integer, real, Z-curve, and EIIP) with Fourier transform, besides the PyFeat method with AB technique as a feature selection. Second, these features are fed to the GBDT technique to diagnosis different positive or negative cases. Third, we evaluate the results using seven performance measures. It showed that the proposed PyFeat method with AB feature selection technique presented the best method for extracting the most significant and critical features based on proteins and lncRNAs datasets.

Proteins Dataset: Table 7 shows the performance evaluation of the proposed prediction system with five Fourier representation and the proposed PyFeat feature extraction methods. The proposed PyFeat feature extraction achieved the best results with the seven performance measures with 4-fold and 10-fold cross-validation based on the proteins dataset. For 4-fold cross-validation,

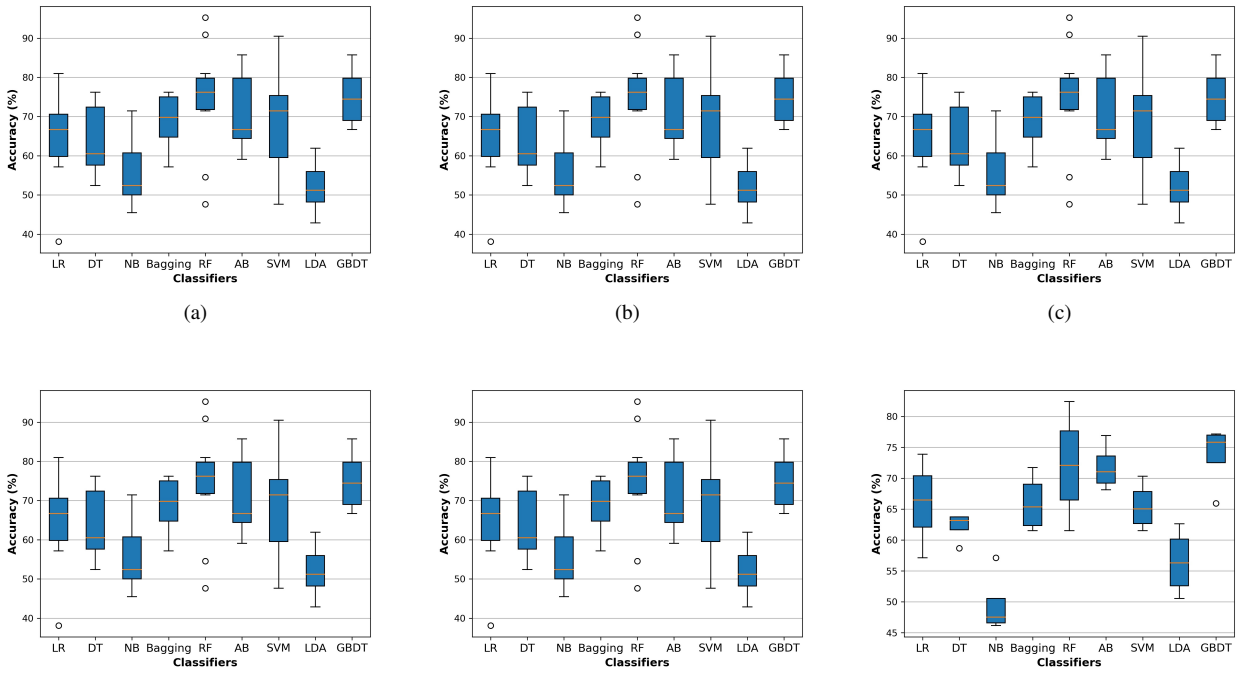

Figure 2. The classifiers' accuracy using 4-fold cross validation technique with feature extraction methods based on proteins dataset (a) Binary, (b) Integer, (c) Real, (d) Z-curve, (e) EIIP, and (f) PyFeat.

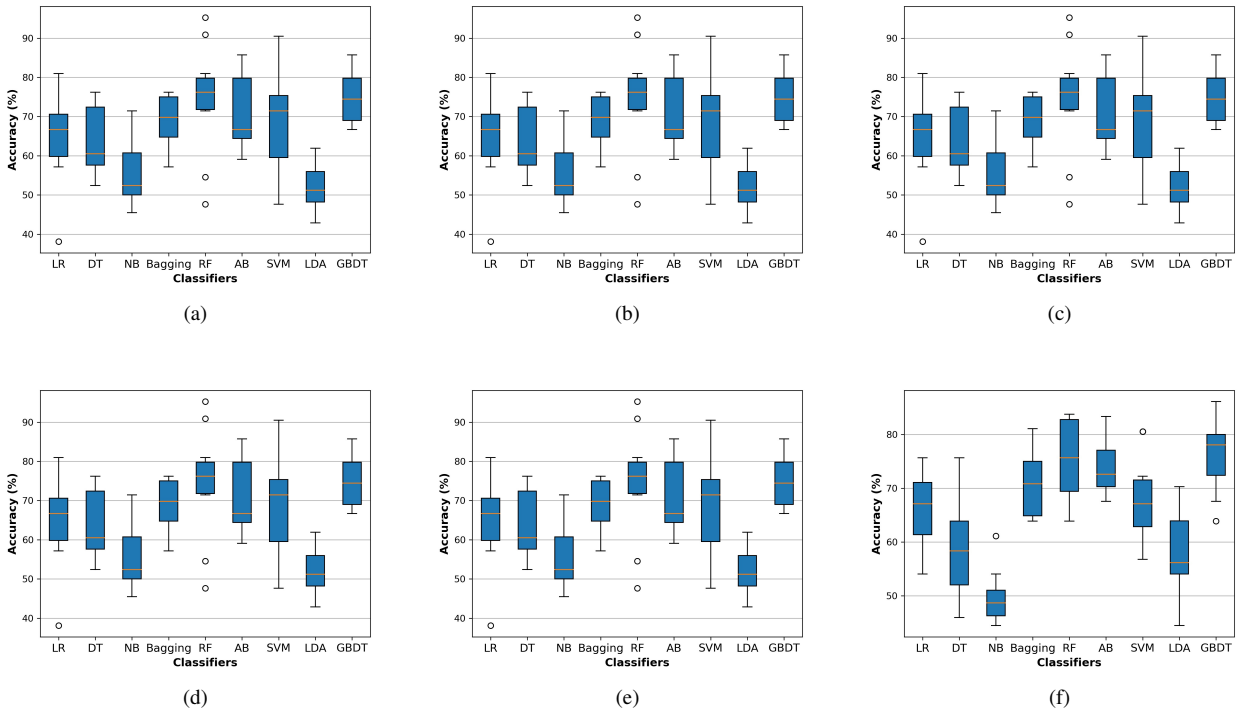

Figure 3. The classifiers' accuracy using 10-fold cross-validation technique with feature extraction methods based on proteins dataset (a) Binary, (b) Integer, (c) Real, (d) Z-curve, (e) EIIP, and (f) PyFeat.

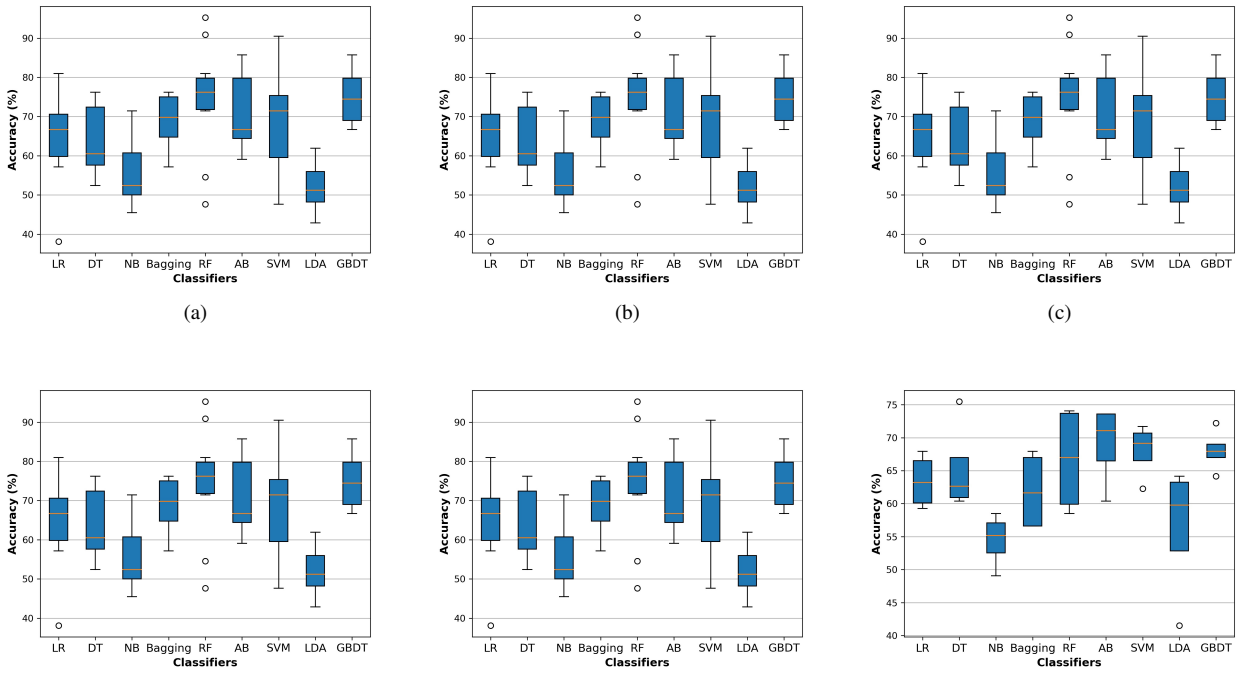

Figure 4. The classifiers' accuracy using 4-fold cross-validation with feature extraction methods based on lncRNAs dataset (a) Binary, (b) Integer, (c) Real, (d) Z-curve, (e) EIIP, and (f) PyFeat.

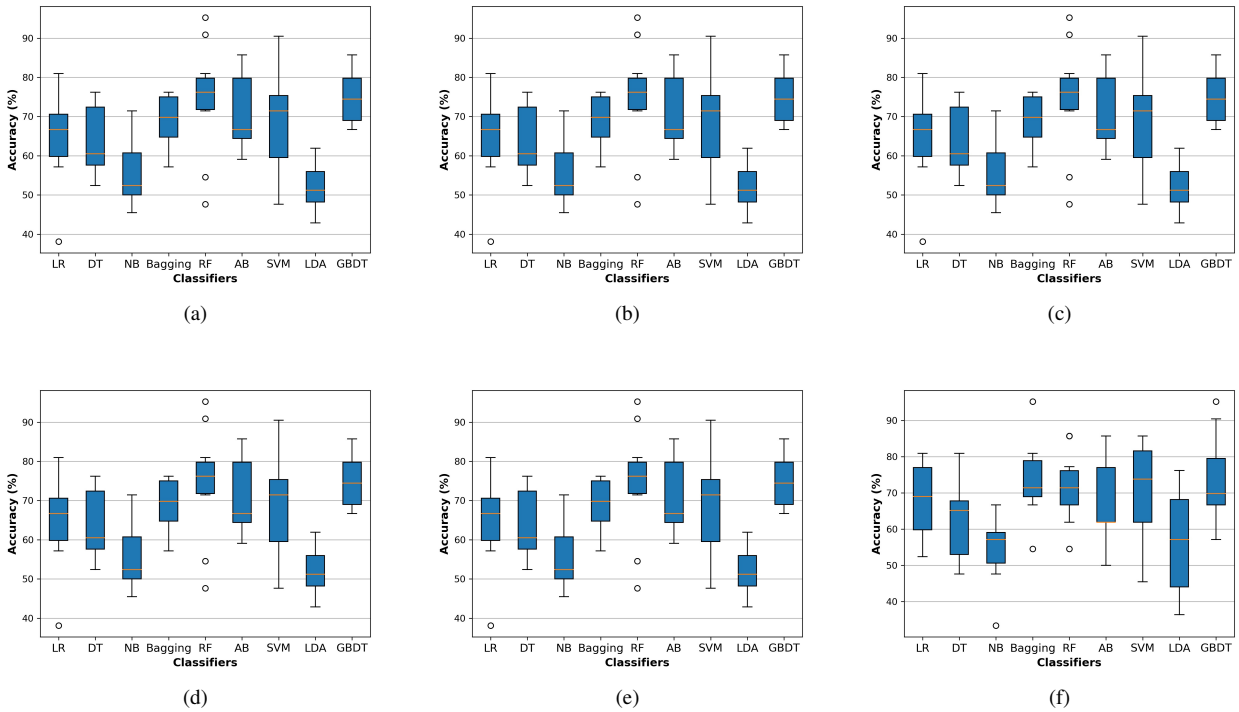

Figure 5. The classifiers' accuracy using 10-fold cross-validation technique with feature extraction methods based on proteins dataset (a) Binary, (b) Integer, (c) Real, (d) Z-curve, (e) EIIP, and (f) PyFeat.

Method	K-Fold	LR (%)	DT (%)	NB (%)	Bagging (%)	RF (%)	AB (%)	SVM (%)	LDA (%)	GBDT (%)
Binary	4	57.5	60.3	52.4	51.4	45.7	59.3	53.8	59.3	**61.7**
	10	57.8	53.0	51.9	52.2	51.5	56.1	55.5	58.9	**59.0**
Integer	4	58.4	51.5	52.8	57.5	58.4	52.4	**62.2**	56.6	60.2
	10	62.3	54.0	53.3	59.8	59.2	56.6	**62.6**	56.5	58.4
Real	4	58.9	55.6	52.8	54.7	57.0	56.6	56.1	49.5	**59.4**
	10	56.0	57.2	53.3	55.4	**58.5**	51.9	55.1	54.2	50.9
Z-curve	4	56.1	54.2	55.6	59.3	55.6	**59.4**	50.5	56.0	53.8
	10	**57.8**	54.1	52.8	53.8	52.4	57.1	56.4	53.7	55.9
EIIP	4	**58.4**	51.9	54.7	52.3	51.4	49.1	52.4	51.8	50.5
	10	54.3	48.6	53.3	52.3	52.9	46.3	56.5	**57.1**	56.1
PyFeat	4	64.8	61.4	54.0	67.6	72.4	75.1	70.0	61.0	**78.4**
	10	64.3	64.3	55.0	69.1	74.2	70.9	69.0	52.1	**74.7**

Table 6. The classifiers' accuracy using 4-fold and 10-fold cross-validation with 5 Fourier representation and the PyFeat based on lncRNAs dataset.

ACC equals 77.8%, AUC equals 83.8%, AUPR equals 77.8%, F1-score equals 77.8%, MCC equals 0.557, SEN equals 79.0%, and SPC equals 76.6%. For 10-fold cross validation, ACC equals 76.2%, AUC equals 84.4%, AUPR equals 83.9%, F1-score equals 76.5%, MCC equals 0.530, SEN equals 79.0%, and SPC equals 73.4%. Fig. 6 summarize the AUC for the proposed PyFeat and five Fourier representation with 4-fold and 10-fold cross-validation based on the proteins dataset.

Metric	K-Fold	ACC (%)	AUC (%)	AUPR (%)	F1-score (%)	MCC	SEN (%)	SPC (%)
Binary	4	47.5	51.2	52.9	49.0	0.050	50.3	44.8
	10	55.8	55.7	58.7	59.4	0.120	57.9	53.6
Integer	4	61.8	63.3	63.5	63.3	0.236	66.1	57.4
	10	62.0	64.9	67.1	59.4	0.247	56.8	67.2
Real	4	57.1	62.0	63.0	57.6	0.142	58.5	55.7
	10	55.8	55.7	58.7	59.4	0.120	57.9	53.6
Z-curve	4	59.0	61.5	62.2	58.6	0.181	57.9	60.1
	10	59.0	64.2	67.3	60.1	0.183	62.3	55.7
EIIP	4	61.5	63.7	66.4	60.0	0.230	57.9	66.0
	10	56.6	61.6	64.3	55.0	0.135	55.2	57.9
PyFeat	4	**77.8**	**83.8**	**83.0**	**77.8**	**0.557**	**79.0**	**76.6**
	10	**76.2**	**84.4**	**83.9**	**76.5**	**0.530**	**79.0**	**73.4**

Table 7. The performance evaluation of the prediction system with the proposed PyFeat and 5 Fourier representation methods based on the proteins dataset.

LncRNAs Dataset: Similarly for lncRNAs dataset, 4-fold cross-validation technique achieved ACC equals 78.4%, AUC equals 86.0%, AUPR equals 86.9%, F1-score equals 78.5%, MCC equals 0.571, SEN equals 79.2%, and SPC equals 77.6%. For 10-fold cross-validation, ACC equals 74.7%, AUC equals 83.1%, AUPR equals 84.7%, F1-score equals 74.6%, MCC equals 0.507, SEN equals 75.5%, and SPC equals 73.8%, as shown in Table 8. Fig. 7 summarize the AUC for the proposed PyFeat and five Fourier representation with 4-fold and 10-fold cross-validation techniques based on the lncRNAs dataset.

Based on Tables 7 and 8, the 4-fold and 10-fold cross-validation represented that results are very close to each other based on the proteins and lncRNAs dataset. In Table 9, the performance evaluation of the proposed prediction system based on the two datasets is shown with 4-fold cross-validation technique. We noticed that the proteins dataset achieved ACC equals 77.8%, AUC equals 83.8%, AUPR equals 77.8%, F1-score equals 77.8%, MCC equals 0.557, SEN equals 79.0%, and SPC equals 76.6%. Based on lncRNAs dataset, ACC equals 78.4%, AUC equals 86.0%, AUPR equals 86.9%, F1-score equals 78.5%, MCC equals 0.571, SEN equals 79.2%, and SPC equals 77.6%. Fig. 8 summarizes the results based on the proteins dataset, lncRNAs dataset, and the average results of the proposed prediction system as demonstrated in Table 9.

We compared the proposed system with Peng et al.[12], Peng et al.[21], and Lei et al.[10]. The proposed system achieved better results. The AUC for Peng et al.[12], Peng et al.[21], and Lei et al.[10] equals%72.9, 78.6%, and 79.0%, respectively. On the other hand, the proposed system achieved AUC equals 84.9%, as shown in Table 10. Fig. 9 represents the chart of the comparison among the AUC of our proposed prediction system and other systems.

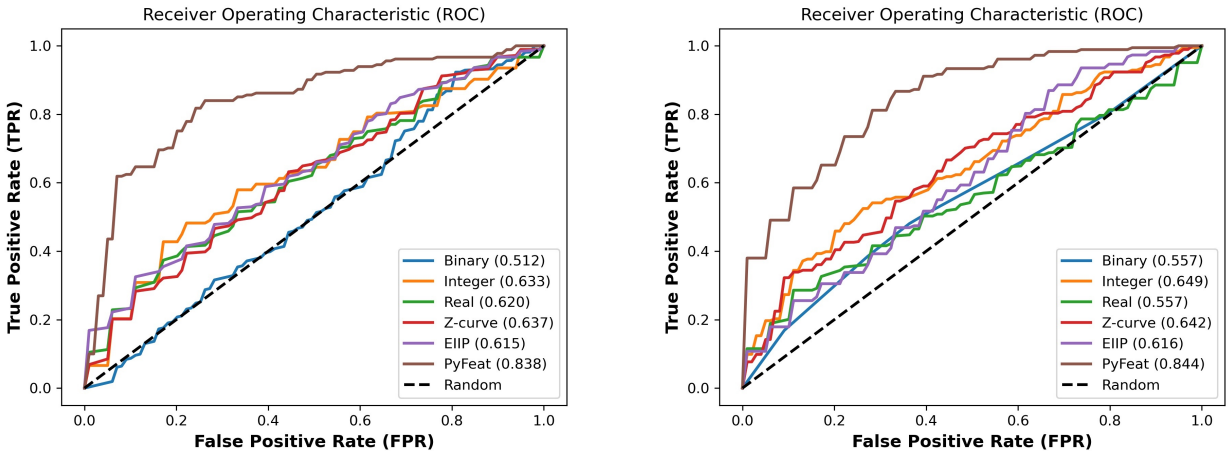

Figure 6. The AUC for the proposed PyFeat and 5 Fourier representations based on the proteins dataset with (a) 4-fold and (b) 10-fold cross-validation techniques.

Metric	K-Fold	ACC (%)	AUC (%)	AUPR (%)	F1-score (%)	MCC	SEN (%)	SPC (%)
Binary	4	61.7	60.3	58.3	61.7	0.235	62.6	60.7
	10	59.0	62.1	62.9	58.7	0.183	58.9	58.9
Integer	4	60.2	64.0	67.3	61.9	0.206	64.5	56.1
	10	58.4	60.9	64.7	59.7	0.180	61.7	55.1
Real	4	59.4	60.5	62.2	57.3	0.194	55.1	63.6
	10	50.9	56.1	59.0	50.6	0.021	51.4	50.5
Z-curve	4	53.8	54.8	57.0	55.9	0.077	59.8	47.7
	10	55.9	60.0	63.1	57.2	0.127	59.8	52.3
EIIP	4	50.5	53.7	59.2	49.8	0.009	49.5	51.4
	10	56.1	54.9	62.2	54.5	0.130	53.3	58.9
PyFeat	4	**78.4**	**86.0**	**86.9**	**78.5**	**0.571**	**79.2**	**77.6**
	10	**74.7**	**83.1**	**84.7**	**74.6**	**0.507**	**75.5**	**73.8**

Table 8. The performance evaluation of the prediction system with the proposed PyFeat and 5 Fourier representation methods based on the lncRNAs dataset.

Discussion

PD is considered the most common movement disease and the second most common neurodegenerative disease after AD. There are several cardinal signs associated with PD, which are tremor, rigidity, bradykinesia, and post instability. So that, we need to diagnose the disease early to avoid these symptoms. Identifying and predicting genes related to disease has biological significance in most biomedical studies, which aid in an early diagnosis and treatment of disease. As a result, identifying genes related to PD is crucial to the disease's diagnosis and treatment. The most recent studies for PD-genes prediction utilize the genes of the proteins and discard lncRNA genes related to the PD. However,lncRNAs are essential in the metastasis and progression of various diseases. As a result, we build our proposed prediction system for identifying protein as well as lncRNA genes related to PD.

In this study, we utilized two datasets for the protein and lncRNA genes, then we represented all genes as DNA FASTA sequences and removed the sequences redundancies. We used 4-fold and 10-fold cross-validation to evaluate the proposed system. The most critical features are extracted by using the PyFeat method with AB as a feature selection technique. This proposed method is achieved the best results compared with the other five numerical representations with Fourier transform. The most significant features are fed to the GBDT technique to diagnose different test cases and build our model to identify genes related to PD.

We evaluate the results using seven performance evaluation measures. For 4-fold cross-validation technique, the proteins

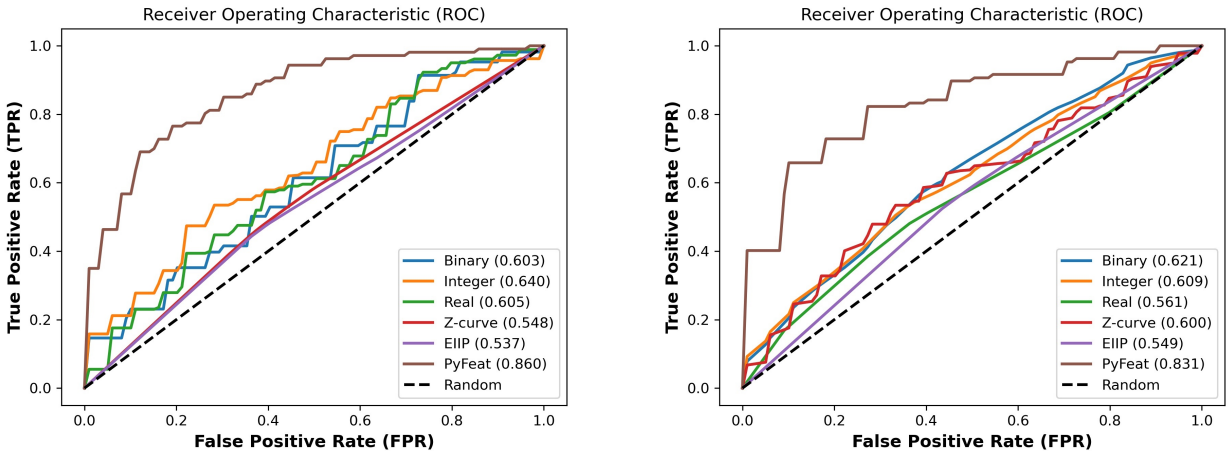

Figure 7. The AUC for the proposed PyFeat and 5 Fourier representations based on the lncRNAs dataset with (a) 4-fold and (b) 10-fold cross-validation techniques.

Datasets	ACC (%)	AUC (%)	AUPR (%)	F1-score (%)	MCC	SEN (%)	SPC (%)
Proteins	77.8	83.8	83.0	77.8	0.557	79.0	76.6
LncRNAs	78.4	86.0	86.9	78.5	0.571	79.2	77.6
Average	78.1	84.9	85.0	78.2	0.564	79.1	77.1

Table 9. Average performance of the proposed prediction system based on the proteins and lncRNAs datasets.

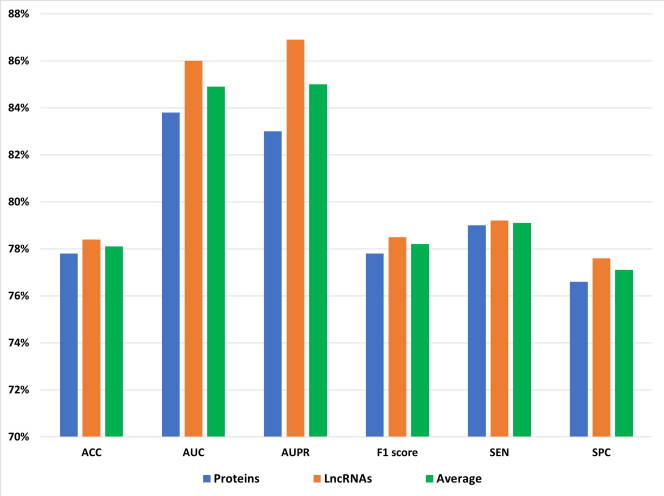

Figure 8. The average performance evaluation of the proposed prediction system based on proteins and lncRNAs datasets.

dataset achieved ACC equals 77.8%, AUC equals 83.8%, AUPR equals 77.8%, F1-score equals 77.8%, MCC equals 0.557, SEN equals 79.0%, and SPC equals 76.6%. On lncRNAs dataset, ACC equals 78.4%, AUC equals 86.0%, AUPR equals 86.9%, F1-score equals 78.5%, MCC equals 0.571, SEN equals 79.2%, and SPC equals 77.6%. The average results based on the proteins and lncRNAs dataset are ACC equals 78.1%, AUC equals 84.9%, AUPR equals 85.0%, F1-score equals 78.2%, MCC equals 0.564%, SEN equals 79.1%, and SPC equals 77.1%.

Finally, we use the proposed prediction system to predict new protein and lncRNA genes related to PD, which is not founded in the databases. These genes are ranked according to the probability predicted by the training model. Then the top ten protein and lncRNA genes are selected, and the literature review is utilized to verify these genes. For proteins, the ten genes are

Studies	AUC (%)
Peng et al.[21]	72.9
Lie et al.[10]	78.6
Peng et al.[12]	79.0
The Proposed System	**84.9**

Table 10. The comparison between our proposed system and some current systems based on AUC.

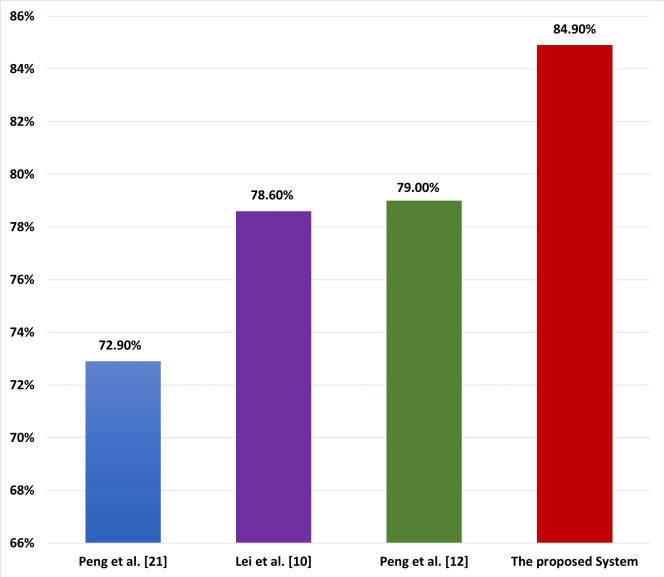

Figure 9. The comparison between our proposed system and some current studies based on AUC.

extracted: PACRG, GIA5, TH, LRRK2, TNR, VCP, KCNJ2, SETX, APBB1, and DCTN1. Based on the literature review, we discover that some of these genes have been reported to be associated with PD. PACRG, TH, LRRK2, TNR, and VCP are reported in[35–41]. In addition, KCNJ2, APBB1, and DCTN1 genes are associated with neurodegenerative diseases[42–44]. The GJAS gene is related to a gene associated with PD[45]. Finally, the SETX gene is related to the tremor which is determined as a sign of PD[46].

For LncRNAs, the ten genes are extracted: PDZRN3, NEAT1, DAOA-AS1, TUG1, PPP3CB, DAPK1, H19, MAPT-AS1, MESTIT1, and PCA3. Based on the literature review, we discover that some of these genes have been reported to be associated with PD. NEAT1, TUG1, DAPK1, H19, MATP-AS1, and PCA3 genes are reported in[47–52]. In addition, PDZRN3 and PPP3CB genes are associated with neurodegenerative diseases as reported in[53,54]. The MESTIT1 gene is associated with the cognitive disease as reported in[55]. Finally, the DAOA-AS1 gene is extracted for bipolar disorder as reported in[56].

Conclusion

We developed a novel prediction system for identifying genes related to PD that genes not only proteins but also lncRNAs. We used two public databases, which are ClinVar for proteins and LncRNADisease V2.0 for lncRNAs. The proposed prediction system consists of four steps. First, we represented the genes as DNA FASTA sequences from the UCSC genome browser and removed the redundancies as a preprocessing step. Second, we extract the most significant features of DNA FASTA sequences using the proposed PyFeat method with AB as a feature selection technique. Then, the selected features are fed to the GBDT technique to diagnosis different test cases. Finally, seven performance metrics are used to evaluate the results of the proposed system. In the future, we aim to apply our proposed prediction system to identify and predict other diseases, which have related genes.

References

1. Delenclos, M., Jones, D. R., McLean, P. J. & Uitti, R. J. Biomarkers in parkinson's disease: Advances and strategies. *Park. & related disorders* **22**, S106–S110 (2016).

2. Bazazeh, D., Shubair, R. M. & Malik, W. Q. Biomarker discovery and validation for parkinson's disease: A machine learning approach. *2016 Int. Conf. on Bio-engineering for Smart Technol. (BioSMART)* 1–6 (2016).
3. Krishnagopal, S., Coelln, R. v., Shulman, L. M. & Girvan, M. Identifying and predicting parkinson's disease subtypes through trajectory clustering via bipartite networks. *PloS one* **15**, e0233296 (2020).
4. Klein, C. & Westenberger, A. Genetics of parkinson's disease. *Cold Spring Harb. perspectives medicine* **2**, a008888 (2012).
5. Redenšek, S., Trošt, M. & Dolžan, V. Genetic determinants of parkinson's disease: can they help to stratify the patients based on the underlying molecular defect? *Front. aging neuroscience* **9**, 20 (2017).
6. Babu, G. S. & Suresh, S. Parkinson's disease prediction using gene expression–a projection based learning meta-cognitive neural classifier approach. *Expert. Syst. with Appl.* **40**, 1519–1529 (2013).
7. Adler, C. H. *et al.* Low clinical diagnostic accuracy of early vs advanced parkinson disease: clinicopathologic study. *Neurology* **83**, 406–412 (2014).
8. Santaella, A. *et al.* Inflammation biomarker discovery in parkinson's disease and atypical parkinsonisms. *BMC neurology* **20**, 1–8 (2020).
9. He, R. *et al.* Recent advances in biomarkers for parkinson's disease. *Front. aging neuroscience* **10**, 305 (2018).
10. Lei, X. & Zhang, Y. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network. *Inf. Sci.* **479**, 386–400 (2019).
11. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of parkinson's disease. *The Lancet Neurol.* **19**, 170–178 (2020).
12. Peng, J., Guan, J. & Shang, X. Predicting parkinson's disease genes based on node2vec and autoencoder. *Front. genetics* **10**, 226 (2019).
13. Xuan, P., Cao, Y., Zhang, T., Kong, R. & Zhang, Z. Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncrna genes. *Front. genetics* **10**, 416 (2019).
14. Ding, L., Wang, M., Sun, D. & Li, A. Tpglda: Novel prediction of associations between lncrnas and diseases via lncrna-disease-gene tripartite graph. *Sci. reports* **8**, 1–11 (2018).
15. Muhammod, R. *et al.* Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences. *Bioinformatics* **35**, 3831–3833 (2019).
16. Radivojac, P. *et al.* An integrated approach to inferring gene–disease associations in humans. *Proteins: Struct. Funct. Bioinforma.* **72**, 1030–1037 (2008).
17. Yang, P., Li, X., Chua, H.-N., Kwoh, C.-K. & Ng, S.-K. Ensemble positive unlabeled learning for disease gene identification. *PloS one* **9**, e97079 (2014).
18. Hwang, W.-Y. Biological feature selection and disease gene identification using new stepwise random forests. *Ind. Eng. Manag. Syst.* **16**, 64–79 (2017).
19. Bonidia, R. P., Sampaio, L. D. H., Lopes, F. M. & Sanches, D. S. Feature extraction of long non-coding rnas: A fourier and numerical mapping approach. *Iberoamerican Congr. on Pattern Recognit.* 469–479 (2019).
20. Zhang, J., Ni, S., Parvin, J., Yang, Y. & Huang, K. Predicting parkinson's disease related genes using frequent gene co-expression analysis. *2011 IEEE Int. Conf. on Bioinforma. Biomed. Work. (BIBMW)* 1042–1044 (2011).
21. Peng, J. *et al.* Predicting disease-related genes using integrated biomedical networks. *BMC genomics* **18**, 1–11 (2017).
22. Yang, K. *et al.* Pdgnet: Predicting disease genes using a deep neural network with multi-view features. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* (2020).
23. Bi, X.-a., Hu, X., Xie, Y. & Wu, H. A novel cernne approach for predicting parkinson's disease-associated genes and brain regions based on multimodal imaging genetics data. *Med. Image Analysis* **67**, 101830 (2021).
24. Bonidia, R. P. *et al.* Feature extraction approaches for biological sequences: a comparative study of mathematical features. *Briefings Bioinforma.* (2020).
25. Chakravarthy, N., Spanias, A., Iasemidis, L. D. & Tsakalis, K. Autoregressive modeling and feature analysis of dna sequences. *EURASIP J. on Adv. Signal Process.* **2004**, 1–16 (2004).
26. Zhang, R. & Zhang, C.-T. Z curves, an intutive tool for visualizing and analyzing the dna sequences. *J. Biomol. Struct. Dyn.* **11**, 767–782 (1994).

27. Nair, A. S. & Sreenadhan, S. P. A coding measure scheme employing electron-ion interaction pseudopotential (eiip). *Bioinformation* **1**, 197 (2006).

28. Wang, J., Kuang, Z., Ma, Z. & Han, G. Gbdtl2e: Predicting lncrna-ef associations using diffusion and hetesim features based on a heterogeneous network. *Front. genetics* **11**, 272 (2020).

29. Qiu, W., Lv, Z., Hong, Y., Jia, J. & Xiao, X. Bow-gbdt: A gbdt classifier combining with artificial neural network for identifying gpcr–drug interaction based on wordbook learning from sequences. *Front. Cell Dev. Biol.* **8**, 1789 (2021).

30. Yu, Z. *et al.* Predicting adverse drug events in chinese pediatric inpatients with the associated risk factors: a machine learning study. *Front. Pharmacol.* **12**, 516 (2021).

31. Landrum, M. J. *et al.* Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862–D868 (2016).

32. Chen, G. *et al.* Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids research* **41**, D983–D986 (2012).

33. Vihinen, M. How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis. In *BMC genomics*, vol. 13, 1–10 (BioMed Central, 2012).

34. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1–13 (2020).

35. Stephenson, S. E. *et al.* Generation and characterisation of a parkin-pacrg knockout mouse line and a pacrg knockout mouse line. *Sci. reports* **8**, 1–11 (2018).

36. Nagatsu, T., Nakashima, A., Ichinose, H. & Kobayashi, K. Human tyrosine hydroxylase in parkinson's disease and in related disorders. *J. Neural Transm.* **126**, 397–409 (2019).

37. Bryant, N. *et al.* Identification of lrrk2 missense variants in the accelerating medicines partnership parkinson's disease cohort. *Hum. molecular genetics* (2021).

38. Castro, S. L. *et al.* The industrial solvent trichloroethylene induces lrrk2 kinase activity and dopaminergic neurodegeneration in a rat model of parkinson's disease. *bioRxiv* (2020).

39. Sáenz-Farret, M., Munhoz, R. P., Fasano, A. & Zúñiga-Ramírez, C. Tnr gene mutation in familial parkinson's disease: Possible implications for essential tremor. *J. Mov. Disord.* (2020).

40. Alieva, A. *et al.* Vcp expression decrease as a biomarker of preclinical and early clinical stages of parkinson's disease. *Sci. reports* **10**, 1–9 (2020).

41. Majounie, E. *et al.* Mutational analysis of the vcp gene in parkinson's disease. *Neurobiol. aging* **33**, 209–e1 (2012).

42. Ferraris, C. *et al.* Association between sour taste snp kcnj2-rs236514, diet quality and mild cognitive impairment in an elderly cohort. *Nutrients* **13**, 719 (2021).

43. Groh, M., Albulescu, L. O., Cristini, A. & Gromak, N. Senataxin: genome guardian at the interface of transcription and neurodegeneration. *J. molecular biology* **429**, 3181–3195 (2017).

44. Konno, T. *et al.* Dctn1-related neurodegeneration: Perry syndrome and beyond. *Park. & related disorders* **41**, 14–24 (2017).

45. Kelm-Nelson, C. A. & Gammie, S. Gene expression within the periaqueductal gray is linked to vocal behavior and early-onset parkinsonism in pink1 knockout rats. *BMC genomics* **21**, 1–13 (2020).

46. Oyama, G. *et al.* Deep brain stimulation for tremor associated with underlying ataxia syndromes: a case series and discussion of issues. *Tremor other hyperkinetic movements* **4** (2014).

47. Simchovitz, A. *et al.* Neat1 is overexpressed in parkinson's disease substantia nigra and confers drug-inducible neuroprotection from oxidative stress. *The FASEB J.* **33**, 11223–11234 (2019).

48. Cheng, J. *et al.* The role of lncrna tug1 in the parkinson disease and its effect on microglial inflammatory response. *NeuroMolecular Medicine* 1–8 (2020).

49. Lu, Y. *et al.* Lncrna malat1 targeting mir-124-3p regulates dapk1 expression contributes to cell apoptosis in parkinson's disease. *J. cellular biochemistry* **121**, 4838–4848 (2020).

50. Zhang, Y., Xia, Q. & Lin, J. Lncrna h19 attenuates apoptosis in mptp-induced parkinson's disease through regulating mir-585-3p/pik3r3. *Neurochem. research* **45**, 1700–1710 (2020).

51. Coupland, K. G. *et al.* Role of the long non-coding rna mapt-as1 in regulation of microtubule associated protein tau (mapt) expression in parkinson's disease. *PLoS One* **11**, e0157924 (2016).

52. Boros, F. A., Maszlag-Török, R., Vécsei, L. & Klivényi, P. Increased level of neat1 long non-coding rna is detectable in peripheral blood cells of patients with parkinson's disease. *Brain research* **1730**, 146672 (2020).

53. Lv, Q., Wang, Z., Zhong, Z. & Huang, W. Role of long noncoding rnas in parkinson's disease: Putative biomarkers and therapeutic targets. *Park. Dis.* **2020**, 5374307–5374307 (2020).

54. Ding, M. & Shen, K. The role of the ubiquitin proteasome system in synapse remodeling and neurodegenerative diseases. *BioEssays: news reviews molecular, cellular developmental biology* **30**, 1075 (2008).

55. Peter, C. J. *et al.* Dna methylation signatures of early childhood malnutrition associated with impairments in attention and cognition. *Biol. psychiatry* **80**, 765–774 (2016).

56. Sayad, A., Badrlou, E., Ghafouri-Fard, S. & Taheri, M. Association analysis between the rs1899663 polymorphism of hotair and risk of psychiatric conditions in an iranian population. *J. Mol. Neurosci.* 1–6 (2020).