Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Speech signals-based Parkinson's disease diagnosis using hybrid autoencoder-LSTM models

Ayşe Nur Tekindor [a],[*], Eda Akman Aydın [b]

[a] Graduate School of Natural and Applied Sciences, Department of Electrical and Electronics Engineering, Gazi University, Ankara, Turkiye
[b] Faculty of Technology, Department of Electrical and Electronics Engineering, Gazi University, Ankara, Turkiye

ARTICLE INFO

Source code link: https://github.com/hmbci-research-group/pd-diagnosis-using-autoencoder-lstm-models

Keywords:
Parkinson's disease
Speech analysis
Deep learning
Autoencoder
Long short-term memory

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that occurs as a result of a decrease in the chemical called dopamine in the brain. There is no definitive treatment for PD, but some medications used to control symptoms in the early stages have a critical effect on the progression of the disease. Approximately 90% of patients with PD have vocal problems, and although voice disorders seen in the early stages are not apparent in the patient's speech, they can be detected by acoustic analysis. In this study, a decision support system was proposed for the diagnosis of PD utilizing the feature extraction power of autoencoder (AE) & long short-term memory (LSTM) models by using speech signals as input data. Firstly, simple (SAE), convolutional (CAE), and recurrent (RAE) AE models were created for the ablation analysis. Then, the effect of hybridization and deepening of these models with LSTM layers on the classification performance was observed. Within the scope of the study, RAE achieved the highest accuracy among the base models while CAE & LSTM hybrid model provided the highest performance among all models with 95.79% accuracy for PD diagnosis based on audio signals. It was concluded that hybridization of the AE and LSTM models significantly improved the performance of simple and convolutional AE, and deepening the network to a certain extent improves the classification performance according to the type of AE.

1. Introduction

Parkinson's disease (PD) occurs as a result of a decrease in the brain chemical dopamine, which works as a neurotransmitter [1]. PD is the second most common neurodegenerative disease after Alzheimer's disease, and it is known that approximately ten million people worldwide are affected by PD [1,2]. The cause of PD is still unknown, but genetic and environmental factors are thought to be effective [3,4]. Individuals with PD may have difficulty walking, speaking, or completing other simple tasks. There is no definitive cure for the disease, but some treatment methods are applied to control symptoms in the early stages have a significant effect on the progression of the disease.

Hoehn and Yahr scale and Unified Parkinson's disease rating scale (UPDRS) are the common scales that are used in diagnosis of PD. However, these tests are subjective and depend on experience of the specialists. For this reason, more objective methods are needed for PD diagnosis.

Electroencephalography (EEG) serves as a powerful tool for directly reflecting brain activity, making it particularly valuable in understanding brain disorders and neuromodulation effects [5,6]. In addition to EEG [7], dopamine transporter (DAT) imaging [2], magnetic resonance imaging (MRI) [8], cardiac scintigraphy [9], cerebrospinal fluid (CSF) examination [10], analysis of handwriting patterns [11], and analysis of gait signals [12] are used to diagnose PD.

Some methods provide definitive results only in the advanced stages of the disease; however the amount of dopamine in the brain has decreased significantly in this stage. For this reason, early diagnosis of the disease is of great importance for the successful treatment process. One of the most important symptoms seen in the early stages of the disease in approximately 90% of people with PD is vocal problems [13]. For this reason, the use of speech signals in the detection of PD is a guiding method to manage diagnosis and treatment processes. Since the voice disorders seen in the early stages of PD are not very obvious in the patient's speech, it is very difficult for the human ear to distinguish these symptoms. Therefore, the development of decision support systems that enable the diagnosis of PD is an important tool for early diagnosis of the disease. In addition, the fact that collecting speech data from patients is easier and less costly than other methods used in diagnosis makes voice signals advantageous for use in these systems.

Studies conducted for speech signals based PD diagnosis can be examined specifically in machine learning (ML) and deep learning (DL).

[*] Corresponding author.
E-mail addresses: anur.tekindor@gazi.edu.tr (A.N. Tekindor), edaakman@gazi.edu.tr (E. Akman Aydın).

https://doi.org/10.1016/j.compbiomed.2025.110334
Received 18 December 2024; Received in revised form 24 April 2025; Accepted 2 May 2025
Available online 25 May 2025

In traditional ML-based methods, firstly features are extracted from signals. Then, these extracted features are applied to the input of a classifier. Polat and Nour [3] proposed a new data sampling method for the classification of PD based on acoustic features from speech signals. In this study, where the data set was divided into five equal parts with the one-against-all method, it was seen that the weighted K nearest neighbor (KNN) model used to classify each data set achieved 89.46% overall classification accuracy and gave more successful results than the classification process performed with acoustic features. In the study where feature selection with genetic algorithm and random forest model were used as a hybrid [1], the imbalance of classes in the data set was eliminated by increasing the number of samples of the minority class using the Synthetic Minority Upsampling Technique (SMOTE), and 95.58% accuracy was achieved with the proposed method. In the study where the performances of different ML and DL models were compared [14], the highest accuracy value of 97% was obtained with extreme gradient boosting. In another study [15] where synthetic examples were created for the minority class with the SMOTE method, the new dataset created was tested in the random forests model and 94.89% accuracy was obtained. It was shown that eliminating the imbalance between classes in the data is important for the accuracy of the study. In the study [16] where the features selected with the relief method were recreated with the variational autoencoder (VAE) and classified with SVM, 91.6% accuracy was obtained. In studies based on machine learning techniques, features such as baseline, acoustic, mel-frequency cepstral coefficients (MFCC), vocal fold, wavelet transform, tunable Q-factor wavelet transform were used [1,16]. However, these studies were limited to these features that were classified in the dataset, and mostly feature selection and classifier performance were evaluated. However, machine learning methods require more processing than deep learning techniques because they require feature engineering.

Deep learning (DL) is a form of ML that uses multilayer neural networks to model and understand complex patterns in data [17]. One of the biggest differences between ML and DL is that while various feature extraction and selection techniques are used for ML, DL does not require a feature extraction stage. Feature extraction in DL involves automatically identifying and extracting relevant features or representations from raw data such as images, audio or text. In recent years, artificial intelligence and DL techniques have been widely used in the diagnosis of neurodegenerative diseases and provide successful results [18,19].

Most of the studies where PD diagnosis with audio signals used time–frequency images of the signals such as spectrogram, mel-spectrogram, scalogram. In a study [20], where LSTM and autoencoder (AE) hybrid model was fed with spectrograms obtained from speech signals separated according to sentences in the text, 88.8% accuracy was achieved. In another study where features were extracted from spectrograms obtained from speech signals using convolutional autoencoder (CAE) [21], 84% accuracy was achieved using leave one subject out (LOSO) technique for both logistic regression and SVM models. In a study where spectrogram and scalogram images were used as input to the AE [22], when the softmax layer was used as a classifier, 87% classification accuracy was achieved for spectrogram and 82% for scalogram. In another study where mel-spectrograms were used [23], 84% classification accuracy was achieved with the recurrent autoencoder (RAE) and SVM hybrid model. Another study [24] used different pre-trained models to evaluate the effect of segment length and number on classification performance. By utilizing log mel-spectrograms obtained from speech signals, 91.80% accuracy was achieved. CNN&LSTM hybrid model was proposed in [25] by using spectrograms as input and 97.01% accuracy was obtained by vowel /i/. In another study [26], spectrograms obtain from vowel /a/ was used as input of the CNN & LSTM hybrid model and 95.67% accuracy was obtained. In the methods proposed in these studies, signals were converted into images and used as input data to two-dimensional models. Signal to image conversion requires more processing steps and high transaction cost. So, models that provide high diagnostic accuracy on 1D signals are needed.

There are studies in which sound signals are used directly as input data to models for PD diagnosis. In the study where the variational mode decomposition technique was used [27], 1D CNN model was fed with the intrinsic mode functions extracted from the signals, and 85.8% accuracy was achieved. Another study aimed to detect PD by decomposing speech signals into simpler intrinsic mode functions in the time domain with the empirical mode decomposition technique [28]. The 1D CNN model was fed with the extracted intrinsic mode functions. As a result of the study, an average accuracy of 73.76% was obtained. In the study where U-Lossian CNN model was fed directly with audio signals [29], 89.64% accuracy was achieved. The results of the studies show that the performance of the models using 1D signals was limited and no process was applied to balance the dataset.

Although the voice disorders seen in the early stages in people with PD are at a level that cannot be perceived by the listeners, these disorders can be detected by acoustic analysis. In the vast majority of studies conducted for the diagnosis of PD using audio signals, the detection system has been created by using specific feature vectors extracted from audio signals or time–frequency representations obtained from these features. Systems created in this way evaluate manually extracted numerical descriptors instead of evaluating the general structure of the sound. For this reason, models can only make predictions in line with the feature groups defined by the system designer. In studies where speech signals are used directly, no process has been performed to eliminate the imbalance of the data set. AE is an unsupervised learning method that has the power to extract features from unlabeled data. In this study, models that allow direct use of audio signals is proposed by taking advantage of the superiority of AE in feature extraction. Firstly, an ablation analysis was conducted to evaluate the effectiveness of simple, convolutional and recurrent AE models in detecting PD from audio signals. Then, the effect of hybridization and deepening of these models with LSTM layers on the classification performance was observed.

The paper is organized as follows. In Section 2, the proposed methodology is given. The results are given in Section 3. Discussion of the results are given in Section 4. Conclusion is outlined in Section 5.

2. Methodology

2.1. Dataset

The Italian Parkinson Voice and Speech (IPVS) dataset was collected by Dimauro and Girardi, and it is available for open access via IEEE Dataport [30]. The data set includes the voice recordings of native Italian speakers of 15 young healthy control (HC) group, 22 elderly HC group and 28 speakers of the PD group. All of these recordings were collected with professional microphones in a quiet, anechoic and ideally heated room at 16 and 44.1 kHz sampling frequencies. The speakers completed the following tasks: voicing the vowels /a/, /e/, /i/, /o/ and /u/ twice, repeated performance of the syllables “ka” and “pa” for 5 s, reading phonemically balanced sentences, reading phonemically balanced words and reading a phonemically balanced text twice [31]. For the balance in age, recordings of young HC group were not included, and only the vowel pronunciations of the elderly HC and PD groups were included in the experiments conducted within the scope of this study. The audio recordings of two speakers from the HC group were not included in the study because it was stated in the metadata provided in the dataset that they were of low quality. Experimental procedure of the proposed method in this study is given in Fig. 1.

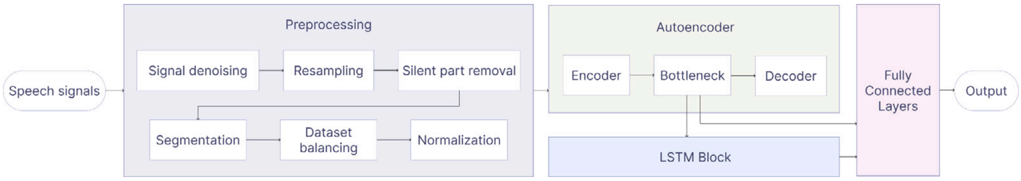

Fig. 1. Experimental procedure of the proposed method.

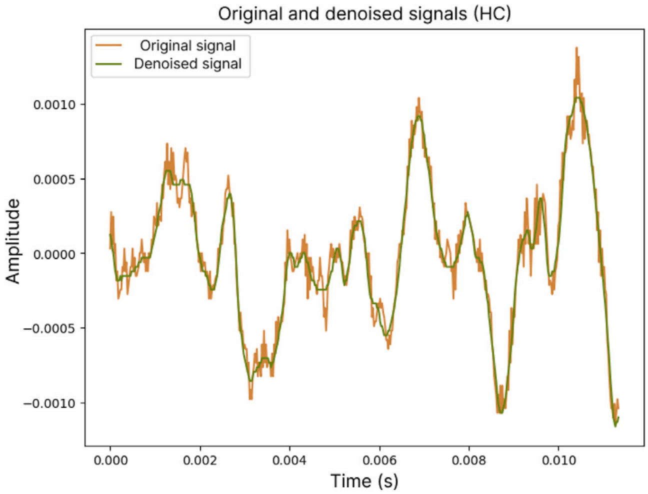

Fig. 2. A part of the original and denoised signals.

2.2. Preprocessing

All signals collected from analytical devices are affected by noise, which affects the accuracy and precision of the analysis. For this reason, it is important for the model to make successful predictions by removing noise from these signals. This process was carried out using the discrete wavelet transform method, which is based on the decomposition of the signal into wavelet coefficients at different scales. Considering that the audio signals were recorded in a quiet room, the decomposition level was applied with the Daubechies mother wavelet as 2 for signals with a sampling frequency of 16 kHz and 3 for signals with a sampling frequency of 44.1 kHz. A part of the original and denoised signal is shown in Fig. 2. After, since the voice recordings in the data set had two different sampling frequencies, 16 kHz and 44.1 kHz, the recordings with a sampling frequency of 44.1 kHz were down sampled to 16 kHz. Then, in order to make sure that the speech signals consisted of the voiced parts containing the main information, the first 2 s and the last 0.5 s of the signal were removed as considering that they are silent parts. Later, The signals were segmented into 25 ms (400 samples) non-overlapped windows. Since there were 20 subjects in the HC group and 28 subjects in the PD group in the data set, a large imbalance emerged in the class distribution when the recordings were segmented. Considering that some individuals in the PD group had longer voice recordings than those in the HC group, the imbalance between classes increased even more after segmentation. In order to eliminate this imbalance, random down sampling was performed to equalize the number of samples in the majority class to the number of samples in the minority class. Then, the input data was scaled between 0 and 1 by applying the min–max normalization method to make the classification process unbiased against any variable and to make the learning process more efficient.

2.3. Autoencoder (AE)

Autoencoder (AE) is an unsupervised deep learning technique that uses ANNs to learn data representations based on data compression and reconstruction [32]. AE can have a shallow design containing input and output layers and a hidden layer, or a deep structure containing a larger number of layers and neurons [33]. The process of transforming input data into a condensed representation in lower dimensions is called encoding, while the process of reconstructing the original data from these encoded data is called decoding. The coding process takes place in the encoder part of the AE structure, and the decoding process takes place in the decoder part. Latent space, also known as the bottleneck, is the space where the low-dimensional data representations obtained as the output of the encoder are defined. The representation in the latent space contains the most salient features of the original data. The structure of a basic AE is shown in Fig. 3. The function used to obtain a low-dimensional data array by making a non-linear matching of the input data array $x = [x_1, x_2, \dots, x_n]$ in the encoder is given in Eq. (1) [34]:

$$m = f(Wx + b) \tag{1}$$

where m is the encoded data sequence, f is the nonlinear activation function, W is the weight matrix, x is the input sequence, and b is the deviation vector. The function used in the decoder part to reconstruct the low-dimensional data array obtained after the coding process is given in Eq. (2).

$$x' = f(W'm + b') \tag{2}$$

where, x' refers to the data sequence reconstructed to be similar to the original input, f', W' and b' refer to the activation function, weight matrix and deviation vector of the decoder, respectively. A simple loss function for the nth sample is given in Eq. (3) [35].

$$L(x_n, x'_n) = \|x_n, f'(W'(f(Wx_n + b) + b'))\|2 \tag{3}$$

When training AEs, it is aimed to minimize the difference between the original input data and the reconstructed data [36]. As the model is trained, the extracted features contain the most important information, so the reconstructed data becomes free from irrelevant features. Therefore, classification results with higher accuracy can be obtained with encoded or reconstructed data. The reason why AE is a unsupervised learning method is that the focus of the models is to learn the input data in depth on its own through non-linear mapping functions and reconstruct it from a low-dimensional representation of this data, rather than labeling the data [37]. The type of AEs varies according to the architecture and purpose of use.

2.3.1. Simple autoencoder (SAE)

SAE is a type of AE that is considered a simple architecture because the encoder and decoder parts consist of fully connected layers [38,39]. SAEs provide size reduction by performing basic compression operations on data.

2.3.2. Convolutional autoencoder (CAE)

CAE is a type of AE in which the encoder and decoder parts are built based on CNN architecture [38,39]. These parts include structures found in a basic CNN model, such as the convolutional layer, activation function, and pooling layer. CNN is widely used in various biomedical signal processing applications as it can learn and extract all meaningful features from time series data [37]. The convolutional layer performs complex matrix multiplication, which uses filters on all

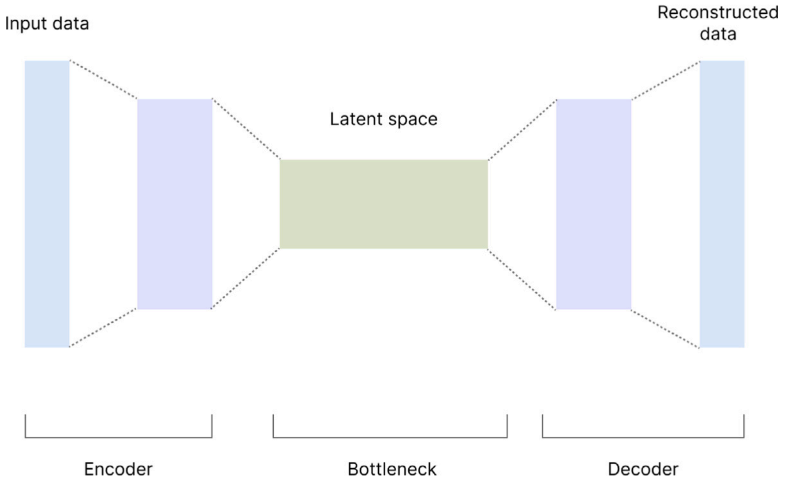

Fig. 3. The structure of autoencoder.

input data rather than using neurons, which can reduce the number of connections between layers. The functions used for encoding and decoding operations in the basic AE structure, respectively, in the CAE structure are given in Eqs. (4) and (5):

$$m = f(W * x + b) \tag{4}$$

$$x' = f(W' * m + b') \tag{5}$$

where W and W' refer to a set of filter matrices. The encoder creates feature maps by applying filters into input data. Upsampling layers are used in the decoder. To calculate the size of the output obtained after the convolution process, Eq. (6) is used:

$$O_s = x_s - f_s + 1 \tag{6}$$

where, O_s refers to the size of the output, x_s refers to the size of the input data, and f_s refers to the size of the filter.

2.3.3. Recurrent autoencoder (RAE)

RAE is a type of AE in which the encoder and decoder parts are built based on the recurrent neural network (RNN) architecture [40,41]. RNNs are powerful neural networks that can process sequential data; however, classical RNN-based AEs are difficult to train due to their long-term temporal dependency. For this reason, the LSTM structure and its variants that can overcome this problem are frequently used in RAEs [42]. The functions used for encoding and decoding operations in the basic AE structure, respectively, in the RAE structure are given in Eqs. (7)–(12) [43]:

$$i_t = f(W_i x_t + U_i h_{t-1} + b_i) \tag{7}$$

$$o_t = f(W_o x_t + U_o h_{t-1} + b_o) \tag{8}$$

$$f_t = f(W_f x_t + U_f h_{t-1} + b_f) \tag{9}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{10}$$

$$h_t = o_t \cdot tanh(c_t) \tag{11}$$

$$x'_t = W' h_t + b' \tag{12}$$

where, W is the weight matrix for the input data, U is the weight matrix for the hidden state, i_t is the vector of the input gate, o_t is the vector of the output gate, f_t is the vector of the forgetting gate, c_t is the cell state vector, and h_t is the hidden state vector.

Table 1
The names of the created models.

Model name	*Layers*	*Model name*	*Layers*
A1	SAE	AL4	CAE & 1 LSTM layer
A2	CAE	AL5	CAE & 2 LSTM layer
A3	RAE	AL6	CAE & 3 LSTM layer
AL1	SAE & 1 LSTM layer	AL7	RAE & 1 LSTM layer
AL2	SAE & 2 LSTM layer	AL8	RAE & 2 LSTM layer
AL3	SAE & 3 LSTM layer	AL9	RAE & 3 LSTM layer

2.4. Long short-term memory (LSTM)

Long short-term memory is a type of RNN designed to overcome the limitations of standard RNNs such as long-term dependency, vanishing and exploding gradients [44]. In this structure, it uses special units called memory cells to store information for a long time. These cells are controlled by input, output and forget gates that regulate the flow of information. The LSTM architecture has a chain-like structure [45]. When creating an LSTM network, the information to be collected from the cell is first determined. The input gate uses the sigmoid function to decide which values from the input will be activated and change the memory or which ones will be excluded [45]. The forget gate decides which information should be discarded from the cell state. This gate takes the previous hidden state and the current input, and outputs a number between 0 and 1 for each number in the cell state. The value 1 means “keep this completely”, while the value 0 means “forget this completely”. In this way, the information to be extracted from the cell is determined. The architecture of an LSTM cell is shown in Fig. 4.

2.5. Proposed hybrid AE-LSTM models

In this study, the AE architecture was used in three different ways as SAE, CAE and RAE. It was observed that the classification performance of hybridizing the models by adding the LSTM block to the AEs, and then increasing the number of this block in order to deepen the network. The models created vary depending on the type of AE and the number of LSTM layers.

The names of the created models are given in Table 1. “A” models are AE models, and “AL” models represent AE & LSTM hybrid models.

In the SAE model, there are fully connected layers with 256 and 128 units in the encoder part, 64 units in the bottleneck and 128 and 256 units in the decoder part, respectively. Architecture of the SAE-based models is shown in Fig. 5-a.

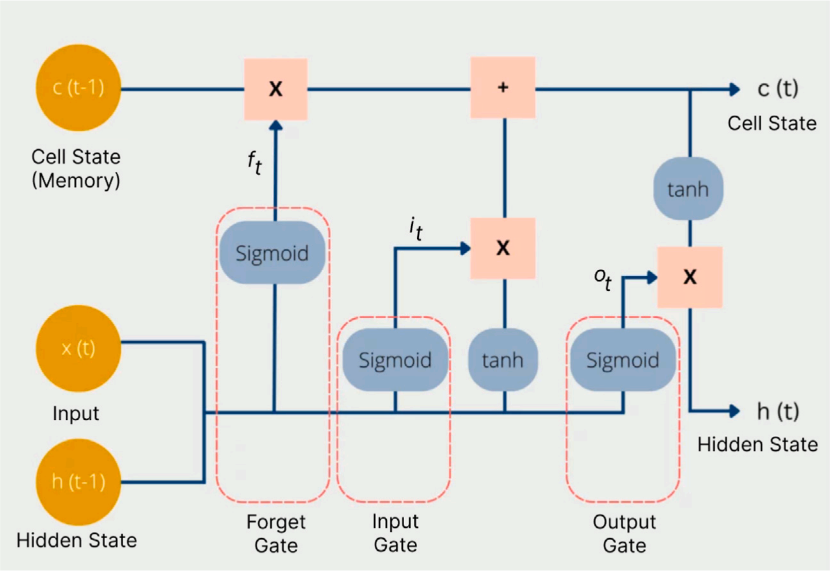

Fig. 4. LSTM cell architecture [32].

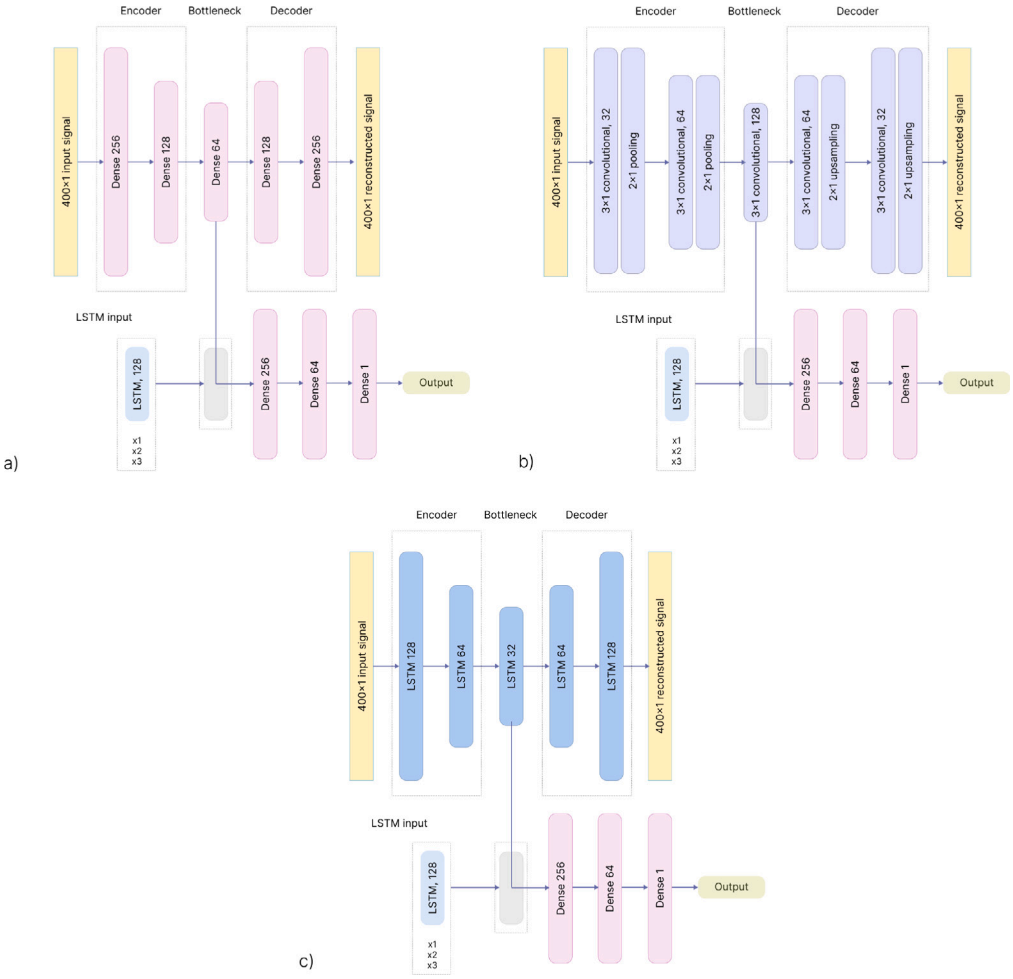

Fig. 5. Proposed AE & LSTM hybrid models: a) SAE-based model architecture, b) CAE-based model architecture, c) RAE-based model architecture.

In the encoder part of CAE, there are two convolutional layers with 32 and 64 filters with a kernel size of 3 × 1, respectively, and pooling layers with a window size of 2 × 1 following each layer, a convolutional layer with 128 filters in the bottleneck, and two convolutional layers

Table 2
Classification performance of the base SAE, CAE, and RAE models.

Model	*Accuracy (%)*	*Precision (%)*	*Sensitivity (%)*	*F1 Score (%)*
SAE (A1)	59.40	94.33	20.00	33.00
CAE (A2)	78.80	79.15	78.20	78.67
RAE (A3)	93.53	94.47	92.48	93.47

with 64 and 32 filters with a kernel size of 3 × 1, respectively, and 2 × 1 up-sampling layers following each layer. Architecture of the CAE-based models is shown in Fig. 5-b.

There are LSTM layers with 128 and 64 units in the encoder part of RAE, 32 units in the bottleneck and 64 and 128 units in the decoder part, respectively. Architecture of the RAE-based models is shown in Fig. 5-c.

All base models have a fully connected layer with 256 units, which is fed by the output of the bottleneck. Then it is followed by two fully connected layers with 64 and 1 unit. To hybridize the models, the LSTM block was added between the bottleneck and the 256-unit fully connected layer, and then the network was deepened by increasing the number of LSTM blocks there.

3. Results

In this study, the success of AE models in detecting PD from audio signals was examined, AE & LSTM hybrid models were created, and the number of layers were changed in order to observe the effect of model depth on the classification performance. For the classification process, firstly, the data set was divided into training and test data at a ratio of 80%–20% on a per-subject basis. This split prevents data leakage that occurs when records belonging to a subject are included in both the training and test sets. Similarly, the training data was divided into training and validation data at a ratio of 80%–20%. NVIDIA L4 Tensor Core GPU in Google Colab was used for training the models. The computer where the experiments were carried out had 16 GB RAM and an Intel Core i7-8565U processor with a speed of 1.80 GHz. Adam optimization was used to make the training faster and more efficient. Batch size was set to 32, learning rate to 0.001, and epoch to 50; however, to prevent over-learning and increase the generalization ability of the model, the early stopping parameter was used. The early stopping condition was met if the validation accuracy remained constant in 7 training rounds. Cross entropy was used as the loss function.

The classification performances of the base AE models are given in Table 2. With the SAE model (A1), 59.40% accuracy, 94.33% precision, 20% sensitivity and 33.00% F1 score were obtained. The CAE model (A2) provided more successful and balanced predictions than the SAE model (A1) with 78.80% accuracy, 79.15% precision, 78.67% sensitivity and 78.67% F1 score. Similarly, the RAE model (A3) achieved 93.53% accuracy, 94.47% precision, 92.48% sensitivity and 93.47% F1 score.

Comparative results of AE & LSTM hybrid models in PD detection are given in Table 3. With the addition of one LSTM layer to the SAE model, (AL1), accuracy increased to 89.32%, precision to 97.98%, sensitivity to 80.30% and F1 score to 88.26%. When the number of LSTM layers in the hybrid model was two (AL2), 90.38% accuracy, 88.41% precision, 92.93% sensitivity and 90.62% F1 score were obtained. The model with three LSTM layers (AL3) achieved 88.65% accuracy, 98.31% precision, 78.65% sensitivity and 87.39% F1 score.

With the addition of one LSTM layer to the CAE model, (AL4), accuracy increased to 91.43%, precision to 90.81%, sensitivity to 92.18%, and F1 score to 91.49%. The accuracy, sensitivity, and F1 score improved to 94.51%, 96.84%, and 94.64%, respectively, while precision decreased to 82.53% with the addition of two LSTM layers to the base CAE model (AL5). The model that has three LSTM layers, (AL6), achieved 95.79% accuracy, 83.19% precision, 98.80% sensitivity and 95.91% F1 score.

Table 3
Classification performance of AE & LSTM hybrid models.

Model	Metric	LSTM (1)	LSTM (2)	LSTM (3)
SAE	*Model*	AL1	AL2	AL3
	Accuracy (%)	89.32	90.38	88.65
	Precision (%)	97.98	88.41	98.31
	Sensitivity (%)	80.30	92.93	78.65
	F1 Score (%)	88.26	90.62	87.39
CAE	*Model*	AL4	AL5	AL6
	Accuracy (%)	91.43	94.51	95.79
	Precision (%)	90.81	82.53	83.19
	Sensitivity (%)	92.18	96.84	98.80
	F1 Score (%)	91.49	94.64	95.91
RAE	*Model*	AL7	AL8	AL9
	Accuracy (%)	88.12	88.65	86.82
	Precision (%)	86.79	89.06	89.79
	Sensitivity (%)	89.92	88.12	81.95
	F1 Score (%)	88.33	88.59	85.69

When an LSTM layer was added to the RAE model, (AL7), the accuracy decreased to 88.12%, precision to 86.79%, sensitivity to 89.92% and F1 score to 88.33%. The addition of two LSTM layers to the same model, (AL8), accuracy increased to 88.65%, precision to 89.06%, sensitivity to 88.12% and F1 score to 88.59%. With the addition of three LSTM layers to the model, (AL9), precision increased to 89.79% while accuracy, sensitivity and F1 score decreased to 86.82%, 81.95% and 85.69%, respectively. Fig. 6 shows the graph of the accuracy values of the AE & LSTM models.

Among the models using the base SAE model, the two-layer LSTM hybrid model (AL2) showed the highest performance with 90.38% accuracy, 88.41% precision, 92.93% sensitivity, and 90.62% F1 score. The model which showed the highest performance among CAE-based models is the three-layer LSTM hybrid model (AL6) with 95.79% accuracy, 83.19% precision, 98.80% sensitivity and 95.91% F1 score. This model was also the most successful among the twelve models created. The base RAE model (A3) showed the highest performance among the models where RAE was used with 93.53% accuracy, 94.47% precision, 92.48% sensitivity and 93.47% F1 score. Fig. 7 shows the confusion matrices of A3 and AL6 models.

4. Discussion

In this study, the performance of SAE, CAE, and RAE models for PD diagnosis based on audio signals was first evaluated as an ablation analysis. Then, by adding the LSTM layers to the AE models, the effect of hybridization and network depth on classification performance was examined. Within the scope of the study, SAE consisting of fully connected layers, CAE consisting of convolutional layers and RAE consisting of LSTM layers were hybridized with LSTM layers and the network depth was changed.

Accordingly, while the accuracy of SAE (A1) was limited to 59.40%, the accuracies of the CAE (A2) and RAE (A3) models, which made more successful predictions compared to SAE, reached to 78.80% and 93.53%. Then, by adding an LSTM layer to the A1 model, a significant increase in the performance of the hybridized model (AL1) was observed and 89.32% accuracy was obtained. While two LSTM layers (AL2) improved the performance of the model, the three LSTM layer model (AL3) showed lower performance than the one LSTM layer model (AL1) and higher performance than A1. Among SAE-based models, the AL2 model with two LSTM layers showed the highest performance with 90.38% accuracy. Similarly, the AL4 hybrid model, obtained by adding an LSTM layer to the A2 model consisting of CAE, showed high performance with a significant increase in accuracy to 91.43%. With the addition of two LSTM layers to this model, the AL5 model obtained 94.51% accuracy. The highest performance in CAE-based models was achieved by three-layer LSTM model AL6 with

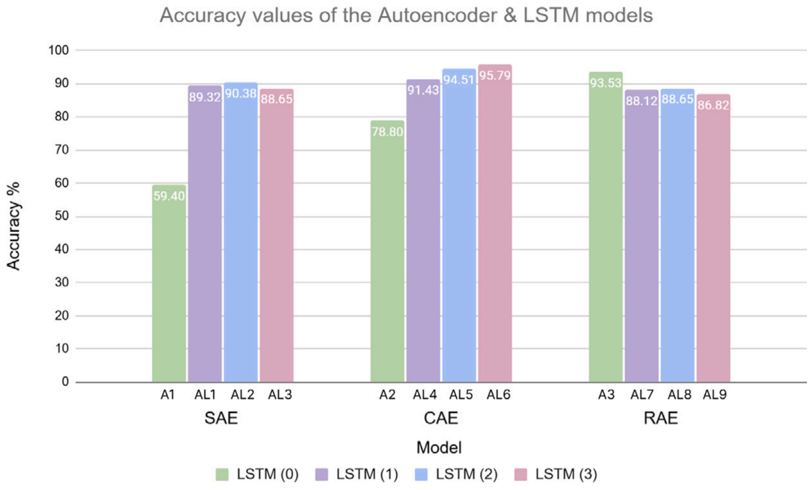

Fig. 6. Graph of the accuracy values of the AE & LSTM models.

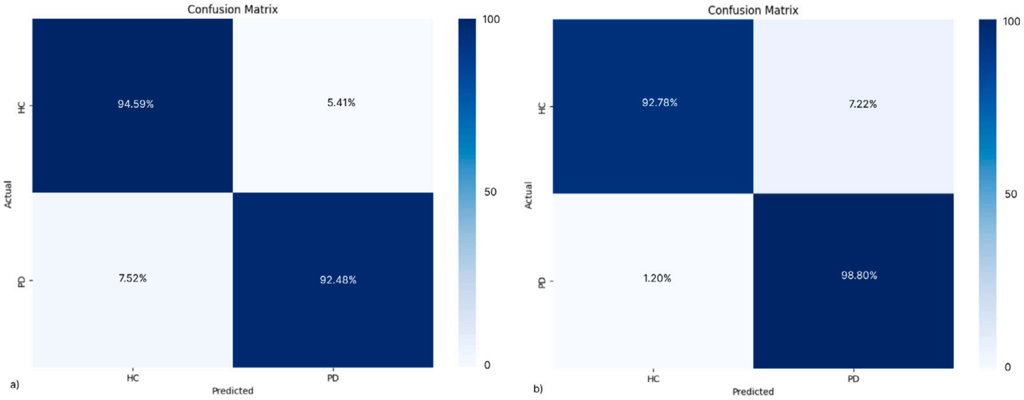

Fig. 7. Confusion matrices of the models showing the highest performance: a) RAE (A3), b) CAE with three LSTM layers (AL6).

95.79% accuracy. When the RAE model was hybridized with the LSTM layer (AL7), a decrease in the performance of the model was observed. Also, while deepening the model with two LSTM layers (AL8) made a slight improvement, three LSTM layers (AL9) caused a decrease in the performance. Among the RAE-based models, the A3 model showed the highest performance with 93.53% accuracy. Among the twelve AE & LSTM models created, the one which showed the highest performance was AL6 having CAE & three LSTM layers.

In hybrid models created with SAE, CAE and RAE, CAE & LSTM models generally showed the highest performance. The encoder and decoder parts that make up the SAE model consist of fully connected layers. Although the hybrid use of these layers with LSTM layers in the autoencoder continuation enabled the models to show significant prediction success, lower performance was observed than the convolutional layers. The reason why RAE & LSTM hybrid models show lower success than SAE & LSTM and CAE & LSTM hybrid models is that the encoder and decoder parts of the RAE model consist of LSTM layers. The data coming out of the AE provides maximum information for the model to learn, therefore deepening the model with LSTM layers could cause the model to become more complex. On the other hand, the maximum information obtained with the LSTM layers used in RAE enabled A3 to show the highest performance among the base models. This also shows the power of the LSTM structure alone. In summary, while hybridization of models positively affects classification success of the SAE and CAE models, it negatively affects the RAE model, deepening the network with LSTM layers improves the performance of the model to a certain extent according to other layers used in the architecture.

Comparative analysis of the studies using AE techniques in the literature is given in Table 4. When studies using the same data set were examined, the highest accuracy obtained from models using speech signals as input data, 89.64%, was obtained with the U-lossian CNN architecture [29]. In this study, no information is given about balancing the data set, cross-validation method and window size. Within the scope of this study, 95.79% accuracy was achieved with the CAE & LSTM hybrid model. Other studies using AE architecture have shown lower performance than the proposed method by using different data sets and different input data [20–23]. Studies using PC-GITA dataset [46] not only included vowels, but also included other recording groups such as word, syllable, sentence and text [22,23]. In addition, it should be taken into consideration that since no action was taken to eliminate the imbalance in the class distribution in the data set in the relevant study, problems such as overfitting may have occurred. For the other study [42] that reached 97.01% accuracy obtained that value only with spectrograms of vowel /i/, which limits the performance and causes overfitting with an F1 score of 58.80%.

The key highlights of this work is as follows:

- The diagnosability of PD based on sound signals with 1D autoencoders has been demonstrated.
- An accuracy of 93.53% was achieved with RAE from speech signals.
- CAE and LSTM together were able to diagnose PD from speech signals with 95.79% accuracy.
- CAE based hybrid models provided the highest accuracy among all hybrid models.

Table 4
Comparison of proposed methods with the literature.

Study	*Method*	*Dataset*	*Voice group*	*Test Accuracy %*	*F1 score %*	*Window/Overlap*	*Model Input Data*	*Cross Validation*	*Dataset Balance*
[20]	LSTM & AE	Hungarian	Sentence	88.8	–	10 ms	Spectrogram	5-Fold CV	Random person selection
[22]	AE & Softmax	PC-GITA	Vowel & Word	87 & 82	–	90 ms	Spectrogram & scalogram	Hold out CV	–
[23]	CAE & RAE & SVM	PC-GITA	Syllable Sentence Text Vowel	84	–	500 ms/250 ms	Mel-spectrogram	10-Fold CV	–
[21]	CAE & LR-SVM	TIMIT & IPVS	–	84	–	–	Spectrogram	LOSO	–
[24]	VGG16	IPVS	Vowel	91.80	–	5 s/ 2.5 s	Log mel-spectrogram	3-Fold CV	Over sampling
[25]	CNN & LSTM	IPVS	Vowel	97.01	58.8	25 ms/ 10 ms	Spectrogram	–	–
[27]	CNN	IPVS	Vowel	85.8	–	1 s	Speech signal	10-Fold CV	–
[28]	CNN	IPVS	Vowel	73.76	–	1 s	Speech signal	LOSO CV	–
Proposed Method	RAE	IPVS	Vowel	93.53	93.47	25 ms	Speech signal	Hold-out CV	Down sampling
[29]	U-Lossian (CNN)	IPVS	Vowel	89.64	89.74	–	Speech signal	–	–
Proposed Method	RAE	IPVS	Vowel	93.53	93.47	25 ms	Speech signal	Hold-out CV	Down sampling
Proposed Method	CAE & LSTM	IPVS	Vowel	95.79	95.91	25 ms	Speech signal	Hold-out CV	Down sampling

- Deepening the network of the hybrid models to a certain extent improved the classification performance and provided more balanced results.
- Due to the LSTM layers in AE structure, hybridization and deepening RAE with LSTM layers makes the model more complex for classification.

5. Conclusion

Parkinson’s disease is a neurodegenerative disease that occurs as a result of decreased dopamine levels in the brain. Although there is no cure of PD, early diagnosis is important for the success of symptomatic treatment. Speech disorders begin to appear in Parkinson’s patients in the early stages. These disorders can be detected with artificial intelligence based decision support systems, even if they are not perceived by listeners. Within the scope of this study, it was aimed to diagnose PD with speech signals by taking advantage of the feature extraction capability of deep learning models. In the study, the effect of hybridization and deepening the network on the classification performance was observed by adding LSTM blocks to AE models. Twelve different AE-based models were evaluated within the scope of the experiments; base RAE and hybrid CAE&LSTM models showed the highest performance. The proposed method was more successful than studies where speech signals were used directly. Using speech signals directly in deep learning models without converting them to time–frequency representations allows models to learn features automatically without the need for feature engineering. Unlike most of the studies in the literature, this study emphasizes the ability of deep learning models to automatically extract features from the original signal. However, it has been shown that hybrid models provide more successful results; deepening the network improves performance to a certain extent. The data preprocessing stage is of great importance for deep learning models to provide successful results. Good preprocessing allows the model to capture deeper information and make better classifications. For future studies, it is aimed to conduct studies that will increase the generalization performance of the models by trying different methods in this step.

Speech data is a data that can be easily collected with simple recording devices in real time or remotely. In this way, the diagnosis of PD based on speech analysis offers a non-invasive, easily accessible, and cost-effective technique. Especially recently, some technologies such as Artificial Intelligence of Things (AIoT) and Internet of Medical Things (IoMT) aim to integrate artificial intelligence and the Internet of Things or Internet of Medical Things to provide an intelligent remote diagnosis for enhancing medical services. With the rapid developments in AIoT and IoMT systems, they can be quickly adapted to clinical systems [47].

This study has provided an important solution that can be used with high accuracy for the diagnosis of Parkinson’s disease based on speech data. On the other hand, it is known that the staging of PD is an important requirement for the planning of treatment processes. Therefore, it is planned to focus on these limitations in future studies.

CRediT authorship contribution statement

Ayşe Nur Tekindor: Writing – original draft, Visualization, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Eda Akman Aydın:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No new data was created in this study. All code for the experiments is publicly available at https://github.com/hmbci-research-group/pd-diagnosis-using-autoencoder-lstm-models.

References

[1] R. Lamba, T. Gulati, H.F. Alharbi, A. Jain, A hybrid system for Parkinson’s disease diagnosis using machine learning techniques, Int. J. Speech Technol. 25 (3) (2022) 583–593.
[2] A. Keles, A. Keles, M. Keles, A. Okatan, PARNet: Deep neural network for the diagnosis of Parkinson’s disease, Multimedia Tools Appl. 83 (1) (2023) 1–13.
[3] K. Polat, M. Nour, Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals, Med. Hypotheses 140 (2020) 109678.
[4] E. Tolosa, A. Garrido, S.W. Scholz, W. Poewe, Challenges in the diagnosis of Parkinson’s disease, Lancet Neurol. 20 (5) (2021) 385–397.

[5] H. Yu, X. Wu, L. Cai, B. Deng, J. Wang, Modulation of spectral power and functional connectivity in human brain by acupuncture stimulation, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (5) (2018) 977–986.

[6] H. Yu, X. Li, X. Lei, J. Wang, Modulation effect of acupuncture on functional brain networks and classification of its manipulation with EEG signals, IEEE Trans. Neural Syst. Rehabil. Eng. 27 (10) (2019) 1973–1984.

[7] S.A.A. Shah, L. Zhang, A. Bais, Dynamical system based compact deep hybrid network for classification of Parkinson disease related EEG signals, Neural Netw. 130 (2020) 75–84.

[8] J. Sun, C. Cong, X. Li, W. Zhou, R. Xia, H. Liu, Y. Wang, Z. Xu, X. Chen, Identification of Parkinson's disease and multiple system atrophy using multimodal PET/MRI radiomics, Eur. Radiol. 34 (1) (2023) 662–672.

[9] R. Sakakibara, F. Tateno, M. Kishi, Y. Tsuyusaki, H. Terada, T. Inaoka, MIBG myocardial scintigraphy in pre-motor Parkinson's disease: A review, Parkinsonism Rel. Disord. 20 (3) (2014) 267–273.

[10] L. Parnetti, A. Castrioto, D. Chiasserini, E. Persichetti, N. Tambasco, O. El-Agnaf, P. Calabresi, Cerebrospinal fluid biomarkers in Parkinson disease, Nat. Rev. Neurol. 9 (3) (2013) 131–140.

[11] S. Hadadi, S.P. Arabani, A novel approach for Parkinson's disease diagnosis using deep learning and Harris Hawks optimization algorithm with handwritten samples, Multimed Tools Appl. 83 (2024) 81491–81510.

[12] B. Vidya, P. Sasikumar, Parkinson's disease diagnosis and stage prediction based on gait signal analysis using EMD and CNN–LSTM network, Eng. Appl. Artif. Intell. 114 (2022) 105099.

[13] C.O. Sakar, G. Serbes, A. Gunduz, H.C. Tunc, H. Nizam, B.E. Sakar, M. Tutuncu, T. Aydin, M.E. Isenkul, H. Apaydin, A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform, Appl. Soft Comput. 74 (2019) 255–263.

[14] P. Varalakshmi, R. Bhuvaneaswari, Parkinson disease detection based on speech using various machine learning models and deep learning models, in: 2021 International Conference on System, Computation, Automation and Networking, ICSCAN, Puducherry, 2021, pp. 1–6.

[15] K. Polat, A hybrid approach to Parkinson disease classification using speech signal: The combination of SMOTE and random forests, in: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science, EBBT, 2019, pp. 1–3.

[16] H. Gunduz, An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification, Biomed. Signal Process. Control. 66 (2021) 102452.

[17] A. Kamilaris, F.X. Prenafeta-Boldú, Deep learning in agriculture: A survey, Comput. Electron. Agric. 147 (2018) 70–90.

[18] H. Yu, X. Lei, Z. Song, C. Liu, J. Wang, Supervised network-based fuzzy learning of EEG signals for Alzheimer's disease identification, IEEE Trans. Fuzzy Syst. 28 (1) (2020) 60–71.

[19] M. Nour, U. Senturk, K. Polat, Diagnosis and classification of Parkinson's disease using ensemble learning and 1D-PDCovNN, Comput. Biol. Med. 161 (2023) 107031.

[20] P. Rozmán, D. Sztahó, G. Kiss, A.Z. Jenei, Automatic recognition of depression and Parkinson's disease by LSTM networks using sample chunking, in: 2021 IEEE 12th International Conference on Cognitive Infocommunications (CogInfoCom), Online, 2021, pp. 1–6.

[21] S.B. Appakaya, R. Sankar, E. Sheybani, Novel unsupervised feature extraction protocol using autoencoders for connected speech: Application in Parkinson's disease classification, in: 2021 Wireless Telecommunications Symposium (WTS), California, 2021, pp. 1–5.

[22] B. Karan, S.S. Sahu, K. Mahto, Stacked auto-encoder based time-frequency features of speech signal for Parkinson disease prediction, in: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), Amaravati, 2020, pp. 1–4.

[23] J.C. Vasquez-Correa, T. Arias-Vergara, M. Schuster, J.R. Orozco-Arroyave, E. Nöth, Parallel representation learning for the classification of pathological speech: Studies on Parkinson's disease and cleft lip and palate, Speech Commun. 122 (2020) 56–67.

[24] H.S. Malekroodi, N. Madusanka, B.I. Lee, M. Yi, Leveraging deep learning for fine-grained categorization of Parkinson's disease progression levels through analysis of vocal acoustic patterns, Bioeng. (2024) PMID: 38534569; PMCID: PMC10968564.

[25] P.V.K. Pandey, S.S. Sahu, B. Karan, S.K. Mishra, Parkinson disease prediction using CNN-LSTM model from voice signal, SN Comput. Sci. 5 (381) (2024).

[26] V. Shibina, T.M. Thasleema, A hybrid approach to detecting Parkinson's disease using spectrogram and deep learning CNN-LSTM network, Int. J. Speech Technol. 27 (2024) 657–671.

[27] I.K. Veetil, V. Sowmya, J.R. Orozco-Arroyave, E.A. Gopalakrishnan, Robust language independent voice data driven Parkinson's disease detection, Eng. Appl. Artif. Intell. 129 (2024) 107494.

[28] A. Tripathi, S.K. Kopparapu, CNN based Parkinson's disease assessment using empirical mode decomposition, in: 29th Association for Computing Machinery (ACM) International Conference on Information and Knowledge Management (CIKM 2020), Online, 2020, pp. 1–7.

[29] R. Maskeliūnas, R. Damaševičius, A. Kulikajevas, E. Padervinskis, K. Pribuišis, V. Uloza, A hybrid U-lossian deep learning network for screening and evaluating Parkinson's disease, Appl. Sci. 12 (22) (2022) 11601.

[30] G. Dimauro, F. Girardi, Italian Parkinson's voice and speech. IEEE DataPort, 2019, URL: http://dx.doi.org/10.21227/aw6b-tg17, Son Erişim Tarihi: 02.07.2024.

[31] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of speech intelligibility in parkinson's disease using a speech-to-text system, IEEE Access 5 (2017) 22199–22208.

[32] K. Zaman, M. Sah, C. Direkoğlu, M. Unoki, A survey of audio classification using deep learning, IEEE Access 11 (2023) 106620-106649.

[33] W.H.L. Pinaya, S. Vieira, R. Garcia-Dias, A. Mechelli, Autoencoders. Machine Learning, Academic Press, 2020, pp. 193–208.

[34] P. Liu, X. Sun, Y. Han, Z. He, W. Zhang, C. Wu, Arrhythmia classification of LSTM autoencoder based on time series anomaly detection, Biomed. Signal Process. Control. 71 (B) (2022) 103228.

[35] S. Mirzaei, P. Ghasemi, EEG motor imagery classification using dynamic connectivity patterns and convolutional autoencoder, Biomed. Signal Process. Control. 68 (2021) 102584.

[36] K. Jun, D.-W. Lee, K. Lee, S. Lee, M.S. Kim, Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition, IEEE Access 8 (2020) 19196–19207.

[37] E. Dasan, I. Panneerselvam, A novel dimensionality reduction approach for ECG signal via convolutional denoising autoencoder with LSTM, Biomed. Signal Process. Control. 63 (2021) 102225.

[38] M. Farina, Y. Nakai, D. Shih, Searching for new physics with deep autoencoders, Phys. Rev. 101 (2020) 075021.

[39] Y. Zhang, A better autoencoder for image: Convolutional autoencoder, in: International Conference on Neural Information Processing ICONIP17-DCEC (2018) Siem Reap, 2018, pp. 1–7.

[40] A. Taheri, Learning graph representations with recurrent neural network autoencoders, Deep. Learn. Day KDD (2018) 1–8.

[41] J. Cowton, I. Kyriazakis, T. Plötz, Bacardit J., A combined deep learning GRU-autoencoder for the early detection of respiratory disease in pigs using multiple environmental sensors, Sensors 18 (8) (2018) 2521.

[42] W. Yu, I.Y. Kim, C. Mechefske, Analysis of different RNN autoencoder variants for time series classification and machine prognostics, Mech. Syst. Signal Process. 149 (2021) 107322.

[43] Z. Ke, H. Vikalo, Real-time radio technology and modulation classification via an LSTM auto-encoder, IEEE Trans. Wirel. Commun. 21 (1) (2022) 370–382.

[44] N. Boualoulou, T.B. Drissi, B. Nsiri, CNN and LSTM for the classification of Parkinson's disease based on the GTCC and MFCC, Appl. Comput. Sci. 19 (2) (2023) 2.

[45] P.V.K. Pandey, S.S. Sahu, B. Karan, S.K. Mishra, Parkinson disease prediction using CNN-LSTM model from voice signal, Springer Nat. (SN) Comput. Sci. 5 (4) (2024) 381.

[46] J.R. Orozco-Arroyave, J.D. Arias-Londoño, J.F. Vargas-Bonilla, M.C. González-Rátiva, E. Nöth, New spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC'14, 2014, pp. 342–347.

[47] S. Saleh, Z. Alouani, O. Daanouni, S. Hamida, B. Cherradi, O. Bouattane, AIoT-based embedded systems optimization using feature selection for Parkinson's disease diagnosis through speech disorders, Intelligence- Based Med. 10 (2024) 100184.