applied sciences

Article

Multi-Label Extreme Learning Machine (MLELMs) for Bangla Regional Speech Recognition

Prommy Sultana Hossain [1], Amitabha Chakrabarty [1], Kyuheon Kim [2] and Md. Jalil Piran [3,*]

[1] Department of Computer Science and Engineering, Brac University, Dhaka 1212, Bangladesh; prommy.sultana.ferdawoos.hossain@g.bracu.ac.bd (P.S.H.); amitabha@bracu.ac.bd (A.C.)
[2] Media Laboratory, Kyung Hee University, Yong-in 17104, Korea; kyuheonkim@khu.ac.kr
[3] Department of Computer Science and Engineering, Sejeong University, Seoul 05006, Korea
* Correspondence: piran@sejong.ac.kr

Abstract: Extensive research has been conducted in the past to determine age, gender, and words spoken in Bangla speech, but no work has been conducted to identify the regional language spoken by the speaker in Bangla speech. Hence, in this study, we create a dataset containing 30 h of Bangla speech of seven regional Bangla dialects with the goal of detecting synthesized Bangla speech and categorizing it. To categorize the regional language spoken by the speaker in the Bangla speech and determine its authenticity, the proposed model was created; a Stacked Convolutional Autoencoder (SCAE) and a Sequence of Multi-Label Extreme Learning machines (MLELM). SCAE creates a detailed feature map by identifying the spatial and temporal salient qualities from MFEC input data. The feature map is then sent to MLELM networks to generate soft labels and then hard labels. As aging generates physiological changes in the brain that alter the processing of aural information, the model took age class into account while generating dialect class labels, increasing classification accuracy from 85% to 95% without and with age class consideration, respectively. The classification accuracy for synthesized Bangla speech labels is 95%. The proposed methodology works well with English speaking audio sets as well.

Keywords: Bangla regional speech classification; Stacked Convolution Autoencoder (SCAE); Multi-Label Extreme Learning machine (MLELMs); Mel Frequency Energy Coefficients (MFECs)

Citation: Hossain, P.S.; Chakrabarty, A.; Kim, K.; Piran, M.J. Multi-Label Extreme-Learning Machine (MLELMs) for Bangla Regional Speech Recognition. *Appl. Sci.* **2022**, *12*, 5463. https://doi.org/10.3390/app12115463

Academic Editor: Lijiang Chen

Received: 25 March 2022
Accepted: 16 May 2022
Published: 27 May 2022

1. Introduction

The Bangla language is well-known around the world, and it is the fifth most spoken language on the planet [1]. The population of Bangladesh speak two different varieties of Bangla. Few people speak the local language of their region. The mainstream Bangla language, which is spoken by about 290 million people, is another variety. There are 55 regional languages spoken in Bangladesh's 64 districts. A regional language, also known as a dialect, is a language a child learns organically without the use of written grammar, and that varies by region. It is a characteristic of languages that are widely spoken in a given location [1] that causes morphological differences in the sounds of the ideal language or literary language. Despite regional variations, the Bangla language can be divided into six classes: Bangla, Manbhumi, Varendri, Rachi, Rangpuri, and Sundarbani. This study primarily focused on seven regional languages; Khulna, Bogra, Rangpur, Sylhet, Chittagong, Noakhali, and Mymensingh divisions, which all belong to one of these classes, and one was chosen at random. A person's regional language can be identified by the wave frequency (pronunciation) of a word pronounced in Bangla.

On a global scale, the human voice is the most widely used method of communication between machines and humans. Dialects are what bind individuals together. Dialects help people convey their ideas more effectively. As stated by Honnet et al. [2], speech recognition is the ability of a machine or computer software to recognize phrases and words in spoken language and translate them into machine-readable form. Voice signals contain not only information about the content of speech, but also about the speaker's identity, emotions, age, gender, and geographical location. In human-computer interaction (HCI), voice signals are also crucial.

Identifying characteristics of audio streams automatically has recently been a hot topic of research for the Bangla language [1,3–16]. Application development based on consumer regional language, age, or gender, as well as a caller-agent coupling in contact centers that correctly allocates agents according to the caller's identity, are made possible by an accurate and efficient extraction of speaker identification from a voice signal. However, in most of these studies, the speaker's regional language is not categorized or a method for detecting synthetic Bangla speech audio signals is not presented. Several works have been done on creating synthesized Bangla speech. According to Rahut [13], Sharma [17], and Gutkin [12], the result is a deficient automatic speech recognition (ASR) and text-to-speech (TTS) system for the Bangla language. Bangla sentence grammar and phonetic construction differ from English sentence grammar and phonetic construction. A typical Bangla sentence has a subject followed by an object and then a verb. Bangla sentences do not use the auxiliary verb, and the subject comes before the object, as it has been pointed out by Ohi [6]. Additionally, the preposition must be placed before a noun or else noun-equivalent words must be used when constructing regional Bangla language sentences. It takes a large vocabulary and a lot of phoneme patterns to recognize speech in Bangla.

In a detailed work published on the challenges and opportunities for Bangla language speech recognition [18], it was noted that, in order for any system to recognize the features in Bangla speech, it must understand the structure of Bangla language grammatically and phonetically to build a flawless ASR and TTS system. Previous researchers have also encountered language-dependent and language-independent challenges when trying to recognize Bangla speech.

For creating a flawless ASR system that produces clean artificial Bangla speech, Mridha et al. [18] emphasized the importance of a large grammatical and phonetic database. One has to use a database with a large vocabulary and phoneme patterns in the Bangla language in order to build an ASR [7,19] system that works flawlessly for the Bangla language. There are no public or private databases available for Bangla speech that contain extensive jargon. This is one of the many reasons why research carried out over the past decade on recognition features in Bangla speech has failed to investigate regional language during Bangla speech feature classification because of database limitations.

There is a comprehensive study of Bangla speech recognition, presented in [14] classification of Bengali accent using deep learning (DL). Hence, the authors failed to build or use the broad corpus necessary to distinguish regional dialects, receiving an accuracy score of 86%. The authors in [20] used neural networks to detect hate speech in social media and achieved an accuracy rate of 77 percent. Rahut, Riffat, and Ridma in [13] reported an accuracy score of 98.16% for the classification of abusive words in Bangla speech by using VGG-16. The Deep CNN model developed by Sharmin et al. in [5] was able to classify Bangla spoken digits with 98.3% accuracy.

For unsupervised learning of sparse and temporal hierarchical features to audio signals, autoencoder neural networks were used [21–28]. In [21] , the authors used a convolutional autoencoder along with Mel Frequency Energy Coefficients (MFEC) data to identify anomalies in different types of machine sounds. Their method produced better results than the baseline. Researchers previously developed a variety of methods for recognizing the gender and age of speakers in Bangla. They focused on two key parts: identifying the optimal features and building a model that is able to recognize those features across various types of speech communication. They achieved an accuracy of 90% for gender and age classification. Anvarjon et al. [29] built an end-to-end CNN with a multi-attention module, able to extract the salient spatial and temporal features and recognize age and gender from speech signals. Due to the use of MFECs input data, the authors in [30] used multi-layers perceptron (MLP) to learn the gender and speaker identity features and it outperformed the CNNs model [29]. To the best of our knowledge, the classification of dialects in Bangla language and detection of synthesized Bangla speech has never been done.

Therefore, the previous work [13] influenced our proposed model for Bangla audio classification of regional language in Bangla speech based on acoustic features and for detecting artificial Bangla speech from audio signals in this paper. The understanding of how to approach this problem was gained from the work done by authors in [31,32] to classify regional languages from American English and Chinese speech.

The proposed method uses convolutional autoencoders stacked with MLELMs to locate the important information from MFECs and recognize it efficiently from synthesized audio and dialect from Bangla speech. The model also outperforms accuracy scores for dialect, age, and gender achieved by previous researchers. Furthermore, the model addresses the issue of a limited Bangla speech dataset. From the seven dialects of the division above, we create a database containing extensive jargon and phoneme patterns. Over 100,000 Bangla spoken utterances were recorded within the institution. There is no ambiguity in the Bangla statement made by the speakers. The input signals are labeled according to the class to which they belong. Afterward, we discuss the proposed approach for recognizing the speaker's regional language and audio authenticity (machine or actual voice), as well as extensive testing to confirm the recommended system's performance on current models and publicly available datasets, as well as future work.

The rest of this paper is organized as follows. In Section 2, we describe the detailed construction of the dataset build and used in this research paper, the architecture of the proposed model and the various experimentation conducted on the dataset, before achieving the accuracy results shown in the following section. Section 3, contains the results obtained for classifying the regional language, synthesized speech, age and gender. Additionally, the results obtained during the comparison with existing models and dataset. Last but not the least the conclusion.

2. Materials and Methods

2.1. Dataset Collection and Prepossessing

A dataset of recorded Bangla speech from seven regional languages is constructed with the help of the TTS system [3] to create a system that is capable of classifying dialect of the speaker in Bangla speech and distinguishing synthesized speech from original speech. According to recent research [6], a robust TTS system requires at least 20 h of speech data. At BRAC University, 30 h of Bangla speech datasets have been compiled with seven different regional languages used in Bangladesh. The dataset is described in detail in the following sections. To test and train, we combine 13 h of Bangla voice data previously published by Brac University. As the dataset consists of large speech data with corresponding transcriptions, TTS is effectively able to produce the synthesized audio Bangla regional speech.

2.1.1. Text Composition

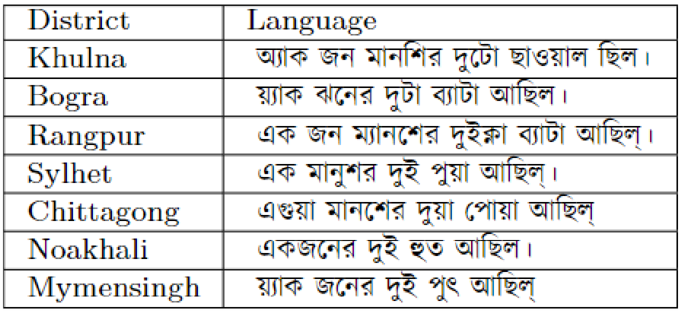
Datasets are constructed with a phonetically balanced text corpus, as suggested by the authors [1]. Text data is collected from a variety of sources. We make sure to include every possible Bangla punctuation for each of the seven regional languages in Bangladesh; Khulna, Bogra, Rangpur, Sylhet, Chittagong, Noakhali, and Mymensingh. Lastly, the dataset for Bangla speech contains almost 2,759,421 utterances. Table 1 shows a sample of the Bangla speech data. To reduce ambiguities in Bangla synthetic speech, nonstandard terms must be translated into their standard pronunciation utilizing the text normalization procedure. Creating the dataset takes into account the process of transforming an unpronounceable text into one that can be pronounced. The following section discusses the process of composing the regional language.

Table 1. Summary of Bangla Speech Data.

Total duration of speech (hours)	21:16:19
Total number of sentences	100,057
Average duration of words each sentence (seconds)	7.81
Total number of words	2,759,421
Number of words of Khulna region	385,714
Number of words of Bogra region	265,482
Number of words of Rangpur region	276,348
Number of words of Sylhet region	348,788
Number of words of Chittagong region	475,428
Number of words of Noakhali region	425,482
Number of words of Mymensingh region	582,179
Total unique words	85,500
Maximum words in a sentence	25
Minimum words in a sentence	5
Average words in a sentence	5.45

2.1.2. Speech Data

Following the preparation of the text corpus, 50 subjects, 25 males and 25 females between the ages of 17 and 54 are asked to record their voices using their mobile phones. Each audio file collected from participants is preprocessed using Audacity software in order to improve the quality of the speech for the TTS synthesis model. Audio data captured at 22 kHz ranged in duration from 0.7 to 40 s. Any audio clips with words more than 15 or less or equal to 3 are removed. The audio files are then stripped into one wav file. After this, the three steps described by Jia et al. [19] are used to synthesize the Bangla language audio files. We use 5760 original Bangla speech wave files for the TTS system, which generated a total of 7482 synthesis voices. Therefore, a total of 13,242 audio clip datasets were used to detect the synthesis signal and dialect of each audio file. In order to control dialects, Bangla words are pronounced in accordance with the dialect of the specific region. Figure 1 shows a sample of the speech data used to create the Bangla dataset with variation for dialect detection. In Bangladesh, some regions have few words that are pronounced similarly, while others have very different pronunciations.



District	Language
Khulna	অ্যাক জন মানশির দুটো ছাওয়াল ছিল।
Bogra	য়্যাক ঝনের দুটা ব্যাটা আছিল।
Rangpur	এক জন ম্যানশের দুইক্না ব্যাটা আছিল্।
Sylhet	এক মানুশর দুই পুয়া আছিল্।
Chittagong	এগুয়া মানশের দুয়া পোয়া আছিল্
Noakhali	একজনের দুই হুত আছিল।
Mymensingh	য়্যাক জনের দুই পুৎ আছিল্

Figure 1. Representation of a sample English sentence 'A man has two sons' in the seven regional Bangla language used in this study.

2.1.3. Prepossessing Dataset

This paper uses unlabeled datasets during the training stage to determine whether the proposed method is capable of detecting audio features correctly. Noise was removed from audio clips, and silent sections of the audio signal were trimmed. A zero-padding was added in order to have a static input. We trimmed each audio file to a length of 10 s. On the time axis of a speech spectrogram, different frequencies can be observed that are received by the human ear. Mel-frequency cepstral coefficients (MFCCs) spectrogram are widely used as input data to process audio features. The mel-cepstrum representation of a sound is widely used to analyze significant audio aspects [17].

Since MFCC applies the discrete cosine transform (DCT) to the logarithm of the filter banks' outputs, the decor-related MFCC features are unsuitable for non-local feature processing. Therefore, MFECs are employed in this study since they do not require the DCT technique and calculate log-energies directly from the filter-bank energies. They produce high-accuracy audio categorization results. For each MFEC data, 130 log mel-band energy characteristics are obtained from the magnitude spectrum using 65 ms analysis frames with a 50% overlap. Finally, the mel-spectrogram is segmented into 35 columns with a hop size of roughly 100 milliseconds every second. Speech data represented in two dimensions is a good candidate for convolutional models. Twenty percent of the dataset is used for testing, ten percent for validation, and seventy percent for training.

2.2. Proposed Model

The proposed method uses the stacked convolutional autoencoder (SCAE) and Multi-label Extreme Learning Machine (MLELM) framework to detect dialect and synthesize Bangla speech from MFEC speech input data. After experimentation with various types of Deep Learning models, the best model was a fully connected SCAE with MLELM for soft classification and score approximation for classes. Convolutional autoencoders are used to handle spatial structure in an audio signal. It benefits in terms of computational complexity, performance, and retraining hidden relationships within the data. The features are then transferred to two MLELM, where the first machine predicts the soft labels and the second machine connects the hard labels to the soft labels. Hard labels are assigned to the unseen data based on the predicted scores for the classes. The following sections and Figure 2 provide a detailed description of the proposed model.

2.2.1. Multi-Label Data Representation

For the given number of classes C, each data instance $F_i = f_{i1}, f_{i2}, \ldots, f_{in}, i = 1, \ldots,$ N is paired with a vector of multiple outputs $O_i = o_{i1}, o_{i2}, \ldots, O_{iC}$ [28]. Instances may belong to more than one class at any given time. In a vector, the values are given in binary format, 1 if the sample falls into the category and 0 if it does not. Consequently, several class labels can be applied simultaneously, which is not possible with signal-label data. The class category combination label is called a label-set. In the following sections, we discuss further the representation of multiple labels.

2.2.2. Stacked Deep Convolutional Auto-Encoder

An autoencoder (AE) is an artificial neural network (ANN) with many hidden layers with fewer nodes than the input and the output is expected to have a similar number of inputs. Convolutional Neural Networks, on the other hand, consist of three parts; convolutional layers, pooling layers, and fully connected layers. This is in contrast to the AE structure, which has input, hidden, and output layers. The input and output layers have n nodes for S samples in the autoencoder, each with feature vector $\mathbf{F}_i$ for $i = 1, \ldots, S$ and $\mathbf{F}_i \epsilon R^n$. As opposed to feed forward networks, autoencoders operate unsupervised. An autoencoder output is, $\mathbf{F} = f_1, f_2, f_3, \ldots, f_n$. In the encoder potion n-dimensional input data is converted to n'-dimension. Where n' is smaller to n, to compress the input data to smaller dimension. Later in the decoded part of autoencoder, the decoded features in the n'-dimensional form is converted back to n-dimension, which is decompressing the decoded features for the output nodes. The encoder maps the input $\mathbf{F}$ to a set of hidden nodes $H = h_1, h_2, h_3, \ldots, h_n$. The node h_j output is computed as

$$h_j = \varphi(\sum_{i=1}^{n} w_{ij} f_i + b_j), \tag{1}$$

where φ represents the transfer function in the encoder section, i starts from 1, w_{ij} is the weight between f_i and h_j, and b_j stands for the bias.

$$f'_k = \varrho(\sum_{j=1}^{n'} w_{jk} h_j + b'_k). \tag{2}$$

In the decoded position function maps the H from encoded representation to estimated $\mathbf{F}'$. Hence, the output of the node f'_k for the k^{th} position is as stated in Equation (2). ϱ act as the transfer function in the decoded side, j begins from 1, w_{ij} is the weight connection value for the h_j and f'_k nodes and b'_k is the bias for the k^{th} node in the decoder. Similar to the multi-layer perceptron the weights are updated through the iterative training of the AE through backpropagation.

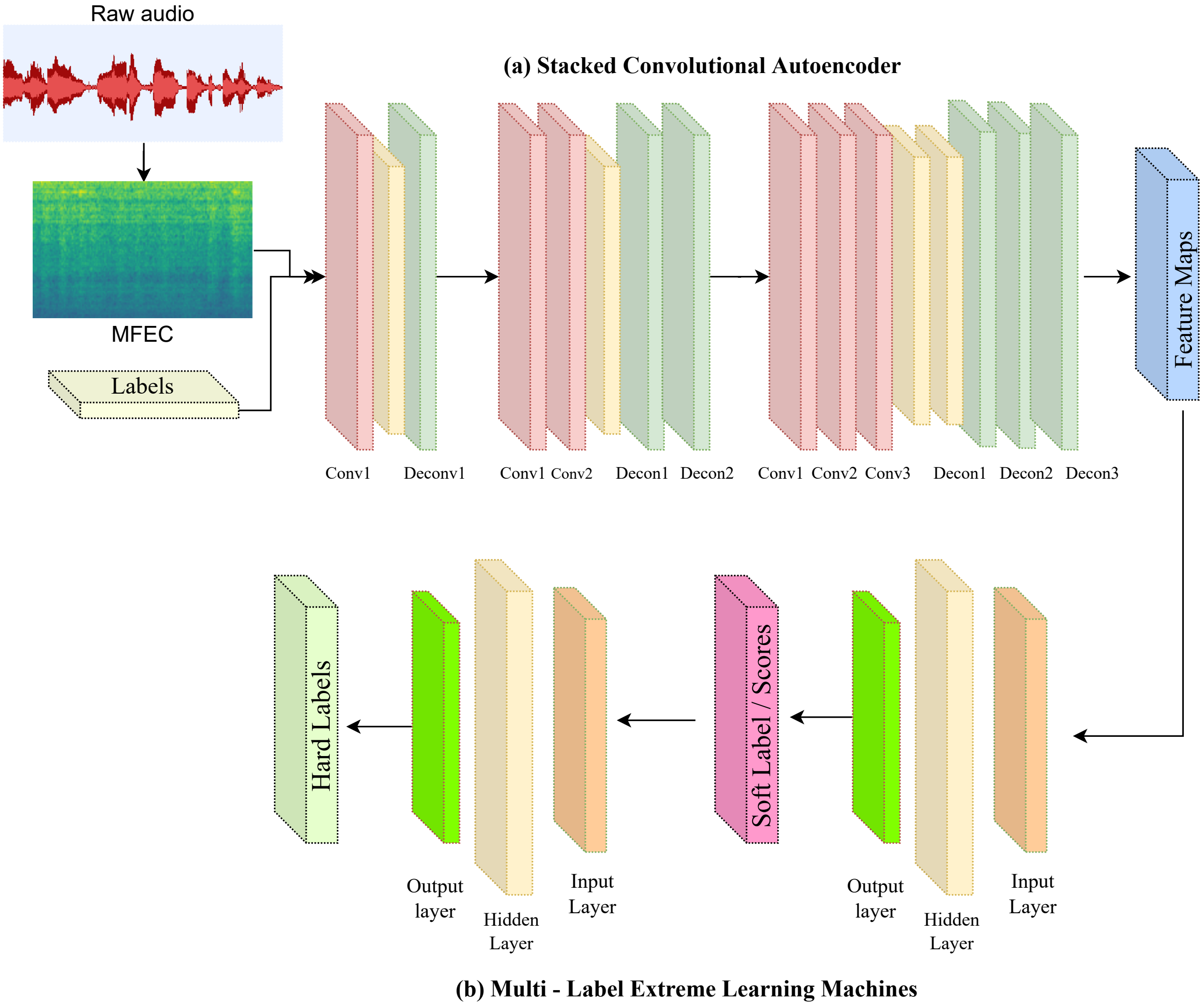


Figure 2. The architecture of Proposed Method consists of two parts. (**a**) SCAE (**b**) MLELMs.

In the encoder section, convolutional layers with ReLU activation functions are followed by max pooling layers. In the encoder part, nonlinear transforms are used to map the input vector to the hidden representation. Next, the reverse transform is reconstructed from the hidden representation to the original audio input signal in the decoded part of the model. SCAE architecture is formed by using the reverse transform as the new representation of the input sample for another convolutional autoencoder, etc. Stacked Autoencoders (SAE) work similarly. For both encoded and decoded parts of the model, all the structures are kept symmetrical to identify low-dimensional hierarchical features in the data.

A convolutional layer is applied in the bottleneck position to obtain a vector by flattening all the units and passing it to an embedded layer that is a low-dimensional, fully connected unit. This resulted in the 2D input data being transformed into lower dimensional features. Later, the multi-label extreme learning machines use the feature vector. The parameters of the decoder and encoder were updated to reduce re-construction error. Table 2. provides the architecture of SCAE.

Table 2. Proposed method detailed architecture.

Layer Names	Architecture	Feature Map Size	Parameters
Conv1	32 × 128 × 1	32 × 64 × 32	29 K
Maxpool	(Max 2 × 2)	32 × 32 × 32	
Bottleneck-Conv	32 × 32 × 32	8 × 8 × 16	768 K
Deconv1	8 × 8 × 16	32 × 64 × 32	
unpool	Max (2 × 2)	32 × 128 × 1	
Conv1	32 × 128 × 1	32 × 32 × 64	228 K
Maxpool	(Max 2 × 2)	16 × 16 × 128	
Conv2	16 × 16 × 128	8 × 8 × 256	12 K
Maxpool	(Max 2 × 2)	4 × 4 × 512	
Bottleneck-Conv	4 × 4 × 512	2 × 2 × 124	221 K
Deconv1	2 × 2 × 124	8 × 8 × 256	
unpool	Max (2 × 2)	16 × 16 × 128	
Deconv2	16 × 16 × 128	32 × 32 × 64	
unpool	Max (2 × 2)	32 × 128 × 1	
Conv1	32 × 128 × 1	32 × 64 × 32	30 K
Maxpool	(Max 2 × 2)	32 × 32 × 64	
Conv2	32 × 32 × 64	16 × 16 × 128	250 K
Maxpool	(Max 2 × 2)	8 × 8 × 128	
Conv3	8 × 8 × 128	4 × 4 × 256	12 K
Maxpool	(Max 2 × 2)	4 × 4 × 512	
Bottleneck-Conv1	4 × 4 × 512	4 × 4 × 256	885 K
Bottleneck-Conv2	4 × 4 × 256	2 × 2 × 124	885 K
Deconv1	2 × 2 × 124	4 × 4 × 256	
unpool	Max (2 × 2)	8 × 8 × 128	
Deconv2	8 × 8 × 128	16 × 16 × 128	
unpool	Max (2 × 2)	32 × 32 × 64	
Deconv3	32 × 32 × 64	32 × 64 × 32	
unpool	Max (2 × 2)	32 × 128 × 1	

2.2.3. Extreme-Learning Machine (ELM)

This is an efficient, compact and sophisticated single layer feed forward neural network that performs classification in a systematic and effective manner [33]. The MLELM architecture is composed of 3 layers: input, hidden, and output. Here the input samples are denoted as $\mathbf{F}$ and $\mathbf{F}\, \epsilon R^n$, class labels are represented as $\mathbf{Y}$ where $\mathbf{Y}\, \epsilon R^E$. The input layer have I number of nodes, hidden layers have D and output layers have E number of nodes. The weights for the input layer to hidden nodes are represented by ω, while the weights from hidden layer to output is denoted by ϖ

It is unlike most ANNs in that the weights associated with the input layer and the biases are randomly initialized and are not updated later. Only the hidden layer in the network learns from the data input, which is reflected in the hidden layer's weights. ϑ is the activation function for the hidden nodes. h_j hidden node, output from the hidden layers are calculated as follows:

$$h_j = \vartheta(\sum_{i=1}^{n} \omega_{ij} f_i + b_j), \tag{3}$$

where, ω_{ij} represents the connection weight between f_i and h_j, and b_j is the bias. As a result the output node o_k

$$o_k = \sum_{D}^{j=1} h_j \varpi_{jk}, \tag{4}$$

ϖ_{jk} represent the weight between the h_j and o_k. Once the MLELM model obtains the weight matrix of ϖ, it is considered to have learned iteratively from the training phase. The model than undergoes through testing phase and later class predication is calculated through the aid of second MLELM in the architecture.

The proposed system uses the topology of the ELM network [28] to perform multi-label classification and score prediction. The encoded features from the SCAE are used as input and the class labels are output from the multi-label extreme learning machines.

In a data set with a signal-label, the sample is assigned the highest values that correspond to the class-label. In multi-label data, however, multiple class labels may be assigned to one sample based on the score achieved. The threshold setting determines the hard multi-labeling. In case the anticipated value exceeds the threshold, the class is considered relevant, and the label is 1; otherwise, it is 0. A low threshold may result in a large number of labels being assigned, while a high threshold might result in misclassification of data instances.

The ELM requires a greater number of hidden nodes to learn efficiently from the data than MLELM. As a result of using SCAE, the number of features in input data to MLELM has decreased. As a result, the weight matrix will be compact, and the concealed layer will be small. Using the weight matrix, the soft classification label scores for that particular class are derived. The next MLELM model predicts the original target labels based on the inputs. In the second MLELM, input weights and biases are randomly initialized. To avoid using a specific threshold for forecasting classes, as in a standard ELM, we use a second MLELM. Based on a calibrated threshold, the final score is transformed into hard class labels.

2.3. Experimentation

In this section, we explain how the training data has been processed and how each of these networks has been repeatedly trained.

2.3.1. Feature Extraction

MFEC features are extracted from audio files at the beginning of this study. To better understand the feature that distinguishes different regional languages, we extract separate data features from the .wav file using standard audio feature extraction metrics. Librosa and the mat-plot python library are used to visualize these values. The metrics used to extract audio features are;

- Amplitude Envelope - displays the maximum amplitude of all the samples in the frame. It helps us determine how loud the signal is. It has although, proven [13] sensitive to outliers but extremely useful for onset detection.
- Zero Crossing Rate - defines the percussive and pitched sound by calculating the number of times a signal crossed the horizontal axis.
- Root Mean Square Energy - calculates the root mean square value of the energy of all sample in single time frame.
- Spectral Centroid - "Brightness" of the sound. This provides us the frequency bins. Where most of the energy in a given sample is stored. In order to extract information from spectrogram, Short-time Fourier transform (STFT) was applied first before spectral centroid could be performed. This feature can help us determine the difference

between the regional language using the variety of frequency bins that can be found in each dialect.

- Spectral Bandwidth - provides a spectral range around the spectral centroid. If we think of the spectral centroid as the mean of the spectral magnitude distribution then spectral bandwidth can be thought of as the 'Variance' of that mean. The spectral bandwidth gives the idea of how the energy of the given sample is spread throughout all the frequency bands.

The distribution of the data features across the seven regional language can be seen in Figures 3 and 4. Where if the energy is spread across the frequency bins, then the value of Spectral Bandwidth will be higher. On the other hand, if the energy is focused on specific frequency bins, then the value of Spectral Bandwidth will be lower.

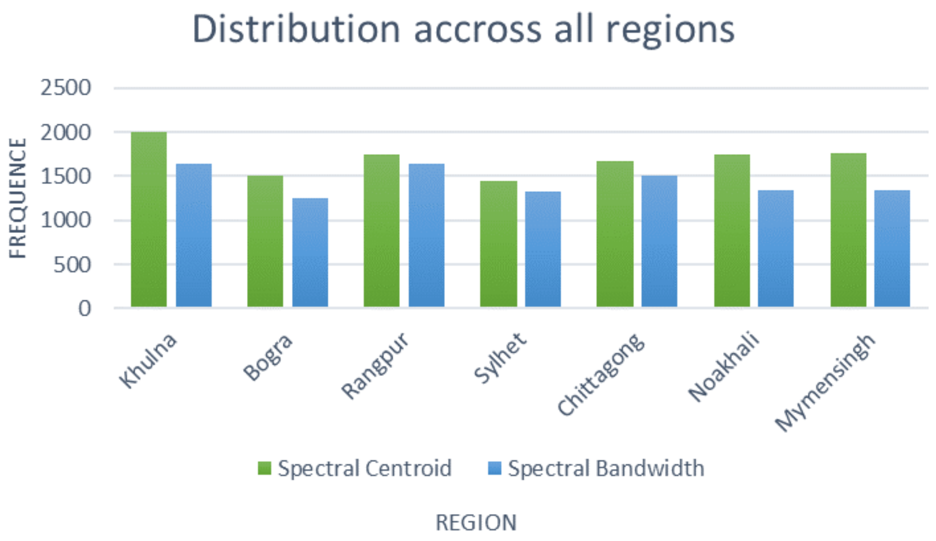


Figure 3. Distribution of audio features; amplitude envelope, zero-crossing rate, and root mean square error across the seven regional languages studied in this research.

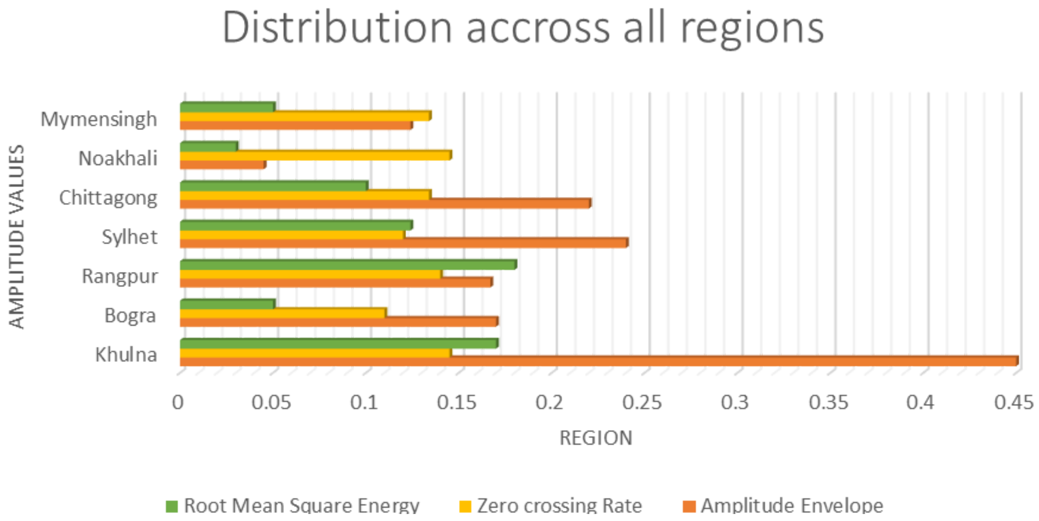


Figure 4. Distribution of audio features; spectral centroid and spectral bandwidth across the seven regional languages studied in this research.

MFEC, however, uses mel-scale, which is a perceptually relevant scale for pitch. The implementation can give up to 39 features, but we only use 13 of them for our work. Because we only have 10,000 samples, there was a chance of over-fitting the neural network model.

In order to get more detailed features from the audio sample, segmentation is necessary. Therefore, each 5-s audio segment is divided into 1-second audio segments containing 22,050 quantized samples. The total number of audio files is 50,000 after segmenting the audio files into 1-second segments. MFCC features are extracted by applying Fast Fourier Transform 2048 times and using a hop length of 512 samples from a total of 22050 samples of each 1-second audio segment. From 1 s of audio, 44 segments of 13 MFCC features are extracted. Hence, there are 572 MFEC features per 1 s of audio. These MFEC data features contain all the information that is visualized separately by the audio feature extraction metrics, as described above. We chose to not use them during the training process of the proposed model but only fed the MFEC data input to the model.

During the training phase, MFEC and audio labels are fed into the SCAE for feature extraction. In order to minimize the Mean Squared Error (MSE) between the input data and the reconstructed output, the encoder and decoder were trained with a learning rate of 0.001 and Adam Optimizer. The training process is scheduled for 200 epochs; however, after no change in validation loss occurs for 10 epochs, the training process is hauled and the best saved model is used. The convolutional autoencoder specifies the number of layers and decreased the number of features. Iteratively, the model is trained until it recognizes the input completely. After the encoded feature is retrieved from the SCAE network, it is delivered to the next step.

2.3.2. Prediction of Soft Class

SCAE encoded characteristics are fed into a multi-label ELM network for soft class prediction. The number of hidden layers in MLELM is determined by the number of input nodes. It operates in batch mode, which means it takes in all of the input instances at once and learns them all at once. Once the MLELM network has learned the weights ϖ from the hidden layer, feature encoded training data is fed back into the MLELM network to create class scores.

$$o'_k = \sum_{D}^{j=1} h_j \varpi_{jk}, \tag{5}$$

the predicted score is calculated through the above equation, ϖ_{jk} weight matrix between hidden node h_j and output node o_k. The result is a layer where all nodes contain the soft classification score for the respective class. As soon as the projected score is obtained, it is transferred to the second MLELM network, which improves the prediction by matching the class scores to the actual class labels. In a single run, the second MLELM network also learns the hidden layer's weights.

2.3.3. Testing Stage

The test data is fed independently into the SCAE and multi-label extreme learning machine networks after they have been trained. Unsupervised sets of encoded features are constructed and then fed to the MLELM networks. The first MLELM model creates individual class scores for each test pattern, which are then input into the second MLELM model. As a result, the test data's hard class labels are determined based on the soft class scores.

3. Results and Discussion

Python is used to implement all networks on three different GPUs; GeForce RTX 3090, Radeon RX 590, and NVidia GeForce RTX 3070. VoxCeleb [34] dataset is used to synthesize English speech for the proposed model. In the VoxCeleb dataset, there are approximately 2000 h of 100,000 phrases taken from YouTube videos narrated by 1,251 celebrities with American, European, and Asian dialects. We compare the proposed model SCAE-MLELMs to two existing models, dense autoencoder (DAE) and convolutional autoencoder (CAE), and used Area Under the Curve (AUC) and partial Area Under the Curve (pAUC) metrics. The AUC provides an aggregate measure of performance across all possible classification thresholds. While pAUC would consider the areas of Receiver Operating Characteristic (ROC) space where data is observed or is a clinically relevant value of test sensitivity or specificity.

The confusion matrix displays the true positive (TP) value, which indicates the number of positively predicted samples that match the true positive labels, and the false negative (FN) value, which indicates the number of positively predicted samples that do not match the positive ground truth labels. TN samples are those that were accurately predicted as negative and had positive values, whereas FP samples are those that were predicted as negative but had positive labels. The accuracy score is determined by the number of positively predicted labels for test data. The rows in the confusion matrices represent genuine labels, while the columns represent anticipated labels. Only the diagonal cells will be the darkest in a perfect confusion matrix, while the rest will be white. The confusion matrix can also be used to determine which classes are most and least mislabeled. It can help identify certain issues in the dataset. We can see this from the confusion matrix below. As the number of recorded and synthesized audio samples for each region is not equal. As a result of the imbalanced distribution of classes in the dataset, there is a slight failure in classification accuracy in each cell of the confusion matrix. As can be seen from the confusion matrix below, the column and row do not add up to 100%.

Moreover, recall (R), precision (P), and F1-score (FS) measurements were used to evaluate the effectiveness of the model, since accuracy alone cannot measure the effectiveness and performance of a model. The following sections provide discussion, confusion matrix and tables with values of the matrix obtained from the models for each specific type of class and its categories, as well as correlation of age with dialect class classification. As previously, researchers used MFCC input format to test their models. In order to compare available datasets and existing algorithms, we use both MFCC and MFEC input data. The model is trained and tested using Bangla and English speech datasets.

3.1. Type of Audio

For the Bangla speech datasets, audio classification is a two-class category problem; original or synthetic. By using the proposed method, actual Bangla voices can be distinguished from generated voices at a high rate. The highest values for precision, recall, and F1-score are obtained for Bangla speech at 91%, 94%, and 93% respectively, with a mean accuracy of 93%. For the English speech dataset, the model produced precision, recall, F1-score, and mean accuracy of 94%, 97%, 94%, 96%, respectively, as shown in Table 3. Figure 5 shows the confusion matrix obtained from the model prediction for both Bangla and English speech. For Bangla speech, the best category-wise accuracy is 94% for original voices and 97% for synthesized voices.

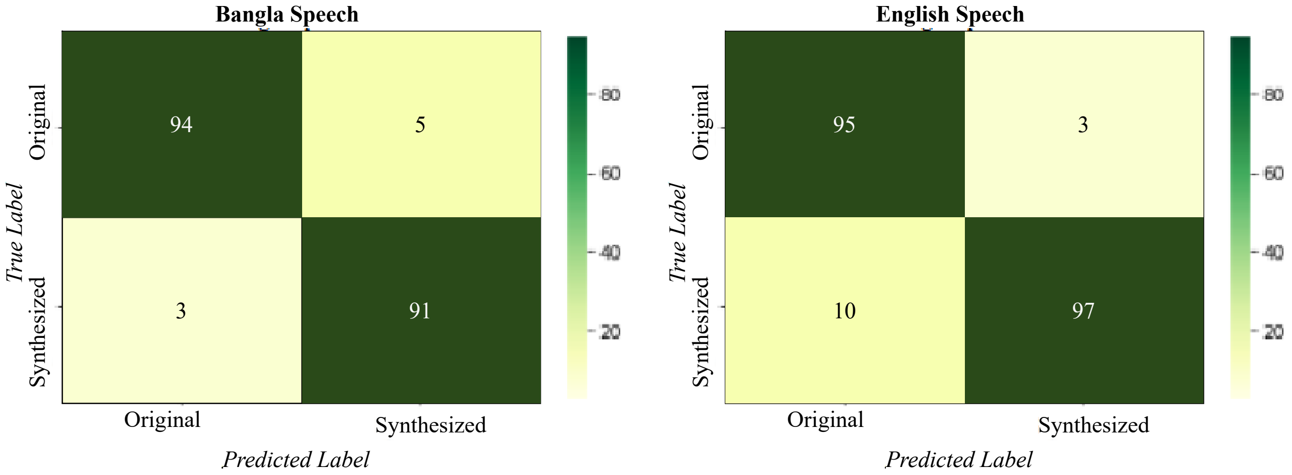


Figure 5. Confusion Matrices of Type of Audio for Bangla and English Speech.

Table 3. Classification Results of Type of Audio for Bangla and English Speech Precision (P), Recall (R), and F1-score (FS) using the SCAE-MLELMs model.

Bangla Speech				English Speech			
Class Group	**P**	**R**	**FS**	**Class Group**	**P**	**R**	**FS**
Original	0.90	0.94	0.93	Original	0.94	0.95	0.94
Synthesized	0.89	0.91	0.90	Synthesized	0.93	0.97	0.93
Accuracy		**0.93**				**0.96**	

3.2. Dialect

Bangla dialects are classified into seven categories: Khulna, Bogra, Rangpur, Sylhet, Chittagong, Noakhali, and Mymensingh. According to Table 4, the highest values obtained for Bangla speech were 78% precision, 78% recall, and 72% F1-score. For Bangla speech, the mean recognition accuracy is 75%. English is a three-category classification problem; Asian (Bangladesh, India, Pakistan, China, Korean), American and European (United Kingdom, Germany, Russia) and the precision, recall, F1-score and mean accuracy were 81%, 88%, 85%, and 81%, respectively.

Table 4. Classification Results of Dialect for both Bangla and English Speech Precision (P), Recall (R), and f1-score (FS) using the SCAE-MLELMs model.

Bangla Speech				English Speech			
Class Group	**P**	**R**	**FS**	**Class Group**	**P**	**R**	**FS**
Khulna	0.67	0.55	0.64	Asian	0.61	0.57	0.70
Bogra	0.78	0.66	0.72	American	0.76	0.88	0.76
Rangpur	0.66	0.52	0.65	European	0.81	0.44	0.85
Sylhet	0.50	0.54	0.58				
Chittagong	0.66	0.70	0.57				
Nokhali	0.72	0.78	0.70				
Mymensingh	0.45	0.55	0.64				
Accuracy		**0.75**				**0.81**	

The larger the variation in dialect type, the better the recognition rate. Because all of the regional languages are spoken in Bangla in the Bangla speech dataset, it can be difficult to distinguish the input audio at a high rate. The situation is different with English, where the dialects are quite diverse, making the process of recognition relatively straightforward. Figure 6 provides the confusion matrix obtained from the model prediction for dialect classification. M, N, C, S, R, B, and K stand for Mymensingh, Noakhali, Sylhet, Rangpur, Bogra, and Khulna, respectively. In English, European, American, and Asian are denoted with the following keywords: EU, AM, AS, respectively. In terms of accuracy, Noakhali (78%) is the best followed by Bogra (66%). As a result, the proposed model confused Bogra with Rangpur and Sylhet with Chittagong, 23% and 32%, respectively, incorrectly predicted. There is a significant amount of confusion due to the similarity of the acoustic features, frequency, and intensity between the words used in those regional parts [9]. There are also similarities between American and European dialects when it comes to English speech dialect prediction since they have a few similar words.

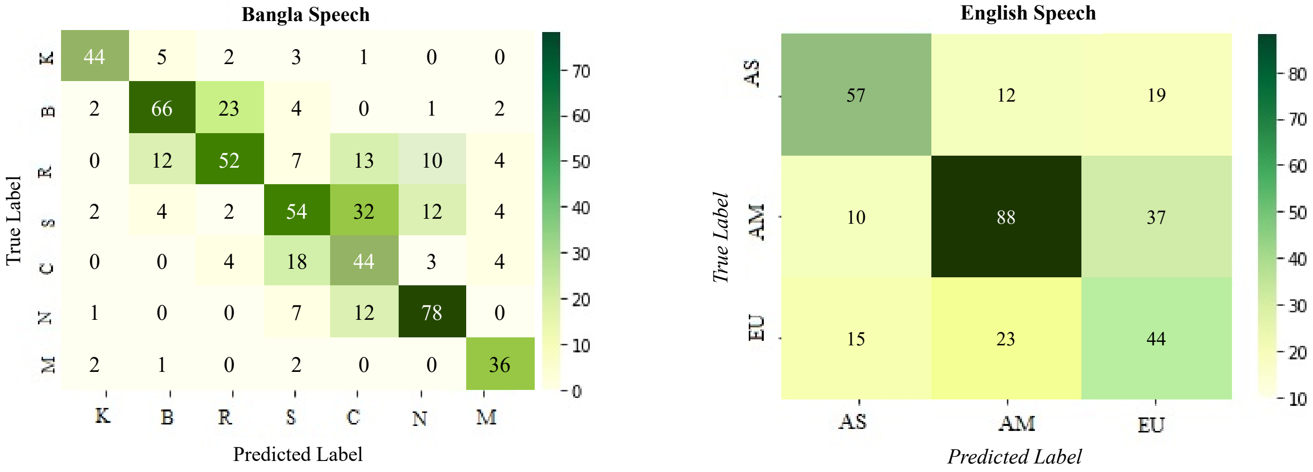


Figure 6. Confusion matrices of dialects for Bangla and English speech.

3.3. Dialect and Age Correlations Classification

Findings indicate that the recognition rate for a speaker's dialect increases with age. Children have a smoother acoustic characteristic compared to adults, who have a high pitch-shift. Hanjun, Nichole, and Charles [18] state that aging has physiological changes that affect how auditory feedback is processed in the brain. Therefore, considering the age feature maps when predicting the dialect of a speaker yields a higher accuracy rate than predicting classes alone. As a result, low false predictions were observed for dialects between Sylhet-Chittagong and Bogra-Rangpur regional languages, as well as American and European speakers. Also, the confusion between close age groups 30s and 40s is reduced when two class labels are considered.

The age and dialect classification for Bangla speech is a fourteen class classification problem; Khulna-Bogra-Rangpur-Sylhet-Chittagong-Noakhali-Mymensingh. There are 12 categories in the English Speech dataset; 20s-30s-40s-50s; Asian-American-European. For Bangla speech, precision, recall, and F1-score are highest at 86%, 78%, 75%, while for English speech they are 83%, 87%, 86% respectively, as shown in Table 5. For Bangla and English speech, the average accuracy is 81 and 87 percent, respectively. Figure 7 shows the confusion matrix obtained from the model prediction for dialect classification. In Bangla speech; CK, CB, CR, CS, CC, CN, CM, AK, AB, AR, AS, AC, AN, AM stands for Child-Khulna, Child-Bogra, Child-Rangpur, Child-Sylhet, Child-Chittagong, Child-Noakhali, Child-Mymensingh, Adult-Khulna, Adult-Bogra, Adult-Rangpur, Adult-Sylhet, Adult-Chittagong, Adult-Noakhali, Adult-Mymensingh, respectively. 2AS, 2AM, 2EU, 3AS, 3AM, 3EU, 4AS, 4AM, 4EU, 5AS, 5AM, 5EU stand for 20s-Asian, 20s-American, 20s-European, 30s-Asian, 30s-American, 30s-European, 40s-Asian, 40s-American, 40s-European, 50s-Asian, 50s-American, 50s-European, respectively.

3.4. Age

The age classification problem for English Speech datasets consists of four classes; 20s, 30s, 40s, and 50s. The age difference between the classes is 10 years. For each age group, Bangla Speech has a different age range. Children's ages range from 12 to 20, while adults' ages range from 30 to 50. The highest values obtained for precision, recall and F1-score for English speech are 88%, 85%, 86% respectively, while for Bangla speech 89%,95%, 92% respectively, as observed in Table 6. The greater the difference in age range, the greater the recognition rate. The mean accuracy for recognition is 82% and 91% for English and Bangla speech respectively.

The confusion matrix obtained from the model prediction for age classification problem is present in Figure 8 for both English and Bangla speech. The best category-wise accuracy for English speech is achieved by twenties (85%) followed by fifties (81%). However, the accuracy rates for thirties and forties 75% and 67%, respectively. The proposed model confuses the prediction for thirties age group with forties, 45% were falsely predicted. One of the main reasons for this confusion is that the acoustic pitch features are similar between the ages [6,18]. According to the MFEC data log-energies of the audio signals, the frequency features for the two age groups are very similar. Due to its increased difference in age range among categories, this similarity cannot be observed in Bangla speech age prediction.

Table 5. Classification Results of Dialect and Age Correlation for Bangla and English Speech Precision (P), Recall (R), and f1-score (FS) using the SCAE-MLELMs model.

Bangla Speech				**English Speech**			
Class Group	**P**	**R**	**FS**	**Class Group**	**P**	**R**	**FS**
Child-Khulna	0.67	0.42	0.57	20s-Asian	0.54	0.53	0.57
Child-Bogra	0.68	0.63	0.72	20s-American	0.71	0.76	0.68
Child-Rangpur	0.56	0.61	0.58	20s-European	0.53	0.56	0.65
Child-Sylhet	0.60	0.56	0.64	30s-Asian	0.63	0.76	0.70
Child-Chittagong	0.66	0.50	0.65	30s-American	0.80	0.83	0.74
Child-Nokhali	0.79	0.71	0.67	30s-European	0.70	0.75	0.86
Child-Mymensingh	0.58	0.45	0.55	40s-Asian	0.61	0.66	0.69
Adult-Khulna	0.63	0.52	0.64	40s-American	0.56	0.72	0.66
Adult-Bogra	0.86	0.78	0.73	40s-European	0.73	0.64	0.85
Adult-Rangpur	0.69	0.64	0.65	50s-Asian	0.79	0.82	0.80
Adult-Sylhet	0.59	0.61	0.58	50s-American	0.81	0.87	0.79
Adult-Chittagong	0.73	0.66	0.75	50s-European	0.83	0.85	0.79
Adult-Nokhali	0.73	0.76	0.67				
Adult-Mymensingh	0.65	0.56	0.68				
Accuracy		**0.81**				**0.87**	

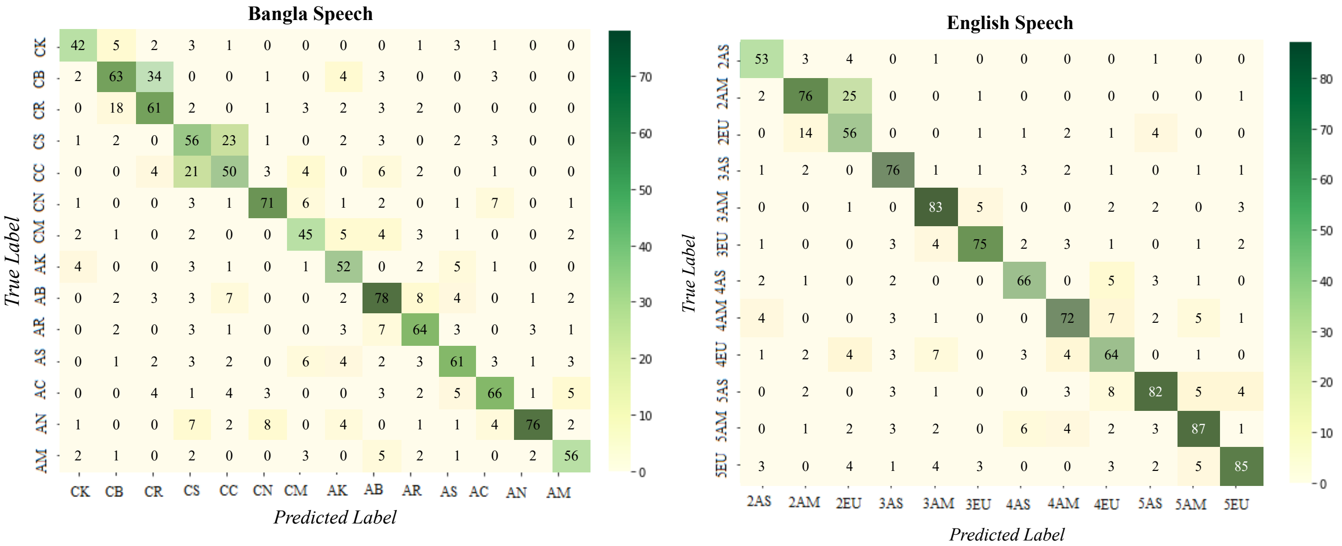


Figure 7. Confusion matrices of dialects for Bangla and English speech.

Table 6. Classification Results of Age for Bangla and English Speech Precision (P), Recall (R), and f1-score (FS) using the SCAE-MLELMs model.

Bangla Speech				**English Speech**			
Class Group	**P**	**R**	**FS**	**Class Group**	**P**	**R**	**FS**
Child	0.89	0.95	0.81	20s	0.76	0.85	0.72
Adult	0.66	0.73	0.92	30s	0.88	0.75	0.87
				40s	0.87	0.67	0.70
				50s	0.68	0.81	0.76
Accuracy		**0.91**				**0.82**	

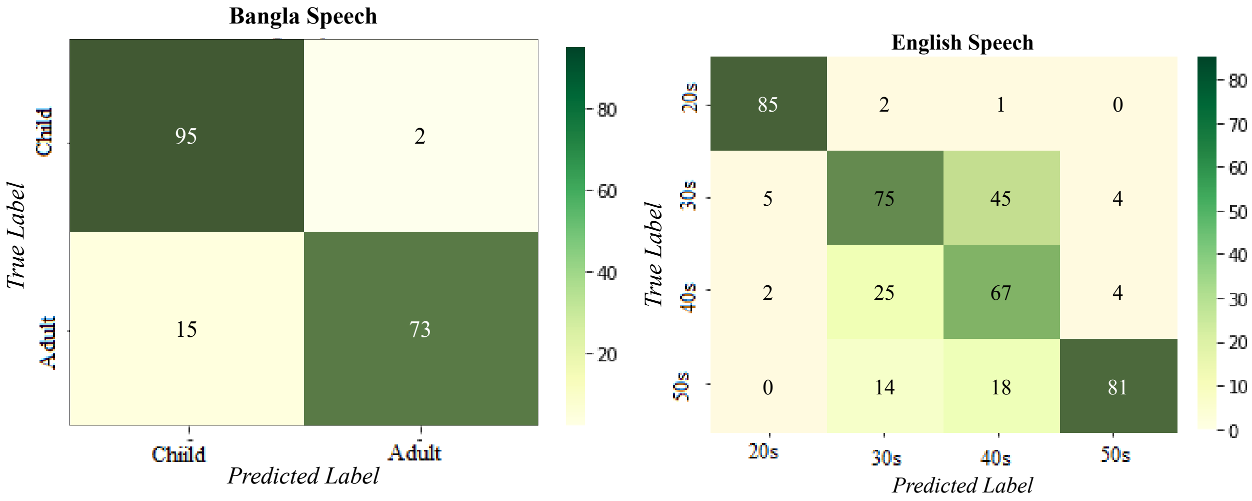


Figure 8. Confusion matrices of Age for Bangla and English Speech.

3.5. Gender

There are two gender categories for the Bangla and English speech datasets; male and female. The highest values obtained for precision, recall and F1-score for for Bangla speech 85%,94%, 93%, while for English speech are 96%, 98%, 96% respectively, as observed in Table 7. The mean accuracy for recognition is 92% and 96% for Bangla and English speech respectively. Confusion matrix obtained from the model prediction for age classification problem is present in Figure 9 for both speeches. The best category-wise accuracy for both speeches is achieved by male category, 87% Bangla and 98% English speech. In comparison to the English speech dataset, the proposed model has more false predictions for Bangla speech.

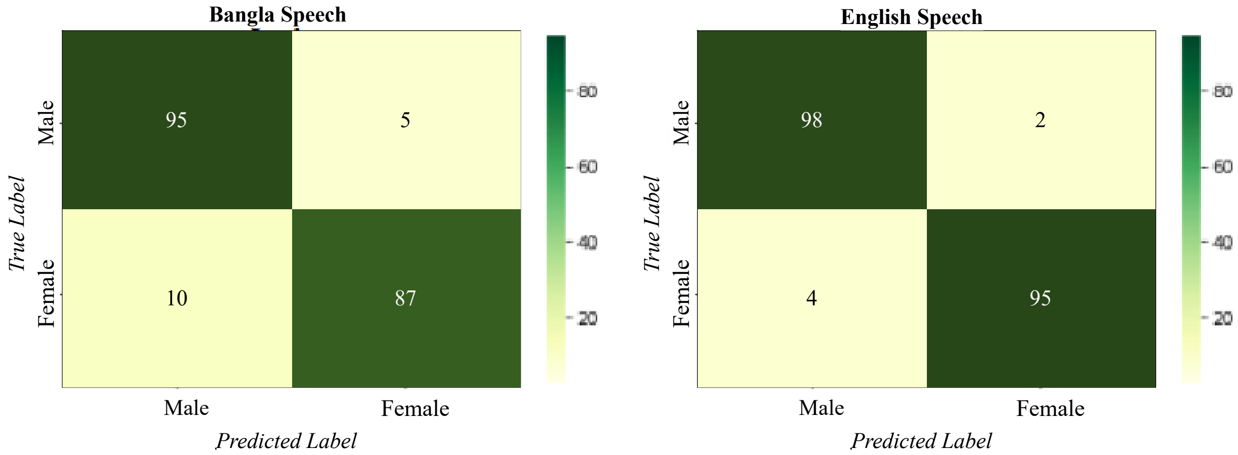


Figure 9. Confusion matrices of Gender for Bangla and English Speech.

Table 7. Classification Results of Gender for Bangla and English Speech precision (P), recall (R), f1-score (FS) using the SCAE-MLELMs model.

Bangla Speech				English Speech			
Class Group	**P**	**R**	**FS**	**Class Group**	**P**	**R**	**FS**
Male	0.82	0.94	0.93	Male	0.96	0.98	0.96
Female	0.85	0.87	0.90	Female	0.94	0.94	0.93
Accuracy		**0.92**				**0.96**	

3.6. Comparison Among Datasets

The number of Convolutional Autoencoders in the SCAE-MLELMs model is varied to evaluate the accuracy of the prediction of the class label for the test data as well as the impact the input data format has on the suggested system. MFCCs and MFECs are selected as input data types. They are the most commonly used data formats in audio recognition studies. The dataset mentioned earlier is used to analyze Bangla's speech. While for English Speech, freely available datasets from Google AudioSet and VoxCeleb is utilized [34]. Table 8 shows the categorization accuracy in the specifics of the experiment with MFCCs as data input format and Table 9 for MFECs data format.

Table 8. Classification Accuracy (%) of the four different SCAE-MLELMs' architecture on different datasets with input format as MFCCs ; Brac University previous and self-built Bangla Speech dataset and Google Audio-Set and VoxCeleb for English speech dataset were used during the experiment. Numbers in bold represent the highest classification accuracies.

Model No.	Bangla Speech			English Speech					
	Audio Type	Brac University Dialect	Gender/Age	Audio Type	Google AudioSet Dialect	Gender/Age	Audio Type	VoxCeleb Dialect	Gender/Age
1	76	75	84	67	79	76	87	78	84
2	74	86	90	78	81	73	90	82	86
3	87	78	89	84	84	90	89	90	92
4	**94**	**93**	**92**	**95**	**94**	**93**	**95**	**94**	**93**

Table 9. Classification Accuracy (%) of the four different SCAE-MLELMs architecture on different datasets with input format as MFECs; Brac University previous and self-built Bangla Speech dataset and Google Audio-Set and VoxCeleb for English speech dataset is used during the experiment. Numbers in bold represent the highest classification accuracies.

Model No.	Bangla Speech			English Speech					
	Audio Type	Brac University Dialect	Gender/Age	Audio Type	Google AudioSet Dialect	Gender/Age	Audio Type	VoxCeleb Dialect	Gender/Age
1	84	87	89	87	88	86	91	92	92
2	**95**	**94**	**94**	**97**	**96**	**94**	**95**	**95**	**93**
3	78	84	90	84	83	91	90	90	91
4	76	79	76	78	76	79	81	82	86

Model 1 employs only one CAE network with MLELMs to evaluate the efficiency of the suggested methods; Model 2 employs three CAE networks with MLELMs. In models 2, 3, and 4, three, four, and six CAE networks are followed by MLELM networks; a comprehensive architectural description can be found in this paper's proposed model section. For all datasets in the MFCCs data format, Model 4 gives the highest classification accuracy for all class labels for both types of speeches. It requires more convolutional autoencoders to detect the prominent aspects of an audio stream in an MFCCs spectrogram. The MFECs data has the highest classification accuracy when using Model 2, since its log-mel-energy properties are easier to discern. As the number of CAE networks for MFECs input data format increases, the model tends to overfit as well. The accuracy of Model 3 and Model 4 is superior to Model 2. Across all datasets of Bangla and English speech, dialect and type of audio have the greatest accuracy in prediction.

3.7. Comparison between Existing Algorithms

For each classification category, it is compared with the models developed by Sharmin et al. [35], a Deep CNN model, and Tursunov [29], a multi-attention module CNN model. We compare the performance and robustness of the techniques using AUC and pAUC measurements. Table 10 shows the AUC and pAUC values for each class category with spectrogram input data format, while Table 11 shows the MFEC data format for both Bangla and English audios. The average AUC and pAUC matrices for each class category for both data formats demonstrate that SCAE-MLELMs model outperforms the current model for both speeches.

Table 10. Performance Results of existing methods; Sharmin et al. [35]: Deep CNN, and Tursunov [29]; Multi-attention module CNN model for MFCCs data type.

Class	**Speech**	**Sharmin et al. [35]**		**Tursunov.A [29]**		**SCAE-MLELMs**	
		AUC (%)	**pAUC (%)**	**AUC (%)**	**pAUC (%)**	**AUC (%)**	**pAUC (%)**
Audio Type	Bangla	67.57	55.24	78.16	64.24	91.24	87.12
	English	87.45	73.15	89.78	82.57	92.75	86.74
Dialect	Bangla	52.47	42.18	61.41	59.78	89.75	83.12
	English	60.42	57.48	68.48	55.46	89.76	86.14
Gender	Bangla	69.40	55.94	84.15	77.42	92.00	87.45
	English	89.73	72.49	91.42	96.87	91.42	89.48
Age	Bangla	78.69	63.54	86.44	78.62	89.76	82.62
	English	83.45	73.44	81.46	77.48	89.41	86.42

In addition, MFECs data formats had greater AUC and pAUC values than MFCCs data formats for all model types. Compared to other classes, the Deep CNN model [36] has the lowest AUC and pAUC performance values for dialect class labels. For dialect class labels, the Deep CNN model [36] has the lowest AUC and pAUC values compared to other classes. In comparison to the approach in [36], the Multi-attention module CNN model [29] produced some of the best results for a few classification labels; age, gender, and audio type. Existing approaches have difficulty distinguishing dialect in speech regardless of language due to their single-label model structure and inability to learn characteristics that include age and dialect in audio frequency patterns. Employing multi-label extreme learning machine networks, as suggested in the model. Moreover, the existing methods do not perform as well with Bangla speech audio input as they do with English speech audio input. The proposed method performs consistently in both languages.

Table 11. Performance Results of existing methods; Sharmin et al. [35]: Deep CNN, and Tursunov [29]; Multi-attention module CNN model for MFECs data type.

Class	**Speech**	**Sharmin et al. [35]**		**Tursunov.A [29]**		**SCAE-MLELMs**	
		AUC (%)	**pAUC (%)**	**AUC (%)**	**pAUC (%)**	**AUC (%)**	**pAUC (%)**
Audio Type	Bangla	76.15	65.73	87.41	78.24	92.11	89.12
	English	79.25	64.65	89.88	72.11	93.70	84.97
Dialect	Bangla	66.48	54.18	76.48	62.18	92.48	93.17
	English	68.42	56.38	81.42	75.40	90.45	82.75
Gender	Bangla	72.11	65.19	86.12	73.14	91.75	83.46
	English	89.73	72.14	88.25	79.48	91.10	80.74
Age	Bangla	83.45	72.38	83.47	78.34	90.29	89.26
	English	85.16	76.34	80.42	79.28	88.49	86.77

4. Conclusions

In this paper, a dataset was presented with seven regional Bangla languages. In addition, a Stacked Convolution Autoencoder followed by MFECs for the classification of synthesized voices and regional Bangla languages was proposed. The propopsed method can extract essential features from unsupervised data and classify them accordingly. From the given input data, the SCAE identifies relevant features for class labeling and produces detailed feature maps. The MLELM networks in the suggested method learn from the training data to produce multi-label classification in a single pass. We used two MLELM networks because the first performs soft classification scores and soft labels. The second MLELM network matches the soft label to hard labels. We conducted extensive training and testing to evaluate the performance, efficiency, and robustness of the system. The suggested method outperforms the existing algorithms (Sharmin et al. [35], a Deep CNN model, and Tursunov [29], a multi-attention module CNN model) with an accuracy score of 91%, 89%, 89%, 92% for synthesized/original audio type, dialect, age, and gender classification, respectively for Bangla Speech, for the spectrogram input data type. While for MFECs input format the accuracy scores are synthesized/original audio type, 92%, dialect, 92%, age 90%, gender 91%. Consequently, MFEC's data input format is more reliable when it comes to recognizing relevant salient features from audio inputs. In addition, the proposed model can improve the classification accuracy score for dialect class to 95% by using detailed feature maps produced from the SCAE, which produces the correlated acoustic feature patterns between dialect class and age. As aging has a physiological change that impacts the processing of auditory feedback in the brain. Hence with the help of MLELM networks, the multi-label data was used to create correlated feature maps of the data. The model also achieves the highest accuracy score against the existing models for the English speech dataset. 93%, 94% 88% and 91% for synthesized/original audio type, dialect, age, gender classification, respectively, for MFECs. The proposed method can be applied to the concept of ASR, TTS and speech recognition and processing in the future, including customer care, health care devices, etc.

Author Contributions: Conceptualization, P.S.H.; Data curation, P.S.H.; Formal analysis, P.S.H.; Funding acquisition, K.K. and M.J.P.; Methodology, P.S.H.; Supervision, A.C.; Writing—original draft, P.S.H.; Writing—review & editing, A.C. and M.J.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by G_ITRC, the Ministry of Science and ICT (MSIT), South Korea, under the Grand Information Technology Research Center Support Program under Grant IITP-2021-2015-0-00742, supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the creating the dataset used in this study.

Data Availability Statement: Raw audio files can be obtained from the following link, https://drive.google.com/drive/folders/1qMYDpfSCUW-fTcUxotnkGHKcl-259AwY?usp=sharing, accessed on 24 March 2022.

Acknowledgments: The authors of this paper would like to thank Golam Mohammad Zilani, Brac University, for helping in creating the Bangla regional language speech dataset.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SCAE	Stacked Convolutional Auto-Encoder
MLELM	Multi-label Extreme-Learning Machine
MFEC	Mel Frequency Energy Coefficients
HCI	Human–Computer Interaction
ASR	Automatic Speech Recognition
TTS	Text-To-Speech
DL	Deep Learning
CNN	Convolutional Neural Network
MLP	Multi-Layer Perceptron
VGG-16	Visual Geometry Group-16

MFCC	Mel Frequency Cepstral Coefficent
DCT	Discrete Cosine Transform
AE	Auto-Encoder
SAE	Stacked Auto-Encoder
ReLU	Rectified Linear Activation Unit
ELM	Extreme-Learning Machine
ANN	Artificial Neural Network
MSE	Mean Squared Error
DAE	Dense Auto-Encoder
CAE	Convolutional Auto-Encoder
AUC	Area under the ROC Curve

References

1. Alam, F.; Habib, S.M.; Sultana, D.A.; Khan, M. Development of annotated Bangla speech corpora. *Proj. Bangla Lang. Process.* **2010**, *9*, 125–132.
2. Honnet, P.-E.; Lazaridis, A.; Garner, P.N.; Yamagishi, J. The siwisfrench speech synthesis database-design and recording of a high quality french database for speech synthesis. *J. Idiap Tech. Rep.* **2017**. Available online: https://www.researchgate.net/publication/315893580_The_SIWIS_French_Speech_Synthesis_Database_-_Design_and_recording_of_a_high_quality_French_database_for_speech_synthesis (accessed on 7 July 2020).
3. Pial, T.I.; Aunti, S.S.; Ahmed, S.; Heickal, H. End-to-End Speech Synthesis for Bangla with Text Normalization. In Proceedings of the 5th International Conference on Computational Science/ Intelligence and Applied Informatics (CSII), Yonago, Japan, 10–12 July 2018; pp. 66–71. [CrossRef]
4. Rahman, S.; Kabir, F.; Huda, M.N. Automatic gender identification system for Bengali speech. In Proceedings of the 2nd International Conference on Electrical Information and Communication Technologies (EICT), Khulna, Bangladesh, 10–12 December 2015; pp. 549–553. [CrossRef]
5. Hassan, F.; Khan, M.S.A.; Kotwal, M.R.A.; Huda, M.N. Gender independent Bangla automatic speech recognition. In Proceedings of the 2012 International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, 18–19 May 2012; pp. 144–148. [CrossRef]
6. Mridha, M.F.; Ohi, A.Q.; Hamid, M.A.; Monowar, M.M. A study on the challenges and opportunities of speech recognition for Bengali language. *Artif. Intell. Rev.* **2021**, *55*, 3431–3455. doi: 10.1007/s10462-021-10083-3 [CrossRef]
7. Gutkin, A.; Ha, L.; Jansche, M.; Pipatsrisawat, K.; Sproat, R. TTS for Low Resource Languages: A Bangla Synthesizer. In Proceedings of the 2016-10th International Conference on Language Resources and Evaluation, Portoroz, Slovenia, 23–28 May 2016; pp. 2005–2010.
8. Sadeque, F.Y.; Yasar, S.; Islam, M.M. Bangla text to speech conversion: A syllabic unit selection approach. In Proceedings of the 2013 International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, 17–18 May 2013; pp. 1–6. [CrossRef]
9. Alam, F.; Nath, P.K.; Khan, M. Text to speech for Bangla language using festival. Project: Bangla Language Processing. Thesis, Brac University Library, Dhaka, Brac Univesity, 2007. Available online: http://hdl.handle.net/10361/675 (accessed on 28 July 2020).
10. Muhammad, G.; Alotaibi, Y.A.; Huda, M.N. Automatic speech recognition for Bangla digits. In Proceedings of the 2009 12th International Conference on Computers and Information Technology, Dhaka, Bangladesh, 21–23 December 2009; pp. 379–383. [CrossRef]
11. Asfak-Ur-Rahman, M.; Kotwal, M.R.A.; Hassan, F.; Ahmmed, S.; Huda, M.N. Gender effect cannonicalization for Bangla ASR. In Proceedings of the 15th International Conference on Computer and Information Technology (ICCIT), Chittagong, Bangladesh, 22–24 December 2012; pp. 179–184. [CrossRef]
12. Gutkin, A.; Ha, L.; Jansche, M.; Kjartansson, O.; Pipatsrisawat, K.; Sproat, R. Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla. *J. Procedia Comput. Sci.* **2016**, *81*, 194–200. [CrossRef]
13. Rahut, S.K.; Sharmin, R.; Tabassum, R. Bengali Abusive Speech Classification: A Transfer Learning Approach Using VGG-16. In Proceedings of the 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), Dhaka, Bangladesh, 21–22 December 2020. [CrossRef]
14. Badhon, S.M.; Rahaman, H.; Rupon, F.R.; Abujar, S. Bengali Accent Classification from Speech Using Different Machine Learning and Deep Learning Techniques. In *Soft Computing Techniques and Applications*; Springer: Singapore, 2021; pp. 503–513. [CrossRef]
15. Alam, T.; Khan, A.; Alam, F. Bangla Text Classification using Transformers. Project: Bangla Language Processing. 2000. Available online: https://www.researchgate.net/publication/345654685_Bangla_Text_Classification_using_Transformers (accessed on 25 March 2022).
16. Das, A.K.; Al Asif, A.; Paul, A.; Hossain, M.N. Bangla hate speech detection on social media using attention-based recurrent neural network. *J. Intell. Syst.* **2021**, *30*, 578–591. [CrossRef]
17. Sharma, G.; Umapathy, K.; Krishnan, S. Trends in audio signal feature extraction methods. *J. Appl. Acoust.* **2020**, *158*, 107020. [CrossRef]

18. Liu, H.; Russo, N.M.; Larson, C.R. Age-related differences in vocal responses to pitch feedback perturbations: A preliminary study. *J. Acoust. Soc. Am.* **2010**, *127*, 1042–1046. [CrossRef] [PubMed]
19. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez, Moreno, I.; et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *arXiv* **2019**, arXiv:1806.04558v4.
20. Jam, M.M.; Sadjedi, H. Identification of hearing disorderly multi-band entropy cepstrum extraction from infant's cry. In Proceedings of the International Conference on Biomedical and Pharmaceutical Engineering, Singapore, 2–4 December 2009; pp. 1–5.
21. Ribeiro, A.; Matos, L.M.; Pereira, P.J.; Nunes, E.C.; Ferreira, A.L.; Cortez, P.; Pilastri, A. Deep Dense and Convolutional Auto-Encoders for Unsupervised Anomaly Detection in Machine Condition Sounds. *arXiv* **2020**, arXiv:2006.10417.
22. Turchenko, V.; Luczak, A. Creation of a deep convolutional auto-encoder in Caffe. In Proceedings of the 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, Romania, 21–23 September 2017; pp. 651–659.
23. Nervana Systems/Neon, Convolutional Auto-Encoder Example Network for MNIST Data Set. 2015. Available online: https://github.com/NervanaSystems//examples/auto-encoder.py (accessed on 10 June 2010).
24. Seyfioğlu, M.S.; Özbayoğlu, A.M.; Gürbüz, S.Z. Deep convolutional auto-encoder for radar-based classification of similar aided and unaided human activities. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 1709–1723. [CrossRef]
25. Guo, X.; Liu, X.; Zhu, E.; Yin, J. Deep Clustering with Convolutional Auto-Encoders. In *Lecture Notes in Computer Science, (Including Subseries Lecture Notes in Artificial Intelligence and Lecture, Notes in Bioinformatics)*; 10635 LNCS:373–382; Springer: Cham, Switzerland, 2017.
26. Ghasedi, Dizaji, K.; Herandi, A.; Deng, C.; Cai, W.; Huang, H. Deep Clustering via Joint Convolutional Auto-Encoder Embedding and Relative Entropy Minimization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5747–5756.
27. Berniker, M.; Kording, K.P. Deep networks for motor control functions. Frontiers in computational neuroscience. *J. Front. Comput. Neurosci.* **2015**, *9*, 2015.
28. Law, A.; Ghosh, A. Multi-label classification using a cascade of stacked auto-encoder and extreme-learning machines. *J. Neurocomput.* **2019**, *358*, 222–234. [CrossRef]
29. Tursunov, A.; Choeh, J.Y.; Kwon, S. Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms. *J. Mdpi Sens.* **2021**, *21*, 5892. [CrossRef] [PubMed]
30. Mamyrbayev, O.; Toleu, A.; Tolegen, G.; Mekebayev, N. Neural architectures for gender detection and speaker identification, Cogent Engineering. *J. Cogent Eng.* **2020**, *7*, 1727168. [CrossRef]
31. Hou, R.; Huang, C. Classification of regional and genre varieties of Chinese: A correspondence analysis approach based on comparable balanced corpora. *Nat. Lang. Eng.* **2021**, *26*, 613–640. [CrossRef]
32. Clopper, C.G.; Pisoni, D.B. Free classification of regional dialects of American English. *J. Phon.* **2021**, *35*, 421–438. [CrossRef] [PubMed]
33. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *J. Neurocomput.* **2006**, *70*, 489–501. [CrossRef]
34. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Dataset Comput. Sci. Lang.* **2019**, *60*, 101027. [CrossRef]
35. Sharmin, R.; Rahut, S.K.; Huq, M.R. Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network. *J. Procedia Comput. Sci.* **2020**, *171*, 1381–1388. [CrossRef]
36. MIT Deep Learning Genomics-Lecture11-PCA, t-SNE, Auto-Encoder Embedings. Youtube, Manolis Kellis. 2020. Available online: https://www.youtube.com/watch?v=Qh6cAXJJxd4 (accessed on 20 June 2020).