Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3597347, 11 pages
https://doi.org/10.1155/2022/3597347

Research Article

Automated English Speech Recognition Using Dimensionality Reduction with Deep Learning Approach

Jing Yu, Nianhua Ye, Xueqin Du, and Lu Han

School of Humanities, Jiangxi University of Chinese Medicine, Nanchang 330004, Jiangxi, China

Correspondence should be addressed to Lu Han; 20030745@jxutcm.edu.cn

Received 29 December 2021; Revised 24 January 2022; Accepted 31 January 2022; Published 7 March 2022

Academic Editor: Deepak Kumar Jain

Speech recognition technology is a multidisciplinary field, comprising signal processing, pattern recognition, acoustics, artificial intelligence, etc. Presently, speech recognition plays a vital role in human-computer interface in information technology. Due to the advancements of deep learning (DL) models, speech recognition system has received significant attention among researchers in several areas of speech recognition like mobile communication, voice recognition, and personal digital assistance. This paper presents an automated English speech recognition using dimensionality reduction and deep learning (AESR-DRDL) approach. The proposed AESR-DRDL technique involves a series of operations, namely, feature extraction, preprocessing, dimensionality reduction, and speech recognition. During feature extraction process, a hybridization of high-dimension rich feature vectors is derived from the speech as well as glottal-waveform signals by the use of MFCC, PLPC, and MVDR techniques. Besides, the high dimensionality of features can be reduced by the design of quasioppositional poor and rich optimization algorithm (QOPROA). Moreover, the Bidirectional Long Short-Term Memory (BiLSTM) technique is employed for speech recognition, and the optimal hyperparameter tuning of the Bidirectional Long Short-Term Memory technique can be chosen using Adagrad optimizer. For the dimensionality reduction technique, the quasioppositional poor and rich optimization algorithm (QOPROA) is applied. The performance validation of the AESR-DRDL technique is carried out against benchmark datasets, and the results reported the better performance of the AESR-DRDL technique compared to recent approaches. The AESR-DRDL technique has shown to be superior in terms of recovery time, with an average of 0.50 days. The AESR-DRDL method's overall performance has been validated using benchmark datasets, and the results show that it outperforms more current technique. Because of this, the AESR-DRDL approach can be used to recognize English speech.

1. Introduction

Voice is widely employed and is considered of the significant data while interacting with people. Voice recognition technique permits machines to translate human voice signals into corresponding commands via understanding and recognition. [1]. When people want to expose some kind of data, they have to use the voice signals that carry the data. The speech signal comprises data about the speaker's personal identity and semantic content. The basis of speech recognition is that all the speakers have distinct features because of their pronunciation and own unique vocal tract features [2]. Speech recognition technology is a cross-discipline including pattern recognition, signal processing, artificial intelligence, hearing mechanism, and sound mechanism. Now, speech recognition has gradually become a basic technique of human-computer exchange in information technology. With the growth of constant speech recognition rate, speech recognition input is becoming an essential form of computer input [3].

Speech identification is the method of translating speech communication into text transcripts. With the growth of computation power and accessibility of transmission with computer through voice, speech detection system has attracted a growing interest. Speech-to-text and text-to-speech schemes are commonly employed in various techniques, namely, personal digital assistance systems, mobile communication, and search engines [4]. But speech recognition has a number of problems. Initially, some languages are existing for speech recognition system. Next, speech

recognition system has to deal with this variation in the data sets, namely, gender and accent [5]. Then, speech signal has several distorting features, namely, microphone quality, environmental, and background noise factor. For dependable voice recognition, speech recognition is considered one of the best capable solutions. However, achieving excellent identification performance necessitates a careful selection of sensory elements. Since deep neural networks can efficiently extract robust latent features that allow various recognition algorithms to validate revolutionary generalization capabilities under a wide range of application conditions, deep learning techniques have recently involved increasing attention in the machine-learning community. It is possible to understand what someone is saying when they use speech recognition software. It is the responsibility of ASR software to recognize human speech and convert it into text.

In order to conquer this challenge, deep learning (DL) architecture has gained much recognition in speech recognition research. The DL method is a subdivision of machine-learning (ML) technique, which employs a group of processes that try to model higher-level abstraction through a deep graph with various processing layers, consisting of many linear and nonlinear conversions [6]. It is the capability of a system or software program to identify words spoken audibly and turn them into legible text that is known as speech recognition or voice-to-text technology. The DL method offers automated selection and ranking of features in the data sets with effective algorithm. In recent times, DL method has gained huge interest and big achievements in the field of applications, like NLP, image processing, speech recognition, and sequence alignment. With the high performance and recent popularity of DL framework, authors begin to adapt this architecture for speech recognition problems. RNN and CNN are the most preferred architectures, i.e., utilized in speech recognition [7, 8].

This research creates a novel automated English speech recognition with dimensionality reduction and deep learning (AESR-DRDL) approach. The proposed AESR-DRDL technique designs the feature extraction process using hybridization of high-dimension rich feature vectors derived from the speech as well as glottal-waveform signals by the use of MFCC, PLPC, and MVDR techniques. In addition, the quasioppositional poor and rich optimization algorithm (QOPROA) is used for the dimensionality reduction technique. Finally, Adagrad optimizer with BiLSTM model is applied for the recognition of speech signals.

2. Related Works

This segment offers a complete review of recently developed speech recognition models. In [9], three methods are examined to enhance speech recognition on Mandarin-English code-switching tasks. Firstly, multitask learning (MTL) is presented, which allows language identity data to enable Mandarin-English code-switching ASR. Obtaining confidence scores from automated speech recognition (ASR) systems is extremely important for downstream applications. A number of recent studies have advocated the use of neural networks to learn confidence scores for words or utterances for end-to-end automatic speech recognition (ASR). The results of those investigations show that word confidence alone does not model deletions, and utterance confidence does not take use of word-level training signals. Next, examine word pieces, in opposition to graphemes, as English modelling units to decrease the modelling unit gaps among English and Mandarin. Even in nonphonetic languages such as English, phoneme-based models consistently beat graphene-based models when it comes to conventional speech detection techniques. In most cases, when more training data is collected, the performance gap between the two gets smaller. Then, it uses TL method for using large number of English and monolingual Mandarin information for compensating the sparsity problem of code-switching tasks. It is used in hybrid automatic speech recognition (ASR) systems to transfer knowledge from one language to another. Encoders and/or prediction networks in the destination language can be pretrained with the source language's models. It is used to initialize the target language AM in hybrid ASR systems. It depends on the initialization model for the encoder and prediction networks.

Weng et al. [10] presented attention-based sequence-to-sequence method for end-to-end speech recognition system. Initially, they proposed an input-feeding framework that feeds previous decoder hidden state data and context vector as input to the decoder. Next, they proposed a hypothesis generation system for consecutive minimal Bayes risk (MBR) training of sequence-to-sequence model in which softmax smoothing into N-best generation in MBR training is introduced. Jiao et al. [11] proposed a DBN-SVM model to detect and classify the error from pronunciation; the model corrects the errors and scores the quality in pronunciation. This method is protracted to speech assessment mode. Then, various researches have been conducted for testing various features, involving the real-timeliness of recognition, the accuracy of pronunciation classification and error detection, recognition rate of distinct vocabularies, and environments.

Sujatha et al. [12] designed a complex system for Speech Analysis using Lexical Analyzer (SALA). This method is utilized in various fields of interest like employment, teaching, one's dexterity in English vocabulary, and communication skills. The presented method takes input audio that comprises English speech. In [13], the linear prediction coding coefficient extraction technique is employed for summing up the information based on English digits pronunciation. After extracting the dataset, it can be employed for an ENN to identify the relations among the linear coding coefficients of audio file with the pronounced digit.

Lin et al. [14] focused on strong speech recognition in air traffic control (ATC) by developing a processing model for integrating multilingual speech recognition into an individual architecture using three cascaded models: pronunciation model (PM), language model (LM), and acoustic model (AM). The AM translates ATC speech into phoneme-based text sequences that the PM later converts to word-based order, i.e., the final objective of this study. In [15], a feature representation learning architecture has been proposed in this study. This method is encompassing the usage of combination of various extracted feature representations

with Compact Bilinear Pooling (CBP), Automated Speech Recognition (ASR), DNN as feature extractor, and last inference through optimized RNN classifiers.

3. The Proposed Model

In this study, an effective AESR-DRDL technique has been developed for the recognition of English speech signals. The proposed AESR-DRDL technique incorporates several stages of operations like preprocessing, feature extraction, QOPROA-based feature selection, Adagrad based hyperparameter optimization, and BiLSTM based speech recognition. The design of QOPROA model is used for reducing the dimensionality of the features and improving the recognition performance. Feature selection is important in text classification since it helps decrease the high dimensional feature space that exists. It is less expensive to compute more accurately to use text classification systems when the feature space is reduced in size. As a result, in text classification, the challenge of identifying the appropriate mix of features is critical. Figure 1 illustrates the overall process of AESR-DRDL technique. The detailed work of each section is suggested in the succeeding sections.

3.1. Preprocessing. The speech input in the figure is a novel voice signal gathered by the voice equipment; the preprocessed technique mostly contains 3 features: sampling the input original voice signals, antialiasing band-pass filter, and eliminating the noise impact caused by numerous features; the feature extraction method was mostly for extracting the reflection from the voice.

3.2. Feature Extraction. Here, the Perceptual Linear Prediction Cepstral Coefficients (PLPC), Perceptual Minimum-Variance Distortionless Response Cepstral Coefficients (PMVDR), Mel-frequency Cepstral Coefficients (MFCC), pitch (F0), and their 1st and 2nd order derivatives are derived as feature vectors of the input speech signals and glottal-waveform signals [16]. As the features are computed in distinct ways, the speech signals can be defined in several ways. Features are combined together for building a complete feature vector. In this study, the extraction of the MFCC and PLPC features takes place using Dan Ellis' toolbox, and pitch feature extraction is carried out by the use of COVAREP toolbox. Every feature extraction level is performed specifically on the input speech signals and glottal waveforms. For the extraction of glottal signals from speech signals, the COVAREP toolbox is applied. When the feature extraction procedure is completed, a feature matrix comprising of rows and columns representing the frames and distinct feature vector components is derived. Speech and glottal-waveform signals provide them. It can make various speaking signals. These are the features of a dataset. COVAREP extracted pitch, while Dan Ellis' toolbox obtained MFCC and PLPC. It targets glottal and vocal waves. Extraction of glottal signals with COVAREP A feature matrix is created for each frame and vector component.

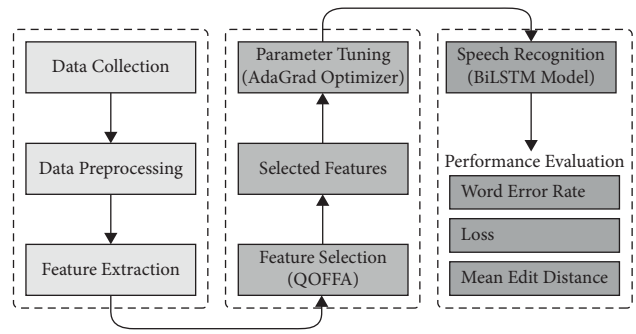

Figure 1: Overall process of AESR-DRDL technique.

3.3. Dimensionality Reduction Using QOPROA. Once the high dimensional features are derived, the QOPROA is utilized to choose an optimal subset of features. The PRO is depending on people's wealthier behavior in society [17]. Generally, they are classified into two financial categories within a society. Initially, it comprises richer people (wealth is greater than normal). Next, it comprises poorer people (wealth is less than regular). All the persons in this group are searching for improving their financial status in society. The poorly economical people try to enhance their financial status and decrease the class gap by learning from richer persons. In the optimization issue, all the solutions in the poorer population move to the global optimum solution in the searching space by learning from the richer solution in the Richer people. In this study, every solution or person in the population is signified as binary vector. The binary code 0 of the individual signifies that the feature was not selected, and the binary code 1 of the individual signifies that the feature was chosen.

The individual χ is signified as $\chi = [\eta_1, \eta_2, \eta_3, \ldots, \eta_n]$, whereas n denotes the amount of features in text corpus. All the positions of the solution or person α are binary values. $\eta_j \in \{0, 1\}$, For instance, a solution or person determined as [0, 1, 0, 1, 0, 0, 1, 1, 1, 0] represents that the feature or term with indexes 2, 4, 7, 8, and 9 is chosen, whereas the others are not selected [18]. The set of solutions in the present generation is named a population. Consider that "N" indicates the population size. Arbitrarily, create 'N' solution with real numbers within zero and one. The process of translating data into a digital representation is known as digitization. The data is arranged into distinct pieces of data (called bit s) that can be delivered individually (typically in multiple-bit groupings called byte s) in this fashion. Information can be preserved, accessed, and shared more easily when it is digitally stored. Afterward, the digitization method is used for all the locations of solution to convert real values into binary values as follows:

$$\chi_{i,j} = \begin{cases} 1, & \chi_{i,j} > \text{rand}, \\ 0, & \text{otherwise}. \end{cases} \quad (1)$$

Now, rand denotes an arbitrary value in the range of zero and one. The candidate solution in the population would be arranged according to the main purpose. The topmost part of the population is mentioned as richer economic class people, and the bottom part of populations are denoted as poorer economic class people as follows:

$$\mathrm{POP}_{\mathrm{Main}} = \mathrm{POP}_{\mathrm{rich}} + \mathrm{POP}_{\mathrm{poor}}. \quad (2)$$

Richer people move to raise the economic class gap by observing the people in poorer economical groups. The poorer economical group of people are moving to decrease the economic class gap by learning from the people in richer economical groups to improve their financial position. This natural behavior of poor and rich peoples is utilized for generating a novel solution. The movement of richer solution is determined in

$$\chi^{\mathrm{new}} = \chi^{\mathrm{old}}_{\mathrm{rich},i,j} + \alpha * \left[\chi^{\mathrm{old}}_{\mathrm{rich},i,j} | - \chi^{\mathrm{old}}_{\mathrm{poor,best},j}\right]. \quad (3)$$

The movement of poorer solutions is determined in

$$\chi^{\mathrm{new}}_{\mathrm{poor},i,j} = \chi^{\mathrm{old}}_{\mathrm{poor},i,j} + \alpha * \left[\left(\frac{\chi^{0}_{\mathrm{rich,best},j} + \chi^{0}_{\mathrm{rich,mean},j} + \chi^{0}_{\mathrm{rich,worst},j}}{3}\right) - \chi^{0}_{\mathrm{poor},i,j}\right]. \quad (4)$$

Opposition-based learning (OBL) is defined in order to reduce the computational complexity and enhance the convergence capability of distinct evolutionary algorithms (EAs) [19]. With the consideration of every present population and the opposite population depending upon the OBL concept, the candidate solutions can be enhanced. The quasiopposite number is used for the generation of optimal solutions compared to opposite number. The opposite number, opposite point, quasiopposite number, and quasiopposite point can be defined using the following equations. Different evolutionary algorithms (EAs) [19] are used to minimize computing complexity and improve convergence. The proposed solutions can be improved by taking into account the current and opposite populations based on the OBL principle. For ideal solutions, the quasiopposite number is employed. For any arbitrary number $\chi \in [a, b]$, the opposite number χ_0 can be represented as follows:

$$x_0 = a + b - x, \quad (5)$$

where the opposite point for multidimension searching area (d dimension) can be represented using

$$x_0^i = a^j + b^i - x^i, \quad i = 1, 2, \ldots, d, \quad (6)$$

and the quasiopposite number x_{qo} of any arbitrary number $\chi \in [a, b]$ can be denoted using

$$x_{qo} = \mathrm{rand}\left(\frac{a+b}{2}, x_0\right). \quad (7)$$

Likewise, the quasiopposite point for multidimension searching area (d dimension) can be represented by

$$x_{qo}^i = \mathrm{rand}\left(\frac{a^i + b^i}{2}, x_0^i\right). \quad (8)$$

The feature dimensionality reduction process takes place using the QOPROA technique. Each position vector considered the value of "0" or "1" where 0 represents that the features are not selected, and 1 indicates the selected features. The transfer function method implies the probability of modifying position vector elements among 0 and 1 efficiently. A transfer function significantly affects the performance of the FS processes and the outcome of the FS technique at the time of searching process. The fitness function of the QOPROA technique can be derived to determine the solutions in obtaining the tradeoff between two objectives, as defined in the following:

$$\mathrm{fitness} = \alpha \Delta_R(D) + \beta \frac{|Y|}{|T|}, \quad (9)$$

where $\Delta_R(D)$ denotes the classifier error rate, $|Y|$ indicates the number of features chosen, and $|T|$ implies a total number of features involved from the existing datasets. α indicates a variable $\in\in [0, 1]$ used to compare the weight of error rate of classification.

3.4. Speech Recognition Using Optimal BiLSTM Model. Finally, the recognition of speech signals takes place using the BiLSTM model. LSTM is a version of the RNN techniques that solves the issue of gradient vanishing. It helps improve the storage strategy of the NN for receiving input and training data. It is useful to model the time series data such as text. The BiLSTM is the integration of the backward LSTM and forward LSTM [20]. The major benefit of the BiLSTM model is that the sequence details are completely exploited in the network. The LSTM unit includes input gate i_t, forget gate f_t, and output gate 0_t, as well as a memory cell state c_t. They intend to influence the capability of the units in storing and upgrading data. The input gate offers a value in the range of 0 to 1 depending upon the input h_{t-1} and w_t. If the outcome becomes 1, it is defined that the cell state details are entirely sustained, and if the outcomes become zero, it gets entirely discarded. Followed by this, the input gate layer decides that the value needs to be upgraded, and tanh layer generates a novel candidate value vector $\widetilde{c_t}$, which can be appended to the cell state. Then, they can be integrated for updating the cell state c_t. Lastly, the final gate determines the outcome depending upon the cell state [20]. Specifically, W_f, U_f, b_f, W_i, U_i, b_j, W_c, U_c, b_c, and W_0, U_0, b_0 denote the intrinsic variables in the LSTM training process, $\sigma(.)$ is sigmoid activation function, and $\odot$ implies the dot multiplication.

$$
\begin{aligned}
f_t &= \sigma\left(W_f w_t + U_f h_{t-1} + b_f\right),\\
i_t &= \sigma\left(W_i w_t + U_i h_{t-1} + b_i\right),\\
\widetilde{c_t} &= \tanh\left(W_c w_c + U_c h_{t-1} + b_c\right),\\
c_t &= i_t \odot \overline{c_t} + f_t \odot c_{t-1},\\
0_t &= \sigma\left(W_0 w_t + U_0 h_{t-1} + b_0\right),\\
h_t &= o_t \tanh \odot \left(c_t\right).
\end{aligned} \tag{10}
$$

The BiLSTM model encompasses forward as well as backward LSTM. $\overrightarrow{\text{LSTM}}$ in BiLSTM reads input from w_1 to e_n for generating $\overrightarrow{h_t}$ and other $\overleftarrow{\text{LSTM}}$ read the input from e_n to w_1 for generating $\overleftarrow{h_{t-1}}$:

$$
\begin{aligned}
\overrightarrow{h_t} &= \overrightarrow{\text{LSTM}}\left(w_t, \overrightarrow{h_{t-1}}, c_{t-1}\right), \quad t \in [1, m+n],\\
\overleftarrow{h_t} &= \overleftarrow{\text{LSTM}}\left(w_t, \overleftarrow{h_{t-1}}, c_{t-1}\right), \quad t \in [m+n, 1].
\end{aligned} \tag{11}
$$

The forward and inverse sequence representations produced by $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are linked together to generate a long vector, and the integrated outcome defines the present time to the input:

$$
h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}. \tag{12}
$$

At last, the outcome $[h_1, \ldots h_i, \ldots h_m, l_1, \ldots l_j \ldots l_n]$ of the entire series is attained, where h_i and l_j are used for representing the outcome in the hidden layer. The intermittent layers in the BiLSTM model return the entire data and ensure that the outcome of every hidden layer sustains the long-term data. Figure 2 demonstrates the structure of BiLSTM technique.

For optimally tuning the hyperparameters of the BiLSTM model, the Adagrad optimizer is applied to it. In the Adagrad optimizer, the gradient and accumulated squared gradients for every variable can be computed at the round [21]:

$$
G_t = \sum_{\tau=1}^{t} g_\tau \odot g_\tau, \tag{13}
$$

where $\odot$ denotes elementwise multiplication, and $g_\tau \in \mathbb{R}^{|\theta|}$ indicates the gradient of present variable at the τ round. The variables in the Adagrad can be upgraded using

$$
\Delta\theta_t = -\frac{\alpha}{\sqrt{G_t + \varepsilon}} \odot g_\tau, \tag{14}
$$

where α implies learning rate, and ε denotes a smoothing component, which eliminates the division by zero. Since the learning rate is fixed prior to the training process, equation (15) can be determined as follows:

$$
\triangle\theta_t = -\alpha\left(\frac{1}{\sqrt{G_t + \varepsilon}} \odot g_\tau\right). \tag{15}
$$

It is known that G_t denotes the computation of earlier gradient, and g_t' can be represented as follows:

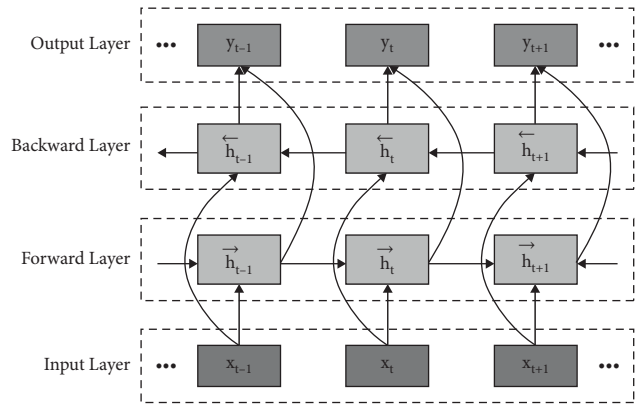

Figure 2: Structure of the BiLSTM model.

$$
g_t' = \frac{1}{\sqrt{G_t + \varepsilon}} \odot g_\tau. \tag{16}
$$

Therefore, the Adagrad can be upgraded using

$$
\Delta\theta_t = -\alpha g_t'. \tag{17}
$$

It is identical to the upgrade procedure of the classical gradient descent. Therefore, Adagrad optimizer can be considered for hyperparameter tuning using the gradient.

4. Experimental Validation

For experimental analysis, the data from the enhanced TEDLIUM release 2 corpus is used [8]. The results are examined in terms of different aspects. Table 1 and Figures 3–5 deal with the performance analysis of the AESR-DRDL technique under varying batch sizes (BS) and epoch counts. The figure offers comprehensive result analysis of the AESR-DRDL technique under BS of 32 and distinct epochs. The experimental results reported that the AESR-DRDL technique has accomplished maximum performance under every epoch. For instance, with 100 epochs, the AESR-DRDL technique has attained WER of 76.17%, loss of 154.63%, and MED of 0.3247. Likewise, with 300 epochs, the AESR-DRDL technique has accomplished WER of 75.22%, loss of 151.19%, and MED of 0.3175. Similarly, with 500 epochs, the AESR-DRDL technique has obtained WER of 76.90%, loss of 154.11%, and MED of 0.3278.

The experimental results provide comprehensive outcome analysis of the AESR-DRDL approach under BS of 64 and distinct epochs. The experimental results reported that the AESR-DRDL system has accomplished maximum performance under every epoch. For instance, with 100 epochs, the AESR-DRDL technique has attained WER of 75.23%, loss of 153.80%, and MED of 0.3230. Following this, with 300 epochs, the AESR-DRDL algorithm has accomplished WER of 75.11%, loss of 155.64%, and MED of 0.3268. Also, with 500 epochs, the AESR-DRDL method has obtained WER of 76.60%, loss of 151.07%, and MED of 0.3173.

A comprehensive analysis of the AESR-DRDL technique under BS of 128 and varying epochs reported that the AESR-

Table 1: Result analysis of AESR-DRDL technique with count of epochs.

Batch size = 32			
No. of epochs	WER (%)	Loss	Mean edit distance
100	76.17	154.63	0.3247
200	76.33	154.95	0.3254
300	75.22	151.19	0.3175
400	75.36	153.73	0.3228
500	76.90	156.11	0.3278
Average	**76.00**	**154.12**	**0.3237**
Batch size = 64			
No. of epochs	WER (%)	Loss	Mean edit distance
100	75.23	153.80	0.3230
200	75.55	152.05	0.3193
300	75.11	155.64	0.3268
400	75.16	151.69	0.3186
500	76.60	151.07	0.3173
Average	**75.53**	**152.85**	**0.3210**
Batch size = 128			
No. of epochs	WER (%)	Loss	Mean edit distance
100	75.18	151.86	0.3189
200	75.31	152.13	0.3195
300	76.16	154.60	0.3247
400	76.15	153.82	0.3230
500	76.49	153.74	0.3229
Average	**75.86**	**153.23**	**0.3218**

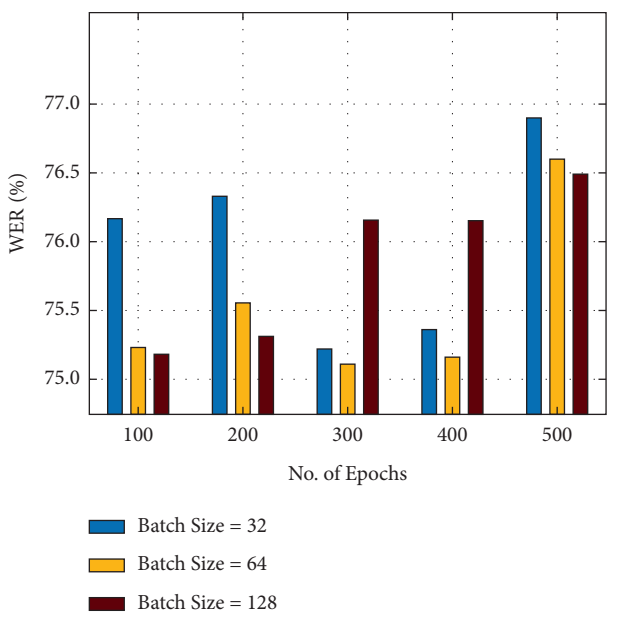

Figure 3: WER analysis of AESR-DRDL technique with varying epochs.

DRDL technique has accomplished maximal performance under every epoch. For instance, with 100 epochs, the AESR-DRDL technique has achieved WER of 75.18%, loss of 151.86%, and MED of 0.3189. In addition, with 300 epochs, the AESR-DRDL system has accomplished WER of 76.16%, loss of 154.60%, and MED of 0.3247. Eventually, with 500 epochs, the AESR-DRDL technique has reached WER of 76.49%, loss of 153.74%, and MED of 0.3229.

Table 2 and Figures 6–8 examine the performance analysis of the AESR-DRDL approach under different BS and layer counts. The figure depicts comprehensive result analysis of the AESR-DRDL approach under BS of 32 and distinct layers.

The experimental outcomes revealed that the AESR-DRDL technique has accomplished maximal performance under every layer. For instance, with 100 layers, the AESR-DRDL algorithm has attained WER of 75.39%, loss of 152.72%, and MED of 0.3207. Similarly, with 300 layers, the AESR-DRDL technique has accomplished WER of 76.67%, loss of 150.97%, and MED of 0.3170. Eventually, with 500 layers, the AESR-DRDL methodology has reached WER of 75.16%, loss of 156.26%, and MED of 0.3282.

The simulation values demonstrate the comprehensive result analysis of the AESR-DRDL system under BS of 64 and distinct layers. The experimental results reported that the AESR-DRDL methodology has accomplished increased performance under every layer. For instance, with 100 layers, the AESR-DRDL algorithm has gained WER of 76.05%, loss of 155.90%, and MED of 0.3274. Besides, with 300 layers, the AESR-DRDL technique has accomplished WER of 76.93%, loss of 155.40%, and MED of 0.3263. Lastly, with 500 layers, the AESR-DRDL technique has obtained WER of 75.10%, loss of 150.95%, and MED of 0.3170.

A comprehensive result analysis of the AESR-DRDL technique under BS of 128 and different layers revealed that the AESR-DRDL technique has accomplished higher performance under every layer. For sample, with 100 layers, the AESR-DRDL approach has attained WER of 75.85%, loss of 153.22%, and MED of 0.3218. Also, with 300 layers, the AESR-DRDL technique has accomplished WER of 76.65%, loss of 156.37%, and MED of 0.3284. Similarly, with 500 layers, the AESR-DRDL method has obtained WER of 74.48%, loss of 151.71%, and MED of 0.3186.

A comparative study of the AESR-DRDL technique with existing techniques takes place in Table 3 [22].

Figure 9 offers the WER analysis of the AESR-DRDL technique with existing techniques. The figure reported that the PPCA and DNN techniques have obtained higher WER of 88.10% and 88.06%, respectively. Following this, the RNN and PQPSO techniques have attained slightly reduced WER of 87.02% and 87.67%, respectively. Moreover, the LSTM and GRU techniques have accomplished reasonable WER of 77.55% and 79.39%, respectively. However, the AESR-DRDL technique has resulted in maximum outcome with a minimal WER of 75.53%.

Figure 10 gives the loss analysis of the AESR-DRDL method with existing techniques. The figure stated that the PPCA and DNN approaches have obtained higher loss of 185.53 and 185.01 correspondingly. Then, the RNN and PQPSO methods have gained slightly lower loss of 186.61 and 179.25, respectively. Furthermore, the LSTM and GRU techniques have accomplished reasonable loss of 160.51 and 162.22 correspondingly. But, the AESR-DRDL algorithm has resulted in maximal outcomes with a minimal loss of 152.85.

Figure 11 provides the MED analysis of the AESR-DRDL system with existing approaches. The figure depicted that the PPCA and DNN techniques have obtained higher MED of

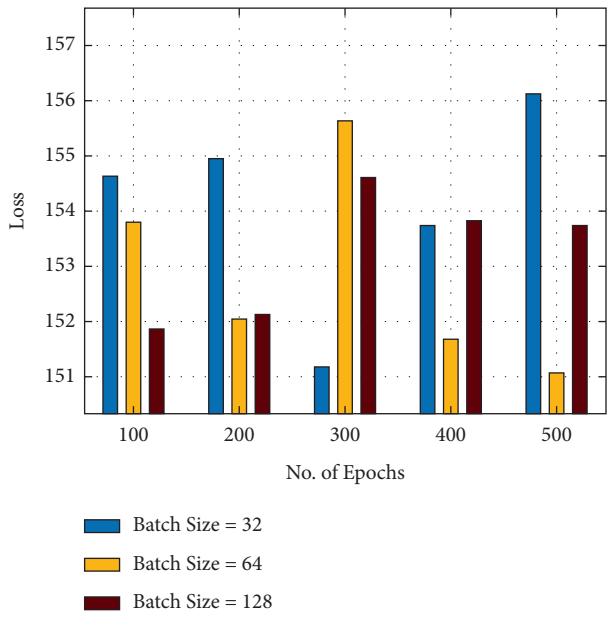

Figure 4: Loss analysis of AESR-DRDL technique with varying epochs.

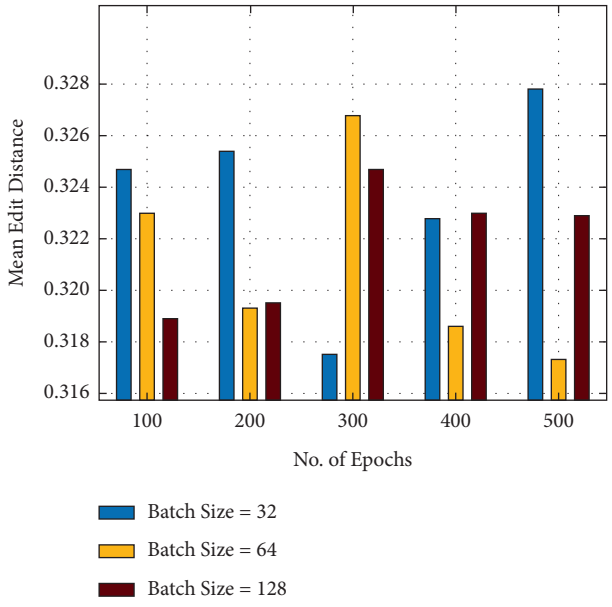

Figure 5: MED analysis of AESR-DRDL technique with varying epochs.

0.4407 and 0.4574, respectively. Afterward, the RNN and PQPSO techniques have attained somewhat lesser MED of 0.4484 and 0.4515, respectively. Moreover, the LSTM and GRU techniques have accomplished reasonable MED of 0.3853 and 0.3939 respectively. However, the AESR-DRDL methodology has resulted in maximal outcomes with the lesser MED of 0.3210.

Table 4 and Figure 12 define the running time (RT) analysis of the AESR-DRDL approach with existing techniques. The figure reported that the PPCA and DNN techniques have obtained higher RT of 2 days and 1.60 days correspondingly. Besides, the RNN and PQPSO techniques have obtained slightly lesser RT of 1.10 days and 1.50 days correspondingly. Moreover, the LSTM and GRU techniques have accomplished reasonable RT of 0.80 days and 1.02 days correspondingly. Finally, the AESR-DRDL technique has resulted in superior outcome with the minimal RT of 0.50 days.

Table 2: Result analysis of AESR-DRDL technique with count of layers.

Batch size = 32			
No. of layers	WER (%)	Loss	Mean edit distance
100	75.39	152.72	0.3207
200	75.27	151.86	0.3189
300	76.67	150.97	0.3170
400	75.47	154.08	0.3236
500	75.16	156.26	0.3282
Average	**75.59**	**153.18**	**0.3217**
Batch size = 64			
No. of layers	WER (%)	Loss	Mean edit distance
100	76.05	155.90	0.3274
200	75.60	154.22	0.3239
300	76.93	155.40	0.3263
400	75.91	154.86	0.3252
500	75.10	150.95	0.3170
Average	**75.92**	**154.27**	**0.3240**
Batch size = 128			
No. of layers	WER (%)	Loss	Mean edit distance
100	75.85	153.22	0.3218
200	76.07	152.90	0.3211
300	76.65	156.37	0.3284
400	74.09	151.88	0.3190
500	75.48	151.71	0.3186
Average	**75.63**	**153.22**	**0.3218**

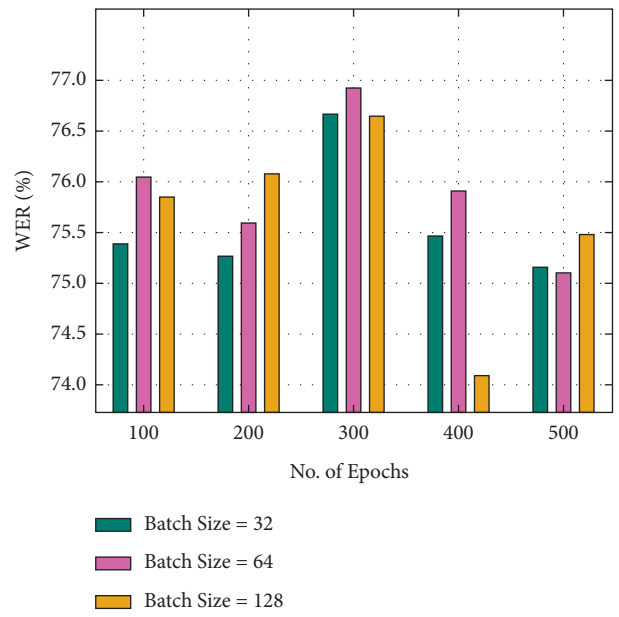

Figure 6: WER analysis of AESR-DRDL technique with varying layers.

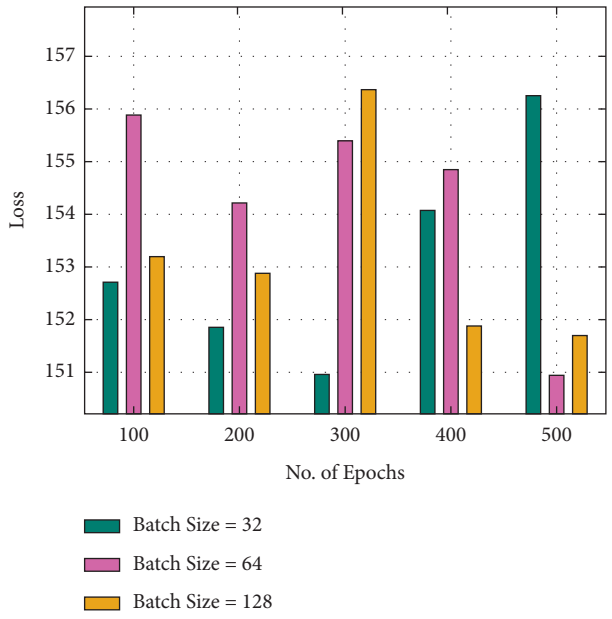

Figure 7: Loss analysis of AESR-DRDL technique with varying layers.

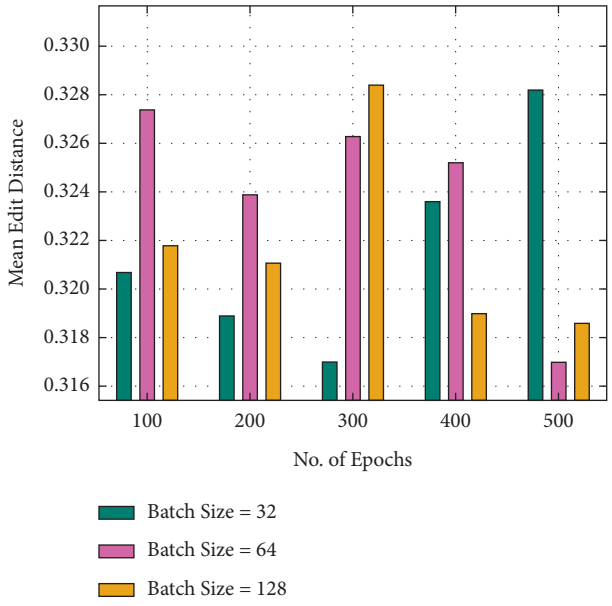

Figure 8: MED analysis of AESR-DRDL technique with varying layers.

Table 3: Comparative analysis of AESR-DRDL technique with existing algorithms.

Methods	WER (%)	Loss	Mean edit distance
RNN	87.02	186.61	0.4484
LSTM	77.55	160.51	0.3853
GRU	79.39	162.22	0.3939
PPCA	88.10	185.53	0.4407
DNN	88.06	185.01	0.4574
PQPSO	87.67	179.25	0.4515
AESR-DRDL	75.53	152.85	0.3210

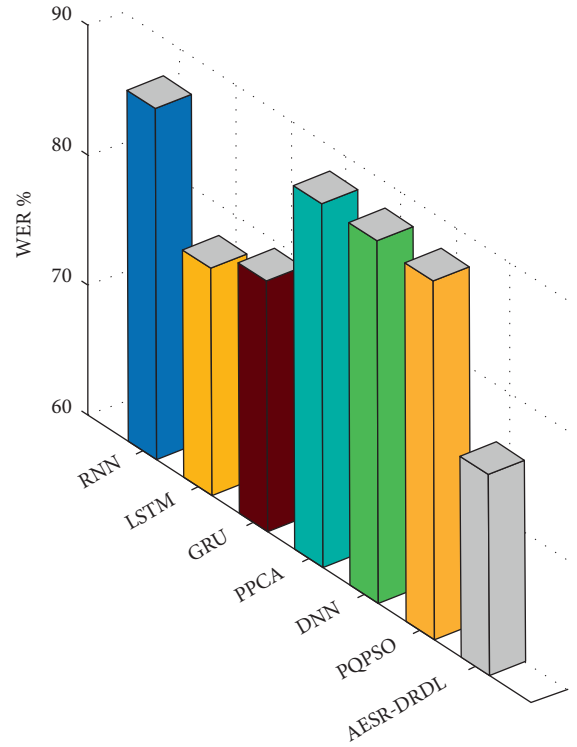

Figure 9: WER analysis of AESR-DRDL technique with existing algorithms.

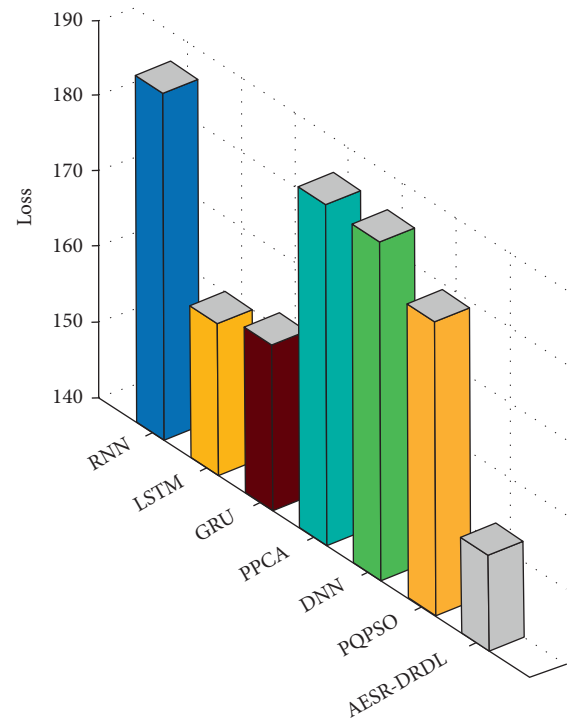

Figure 10: Loss analysis of AESR-DRDL technique with existing algorithms.

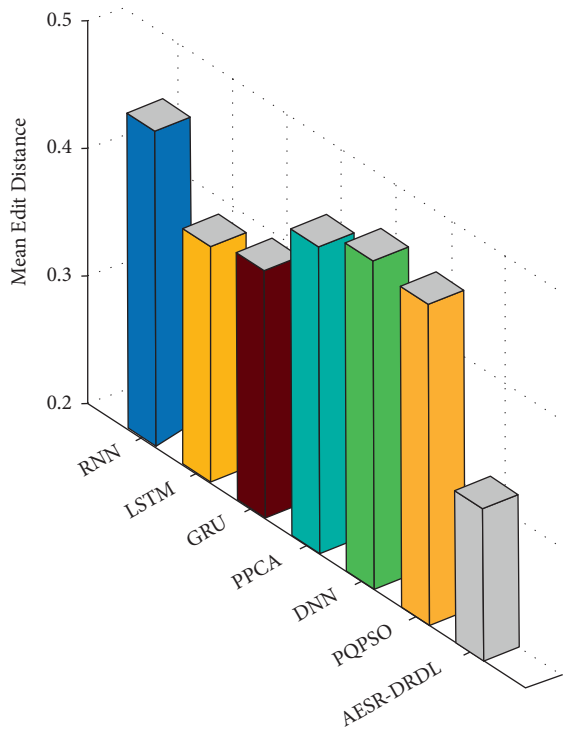

Figure 11: MED analysis of AESR-DRDL technique with existing algorithms.

Table 4: Running time (RT) analysis of AESR-DRDL technique with existing algorithms.

Methods	Running time (days)
RNN	1.10
LSTM	0.80
GRU	1.02
PPCA	2.00
DNN	1.60
PQPSO	1.50
AESR-DRDL	0.50

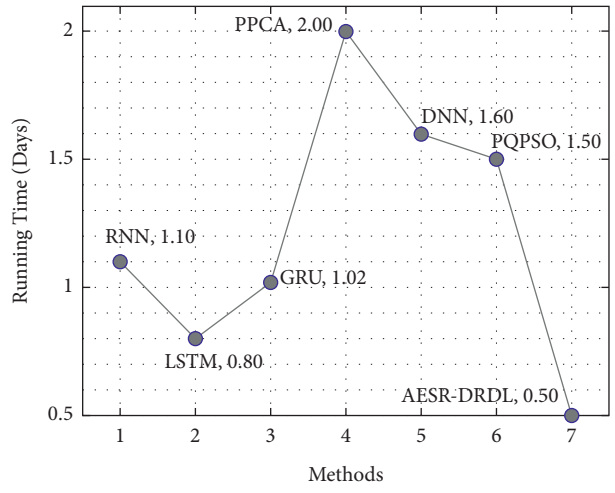

Figure 12: RT analysis of AESR-DRDL technique with existing algorithms.

From the abovementioned tables and figures, it can be ensured that the AESR-DRDL algorithm has resulted in maximum speech recognition performance over the other techniques.

5. Conclusion

In this study, an effective AESR-DRDL technique has been developed for the recognition of English speech signals. The proposed AESR-DRDL technique incorporates several stages of operations, namely, preprocessing, feature extraction, QOPROA-based feature selection, BiLSTM based speech recognition, and Adagrad based hyperparameter optimization. The design of QOPROA technique helps in reducing the dimensionality of the features and improving the recognition performance. Automated English voice recognition (AESR-DRDL) is presented in this paper, using dimensionality reduction and deep learning (AESR). Prior processing, feature extraction, dimension reduction, and speech recognition are all included in this new AESR-DRDL method. High-dimensional rich feature vectors are extracted during feature extraction from the speech and glottal-waveform signals by the employment of MFCC, PLPC, and MVDR approaches. Furthermore, the invention of a quasioppositional poor and rich optimization approach can lower the high dimensionality of features (QOPROA). In order to highlight the enhanced experimental result analysis of the AESR-DRDL technique, a wide range of simulations occur on benchmark datasets, and the results indicated the superior outcomes of the AESR-DRDL technique compared to recent approaches. With an average recovery time of 0.50 days, the AESR-DRDL approach has proven to be superior. The performance validation of the AESR-DRDL technique is carried out against benchmark datasets, and the findings revealed that the AESR-DRDL technique outperformed recent approaches in terms of overall performance. Therefore, the AESR-DRDL technique can be applied as an effective tool for English speech signal recognition. Many simulations were done using benchmark datasets to show the AESR-DRDL technique's superiority. AES-DRDL outperformed others. The AESR-DRDL approach is used to recognize English voice signals. In future, hybrid DL models can be used instead of the BiLSTM model to further improve the recognition performance.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[2] X. Cui, Z. Chen, and F. Yin, "Speech enhancement based on simple recurrent unit network," *Applied Acoustics*, vol. 157, Article ID 107019, 2020.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on International Conference on Machine learning*, pp. 1764–1772, Bejing, China, June 2014.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Wang, X. Wang, and S. Lv, "End-to-end Mandarin speech recognition combining cnn and blstm," *Symmetry*, vol. 11, pp. 1–19, 2019.

[6] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[7] P. Wei and Y. Zhao, "A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep auto-encoder model," *Pers Ubiquit Computation*, vol. 23, no. 3-4, pp. 521–529, 2019.

[8] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer Publishing Company, New York, NY, USA, 2014.

[9] C. Shan, C. Weng, G. Wang et al., "Investigating end-to-end speech recognition for Mandarin-English code-switching," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6056–6060, IEEE, Brighton, UK, May 2019.

[10] C. Weng, J. Cui, G. Wang et al., "Improving attention based sequence-to-sequence models for end-to-end English conversational speech recognition," in *Proceedings of the Interspeech*, pp. 761–765, Hyderabad, India, September 2018.

[11] F. Jiao, J. Song, X. Zhao, P. Zhao, and R. Wang, "A spoken English teaching system based on speech recognition and machine learning," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 14, pp. 68–82, 2021.

[12] B. Sujatha, B. Vanajakshi, and K. N. Nirmala, "SALA-an integrated framework for speech recognition using lexical analyzer," in *Proceedings of the 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, pp. 386–390, IEEE, October 2020.

[13] S. Khajehasani and L. Dehyadegari, "Speech recognition using elman artificial neural network and linear predictive coding," *Recent Advances in Computer Science and Communications*, vol. 13, no. 4, pp. 650–656, 2020.

[14] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, "A unified framework for multilingual speech recognition in air traffic control systems," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[15] B. Sertolli, Z. Ren, B. W. Schuller, and N. Cummins, "Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech," *Computer Speech & Language*, vol. 68, Article ID 101204, 2021.

[16] F. Daneshfar and S. J. Kabudian, "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 1261–1289, 2020.

[17] S. H. Samareh Moosavi and V. K. Bardsiri, "Poor and rich optimization algorithm: a new human-based and multi populations algorithm," *Engineering Applications of Artificial Intelligence*, vol. 86, pp. 165–181, 2019.

[18] R. R. Bhukya, B. M. Hardas, and T. Ch, "An automated word embedding with parameter tuned model for web crawling," *Intelligent Automation & Soft Computing*, vol. 32, no. 3, pp. 1617–1632, 2022.

[19] T. Van Tran, B. H. Truong, T. P. Nguyen, T. A. Nguyen, T. L. Duong, and D. N. Vo, "Reconfiguration of distribution networks with distributed generations using an improved neural network algorithm," *IEEE Access*, 2021.

[20] M. Prakash, A. Harshavardhan, and D. Sivabalaselvamani, "An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM," *Concurrent Engineering Research and Applications*, vol. 29, no. 4, pp. 386–395.

[21] S. Wang, Y. Zhu, W. Gao, M. Cao, and M. Li, "Emotion-semantic-enhanced bidirectional LSTM with multi-head attention mechanism for microblog sentiment analysis," *Information*, vol. 11, no. 5, p. 280, 2020.

[22] C. Zhang, M. Yao, W. Chen, S. Zhang, D. Chen, and Y. Wu, *Gradient Descent Optimization in Deep Learning Model Training Based on Multistage and Method Combination Strategy*, Security and Communication Networks, 2021.