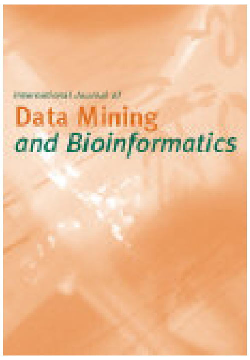
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

Diagnosis of Parkinson's disease genes using LSTM and MLP-based multi-feature extraction methods

Priya Arora, Ashutosh Mishra, Avleen Malhi

DOI: 10.1504/IJDMB.2023.10058573

Article History:

Received:	14 April 2023
Last revised:	24 May 2023
Accepted:	19 June 2023
Published online:	17 October 2023

Diagnosis of Parkinson’s disease genes using LSTM and MLP-based multi-feature extraction methods

Priya Arora

Chitkara University Institute of Engineering and Technology,
Chitkara University,
Punjab, India
Email: priya.sadana@chitkara.edu.in

Ashutosh Mishra*

Department of Computer Science and Engineering,
Thapar Institute of Engineering and Technology,
Patiala, Punjab, India
Email: ashutosh.mishra@thapar.edu
*Corresponding author

Avleen Malhi

Data Science and AI,
Bournemouth University, UK
Email: amalhi@bournemouth.ac.uk

Abstract: Disease gene identification using computational methods is one of the most challenging issues to improve the treatment and diagnosis of Parkinson’s disease (PD). Various intelligent computing techniques have been introduced to predict disease associated genes but the major difference among these approaches is in the data type to be used to create a feature vector. In this paper, deep learning methods such as multi-layer perceptron (MLP) and long short-term memory (LSTM) are adopted to identify genes that are responsible for Parkinson’s disease. The proposed method has been optimised on the bases of: 1) amino acid’s physicochemical properties to construct a feature vector; 2) feature extraction method to reduce the effect of noise and to speed up the process; 3) the genes prediction is done by employing deep learning methods. Compared with different type of datasets and other classifiers, the proposed method improves the prediction performance of neurodegenerative diseases. The experimental results indicate that the proposed deep learning approach outperforms the existing gene identification methods with higher recall, precision and F-score of 88.2, 84.5 and 85.0, respectively. The results of proposed system indicate the efficiency and accuracy for Parkinson’s disease gene identification and classification.

Keywords: disease gene; deep learning models; feature extraction; physicochemical properties of amino acid; Parkinson’s disease.

Reference to this paper should be made as follows: Arora, P., Mishra, A. and Malhi, A. (2023) ‘Diagnosis of Parkinson’s disease genes using LSTM and MLP-based multi-feature extraction methods’, *Int. J. Data Mining and Bioinformatics*, Vol. 27, No. 4, pp.326–348.

Biographical notes: Priya Arora is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Chitkara University, Rajpura. Her area of research interests includes data science, machine learning, deep learning, and computational biology.

Ashutosh Mishra is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, India. His research interests include software engineering, data mining, semantic web technologies and cognitive computation.

Avleen Malhi is a Senior Lecturer in Data Science and AI in the Department of Computing and Informatics at Bournemouth University, UK, and visiting researcher in the Department of Computer Science at Aalto University, Finland.

1 Introduction

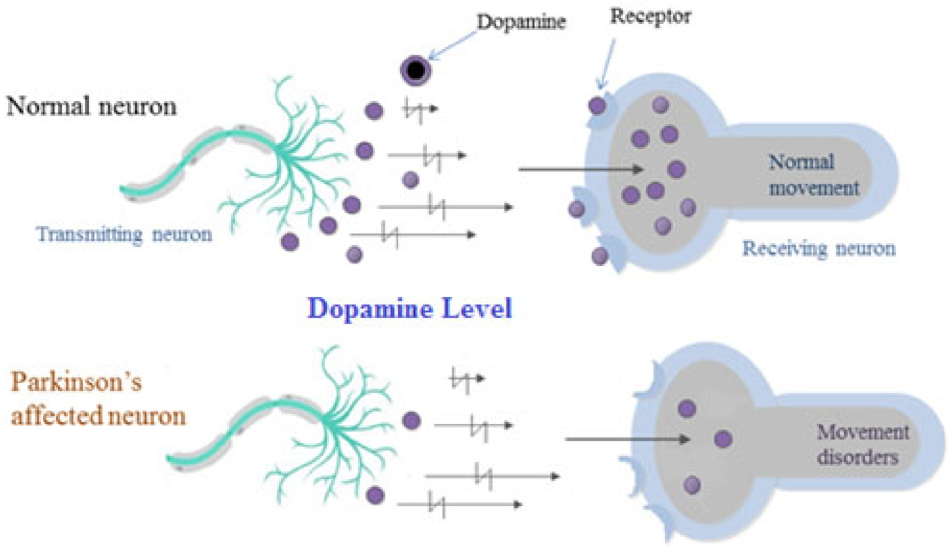
Parkinson's disease (PD) is a progressive neurodegenerative disease associated with central nervous system affected by the loss of a neuro-transmitter called dopamine. The existing neurons in the brain are responsible for the production of dopamine. The level of dopamine is reduced when the neurons die, which causes the movement problems in Parkinson's (Abdukodirov et al., 2022). When the level of dopamine decreases, symptoms such as slowness, tremor, and stiffness occur. People with PD have lower dopamine levels than healthy people. Dopamine level in healthy and Parkinson's affected neuron is shown in Figure 1. James Parkinson was the first person who announced PD as a 'shaking palsy' in 1817 (Draoui et al., 2020). PD is the second most common neurological disorder in adults after Alzheimer's.

The symptoms of PD include muscle rigidity, tremors, slow movement, and abnormalities in speaking and writing (Pereira et al., 2016). PD is more prevalent in the elderly, with an average onset age of 60. Studies reveal that genetic and environmental factors, oxidative stress, and age play a significant role in the progressive mortality of dopaminergic neurons, although the precise causes of PD are unknown.

There are many experimental approaches that have been introduced in recent years to identify disease genes from vast number of candidate genes. These techniques differ in the genomic data type used to generate feature vectors, such as protein-protein interactions (PPIs) (Yang et al., 2011; Zhang et al., 2011; Madeddu et al., 2020), gene expression profiles (Yousef and Charkari, 2015), protein and biological functions. Unfortunately, all these techniques are based on the information of proteins achieved from protein domains, gene ontology, and, PPI data. Hence, might not be implemented accurately as information is incomplete, noisy, and time consuming. Data that can be used for all proteins and has significant role in solving issues such as PPIs (Yu et al., 2010; Yousef and Charkari, 2013), predicting a subcellular locations (Fukasawa et al., 2014) is the protein sequences.

Several methods for anticipating disease genes have been presented to date (Köhler et al., 2008; Miao et al., 2017; Jowkar and Mansoori, 2016; Danaee et al., 2017). However, only a small percentage of them are used to locate the PD gene. Yousef and Charkari (2015) effectively applied their proposed method to a subset of 50 PD-related proteins (genes). They advocated for a greater emphasis on the physicochemical properties of amino acids to improve efficacy. Given the small size of their dataset, their estimates are typically conservatively optimistic. Amino acids must possess a variety of physicochemical properties for broad spectrum analysis. However, an exhaustive investigation of all the genes involved in PD has not yet been conducted. This study extracted a feature vector from 12 amino acid physicochemical parameters. Due to the incorporation of a broader spectrum of physicochemical parameters, we are therefore able to provide more information regarding the interactions.

Figure 1 Healthy and Parkinson's affected neuron (see online version for colours)



1.1 Contributions

This research proposes a method for identifying proteins (genes) that are responsible for PD by using protein sequences. In this paper, the multi-layer perceptron (MLP) and long short-term memory (LSTM) deep learning models are used to identify PD genes.

The following are the main contributions of this paper are as follows.

1 Deep learning methods are proposed for identifying PD genes using protein sequences as prior knowledge.

2 Twelve physicochemical properties of amino acids are used to extract the features from protein sequences.

3 Forward selection and backward elimination (FSBE) feature reduction method is used to extract vital and distinguish features.

4 A comparative study with existing systems is carried out to show the effectiveness of our proposed model.

The structure of the article is as follows. Section 2 briefly summarises research done in the gene identification field. Section 3 introduces the architecture and method of our proposed system. The experimental results obtained after implementing our proposed model are introduced in Section 4. The conclusion of the article and its future scope are summarised in Section 5.

2 Related work

Numerous methods have been introduced to identify genes associated with the disease based on the variety of data such as biological data, gene sequence, functional annotation, evolutionary features, gene expression profile, and PPI data. In this section, an overview of these methods for gene identification is presented.

Adie et al. (2005) applied decision tree based on various genome sequences, such as evolutionary conservation, coding sequence length, etc. They proved that topological information in the PPI network is useful for disease gene prediction. Xu and Li (2006) applied K-nearest neighbour (KNN) to identify disease-associated genes using PPI topological features. Smalter et al. (2007) employed support vector machine (SVM) classifier on data generated from PPI topological properties, sequence-derived, and evolutionary features to predict disease genes whereas Radivojac et al. (2008) presented a combined method with three individual SVM classifiers to predict disease genes. Jiang et al. (2017) presented an identification method to identify Alzheimer's disease genes (ADG) using a two-stage cascading classifier method. This method shows better performance by combining ReliefF, a feature selection method and, developed a two-stage classifier method based on majority voting of three methods including, random forest (RF), SVM and extreme learning machine (ELM). Mordelet and Vert (2011) introduced a method named ProDiGe to prioritise disease associated genes using unlabelled and positive samples. The authors had integrated variety of gene related information to create feature vector such as PPI data, protein functional information, and protein sequences. Finally, SVM classifier is employed -to differentiate positive samples from random samples (RS). Yang et al. (2012) introduced a PUDI method to identify genes and used data from gene ontology, PPI network, and protein domains biological networks. They had separated the unknown genes set into various subsets such as likely positive, weakly negative, likely negative, and reliable negative, based on the similarity measures. Finally, according to the likelihood of whether they are negative or positive, they assume that different weights are presented to the multi-layer weighted SVM. Yang et al. (2014) extended their previous work and proposed an EPU method for disease gene identification. Along with previous data, they included gene expression and phenotype similarity networks data also. Yousef and Charkari (2015) developed a method to identify disease genes using both positive and negative data as one class classification. They used principal component analysis (PCA) to reduce dimensionality and applied support vector data description (SVDD) method to train the model. A comparative analysis of various state-of-art methods for gene identification with their merits and demerits is presented in Table 1.

The methods described in Table 1 are based on protein information obtained from previous knowledge such as gene ontology, protein domains and PPI network that may contain incomplete information or contains some errors, so might not be employed properly. Hence, a universal knowledge is vital to handle this issue. Data that is available for all proteins is the protein sequences and has significant role in solving problems including, PPIs, protein functional and structural classes, etc. In the present work, same has been used.

Table 1 Comparative analysis of state-of-art methods

Authors	*Proposed method*	*Merits*	*Demerits*
Xu and Li (2006)	Introduced a method using KNN to predict genes that are more likely to be involved in disease	Used topological properties in PPI network	Types of function-related or sequence features have been ignored.
Smalter et al. (2007)	Proposed a classification system based on topological features of genes.	Used topological and sequence based features	Data source contains error and missing values
Mordelet and Vert (2011)	Proposed a method named ProDiGe to prioritise disease genes	Randomly selected negative samples from unknown genes	Random negative set still suffers from noisy data
Yang et al. (2014)	Ensemble based method using KNN, naïve Bayes and multi-level support vector machine is proposed to integrate biological sources to identify disease genes.	Applied Euclidean distance to select negative samples from unknown genes	High dimensional feature vectors (4,000 features)
Yousef and Charkari (2015)	A one-class classification method using support vector data description (SVDD) was built to identify disease genes	Protein sequences were used to extract features	Trained the model using only positive samples
Peng et al. (2019)	Introduced a PD gene identification method using Node2vec and auto-encoder with SVM classifier	Node2vec is used to extract genes features based on network	Semantic meaning is not captured
Miao et al. (2017)	Developed an Alzheimer's disease (AD) related genes identification method based on majority voting of support vector machine (SVM), random forest (RF) and extreme learning machine (ELM) methods	Cascading classifiers and majority voting gives higher sensitivity and specificity	High dimensional microarray data

Table 2 shows the comparative analysis of recent machine learning and deep learning methods. Bi et al. (2021) combined data from functional magnetic resonance imaging (fMRI) and single nucleotide polymorphisms (SNPs) to create a realistic multimodal analysis model. The model was comprised of three steps. To begin, they employed correlation analysis to construct the subject's fusion. Second, they used their neural network as a clustering evolutionary random neural network ensemble (CERNNE) to analyse the fusion characteristic. Finally, for optimising the ensemble learner, they merged random neural networks and applied the clustering algorithm. Helmy et al. (2022) created the prediction technique that predominantly finds PD-related genes based on protein and lncRNA genes in order to take advantage of the biological significance of lncRNAs in addition to proteins. To obtain crucial and distinguishing information, the suggested approach depicts all genes as DNA FASTA sequences. Peng et al. (2019) and Gautam and Sharma (2020) utilised the Node2vec tool to generate a vector representation of each gene in a PPI network, followed by an autoencoder to minimise the dimension of the resulting vector. Ultimately, new genes associated with PD were predicted using a SVM classifier. In addition, they utilised N2A-SVM trained on the most recent dataset to predict genes for PD. Gautam and Sharma (2020) and Olah (2015) provide a comprehensive review of deep learning techniques used in the prognosis of eight distinct neuropsychiatric and neurological conditions, including Alzheimer's, stroke, epilepsy, autism, Parkinson's, migraine, multiple sclerosis, and cerebral palsy. These diseases are severe, life-threatening, and, in the majority of cases, can lead to other dangerous human diseases.

Table 2 Comparative analysis of recent methods

Authors	*Proposed method*	*Dataset*	*Methods used*
Bi et al. (2021)	Introduced multimodal data fusion analysis framework for predicting PD genes	PPMI	ANN
Helmy et al. (2022)	Identified PD-related genes: protein and lncRNA using Adaboost as feature selection method	Protein genes	Gradient boosted decision tree
Park et el. (2020)	Predicts Azheimer's disease using multiple heterogeneous omic data	Gene expression	Deep neural network
Gautam and Sharma (2020)	Highlight the research work on early diagnosis of neurological diseases using deep learning techniques	-	Deep neural network (DNN), deep-belief network (DBN), deep autoencoder (DA) and convolutional-neural network (CNN)
Peng et al. (2019)	Used N2A-SVM algorithm to discover new genes associated with PD	genes	Node2vec auto encoder and SVM (N2A-SVM)
Chen et al. (2020)	Identifying potential disease-associated genes and explore the relationships between diseases and genes, and has an important impact on the disease etiology research.	OMIM	Convolutional neural network
Stolfi et al. (2023)	The proposed method for detecting Parkinson's disease is based on the flexible analytic wavelet transform (FAWT).	EEG recordings	SVM, RF, RBF, k-nearest neighbour classifier
Ahn et al. (2020)	proposed a DNN based model to classify cancer and normal samples.	Gene expression	Deep neural network

3 Proposed method

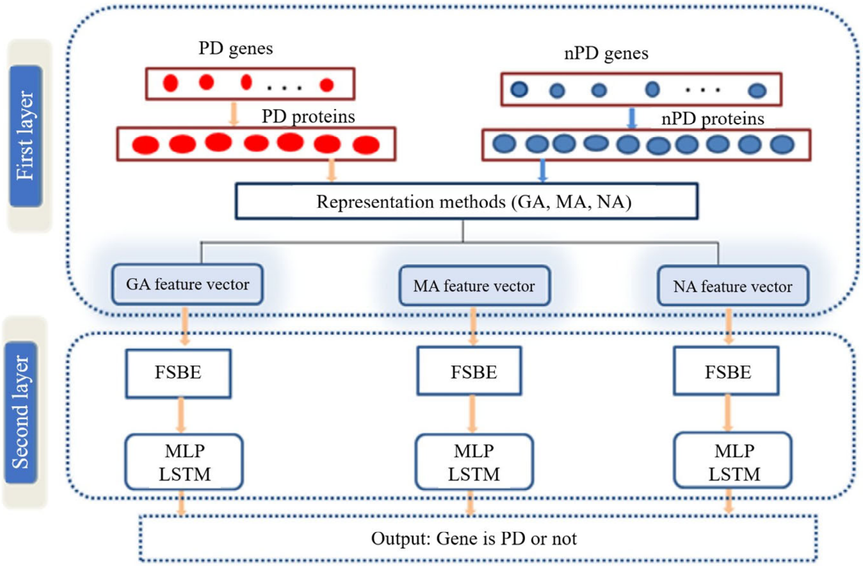
The proposed system model as shown in Figure 2 for identifying PD genes has been defined in this section. The proposed method comprises of three steps:

1 utilise physicochemical properties of amino acids to translate protein sequences into numerical features

2 stepwise FSBE method is used to extract best features and remove the worst from remaining attributes

3 train the models.

Figure 2 Proposed system model for PD identification (see online version for colours)



3.1 Feature extraction

Extracting features for known and unknown genes is one of the most essential stages in identifying disease genes. We used three representation methods, namely normalised Moreau-Broto autocorrelation (NA) (Zulfiqar et al., 2021), Geary autocorrelation (GA) (Chen et al., 2020), and Moran autocorrelation (MA) (Xia et al., 2020), to extract information encoded in protein sequences, which we then used to classify genes. These techniques use the unique physicochemical property of each amino acid to illustrate the effect of their proximity in a sequence separated by a specific number of amino acids. Additionally, it must be possible to identify patterns that transcend the entire sequence. Protein sequences contain extremely valuable information, which is why these techniques are utilised. In addition, these techniques are used in other works (Yousef and Charkari, 2015) and offer advantages over alternative methods. These autocorrelation methods can be defined below as in equations (1)–(4).

Moreau-Broto autocorrelation for protein sequence are defined as:

$$AC(l) = \sum_{i=1}^{N-l} P_i P_{i+l} \qquad l = 1, 2, \ldots\ldots\ldots, nlag \tag{1}$$

where l is lag of auto-correlation, P_i and P_{i+l} are the properties of amino acids, *nlag* is value of lag.

Normalised Moreau-Broto autocorrelation (NA) can be defined as:

$$NAC(l) = \frac{AC(l)}{N-l} \qquad l = 1, 2, 3, ..., nlag \tag{2}$$

GA can be defined as:

$$GA(l) = \frac{\frac{1}{2(N-l)}\sum_{i=1}^{N-l}\left(P_i - p_{i+l}\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \tilde{P}'\right)^2} \qquad l = 1, 2, 3, ..., nlag \tag{3}$$

where l is lag of auto-correlation, P_i and P_{i+l} are the properties of amino acids, *nlag* is value of lag.

MA can be defined as:

$$MA(l) = \frac{\frac{1}{N-l}\sum_{i=1}^{N-l}\left(P_i - \tilde{P}'\right)\left(P_{i+l} - \tilde{P}'\right)}{\frac{1}{N-l}\sum_{i=1}^{N}\left(P_i - \tilde{P}'\right)^2} \qquad l = 1, 2, 3, ..., nlag \tag{4}$$

where l is lag of auto-correlation, P_i and P_{i+l} are the properties of amino acids, *nlag* is value of lag, $\tilde{P}'$ is considered property along sequence, i.e., $\tilde{P}' = \frac{\sum_{i=1}^{N} P_i}{N}$.

These representation methods used their physicochemical properties of amino acids to explain the neighbouring effect between amino acids with other amino acids within a sequence. These methods help to gain relevant information, which is unknown in protein sequences. In this research, we employed a set of 12 physical properties to learn more about amino acid sequences, since all representation approaches are founded on these very same qualities. In order to further our understanding of the amino acid sequence, we have relied on a total of 12 physicochemical features. Many different physicochemical properties are taken into account, including polarity (Grantham, 1974), residue accessible surface area in tripeptide (Chothia, 1976), hydrophilicity (Hopp and Woods, 1981), polarisability (Charton and Charton, 1982), solvation free energy (Eisenberg and McLachlan, 1986), entropy of formation (Chothia, 1992), partition coefficient (Quinlan, 1996), amino acid composition (AAC) (Grantham, 1974), hydrophobicity (Sweet and Eisenberg, 1983), transfer-free energy (Janin, 1979), CC in regression analysis (Prabhakaran and Ponnuswamy, 1982), and graph shape index (Fauchere, 1988). These properties are utilised to obtain the features. Min-max normalisation method is applied to normalise the physicochemical properties as shown in equation (5).

$$P_{xy} = \frac{P_{xy} - P_{y,\min}}{P_{y,\max} - P_{y,\min}} \tag{5}$$

where P_{xy} represents y^{th} descriptor value for x^{th} amino acid, $P_{y,\min}$ is y^{th} descriptor minimum value of amino acids and $P_{y,\max}$ is y^{th} descriptor maximum value of amino acids.

Table 3 Normalised values for 12 physicochemical properties of amino acid

POL	*RAS*	*HY-PHIL*	*POL-ZAB*	*HY-PHOB*	*SFE*	*AAC*	*CC*	*GSI*	*TFE*	*PC*	*EOF*
0.4939	06492	0.6009	0.3118	0.5491	0.3236	0.5640	0.4697	0.5990	0.4121	0.4009	0.2933
0.4498	0.5717	0.4832	0.4401	0.5024	0.3218	0.5284	0.3475	0.5620	0.3236	0.5087	0.2901
0.3728	0.6047	0.4179	0.2146	0.4728	0.1580	0.5186	0.3965	0.5159	0.3995	0.3693	0.2514
0.4425	0.6350	0.2676	0.3707	0.5331	0.3168	0.5892	0.4205	0.5260	0.4682	0.4376	0.3083
0.3960	0.6876	0.4544	0.4539	0.4181	0.3880	0.4401	0.3815	0.5786	0.2829	0.3541	0.2655
0.5671	0.7023	0.5120	0.4337	0.7802	0.2965	0.7777	0.4015	0.4695	0.5605	0.5172	0.3611
0.3364	0.5164	0.5201	0.4416	0.4001	0.3185	0.4208	0.3528	0.6075	0.2744	0.2200	0.3118
0.4286	0.6246	0.6031	0.4445	0.4597	0.2866	0.4661	0.3501	0.5470	0.2916	0.4461	0.2579
0.3155	0.4872	0.5784	0.4930	0.3641	0.3038	0.3742	0.3285	0.4999	0.3519	0.4141	0.2694
0.2813	0.5541	0.5030	0.5562	0.3162	0.3696	0.3180	0.2812	0.4504	0.2919	0.4126	0.2711
0.4521	0.6842	0.5479	0.5200	0.4633	0.5097	0.4918	0.2708	0.5198	0.4121	0.2806	0.2739
0.3942	0.6144	0.7510	0.3679	0.4453	0.2707	0.4713	0.3482	0.4971	0.3672	0.3911	0.2732
0.3528	0.4531	0.5024	0.4300	0.4077	0.2605	0.4627	0.3244	0.3955	0.3330	0.3647	0.2675
0.3470	0.6151	0.5335	0.3759	0.3929	0.2866	0.4136	0.3226	0.5062	0.3617	0.2535	0.2487
0.3506	0.5536	0.4662	0.4221	0.4088	0.2647	0.4368	0.3318	0.5104	0.3212	0.3458	0.2634
0.4163	0.5525	0.5010	0.2969	0.4258	0.2452	0.4561	0.3731	0.3117	0.2534	0.3248	0.2981
0.3936	0.5981	0.4245	0.1972	0.4304	0.2273	0.4353	0.3199	0.5150	0.4013	0.4832	0.2881
0.4470	0.6924	0.4565	0.2857	0.4619	0.2755	0.4881	0.3350	0.5401	0.3413	0.3952	0.2977
0.2983	0.6543	0.4330	0.3315	0.4322	0.3181	0.4399	0.3640	0.5650	0.4133	0.3421	0.2873
0.4578	0.7251	0.4432	0.2991	0.4728	0.3355	0.5119	0.3448	0.4580	0.3203	0.2834	0.2526

The normalised values for physicochemical properties used in this article are given in Table 3.

Algorithm 1 Algorithm to identify PD genes using MLP and LSTM

```
Start
    **Input:** Protein sequences (genes)
    **Identify:** gene is PD or not using feature vector
    **Output:** S={0,1}
    Initialisation
    **foreach** protein sequence **do**
    • extractGeary()
            $GA(l)=\frac{\frac{1}{2(N-1)}\sum_{i=1}^{N-1}\left(P_i-P_{i+l}\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i-\tilde{P}'\right)^2}$
    • extractMoran()
            $MA(l)=\frac{\frac{1}{(N-1)}\sum_{i=1}^{N-1}\left(P_i-\tilde{P}'\right)\left(P_{i+l}-\tilde{P}'\right)}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i-\tilde{P}'\right)^2}$
    • extractMoreauBroto()
    $AC(l)=\sum_{i=1}^{N-1}P_iP_{i+1}$
    $AC(l)=\frac{AC(l)}{N-1}$
    **form:** feature
    factor **foreach**
    method **do** forward
    Selection
    backward elimination
end
        **for** numeric data n
        **do** LSTM train(n)
        **foreach** Params,
        set do epochs = 80
        fully connected layers = 2
        hidden units = 150
        learning rate = 0.01
**end**
        MLP train(n)
            **foreach** Params, set do
            epochs = 80
            fully connected layers = 2
            hidden units = 100
            End
        Calculate Precision, Recall
        Calculate F-score
        LSTM test<-Predict{0,1}
        Accuracy = accuracy, V accuracy
        MLP test<-Predict{0,1}
        Accuracy = accuracy, V accuracy
End
```

3.2 Long short-term memory

LSTM is a special kind of recurrent neural network (RNN), currently become popular in the area of machine learning (Moawad, 2018), introduced by Hochreiter and Schmidhuber (1997). LSTM is designed to learn long-term dependencies and memorise information for a long time. It is organised in the chain-like structure (Olah, 2015) and consists of repeating module as in standard RNN. LSTM is based on three added gates; Input gate, forget gate and output gate as shown in Figure 3. LSTM has four layers to interact instead of a single neural network layer (Erguder, 2018).

LSTM method comprises of storage blocks called memory cells. The two states, i.e., hidden state and cell state are moving to the next cell. Initially, the cell state allows the data to flow forward substationally unchanged. However, certain linear transformations can take place. With sigmoid gates, data can be added to or deleted from the cell state. LSTM aims to avoid long-term dependence issue to control the memory process with the use of gates.

The initial and main step in building an LSTM system is to find and then remove unwanted information from the cell. A sigmoid function helps to identify and excludes unwanted data from cell state, which gets the output of previous timestamp (h_{t-1}) at $t-1$ time and current input (x_t) at t time with bias b_f as shown in equation (6). In addition, the sigmoid function decides which part of the previous output could be excluded. The gate is also known as forget gate. Output of sigmoid layer determines whether to completely retain or completely discard information (0 or 1).

$$f_t = \sigma\left(W_f.\left[h_{t-1}, x_t\right] + b_f\right) \tag{6}$$

where σ denotes sigmoid function, W_f and b_f are weight matrices and bias respectively.

The next step is to determine and store information in cell state from the new input x_t state. This can be done with sigmoid and tanh function. Sigmoid layer decides whether to update or ignore new information (0 or 1) and tanh layer assigns weight to passed value to determine its importance level (–1 or 1). Then multiply these values to update new cell state and add this new memory (N_t) to old memory (c_{t-1}) resulting in ct.

$$i_t = \sigma\left(W_i.\left[h_{t-1}, x_t\right] + b_i\right) \tag{7}$$

$$N_t = \tanh\left(W_n.\left[h_{t-1}, x_t\right] + b_n\right) \tag{8}$$

$$c_t = f_t * c_{t-1} + i_t * N_t \tag{9}$$

where c_t and c_{t-1} are cell states at t and $t - 1$ time, respectively.

The last step determines which cell state information is used as output. This step separates the final memory from the hidden state. As shown in equation (10) sigmoid layer determines which part of cell state makes it an output O_t with last hidden state h_{t-1}. Then, output of sigmoid function is multiplied with new values formed by tanh function from estimated cell state as shown in equation (11).

$$O_t = \sigma\left(Wo.\left[h_{t-1}, x_t\right] + bo\right) \tag{10}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{11}$$

3.3 Multi-layer perceptron

The MLP is a form of multi-layered, deep feedforward network. The accuracy of nonlinear task prediction improves as the number of layers increases (Singh and Kumar, 2020). Figure 4 depicts the essential structure of an MLP network, which consists of two hidden layers.

In a MLP, the outputs of one layer serve as the inputs to the next layer; therefore, the neurons should be arranged in a linear fashion. In MLP, data can be converted at three distinct layers: input, hidden, and output. Through the structure of the neural network, each neuron in the collection is connected to all of the neurons in the stratum above it. Weights within [1, 1] must be utilised to categorise interlayer connections. Summation and activation can be conducted at each node in an MLP (Ramchoun et al., 2016). According to equation (12), the weight (w) and bias (b) are related to the MLP network parameters via a summation function. The weights of the input layer were then multiplied by each neuron.

$$S_k = \sum\nolimits_{i=1}^{n} w_{ik} I_i + B_k \tag{12}$$

where I_i represents the input variable, n represents the number of input, B_k represents the bias and w_{ik} represents the connection weight

The output of equation (11) should be used to activate an activation function. In MLP, the most frequently used activation functions are hyperbolic tangent (tanh), sigmoid, rectifier linear unit (ReLU), sigmoid, and leaky ReLU. In this paper, we applied the ReLU activation function to the hidden layer and the sigmoid activation function to the output layer, as shown in equations (13) and (14).

$$f_k(x) = \max\left(0, S_k\right) \tag{13}$$

$$f_k(x) = \frac{1}{1 + e^{-S_k}} \tag{14}$$

Therefore, the final neuron output can be calculated using equation (15).

$$y_i = f_k\left(\sum\nolimits_{i=1}^{n} w_{ik} I_i + B_k\right) \tag{15}$$

Figure 3 LSTM neural network structure (see online version for colours)

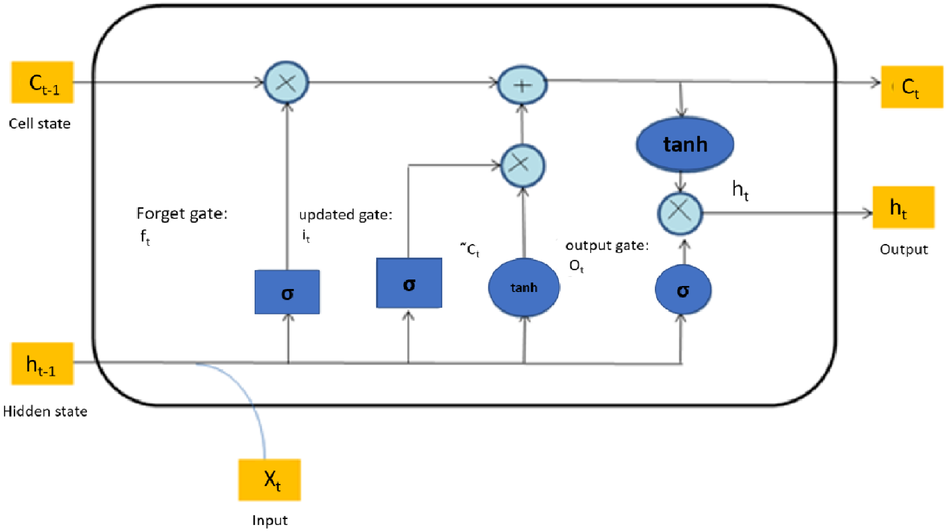
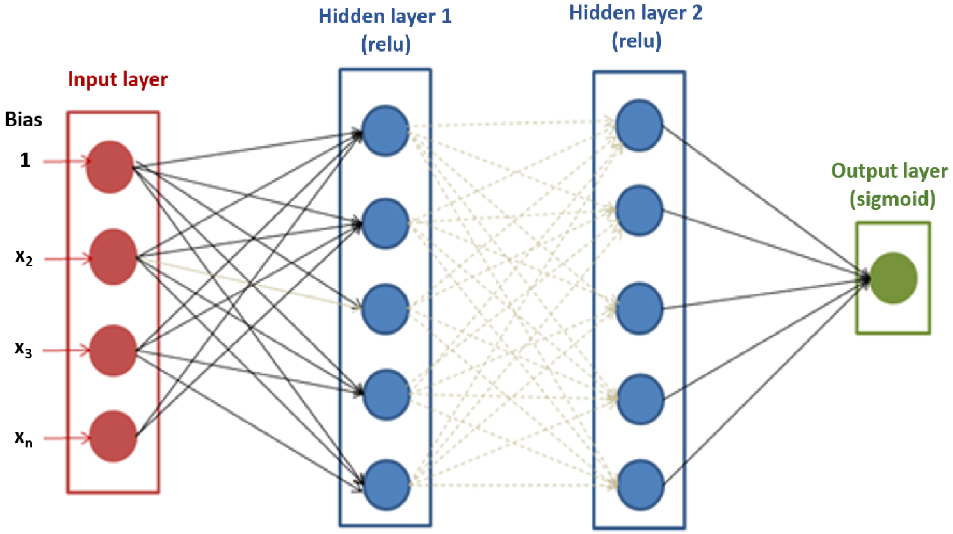


Figure 4 Architecture of MLP (see online version for colours)



3.4 *Experimental setup*

The proposed PD gene identification method has been constructed using two separate methods, one is MLP and other one is LSTM. The Keras (Chollet, 2015) library has been used for implementing deep learning methods due to its user-friendly nature. The different optimisation criteria used in training and testing the proposed model are listed in Table 4.

Three hidden layers are used in the MLP technique's construction, while just one is used in the LSTM approach. For both models, initial weights are balanced over all available layers. The output layer uses the sigmoid activation function to predict probability of having a disease gene or not. Adam optimisation method to update network weights is applied to both the models. Both models use the multi-class logarithmic function of cross entropy as the loss function. The number of training examples applied to the input layer before the weight update is 320 (batch size). The early stopping method (Zhang et al., 2016) is used to determine the number of training epochs. Initially 100 epochs are taken to analyse the performance of model. It has observed that after 80 epochs, the model performance stopped showing any significant improvement because the loss and accuracy value remains almost same till 80 epochs. So, our models are trained for 80 epochs. The next section will discuss the results of the proposed model using configured parameters.

Table 4 Parameters used for MLP and LSTM

Tuning parameters MLP	*LSTM*
Model initialisation	
Hidden layers	3 1
Hidden units	100, 70 and 50 gated memory units 128
Activation function	ReLU, tanh, Sigmoid ReLU, tanh, Sigmoid
Layer type	Dense
Dropout	0.1
Model compilation	
Loss function	Categorical cross-entropy
Optimiser	Adam
Model training	
Batch Size	320
Epoch	80

4 Results and discussion

The performance of MLP and LSTM methods has been studied on the imbalanced dataset in this section. Firstly, the optimal number of features extracted through feature selection and backward elimination method has been reviewed and optimised. Then, the influence of sequence representation methods has been evaluated on the performance of both MLP and LSTM methods. Finally, a comparison between our proposed method and other disease gene identification methods has been done to obtain relatively negative data to confirm the effectiveness of method.

4.1 Experimental data

The dataset used in this research study consists of human PD and non-Parkinson's disease (nPD) protein sequences (genes) are extracted from the NCBI (http://www.ncbi.nlm.nih.gov/geo), Ensembl, and UNIPROT databases (Universal Protein Resource, http://www.uniprot.org). All the sequences are saved as fasta file. The obtained dataset was then cleaned from duplicate and partial protein sequences. There are a total of 2,815 sequences, of which 1,220 are positive data (PD) and 1595 are negative data (nPD). The sequences obtained was then cleaned by eliminating duplicate and partial protein sequences in python. The training and validation samples are randomly selected from the available data. This helps in reducing biases and ensures a representative distribution of data across different classes. To ensure data consistency and quality, a rigorous data pre-processing procedure was followed, including removing duplicate entries, handling missing values, and normalising numerical features. The 70% of the data are randomly selected as training and remaining 30% is used as validation samples. Using GA, MA, and NA sequence representation algorithms, each sequence is converted into three feature vectors. This task was completed using the R protr package.

4.2 Performance of the proposed method

We trained and tested the model with LSTM and MLP based on NA, GA, and MA representation approaches independently to evaluate their respective robustness. There are two perspectives on the outcomes of these procedures. First, we trained the model by using all the 360 features and evaluate the results. Then evaluate the results after using feature reduction methods.

Since there are more unidentified genes than disease genes, we must evaluate the efficacy of this strategy using an incomplete dataset. MLP performs exceptionally well even with a small quantity of data, whereas the LSTM model performs better with a large amount of data. We utilised early stopping to prevent overfitting. As long as the accuracy of the training set continues to improve, the network will not learn and the accuracy of the validation set will remain stable.

Figure 5 Epoch vs. loss for MLP and LSTM methods (see online version for colours)

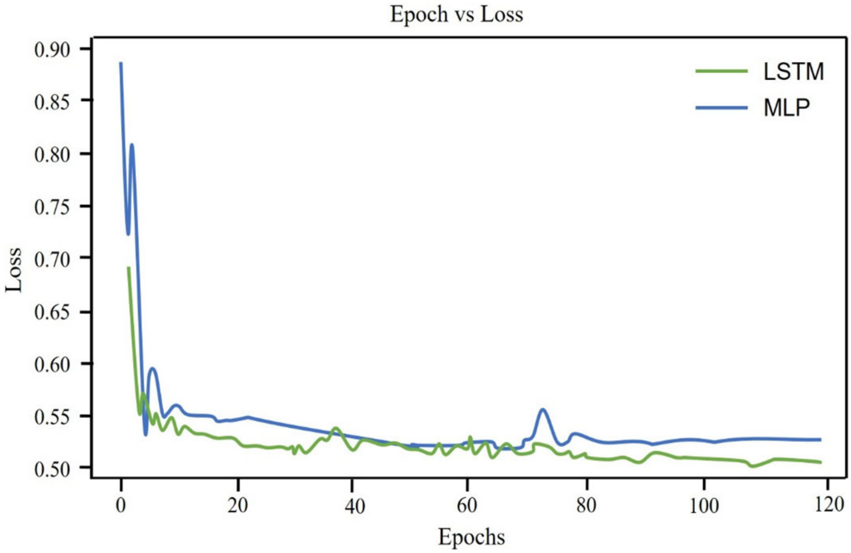
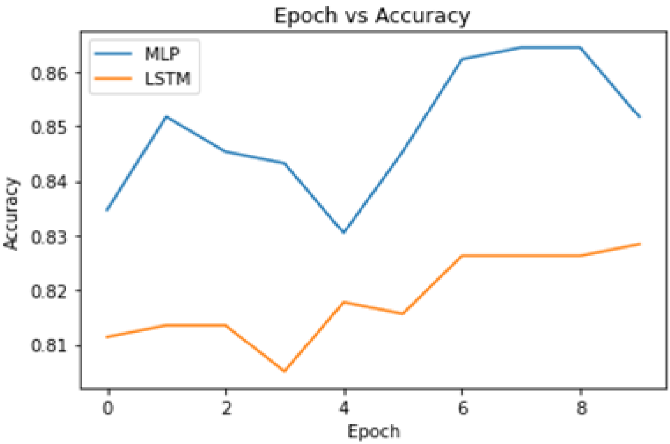


Figure 5 shows the epoch versus loss graph for both MLP and LSTM methods. As it is shown that loss stops decreasing after 80 epochs. So, stopping criteria will be activated at 80th epoch. Figure 6 shows the epoch versus accuracy graph for both MLP and LSTM methods. It can be concluded that MLP performs better than LSTM.

Figure 6 Epoch vs. accuracy for MLP and LSTM methods (see online version for colours)



4.3 Discussion

We have compared the F-score of proposed MLP and LSTM methods on imbalanced dataset. The results obtained from different representation methods for both MLP and LSTM is shown in Table 5. The outcomes of these methods are displayed in two ways. At first, we evaluate the results after using all the features. Secondly, we use feature reduction methods to evaluate the results.

We used an imbalanced dataset to study the performance of this method, as the number of unknown genes is more than that of disease genes. Figure 7 to Figure 10 shows the performances without and with feature selection for both MLP and LSTM methods. It can be observed from the plots that method with reduced features show better performance than methods without feature reduction. For example, the F-score of GA for MLP is improved from 83.4% to 85%, while for LSTM F-score is improved from 81.45% to 83.96%. Also, it has been observed that performance of GA is superior to other representation methods. We can say that MLP performs extremely well even with a small amount of data, and when the amount of data is large, LSTM model performs better. The proposed MLP method for GA representation with reduced features significantly outperforms LSTM method and produces stable and scalable performance.

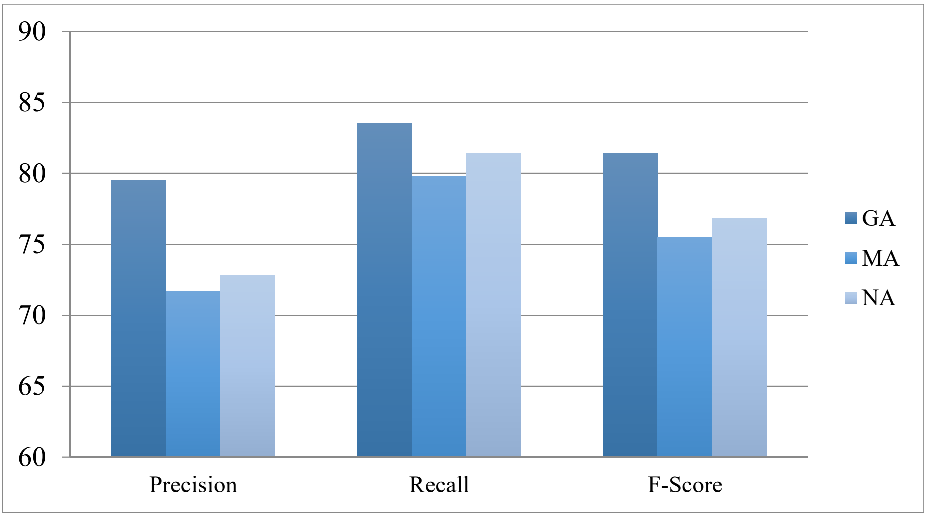
Based on the results obtained, we conclude that MLP method has higher accuracy in the diagnosis and prediction of neurological diseases, and is superior to other classifiers. Due to the complexity of genetic and microarray data, it is difficult to make an accurate diagnosis, so computer-aided advanced machine learning technology is used to improve the prediction accuracy and treatment level of neurological diseases.

In this paper, we used feature selection and backward elimination method in data pre-processing and then applied deep learning methods on reduced features for gene identification.

Table 5 Performances of sequence representation methods

Methods	*No. of features*	*Precision*	*Recall*	*F-score*
		Without feature selection for LSTM		
GA	360	79.5	83.5	81.45
MA	360	71.7	79.8	75.53
NA	360	72.8	81.4	76.86
		With extracted features for LSTM		
GA	65	82.3	85.7	83.96
MA	60	74	81.4	77.52
NA	71	77.2	82.0	79.52
		Without feature selection for MLP		
GA	360	81.7	85.2	83.41
MA	360	73.1	81.5	77.07
NA	360	74.9	82.9	78.69
		With extracted features for MLP		
GA	65	84.5	88.2	85.0
MA	60	76.9	83	79.83
NA	71	79.6	84.5	81.97

Figure 7 Performance (percentage) of representation methods for LSTM (see online version for colours)



According to the results obtained from previous research on different datasets, the performance of deep learning methods is more advanced and better than other traditional machine learning classifiers. This motivation prompted us to propose deep learning model, which produces unbiased and stable classification model. Additionally, deep learning can effectively discover patterns in high-dimensional data. The results obtained shows that MLP method has greater performance where feature selection is done with FSFE method.

Figure 8 Performance (percentage) of representation methods using feature selection for LSTM (see online version for colours)

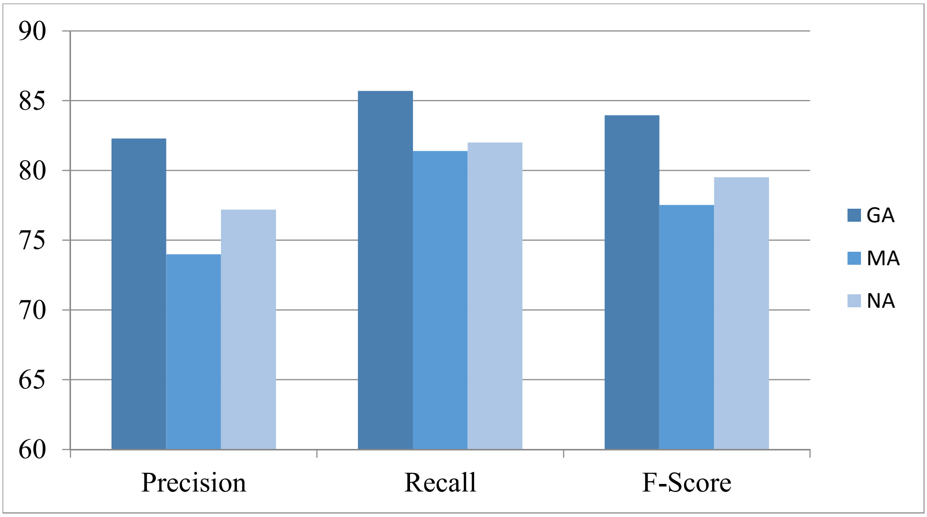
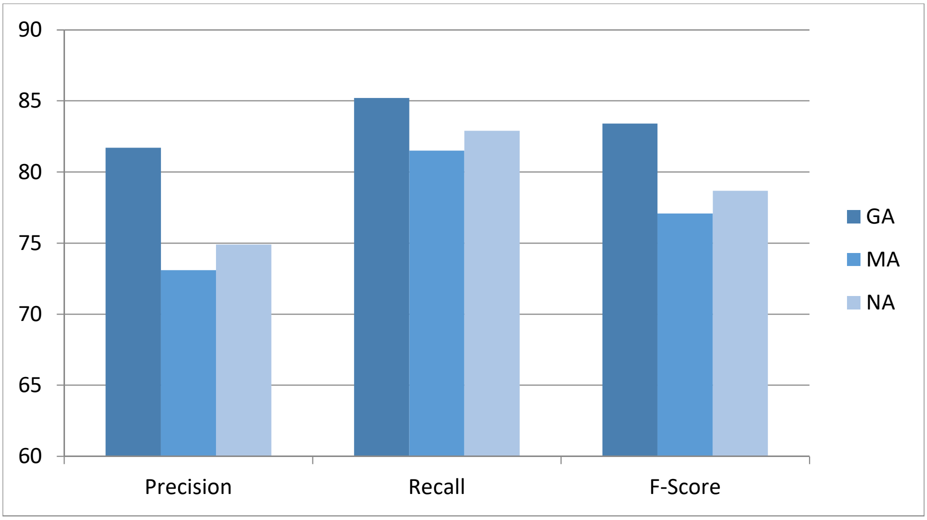


Figure 9 Performance (percentage) of representation methods for MLP (see online version for colours)



4.4 Comparison with existing works

It has been observed from Table 6 that proposed MLP with feature reduction method outperforms other state-of-art methods. The proposed method is compared with six state-of-art methods, including, SFM method (Yousef and Charkari, 2015), EPU (Yang et al., 2012), PUDI (Mordelet and Vert, 2011), ProDiGe (Miao et al., 2017), Smalter's (Xu and Li, 2006), and Xu's (Adie et al., 2005). The protein sequences for both PD and non-PD genes have been collected from NCBI, Ensembl and Uniprot databases. It has been observed that in terms of F-score proposed MLP method on average, is 5.4%, 6.4%, 10.1%, 16.9%, 22.8% and 23.4% higher than EPU, SFM, PUDI, ProDiGe, Smalter's method, Xu's method respectively for imbalanced datasets. The basic difference between the mentioned methods and our proposed method is the prior knowledge used to extract feature vector. In this paper, the sequence of proteins was realised as the most common knowledge while in previous work; prior knowledge was affected by noise. The second issue is about classification algorithm used to identify disease genes. Since our preferred sample is unbalanced, we used a precision-recall (PR) curve to handle highly skewed data. In order to establish the relationship between precision and recall and measure the performance of the classifier, the area under the PR curve is preferred. PR relationship is shown in Figure 11 for proposed and existing methods.

Figure 10 Performance (percentage) of representation methods using feature selection for MLP (see online version for colours)

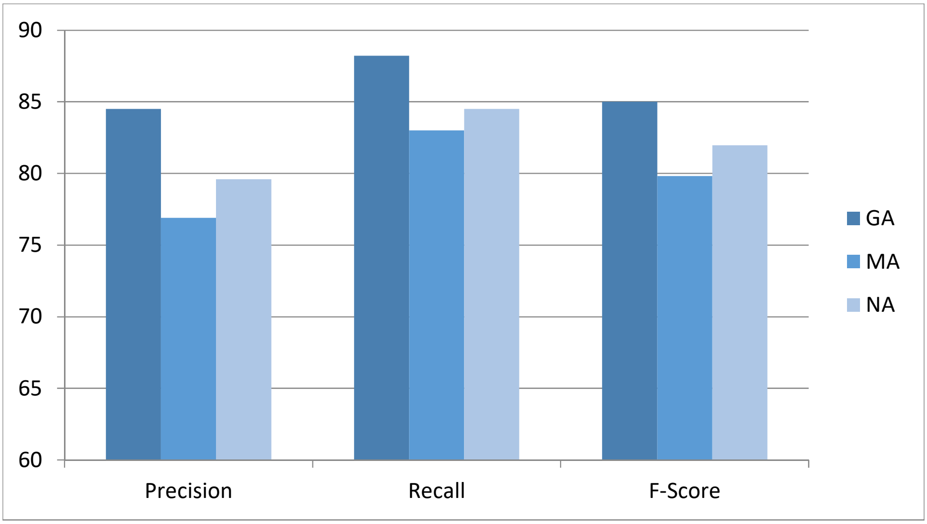
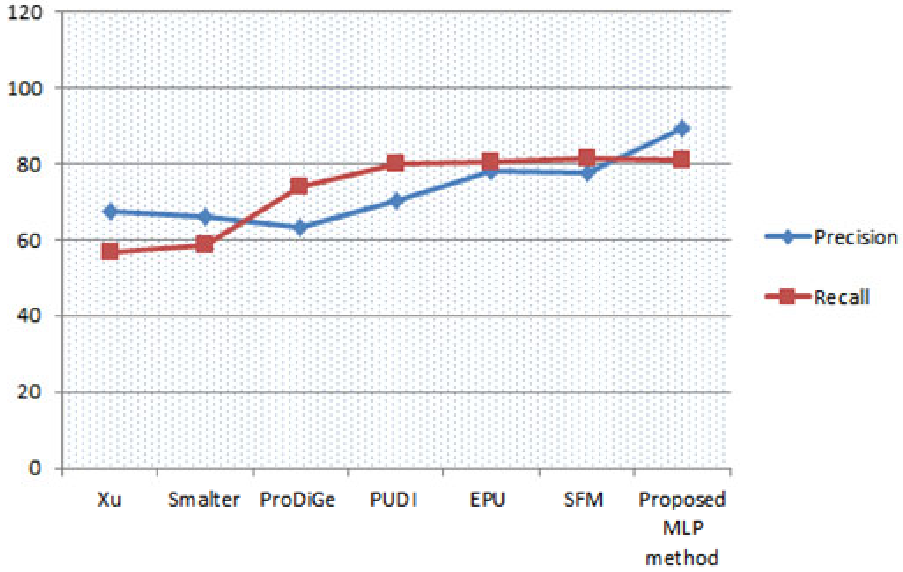


Table 6 Comparative evaluation between proposed and existing methods

Method	*Recall (%)*	*Precision (%)*	*F-score (%)*
SFM (Yousef and Charkari, 2015)	81.4	77.9	79.6
EPU (Yang et al., 2012)	80.4	78.2	78.6
PUDI (Mordelet and Vert, 2011)	80.1	70.3	74.9
ProDiGe (Miao et al., 2017)	74.0	63.1	68.1
Smalter's method (Xu and Li, 2006)	58.7	66.2	62.2
Xu's method (Adie et al., 2005)	56.8	67.4	61.6
Proposed MLP method	*88.2*	*84.5*	*85.0*

Figure 11 PR curve for all methods (see online version for colours)



4.5 Predicting novel disease genes

Given a particular disease class, the set of confirmed disease genes are obtained from UNIPROT. Using all these disease genes as positive training set, we perform experiments by applying our proposed method to predict novel disease genes from all the unlabelled gene set. We first rank all the genes based on the probability predicted by the trained model. Based on literature review, we find that some of these genes have been reported to be associated with PD.

We first applied our method to discover novel disease genes for PD. We then search literature to check whether any of these predicted disease genes are really related to PD. We found that nine predicted genes, namely, DHDDS, PARK2, PICK1, MT-ND4, NDUFB5, NDUFA6, CIC, TRIM63 and ATP5A1 have been reported to be associated with PD.

5 Conclusions

Identification of genes is of great importance for the treatment and diagnosis of PD. In this paper, we introduced deep neural networks for genes identification using the protein sequences. Normalised Moreau-Broto autocorrelation, GA and MA representation methods are used to convert protein sequences (genes) into feature vectors with 12 physicochemical properties of the amino acids. Then, FSBE feature reduction method is applied to extract the important features. The proposed method based on MLP and LSTM improved the F-measure score compared to previous methods on imbalanced datasets. The proposed methods show the better performance with a high F-measure rate of 85% and 83.96% respectively. Compared results of the proposed method show better performance than previous reported works. For the future work, we can consider more physicochemical properties with GA representation method to combine multiple different classifiers to achieve better classification. We can also apply this method to predict other neurological disease genes.

Disclaimer

The authors received no financial support for the research.

References

Abdukodirov, E.I., Khalimova, K.M. and Matmurodov, R.J. (2022) 'Hereditary-genealogical features of Parkinson's disease and their early detection of the disease', *International Journal of Health Sciences*, No. 1, pp.4138–4144, DOI: 10.53730/ijhs.v6nS1.5802.

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2005) 'Speeding disease gene discovery by sequence based candidate prioritization', *BMC Bioinform.*, Vol. 22, No. 6, p.55.

Ahn, T., Goo, T., Lee, C.H., Kim, S., Han, K., Park, S. and Park, T. (2020) 'Deep learning-based classification and interpretation of gene expression data from cancer and normal tissues', *International Journal of Data Mining and Bioinformatics*, Vol. 24, No. 2, pp.121–139.

Bi, X.A., Hu, X., Xie, Y. and Wu, H. (2021) 'A novel CERNNE approach for predicting Parkinson's disease-associated genes and brain regions based on multimodal imaging genetics data', *Med. Image Anal.*, Vol. 67, p.101830.

Charton, M. and Charton, B.I. (1982) 'The structural dependence of amino acid hydrophobicity parameters', *J. Theor. Biol.*, Vol. 99, No. 4, pp.629–644.

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P.J., Ma, Q. and Zhang, Y. (2020) 'Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier', *Computers in Biology and Medicine*, 1 August, Vol. 123, p.103899.

Chen, X., Huang, Q., Wang, Y., Li, J., Liu, H., Xie, Y., Dai, Z., Zou, X. and Li, Z. (2020) 'A deep learning approach to identify association of disease-gene using information of disease symptoms and protein sequences', *Analytical Methods*, Vol. 12, No. 15, pp.2016–2026.

Chollet, F. (2015) 'Keras', *Github Repository*.

Chothia, C. (1976) 'The nature of the accessible and buried surfaces in proteins', *J. Mol. Biol.*, Vol. 105, No. 1, pp.1–12.

Chothia, C. (1992) 'Proteins. One thousand families for the molecular biologist', *Nature*, Vol. 357, No. 6379, pp.543–544.

Danaee, P., Ghaeini, R. and Hendrix, D.A. (2017) 'A deep learning approach for cancer detection and relevant gene identification', in *Pacific Symposium on Biocomputing*, pp.219–229.

Draoui, A., El Hiba, O., Aimrane, A., El Khiat, A. and Gamrani, H. (2020) 'Parkinson's disease: from bench to bedside', *Revue neurologique*, 1 September, Vol. 176, Nos. 7–8, pp.543–559.

Eisenberg, D. and McLachlan, A.D. (1986) 'Solvation energy in protein folding and binding', *Nature*, Vol. 319, No. 6050, pp.199–203.

Erguder, H. (2018) 'Recurrent neural network Nedir', *Deep Learning Turkey* [online] https://medium.com/@hamzaerguder/recurrent-neural-network-nedir-bdd3d0839120 (accessed 21 June 2021).

Fauchere, J.L. (1988) 'Amino acid side chain parameters for correlation studies in biology and pharmacology', *Int. J. Pept. Protein Res.*, Vol. 32, No. 4, pp.269–278.

Fukasawa, Y., Leung, R.K., Tsui, S.K. and Horton, P. (2014) 'Plus ça change-evolutionary sequence divergence predicts protein subcellular localization signals', *BMC Genomics*, Vol. 15, No. 1, p.46.

Gautam, R. and Sharma, M. (2020) 'Prevalence and diagnosis of neurological disorders using different deep learning techniques: a meta-analysis', *Journal of Medical Systems*, Vol. 44, No. 2, pp.1–24.

Grantham, R. (1974) 'Amino acid difference formula to help explain protein evolution', *Science*, Vol. 185, No. 4154, pp.862–864.

Grantham, R. (1974) 'Amino acid difference formula to help explain protein evolution', *Science*, Vol. 185, No. 4154, pp.862–864.

Helmy, M., Eldaydamony, E., Mekky, N., Elmogy, M. and Soliman, H. (2022) 'Predicting Parkinson disease related genes based on PyFeat and gradient boosted decision tree', *Scientific Reports*, Vol. 12, No. 1, p.10004.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Comput.*, Vol. 9, No. 8, pp.1735–1780.

Hopp, T.P. and Woods, K.R. (1981) 'Prediction of protein antigenic determinants from amino acid sequences', *Proc. Natl. Acad. Sci. USA*, Vol. 78, No. 6, pp.3824–3828.

Janin, J. (1979) 'Surface and inside volumes in globular proteins', *Nature*, Vol. 277, No. 5696, pp.491–492.

Jowkar, G.H. and Mansoori, E.G. (2016) 'Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification', *Computational Biology and Chemistry*, Vol. 64, pp.263–270.

Köhler, S., Bauer, S., Horn, D. and Robinson, P.N. (2008) 'Walking the interactome for prioritization of candidate disease genes', *The American Journal of Human Genetics*, Vol. 82, No. 4, pp.949–958.

Madeddu, L., Stilo, G. and Velardi, P. (2020) 'A feature-learning-based method for the disease-gene prediction problem', *International Journal of Data Mining and Bioinformatics*, Vol. 24, No. 1, pp.16–37.

Miao, Y., Jiang, H., Liu, H. and Yao, Y.D. (2017) 'An Alzheimers disease related genes identification method based on multiple classifier integration', *Computer Methods and Programs in Biomedicine*, Vol. 150, pp.107–115.

Moawad, A. (2018) 'The magic of LSTM neural networks', *Deep Learning Turkey* [online] https://medium.com/datathings/the-magic-of-lstm-neural-networks-6775e8b540cd (accessed 8 July 2021).

Mordelet, F. and Vert, J.P. (2011) 'ProDiGe: prioritization of disease genes with multitask machine learning from positive and unlabeled examples', *BMC Bioinformatics*, Vol. 12, No. 1, p.389.

Olah, C. (2015) *Understanding LSTM Networks* [online] http://colah.github.io/posts (accessed 9 August 2021).

Park, C., Ha, J. and Park, S. (2020) 'Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset', *Expert Syst. Appl.*, Vol. 140, p.112873.

Peng, J., Guan, J. and Shang, X. (2019) 'Predicting Parkinson's disease genes based on node2vec and autoencoder', *Frontiers in Genetics*, Vol. 10, p.226.

Pereira, C.R., Pereira, D.R., Silva, F.A., Masieiro, J.P., Weber, S.A., Hook, C. and Papa, J.P. (2016) 'A new computer vision-based approach to aid the diagnosis of Parkinson's disease', *Computer Methods and Programs in Biomedicine*, Vol. 136, pp.79–88.

Prabhakaran, M. and Ponnuswamy, P.K. (1982) 'Shape and surface features of globular proteins', *Macromolecules*, Vol. 15, pp.314–320.

Quinlan, J.R. (1996) *Improved Use of Continuous Attributes in C4.5*, arXiv preprintcs/9603103.

Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., Boyle, S.M. and Mooney, S.D. (2008) 'An integrated approach to inferring gene-disease associations in humans', *Proteins: Structure, Function, and Bioinformatics*, Vol. 72, No. 3, pp.1030–1037.

Ramchoun, H., Idrissi, M.A.J., Ghanou, Y. and Ettaouil, M. (2016) 'Multilayer perceptron: architecture optimization and training', *IJIMAI*, Vol. 4, No. 1, pp.26–30.

Singh, V. and Kumar, P. (2020) 'Word sense disambiguation for Punjabi language using deep learning techniques', *Neural Computing and Applications*, Vol. 32, No. 8, pp.1–11.

Smalter, A., Lei, S.F. and Chen, X. (2007) 'Human disease-gene classification with integrative sequence-based and topological features of protein-protein interaction networks', in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pp.209–216.

Stolfi, P., Mastropietro, A., Pasculli, G., Tieri, P. and Vergni, D. (2023) 'NIAPU: network-informed adaptive positive-unlabeled learning for disease gene identification', *Bioinformatics*, Vol. 39, No. 2, p.btac848.

Sweet, R.M. and Eisenberg, D. (1983) 'Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure', *J. Mol. Biol.*, Vol. 171, No. 2, pp.479–488.

Universal Protein Resource [online] http://www.uniprot.org (accessed 11 February 2020).

Xia, J.F., Han, K. and Huang, D.S. (2010) 'Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor', *Protein Pept. Lett.*, Vol. 17, No. 1, pp.137–145.

Xu, J. and Li, Y. (2006) 'Discovering disease-genes by topological features in human protein-protein interaction network', *Bioinformatics*, Vol. 22, No. 22, pp.2800–2805.

Yang, P., Li, X., Chua, H.N., Kwoh, C.K. and Ng, S.K. (2014) 'Ensemble positive unlabeled learning for disease gene identification', *PloS One*, Vol. 9, No. 5, p.e97079.

Yang, P., Li, X., Wu, M., Kwoh, C.K. and Ng, S.K. (2011) 'Inferring gene-phenotype associations via global protein complex network propagation', *PloS One*, Vol. 6, No. 7, p.e21502.

Yang, P., Li, X.L., Mei, J.P., Kwoh, C.K. and Ng, S.K. (2012) 'Positive-unlabeled learning for disease gene identification', *Bioinformatics*, Vol. 28, No. 20, pp.2640–2647.

Yousef, A. and Charkari, N.M. (2013) 'A novel method based on new adaptive LVQ neural network for predicting protein-protein interactions from protein sequences', *Journal of Theoretical Biology*, Vol. 336, pp.231–239.

Yousef, A. and Charkari, N.M. (2015) 'A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification', *Journal of Biomedical Informatics*, Vol. 56, pp.300–306.

Yu, C.Y., Chou, L.C. and Chang, D.T. (2010) 'Predicting protein-protein interactions in unbalanced data using the primary structure of proteins', *BMC Bioinformatics*, Vol. 11, No. 1, p.167.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016) *Understanding Deep Learning Requires Rethinking Generalization*, arXiv:1611.03530.

Zhang, W., Sun, F. and Jiang, R. (2011) 'Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach', *BMC Bioinformatics*, Vol. 12, No. 1, pp.1–10.

Zulfiqar, H., Yuan, S.S., Huang, Q.L., Sun, Z.J., Dao, F.Y., Yu, X.L. and Lin, H. (2021) 'Identification of cyclin protein using gradient boost decision tree algorithm', *Computational and Structural Biotechnology Journal*, Vol. 1, No. 19, p.4123–4131.