Computing and Informatics, Vol. 44, 2025, 1071–1100, doi: 10.31577/cai_2025_5_1071

# MODIFIED CONVOLUTIONAL NEURAL NETWORK FOR SPEAKER AGE AND GENDER CLASSIFICATION

Laxmi Kantham Durgam*

*Digital Signal Processing Lab*
*Department of Electronics and Communication Engineering*
*National Institute of Technology Warangal*
*506004 Telangana, India*
*&*
*Speech Communications Lab*
*Department of Electronics and Multimedia Communications*
*Faculty of Electrical Engineering and Informatics, Technical University of Košice*
*Letná 9, 042 00 Košice, Slovakia*
*e-mail:* ld712103@student.nitw.ac.in

Ravi Kumar Jatoth

*Digital Signal Processing Lab*
*Department of Electronics and Communication Engineering*
*National Institute of Technology Warangal*
*506004 Telangana, India*
*e-mail:* ravikumar@nitw.ac.in

Daniel Hládek, Stanislav Ondáš, Matúš Pleva, Jozef Juhár

*Speech Communications Lab*
*Department of Electronics and Multimedia Communications*
*Faculty of Electrical Engineering and Informatics, Technical University of Košice*
*Letná 9, 042 00 Košice, Slovakia*
*e-mail:* {daniel.hladek, stanislav.ondas, matus.pleva, jozef.juhar}@tuke.sk

* Corresponding author

**Abstract.** Identifying a person's age and gender from speech signal characteristics poses a significant challenge in personal identity recognition systems, particularly when security considerations are involved. In signal processing applications such as speaker recognition, biometric identification, human-machine interface (HMI), and telecommunication, the estimation of age and gender from voice is a crucial and demanding problem. In several signal processing domains, deep learning models have demonstrated remarkable effectiveness. In this paper, we propose a modified convolutional neural network to identify the age and gender of the speaker using the characteristics of the MFCC speech. We also included techniques to reduce the dimensionality of the speech feature set. We tested modified one-dimensional convolutional neural networks (1D-CNN) and machine learning models such as support vector classification (SVC), decision trees (DT), and random forests (RF). The modified 1D-CNN based on deep learning, along with dimensionality reduction, random seeding, and cross-validation, is proposed for the recognition of age and gender in speech. We applied different dimensionality reduction techniques, such as principal component analysis (PCA) and independent component analysis (ICA), along with random seeding and various sets of cross-validation. In this study, we used the Children Speech Recorning Dataset, Biometric Visions and Computing (BVC), and the Mozilla Common Voice speech datasets for estimating age and gender from speech. The proposed 1D-CNN model exhibits a promising performance compared to the state-of-the-art (SOTA) approaches. The models were evaluated and compared with evaluation metrics, such as accuracy. The dimensionality reduction techniques, selection of speech features, and seeding show a significant impact on the performance of the suggested model.

**Keywords:** Age and gender estimation, speaker recognition, MFCC, modified 1D-CNN, dimensionality reduction, random seed, cross-validation

**Mathematics Subject Classification 2010:** 68-T40

## 1 INTRODUCTION

The ability to identify the age and gender [1, 2] of a speaker from speech signals is essential for speech recognition applications [3] such as biometric identification [4] based on speech, personal voice assistance [5], and human-machine interfaces [6]. The perception of user preferences, content customization [7], and targeted services [8] can benefit from an accurate age and gender classification from speech. Audio signals can be used to estimate age and gender due to their simplicity and quantity, making them a valuable source of information [1, 2].

The primary goal of this research is to use several sets of MFCC speech features [9] to determine the age and gender [1, 2] of the speaker from speech. To estimate the age and gender of the person, we employed fundamental machine learning (ML) [10] and deep learning (DL) [11] techniques. Support Vector Classification

(SVC) [12], Decision Tree (DT) [13], Random Forest (RF) [14] and One-Dimensional Convolutional Neural Network (1D-CNN) [1] techniques were employed.

The several sets of Mel frequency cepstral coefficients (MFCC) [15] based on speech characteristics were chosen with dimensionality reduction methods [16] such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [17, 18]. By using these dimensionality reduction strategies, the number of calculations and the size of the model are decreased. To estimate the age and gender of the speaker, the most useful characteristics were fed to algorithms for machine learning (ML) and deep learning (DL). Real-time voice recognition systems [19] will benefit from this approach, which will reduce the dimensionality and computational complexity of the model.

In addition, in this investigation we used random seeding [20, 21] for our machine learning models. We examined the effectiveness of ML and DL modules for age and gender recognition [1, 2] from speech using a cross-validation approach [22, 23]. Cross-validation is an important component of model evaluation. Cross-validation aids in training the model by utilizing all of the data by breaking it into several folds. We use voice signals to investigate and contrast the significance of the feature, as well as to estimate age and gender. In different machine learning tasks, PCA and ICA [17, 18] have been widely used to convert high-dimensional data into a lower-dimensional space while preserving the most important variance in the data.

Sikder et al. [24] suggested the Modified CNN models to use the Face emotion recognition dataset, Gender recognition, and Age recognition datasets to determine the emotion, age, and gender of children and adults. Kwasny and Hemmerling [25] proposed a transfer learning-based approach for automated estimation of the gender, age, and emotion of speakers. Using the TIMIT and VoxCeleb1 datasets, Maseri and Mamat [26] explain how a voice recognition system that uses the MFCC feature extraction technique for the front end and HMM recognition for the back end performs. The HMM training is done using the Baum-Welch method, while decoding is done using the Viterbi algorithm. Jain et al. [27] suggested that the speech recognition model suggests the use of several feature extraction methods, including MFCC, spectrograms, and Croma. Along with other classification algorithms, dimensionality reduction techniques such as SVM and PCA are also mentioned. Many feature extraction methods, including DWT, MFCC, and LDA, were described by Murugappan et al. [28]. The effectiveness of their model was assessed and compared for different MFCC and LDA values. To predict age, gender, and mood using the MFCC features, Zaman et al. [29] used various machine learning algorithms, including SVM, DT, and RF. Fulop [30] proposed the speech spectrum in addition to the feature extraction and speech spectrum analysis methods. Accurate speech spectrum determination is also discussed in his paper, particularly for short frames, and is commonly sought for in a number of domains, including speech processing, recognition, and acoustic phonetics. The Mozilla Common Voice dataset and other voice datasets were described by Ardila et al. [31] as beneficial to many speech-related research projects. Shafran et al. explained voice-based gender prediction using HMM and SVM in [32]. The authors examined how well SVM and GMM perform when

estimating gender based on speech characteristics. Přibil et al. [33] talked about using the GMM model to identify the gender and age of a speaker based only on their voice. Bocklet et al. [34] described how to use GMM and SVM algorithms to determine the age and gender of a speaker's speech.

The primary contributions of this paper can be stated as follows:

1. Speech data for children and adults were collected from Children Speech Recording Dataset [35], Biometric Visions and Computing (BVC) [36], and the Mozilla Common Voice Speech datasets [37], and a new speech dataset was created to identify the age and gender of the speaker.
2. The modified convolution neural network was proposed for identification of age and gender from speech using MFCC speech features and dimensionality reduction techniques [17, 18].
3. The performance of several machine learning models was compared with the proposed modified convolution neural network including the dimensionality reduction techniques.

This study is organized as follows. Section 1 shows the introduction and brief explanation of the problem statement. The characteristics of the voice corpus in the Children Speech Recording Dataset [35], Biometric Visions and Computing (BVC) [36] and Mozilla Common Voice dataset 5.1 [37] are described. The article's Section 2 also explains the Machine Learning algorithms that use various PCA, ICA [17, 18], and random seeding [20, 21] techniques. The results of several machine learning model strategies including cross-validation [22, 23], are presented in Section 3 of the same text, before the conclusion and future scope are provided in Section 4.

## 2 METHODOLOGY

The machine learning algorithms determine the age and gender of the speaker [1, 2] from speech, based on the MFCC speech features [15]. From the aforementioned voice dataset, the MFCC speech characteristics are extracted using the librosa function [9, 15]. The speaker's age and gender are predicted using the child and adult speech dataset that was previously discussed. The pipeline for the age and gender prediction model is displayed in Figure 1. There is a comparison of the ML and DL methods with various PCA, ICA, cross-validation [22, 23], and random seed [20, 21]. We are obtaining varied levels of accuracy for different PCA and ICA, seed, and CV values.

### 2.1 Datasets

Due to the scarcity and unavailability of child and adult-based speech databases, we created our own data set to estimate the age and gender of the speaker from the

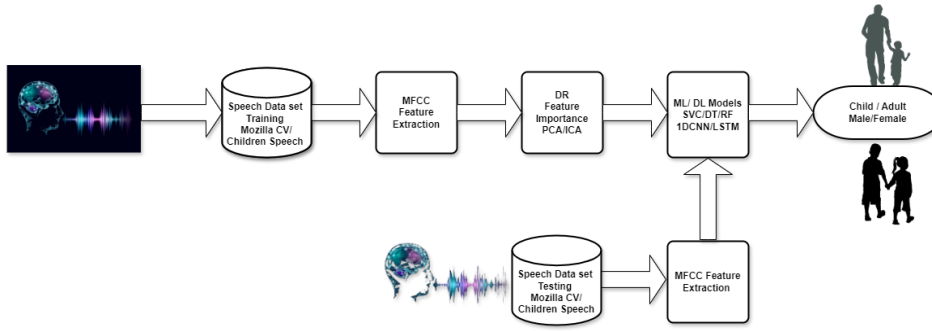


Figure 1. Pipeline for age and gender estimation from speech

speech by combining data from the Children Speech Recording Corpus [35], Biometric Visions and Computing (BVC) [36], and the Mozilla common voice-based speech corpus [37]. Using age and gender identifiers, we collected 3 000 speech samples. Of the three thousand speech samples, fifteen hundred are from the children's speech corpus and biometric visions and computing datasets, while the remaining fifteen hundred are from the adults' class in the Mozilla Common Voice Speech Corpus, as shown in Figure 2. Every speech sample includes audio recordings and age and gender information. Male and female children's voice recordings were gathered from both the Children Speech Recording Dataset [35] and the biometric vision speech (BVC) databases. The acquired speech samples were marked as children and added to the voice corpus. The 39 statistical MFCC features were retrieved from a speech dataset using the librosa-based MFCC feature extraction technique [15]. After the features were extracted, label columns were added to finish the dataset. The MFCC features of the speakers' audio files were utilized to ascertain their age and gender. The age and gender category labels are part of an extracted MFCC speech feature set from our voice dataset.

### 2.1.1 Children Speech Recording Dataset

The Children Speech Recording Dataset [35] contains audio recordings of 11 children. Both male and female children are categorized as children and also included in the voice corpus. There are nine numbers, five phrases, and many spontaneous statements are included in the dataset. Children's vocal utterances from under 14 years old are included in this speech dataset. The dataset was recorded in English, and both native and non-native speakers contributed their voices. In the Human-Robot Interaction (HRI) Laboratory at Plymouth University, two front microphones from the Aldebaran NAO robot, a portable Zoom H1 microphone, and a studio-grade Rode NTI-A microphones were used to record the speech samples.

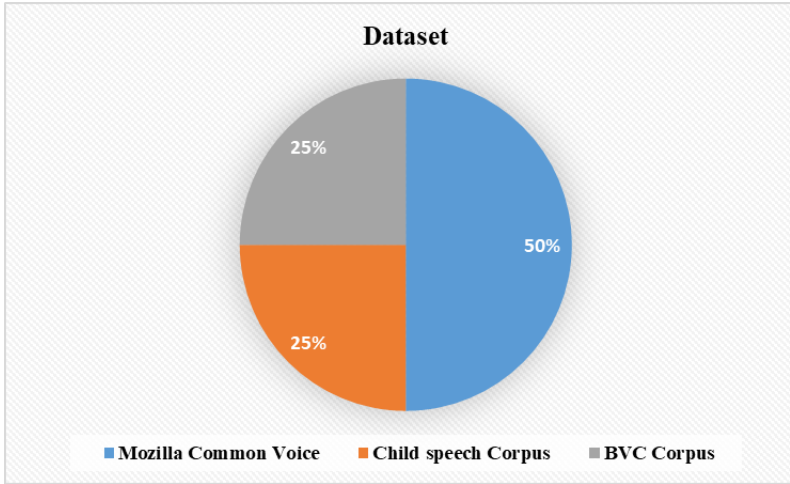


Figure 2. Origin of samples in the experimental dataset

### 2.1.2 Biometrics Visions and Computing (BVC)

The Biometrics Visions and Computing (BVC) dataset [36] of gender and age from Vocal Collection contains vocal utterances from 526 people, with one to five recordings per person (336 males and 190 females). There are 3 964 vocal utterances, including 2 149 male and 1 815 female voices. Five distinct English speeches and their equivalent translated native languages were gathered from the subjects in the first and second sessions. The set of native languages contains 28 different native languages. Speech samples of less than 18 years were collected and added to the child speech dataset.

### 2.1.3 Mozilla Common Voice 5.1

The predictive model for speaker age estimation was developed using the Mozilla Common Voice dataset [37]. This dataset consists of 64 000 speech audio files in MP3 format contributed by 61 528 unique speakers. Each recording is accompanied by metadata fields. A subset of 1 500 audio files was extracted for this study, containing metadata for filename, age, and gender. The age groups represented in the dataset range from teenagers to individuals in their nineties, categorized into decades (e.g., twenties, thirties, etc.). The cleaned dataset, comprising selected speech recordings, was processed to predict speaker age using Mel Frequency Cepstral Coefficient (MFCC) features derived from the audio files.

### 2.1.4 The Experimental Dataset

In addition to the Mozilla Common Voice dataset, our study incorporated 3 000 voice samples, divided equally between child and adult speakers. Specifically, 1 500 samples were obtained from child speakers, sourced from the Children Speech Recording Dataset and Biometrics Vision and Computing (BVC) datasets, while the remaining 1 500 samples were from adult speakers drawn from the Mozilla Common Voice 5.1 dataset. Each subgroup (children and adults) consisted of 750 male and 750 female samples. For this study, individuals under 18 years of age were categorized as children, while those aged 18 and above were classified as adults.

## 2.2 MFCC Feature Extraction

In our study, the raw audio data was processed to extract pertinent information for speech analysis, and the MFCC speech characteristics were extracted using the feature extraction approach [9]. Mel Frequency Cepstral Coefficients (MFCC) [15] are used to generate a perceptually appropriate logarithmic scale from the frequency spectrum, simulating the human auditory system. From the MFCC feature extraction, 39 characteristics were extracted for each audio frame.

### 2.2.1 Mel Frequency Cepstral Coefficients (MFCC) Features

The Mel Frequency Cepstral Coefficients (MFCC) [9, 15] are a ubiquitous technique in automatic speech recognition (ASR) systems and speech processing [3]. It has demonstrated efficacy in extracting pertinent information from voice signals while lowering the data's dimensionality. Since MFCC draws inspiration from the human auditory system, it is a good fit for applications involving speech [38]. Several approaches are used in MFCC extraction, which converts unprocessed audio data into a condensed representation that emphasizes the important acoustic properties of speech signals.

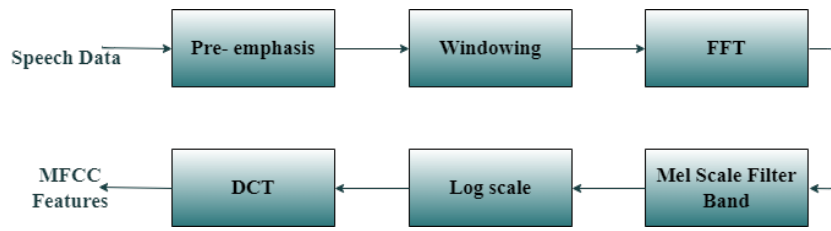


Figure 3. MFCC speech feature extraction

As seen in Figure 3, the MFCC feature extraction method [15] consists of the following steps: Pre-emphasis, framing, windowing, FFT, Mel filter bank, log scale, and DCT. A collection of MFCC coefficients is the result of applying the discrete

cosine transform (DCT) on the log-filter bank energies during the Mel Frequency Cepstral Coefficients (MFCC) extraction procedure.

We set the window duration to 25 ms, the window shift to 10 ms, and the coefficients for 39 MFCC speech characteristics per 10 ms frame. We estimated age and gender using twelve MFCC features, twelve delta MFCCs, and Double Delta MFCC's twelve characteristics, along with one log frame of energy, one delta log frame of energy, and one double delta log frame of energy.

The variability in the size of MFCC matrices brought on by variations in audio file length is challenging for machine learning algorithms, which typically demand fixed-length input vectors. Transforming the MFCC matrix into a fixed-length feature representation as though the MFCC feature set were restricted to 39 is a more sophisticated method. In order to reduce the dimension of speech feature sets, we also used methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA). These techniques produce a fixed-length feature vector that may be entered into machine learning models by converting the MFCC matrix to a lower-dimensional space while preserving important information.

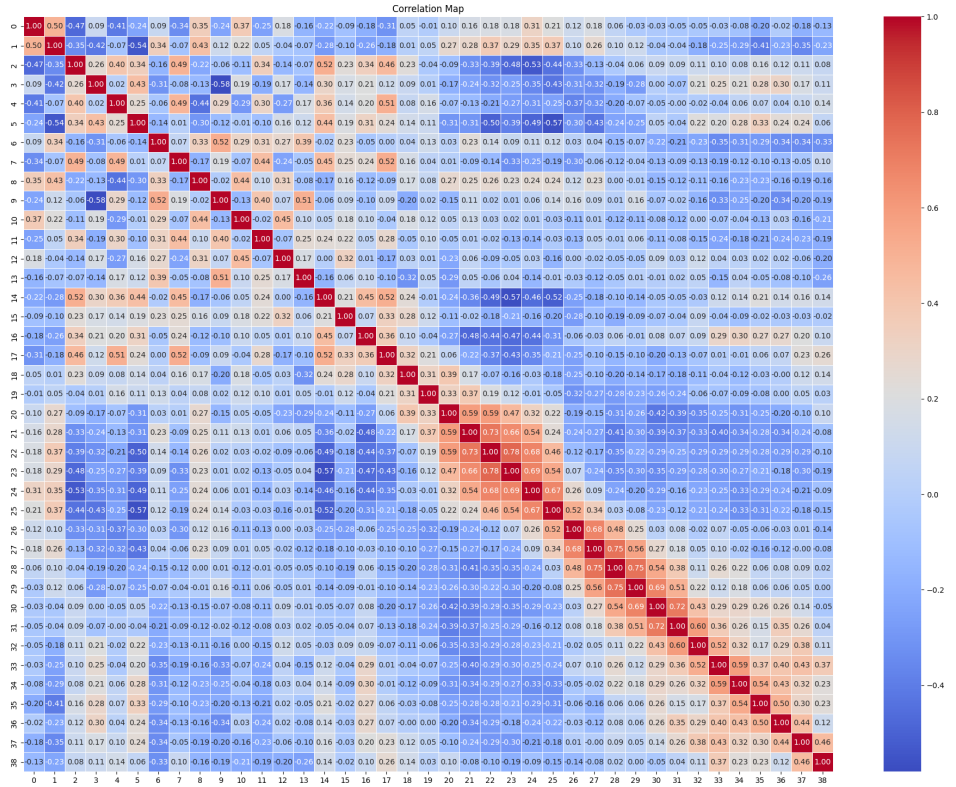


Figure 4. Correlation heat map of speech features

### 2.2.2 Correlation Heat Map

These MFCC coefficients are utilized as features for additional analysis, including speaker identification [1, 2, 3], speech recognition, and other speech-related tasks [4, 5, 6]. They describe the spectral properties of the audio signal in a compact form. Among the 3 000 audio recordings in our dataset, we were able to extract 39 MFCC characteristics for every audio file. The correlation heat map of the extracted 39 MFCC features is shown in Figure 4. The correlation heat map [39] shows the correlation between the MFCC speech features. This will explain the correlation of individual speech features with each other.

## 2.3 Support Vector Classification (SVC)

The SVC [12], one of the supervised machine learning approaches for binary classification and regression issues, is the support vector machine [40]. For age and gender categorization in Support Vector Machines (SVM), the most popular Gaussian radial basis function (RBF) is employed as the kernel function. All MFCC speech samples [15] are normalized using the min-max scalar data preprocessing approach to bring all the features to the same level. The RBF kernel SVM classification technique is used to categorize the generated MFCC feature vectors. Of all the kernel classification methods, the Gaussian RBF kernel function is used most frequently to achieve the superior SVM classification of the kernel [12]. The effectiveness of the algorithm is assessed and contrasted using different PCA, ICA random seeding, and cross-validation training and testing techniques.

## 2.4 Decision Tree (DT)

The DT algorithm [13] is one of the supervised machine learning techniques that is frequently utilized for regression and classification problems. The decision tree algorithm. The DT machine learning method is utilized to classify the speaker's age and gender using MFCC speech characteristics. As a categorization of outcomes, the DT includes leaf nodes, internal subbranches, branch nodes, and root nodes. DT algorithms come in several varieties with varying approaches; they include ID3, CART, and C4.5 [41]. The decision tree method is mostly applied to tasks related to knowledge discovery, data mining, and data cleansing. The effectiveness of the DT algorithm is assessed and contrasted using various PCA and ICA cross-validation techniques and training and testing seeding.

## 2.5 Random Forest (RF)

The RF [14] is one supervised machine learning approach used for regression and classification problems. The RF algorithm is a mixture of many DT algorithms; by adding additional DT algorithms, the machine learning model's accuracy will

increase. Random forest techniques employ a variety of hyperparameters; the algorithm's accuracy is dependent on these hyperparameters. Because the RF algorithm uses a mixture of DT, it produces better outcomes than DT algorithms. The effectiveness of the DT algorithm is assessed and contrasted using various PCA, ICA, cross-validation techniques, and training and testing seeding.

### 2.6 Modified 1D-CNN

Convolutional Neural Networks (1D-CNNs) are a class of deep learning models primarily used for image and signal processing tasks [1, 2]. 1D-CNNs have shown remarkable success in various computer vision applications [42]. However, their effectiveness is not limited to images alone, as they can also be applied to 1D data sequences, such as time series data, audio signals, and text sequences. In our research, we utilized 1D-CNNs to process the extracted MFCC features [9, 15], aiming to efficiently learn relevant patterns for age estimation. As of today we do not have an exact number for hidden layers in 1D-CNN architecture, so we modified the 1D-CNN architecture by changing the number of convolution and max pooling and padding layers by fine-tuning the number of layers the proposed model showed better results compared to all other models. Our 1D-CNN architecture for age and gender estimation comprises several layers to process the extracted features.

**1D-CNN Layer with ReLU Activation Function:** The initial 1D-CNN layer applies convolutional filters to capture local patterns and relationships in the feature data. The Rectified Linear Unit (ReLU) activation function [43] introduces non-linearity, enabling the network to learn complex representations.

**Pooling Layer:** The pooling layer samples the feature maps, reducing the spatial dimensions and retaining important information. Max pooling, commonly used in CNNs, extracts the maximum value within each pooling window.

**Flatten Layer:** The flattening layer reshapes the pooled feature maps into a 1D vector, preparing the data for fully connected layers.

**Dense Layer with ReLU Activation:** The dense layer contains neurons fully connected to the flattened features. ReLU activation is applied to introduce non-linearity and capture complex relationships.

**Output Layer with Sigmoid Activation:** The final output layer contains a single neuron with sigmoid activation, enabling the model to produce a probability score representing the estimated age.

The input shape of 1D-CNN is determined by the MFCC speech features; however, because dimensionality reduction techniques are used, the input shape of the 1DCNN varies depending on the input vector sample.

$$\text{1D CNN Input shape} = (\text{batch size}, \text{sequence length}, \text{number of channels}).$$

Input size is determined by the kernel size, stride, and padding employed in the convolutional layer. Possible input size based on the output shape of (None, 37, 64). For this state, the Kernel Size = 3; Stride = 1; Padding = 0.

The input length can be calculated as follows:

$$\text{Input Length} = (37 - 1) * 1 + 3 = 36 + 3 = 39,$$

where Input shape for Kernel size = 3, stride = 1, padding = 0. The input size is 39.

In our study, we further explored the impact of dimensionality reduction using PCA and ICA on the three types of features (MFCC, first derivative, and second derivative). We applied PCA and ICA [17, 18] with different values of PCA and ICA for each feature type, generating reduced feature sets. For each PCA, ICA, CV, and seed value, we applied a feature set and employed 1D-CNNs to learn and predict age and gender estimation. This process involved feeding the reduced feature sets through the same 1D-CNN architecture described earlier and it is depicted in Figure 5. This is tailored to handle the specific dimensions of the features.

**Model: "sequential_86"**

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv1d_130 (Conv1D) | (None, 37, 64) | 256 |
| max_pooling1d_130 (MaxPooling1D) | (None, 18, 64) | 0 |
| conv1d_131 (Conv1D) | (None, 16, 64) | 12,352 |
| max_pooling1d_131 (MaxPooling1D) | (None, 8, 64) | 0 |
| flatten_86 (Flatten) | (None, 512) | 0 |
| dense_172 (Dense) | (None, 100) | 51,300 |
| dropout_44 (Dropout) | (None, 100) | 0 |
| dense_173 (Dense) | (None, 1) | 101 |

**Total params:** 192,029 (750.12 KB)
**Trainable params:** 64,009 (250.04 KB)
**Non-trainable params:** 0 (0.00 B)
**Optimizer params:** 128,020 (500.08 KB)

Figure 5. 1D-CNN architecture

After training and evaluating the models, we observed better accuracy and correct age estimations. Furthermore, we calculated various metrics, including accuracy, F1 score, precision, recall, and support, by analyzing the confusion matrix [44] generated using a heat map visualization. These metrics provide insights into the model's overall performance and its ability to correctly estimate age using different sets of PCA, ICA, CV, and seed values.

### 2.7 Dimensionality Reduction

Dimensionality reduction is a technique that reduces a dataset's features or dimensions while preserving the most important information [16]. The curse of dimensionality, which occurs when a dataset has more characteristics than it needs, can cause overfitting and processing inefficiencies in high-dimensional datasets. We can reduce the feature set by using the cross-correlation method, but if the features are not correlated heavily, then we cannot use the correlation heat map to eliminate the features in dimensionality reduction. In such cases, we can go with other dimensionality reduction methods like PCA and ICA. The model's dimensionality may be decreased by employing a variety of techniques, including PCA and ICA, without affecting the model's performance.

#### 2.7.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) [17] is a commonly used dimensionality reduction approach that develops a new collection of orthogonal characteristics known as principal components of the original data. The variance of these principal components is used to order them, and the most significant components are those that extract the most information from the original dataset. MFCC speech traits [9, 15] were derived for every speech sample in our investigation. High-dimensional feature sets are produced by extracting the voice characteristics from each sample. We used principal component analysis (PCA) on a per-feature-type basis, varying the number of principal components (K) to achieve the desired reduction in dimensionality. Possible values for K are 5, 10, 15, 20, 25, 30, and 35.

We reduced each feature set to a smaller collection of principal components using PCA, which allowed us to keep the most crucial information while removing the less critical ones. This decrease in dimensionality probably produced models that were less prone to overfitting and more computationally efficient [45]. The various machine learning models were used for the age estimation in the classification process that was followed by PCA. To see how the various reduced feature sets influenced the model's performance, we measured accuracy and other metrics while assessing the model's performance for each PCA dimensionality level.

The fundamental mathematics of Principal Component Analysis (PCA) can be stated in a few straightforward steps. PCA lowers the dimensionality of a dataset by identifying the directions of maximum variance, known as principal components, and projecting the data onto them.

PCA finds a transformation

$$\mathbf{Z} = \mathbf{W}^{\top}\mathbf{F}_{\text{centered}},$$

where,

$$\mathbf{F} \in \mathbb{R}^{(i,j)},$$

where $\mathbf{F}$ is the vector of speech characteristics and $\mathbb{R}^{(i,j)}$ is a matrix with $i$ charac-

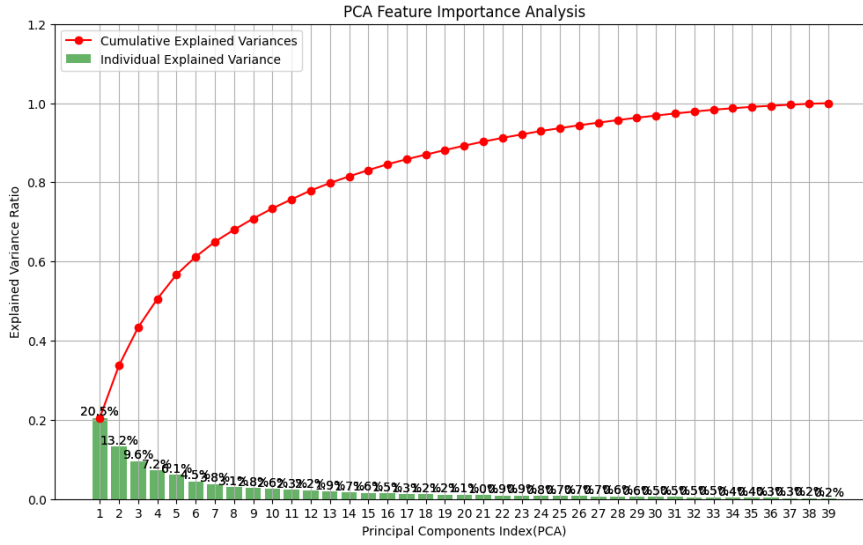


Figure 6. PCA feature importance analysis

teristics (rows) and $j$ samples (columns). $Z$ represents converted data in the new coordinate system (principal components), and W is a transform matrix comprising the covariance matrix's eigenvectors. The primary goal is to identify a set of new axes (principal components) that maximize the variety of the data.

The explained variance ratio quantifies the portion of the original dataset's total variance that can be ascribed to each major component. The explained variance ratio of a principal component is the ratio of its eigenvalue to the sum of the eigenvalues of all the other principal components. From Figure 6, the Cumulative Explained Variance plot is a visual depiction of the proportion of the dataset's variance that can be explained cumulatively by each component. During PCA, the data are converted into a new coordinate system, and the axes are ranked according to how well they capture the variance in the data.

$$\text{Explained Variance Ratio} = \frac{\lambda_x}{\sum_{y=1}^{m} \lambda_y},$$

where the numerator denotes the eigenvalue for the $x^{\text{th}}$ main component. The denominator indicates the total of all eigenvalues between $m$ and $y = 1$.

The cumulative explained variation plot is a graphical depiction that illustrates how much of the dataset's variation is explained cumulatively by each component. Cumulative explained variance is used to choose which dimensions to preserve while minimizing information loss. The explained variance in principal component analysis (PCA) is the percentage of total variance explained by each principal component. The individually explained variance aids in determining which components account

for the majority of the variance. Figure 6 depicts the relationship between cumulative explained variance and individual explained variance values, utilizing 39 PCA data points from our entire child and adult speech dataset. The model's efficiency rises as the cumulative explained variance increases, and a limited number of principal components with high individual explained variance are sufficient for representing the data.

$$\text{Cumulative Explained Variance Ratio} = \frac{\sum_{x=1}^{k} \lambda_x}{\sum_{y=1}^{m} \lambda_y},$$

where the numerator is a representation of the total of the first $k$ principal components' eigenvalues. The denominator symbolizes the sum of all eigenvalues for each of the $m$ components.

### 2.7.2 Independent Component Analysis (ICA)

The method of independent component analysis [18] is used to disentangle mixed signals into their component parts. Applications for Image and Audio Processing, including Biomedical Signal Analysis [46]. A statistical and computational method called independent component analysis (ICA) is used in machine learning to disentangle a multivariate signal into its independent non-Gaussian components. Finding a linear transformation for the data that gets the transformed data as close to statistical independence as feasible is the aim of the ICA. ICA is an effective technique for dissecting combined signals into their constituent parts. Numerous applications, including data compression, image analysis, and signal processing, can benefit from this. Since ICA is a non-parametric technique, it does not need any presumptions on the data's underlying probability distribution. Since ICA is an unsupervised learning method, labeled examples are not necessary for its application to data. Because of this, it can be helpful in circumstances when labeled data is unavailable. By identifying significant characteristics in the data that may be used for other tasks, including classification, ICA can be utilized for feature extraction.

According to ICA, the observed data X is a linear combination of independent sources S:

$$\mathbf{X} = \mathbf{AS},$$

where $\mathbf{X}$ is Observed mixed signals of $i \times j$, $i$ is the number of signals, $j$ is the number of samples, $\mathbf{A}$ is the mixing matrix of $i \times i$. The objective is to identify separate sources S using a separation matrix $\mathbf{W}$, such that

$$\mathbf{S} = \mathbf{WX}.$$

Independent Component Analysis (ICA) is a statistical and computer method for breaking down multivariate signals into statistically independent components. It is widely used in signal processing, blind source separation, and feature extraction.

### 2.8 Random Seeding

To create random integers in Python, we use the random seeding [20, 21] function. The pseudo-random numbers are produced by using random seeding. Using some specific values, the random seed method produces specific random numbers. It is also called a "seed value." The seed function stores the state of a random function such that for a given seed value, it may produce the same random numbers when the code is executed repeatedly on the same system or separate machines. The random number generator may be initialized to a known state by seeding, making it possible to replicate the same random number sequence. This is helpful for testing, troubleshooting, and ensuring repeatability in simulations and experiments. If the same seed is utilized, this makes the number sequence repeatable and predictable. The numbers produced will vary with each run if a seed is not provided. Random seeding was used during data shuffle for train-test splits and initiation in PCA and ICA to achieve consistent findings across numerous runs of the experiment. This method allows reproducibility and enables meaningful comparisons across different settings.

### 2.9 Cross-Validation (CV)

Evaluating a model's performance on a small amount of data (K-folds) in machine learning is called cross-validation (CV) [22, 23]. To train the model, the provided data is divided into many folds, or subsets, of which one is utilized as a validation set. Each time this process is performed, a different fold is used as the validation set. Finally, an average of the results from each validation phase is produced, providing a more trustworthy evaluation of the model's performance. Cross-validation serves primarily to avoid overfitting [47], a phenomenon in which a model performs badly on newly discovered data after being trained excessively well on the training set. Cross-validation yields a more accurate assessment of the model's generalization performance, that is, its capacity to function effectively on novel untested data by testing the model across several validation sets.

The $K$-fold cross-validation [22, 23] is one of the several cross-validation procedures that were used in our research. To compute $K$-fold cross-validation, we divided the dataset into $K$-folds, or subsets. We then trained on all the folds, saving one fold $(K-1)$ for the evaluation of the trained model. Using a distinct subset set aside for testing each time, we iterated $K$ times using this procedure. Due to $K$-fold cross-validation's repetition of the train/test split, it operates $K$ times quicker than Leave One Out cross-validation. Examining the findings of the comprehensive testing procedure is easier.

Cross-validation offers a more reliable assessment of the model's performance on unknown data, which helps to prevent overfitting. You may compare many models using cross-validation and choose the one that performs the best overall. By choosing the values that produce the best results on the validation set, cross-validation may be used to optimize a model's hyperparameters, such as the regularization parameter.

When compared to traditional validation procedures, cross-validation is a more data-efficient method since it makes use of all the available data for both training and validation.

### 2.10 Evaluation Metrics

We further assessed the performance of our age and gender estimation models by evaluating model testing accuracy, cross-validation, and loss. These metrics provide deeper insight into how accurately the models classified or predicted whether the individual was a child or an adult for age and male or female for gender estimation.

The confusion matrix [44] was used to compute the evaluation criteria for the speech-based age and gender estimate algorithms. Figure 7 shows the confusion matrix of the 1D-CNN model for age estimation. The formulas for the evaluation metrics used in the context of binary classification. The prediction of the age (child or adult) of the speakers into two classes based on the confusion matrix is shown.

$$\text{Accuracy} = ((\text{TP} + \text{TN}))/((\text{TP} + \text{TN} + \text{FP} + \text{FN})), \tag{1}$$

$$\text{Loss} = ((\text{FP} + \text{FN}))/((\text{TP} + \text{TN} + \text{FP} + \text{FN})). \tag{2}$$

Here TP: The model properly predicted positive classifications. TN: The model accurately predicted the negative classes. FP: Positive courses are projected wrongly, while negative classes are categorized inaccurately. FN: Positive categories are improperly classified.

Age and gender estimation from speech using various MFCC speech features was analyzed on ML and DL algorithms using various dimensionality relations, seeding, and cross-validation [20, 21, 22, 23]. The proposed 1D-CNN achieves better results compared to all the other existing methods. Based on the confusion matrix in Figure 8, the speaker's gender (male or female) is predicted and divided into two classes. The evaluation metrics, like model training, test accuracy, and loss, are calculated for age and gender estimation models from the confusion matrix using the above-mentioned formulas as Equations (1) and (2).

## 3 RESULTS

The age and gender of the speaker were ascertained using a variety of dimension reduction, seeding, and cross-validation techniques on ML and DL algorithms. We successfully used PCA and ICA to reduce dimensionality in several ML and DL applications, both with and without PCA and ICA feature sets of varied sizes. We evaluated the accuracy and predictability of age and gender estimates. After using ML and DL algorithms with different PCA and ICA levels, the Figure 9 compares the test and train accuracy of age and gender models.

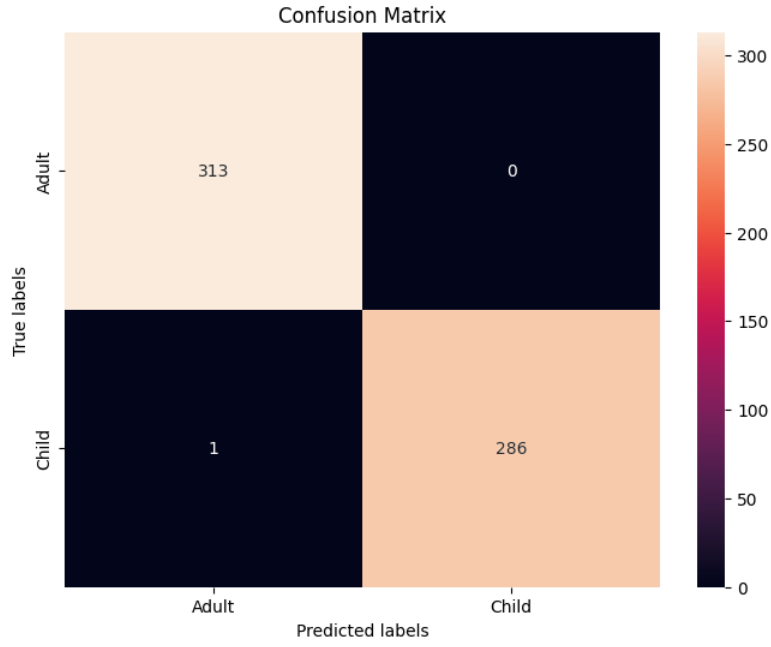


Figure 7. Confusion matrix (Child or Adult classes)

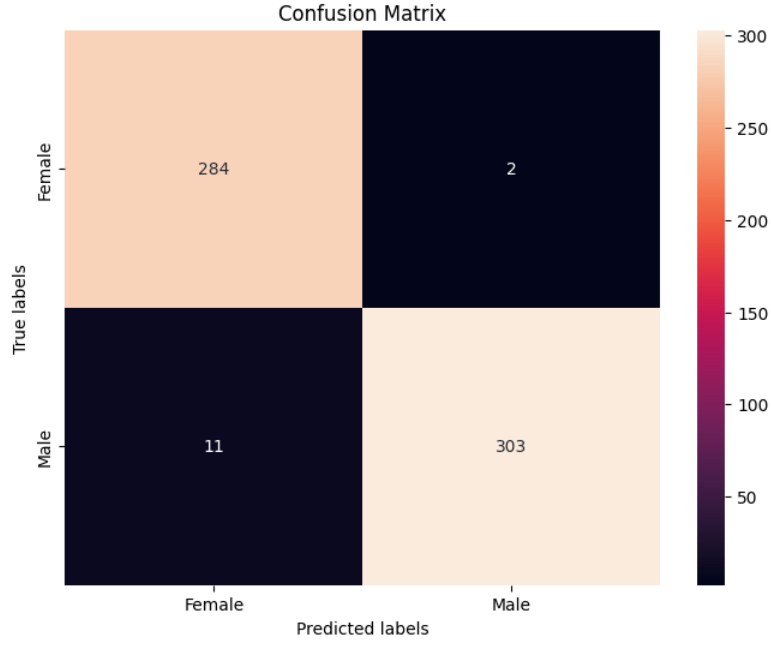


Figure 8. Confusion matrix (Male or Female classes)

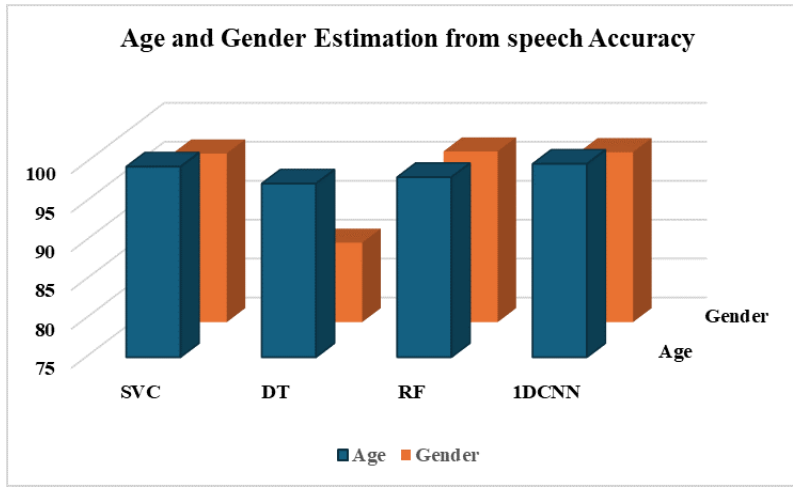


Figure 9. Age and gender identification test accuracy

### 3.1 Age Estimation from Speech

The outcome of the proposed 1D-CCN model was observed and shown in Figure 10. The individual's age and gender were estimated using our own dataset. The proposed 1D-CNN outperforms all other machine learning algorithms in terms of accuracy. It was observed and compared with the train and test accuracy of various machine learning models.

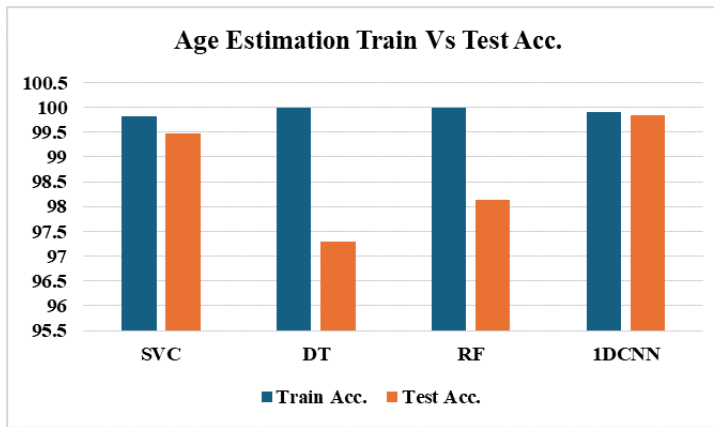


Figure 10. Age estimation train vs test accuracy

When compared the models without PCA, the machine learning models with different PCA levels provide greater accuracy. The train-test split is normally dis-

persed, with 20 percent going toward testing and 80 percent going toward training. When comparing the age estimate model's performance with and without PCA values, the latter produced superior results. According to Table 1, rather than 39, the suggested 1D-CNN model produced satisfactory results for PCA values of 30, 25, 15, and 10.

| **Model** | **PCA35** | **PCA30** | **PCA25** | **PCA20** | **PCA15** | **PCA10** | **PCA5** |
|---|---|---|---|---|---|---|---|
| **SVC** | 99.46 | 99.60 | 99.33 | 99.06 | 99.73 | 99.20 | 98.01 |
| **DT** | 98.13 | 97.73 | 97.60 | 97.20 | 97.60 | 97.46 | 97.46 |
| **RF** | 99.20 | 99.33 | 98.26 | 97.60 | 98.13 | 98.40 | 97.33 |
| **1D-CNN** | 99.83 | 99.54 | 99.39 | 99.83 | 98.52 | 99.34 | 98.83 |

Table 1. Performance analysis of different models with various PCA

The machine learning models obtain better accuracy with different ICA values than those without ICA values. As is typical, 80 percent of the train-test split goes toward training and 20 percent toward testing. The age estimate model performed better when comparing results with and without ICA values. Based on Table 2, instead of 39 ICA values, the suggested 1D-CNN model produced decent results at 20 and 10.

| **Model** | **ICA35** | **ICA30** | **ICA25** | **ICA20** | **ICA15** | **ICA10** | **ICA5** |
|---|---|---|---|---|---|---|---|
| **SVC** | 96.13 | 97.60 | 98.80 | 99.06 | 99.86 | 99.33 | 99.33 |
| **DT** | 89.73 | 91.46 | 94.53 | 91.06 | 94.53 | 98.13 | 98.26 |
| **RF** | 97.06 | 98.80 | 95.86 | 95.33 | 97.20 | 99.06 | 98.53 |
| **1D-CNN** | 98.33 | 98.83 | 97.50 | 99.23 | 97.50 | 99.50 | 92.33 |

Table 2. Performance analysis of different models with various ICA

To estimate age from speech, the seeding strategy [20, 21] was also used and tested with ML and DL algorithms. The different ML and DL algorithms yield varying accuracies depending on the seed value. In comparison to all other seeding strategies, the seed 40 for 1D-CNN yields greater accuracy, as shown in Figure 11.

To use ML and DL approaches, training and testing undergo 5, 10, and 15-fold cross-validation [22, 23]. Table 3 illustrates that, of all the testing and training techniques, 5-fold cross-validation (CV5) using the suggested 1D-CNN produces the best results.

### 3.2 Gender Estimation from Speech

The speakers' gender was estimated using a dataset we compiled, and the outcomes of the suggested 1D-CCN, which included a variety of machine learning models, were noted and displayed in Figure 12. When compared to all other ML and DL algorithms, the suggested 1D-CNN produces superior results. The different machine

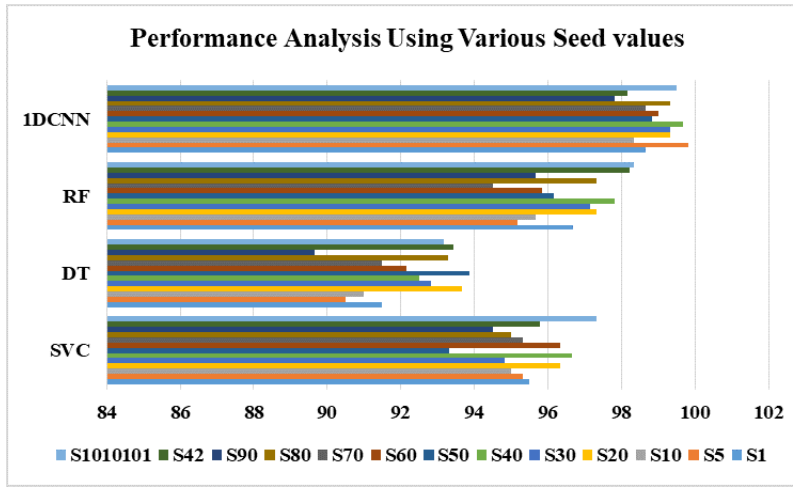


Figure 11. Performance analysis using various seed values

| CV | CV5 | | CV10 | | CV15 | |
|---|---|---|---|---|---|---|
| **Model** | **Test Acc.** | **Loss** | **Test Acc.** | **Loss** | **Test Acc.** | **Loss** |
| **SVC** | 99.1 | 0.006 | 99.3 | 0.050 | 99.4 | 0.070 |
| **DT** | 96.5 | 0.009 | 97.1 | 0.011 | 96.9 | 0.014 |
| **RF** | 97.6 | 0.004 | 97.9 | 0.007 | 98.0 | 0.013 |
| **1D-CNN** | 99.7 | 0.013 | 99.66 | 0.015 | 99.53 | 0.042 |

Table 3. Performance analysis of different models with various cross validation

learning models' training and test accuracy were noted and contrasted with one another.

The machine learning models with varying PCA values produce greater accuracy when compared to models without PCA. With 20 percent going toward testing and 80 percent toward training, the train-test split is regularly distributed. When evaluating the performance of the gender estimation model with and without PCA values, the latter yielded better outcomes. Considering Table 4, the recommended 1D-CNN model yielded good results with PCA values of 35 and 30, instead of 39.

| **Model** | **PCA35** | **PCA30** | **PCA25** | **PCA20** | **PCA15** | **PCA10** | **PCA5** |
|---|---|---|---|---|---|---|---|
| **SVM** | 96.66 | 95.60 | 95.73 | 95.33 | 94.53 | 91.86 | 87.20 |
| **DT** | 88.53 | 86.53 | 87.46 | 89.6 | 89.46 | 88.26 | 86.40 |
| **RF** | 95.20 | 94.93 | 94.13 | 95.60 | 92.67 | 93.20 | 91.86 |
| **1D-CNN** | 97.83 | 97.83 | 97.33 | 96.49 | 97.00 | 97.00 | 88.18 |

Table 4. Gender identification with PCA

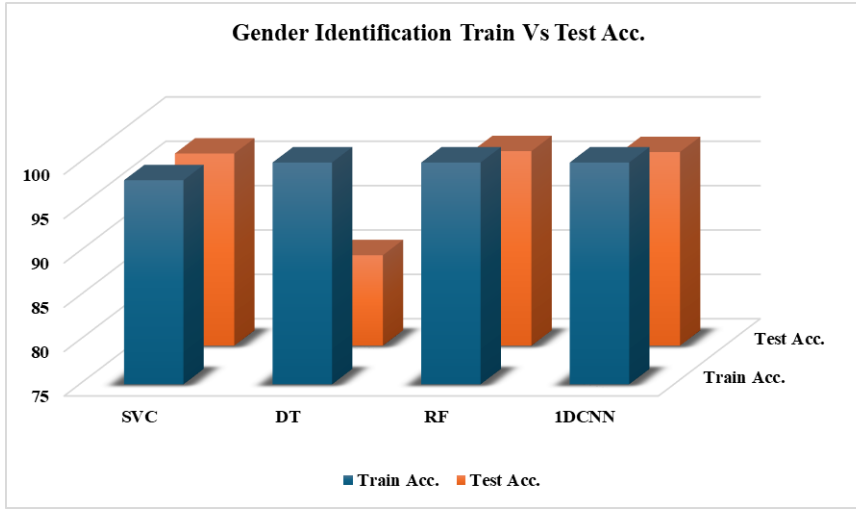


Figure 12. Gender identification train vs test accuracy

Compared to models without ICA, the machine learning models with different ICA values provide improved accuracy. The train-test split is normally dispersed, with 20 percent going toward testing and 80 percent going toward training. When comparing the performance of the age estimation model with and without ICA values, the latter produced inferior results. From Table 5, rather than 39, the suggested 1D-CNN model produced satisfactory results for ICA 20.

| **ICA** | **ICA35** | **ICA30** | **ICA25** | **ICA20** | **ICA15** | **ICA10** | **ICA5** |
|---|---|---|---|---|---|---|---|
| **SVM** | 94.80 | 97.06 | 97.87 | 96.93 | 96.00 | 96.00 | 90.27 |
| **DT** | 84.40 | 83.60 | 86.13 | 83.06 | 87.33 | 88.26 | 87.46 |
| **RF** | 91.86 | 90.40 | 91.46 | 89.20 | 91.60 | 91.20 | 90.40 |
| **1D-CNN** | 95.83 | 97.33 | 96.66 | 97.50 | 96.83 | 95.83 | 88.83 |

Table 5. Gender identification using ICA

The gender estimation of the speech process was also tested using the seeding approach [20, 21] in comparison to ML and DL algorithms. Varying seed values provide varying degrees of accuracy for various machine learning and deep learning methods. Figure 13 illustrates that compared to all other seeding strategies, seed 40 for 1D-CNN yields the highest accuracy results.

Training and testing are subjected to 5-10 and 15-fold cross-validation to use ML and DL techniques [22, 23]. The results of all the testing methods are best achieved with cross-validations of 5-fold (CV5) and 15-fold (CV15) using the recommended 1D-CNN, as shown in Table 6. The average of all k-folds (5, 10, and 15) cross-validation test accuracy and loss parameters was displayed.

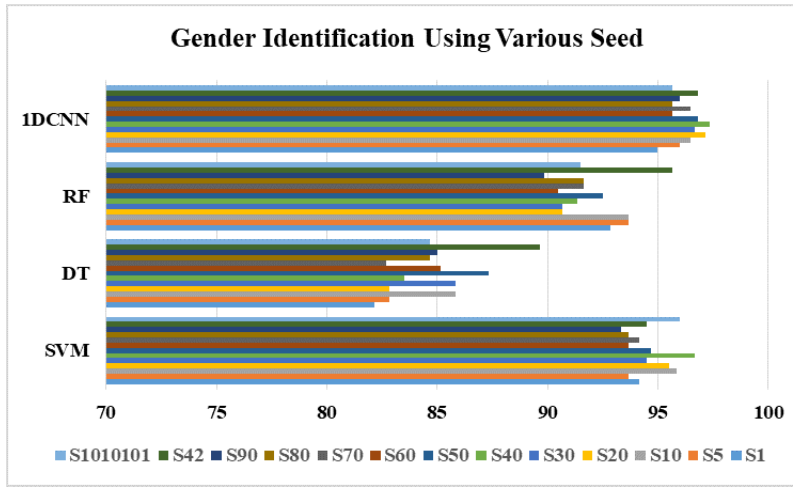


Figure 13. Gender identification using various seed

| CV | CV5 | | CV10 | | CV15 | |
|---|---|---|---|---|---|---|
| **Model** | **Test Acc.** | **Loss** | **Test Acc.** | **Loss** | **Test Acc.** | **Loss** |
| **SVC** | 92.4 | 0.017 | 92.80 | 0.012 | 92.90 | 0.021 |
| **DT** | 87.9 | 0.017 | 88.31 | 0.022 | 88.20 | 0.026 |
| **RF** | 89.6 | 0.017 | 89.30 | 0.022 | 88.90 | 0.027 |
| **1D-CNN** | 97.66 | 0.066 | 97.53 | 0.057 | 97.66 | 0.054 |

Table 6. Gender identification with various cross validation

### 3.3 Comparison of Age and Gender Estimation with SOTA Approaches

The performance of the suggested model is compared with other age and gender estimation models shown in Table 7. The proposed 1D-CNN model performs better when utilizing a variety of speech characteristics to determine the age and gender of the speaker. The results of several age group categories in earlier editions. The suggested work distinguished between adult and child speech categories and voice gender using the Mozilla Common Voice, biometric visions (BVC), and Children's Voice Recognition databases. The accuracy of the speaker's age and gender tasks, as well as a comparison to current literature, are displayed in Table 7. The proposed effort is promising, despite using only 2 classes, and could yield better outcomes than previous investigations.

## 4 CONCLUSIONS

In this paper, we study various machine learning and deep learning architectures with PCA, ICA, cross-validation, and random seeding to estimate the age and

| Studies | Task | No. of Classes | Dataset | Accuracy |
|---|---|---|---|---|
| [48] | Age and Gender | 7 | aGender | 58.98 |
| [49] | Age and Gender | 3 | A dataset of 384 speakers | 77.5/49.52 (M/F) |
| [49] | Age | 6 | Common Voice | 96 |
| [50] | Age and Gender | 12 | Common Voice | 76 |
| [51] | Age and Gender | 3 | TIDIGITS | 92.25 |
| [2] | Age and Gender | 6 | Common Voice | 80 |
| [1] | Age and Gender | 10 | Common Voice | 94.40 |
| This study | Age | 2 | Child and Common Voice | 99.83 |
| This study | Gender | 2 | Child and Common Voice | 98.00 |

Table 7. Performance comparison of proposed method with existing studies

gender of the speaker from speech using various MFCC speech features. In this work, we implemented SVC, DT, and RF ML algorithms, including modified 1D-CNN.

The DL-based age and gender estimation using various dimensionality reduction techniques such as PCA and ICA. The performance of the model was observed and compared with each other for various sets of MFCC speech characteristics. MFCC speech characteristics, along with PCA and ICA, played a major role in identifying the age and gender of the person from various MFCC speech characteristics.

The PCA and ICA dimensionality reduction techniques were implemented for age and gender recognition, and results were observed. Instead of using all the MFCC speech features, we can use highly effective features using PCA and ICA methods for the identification of the age and gender of the speaker from speech. Train-test seed and cross-validation techniques were applied, and the performance was observed with existing literature.

The 5-10 and 15-fold cross-validations were observed, and the results were analyzed. The modified DL-based 1D-CNN model gives better performance compared with all other existing models. The 1D-CNN model gave better results with MFCC speech features along with PCA, ICA, seeding, and cross-validation methods. The age and gender estimation of the person was identified with the proposed method and provides good results compared to existing methods.

The main aim of this study is to analyze the importance of speech features with dimensionality reduction techniques, which are useful for real-time speech recognition applications. This research helps to provide information about the field of study and limitations of the existing ML and DL algorithms for age and gender identification. We can also implement advanced dimensionality reduction techniques for real-time speech processing applications. We will try to implement real-time speaker age and gender from speech using various dimensionality reduction and hybrid deep learning algorithms for human-machine interface applications.

**Acknowledgments**

This work is partially supported by the Science and Engineering Research Board (SERB) govt. of India, under the Grant No. EEQ/2021/001087. The TUKE research was partially supported by the Ministry of Education, Research, and Development and Youth of the Slovak Republic under the projects VEGA 2/0092/25 and KEGA 049TUKE-4/2024, and partly by the Slovak Research and Development Agency under the projects APVV-22-0414 and APVV-22-0261.

## REFERENCES

[1] Yücesoy, E.: Speaker Age and Gender Recognition Using 1D and 2D Convolutional Neural Networks. Neural Computing and Applications, Vol. 36, 2024, No. 6, pp. 3065–3075, doi: 10.1007/s00521-023-09153-0.

[2] Sánchez-Hevia, H. A.—Gil-Pita, R.—Utrilla-Manso, M.—Rosa-Zurera, M.: Age Group Classification and Gender Recognition from Speech with Temporal Convolutional Neural Networks. Multimedia Tools and Applications, Vol. 81, 2022, No. 3, pp. 3535–3552, doi: 10.1007/s11042-021-11614-4.

[3] Kheddar, H.—Hemis, M.—Himeur, Y.: Automatic Speech Recognition Using Advanced Deep Learning Approaches: A Survey. Information Fusion, Vol. 109, 2024, Art. No. 102422, doi: 10.1016/j.inffus.2024.102422.

[4] Guven, G.—Guz, U.—Gürkan, H.: A Novel Biometric Identification System Based on Fingertip Electrocardiogram and Speech Signals. Digital Signal Processing, Vol. 121, 2022, Art. No. 103306, doi: 10.1016/j.dsp.2021.103306.

[5] Zhong, R.—Ma, M.—Zhou, Y.—Lin, Q.—Li, L.—Zhang, N.: User Acceptance of Smart Home Voice Assistant: A Comparison Among Younger, Middle-Aged, and Older Adults. Universal Access in the Information Society, Vol. 23, 2024, No. 1, pp. 275–292, doi: 10.1007/s10209-022-00936-1.

[6] Chen, L.—Su, W.—Wu, M.—Pedrycz, W.—Hirota, K.: A Fuzzy Deep Neural Network with Sparse Autoencoder for Emotional Intention Understanding in Human-Robot Interaction. IEEE Transactions on Fuzzy Systems, Vol. 28, 2020, No. 7, pp. 1252–1264, doi: 10.1109/TFUZZ.2020.2966167.

[7] Sarwar, R.—An Ha, L.—Teh, P. S.—Sabah, F.—Nawaz, R.—Hameed, I. A.—Hassan, M. U.: AGI-P: A Gender Identification Framework for Authorship Analysis Using Customized Fine-Tuning of Multilingual Language Model. IEEE Access, Vol. 12, 2024, pp. 15399–15409, doi: 10.1109/ACCESS.2024.3358199.

[8] Radha, K.—Bansal, M.: Feature Fusion and Ablation Analysis in Gender Identification of Preschool Children from Spontaneous Speech. Circuits, Systems, and Signal Processing, Vol. 42, 2023, No. 10, pp. 6228–6252, doi: 10.1007/s00034-023-02399-y.

[9] Luengo, I.—Navas, E.—Hernáez, I.: Feature Analysis and Evaluation for Automatic Emotion Identification in Speech. IEEE Transactions on Multimedia, Vol. 12, 2010, No. 6, pp. 490–501, doi: 10.1109/TMM.2010.2051872.

[10] Deng, L.—Li, X.: Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, 2013, No. 5, pp. 1060–1089, doi: 10.1109/TASL.2013.2244083.

[11] Zaman, K.—Sah, M.—Direkoglu, C.—Unoki, M.: A Survey of Audio Classification Using Deep Learning. IEEE Access, Vol. 11, 2023, pp. 106620–106649, doi: 10.1109/ACCESS.2023.3318015.

[12] Solera-Urena, R.—Garcia-Moral, A. I.—Pelaez-Moreno, C.—Martinez-Ramon, M.—Diaz-de Maria, F.: Real-Time Robust Automatic Speech Recognition Using Compact Support Vector Machines. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, 2012, No. 4, pp. 1347–1361, doi: 10.1109/TASL.2011.2178597.

[13] Hu, R.—Zhao, Y.: Knowledge-Based Adaptive Decision Tree State Tying for Conversational Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, 2007, No. 7, pp. 2160–2168, doi: 10.1109/TASL.2007.901830.

[14] Phan, H.—Maass, M.—Mazur, R.—Mertins, A.: Random Regression Forests for Acoustic Event Detection and Classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, 2015, No. 1, pp. 20–31, doi: 10.1109/TASLP.2014.2367814.

[15] Abdul, Z. K.—Al-Talabani, A. K.: Mel Frequency Cepstral Coefficient and Its Applications: A Review. IEEE Access, Vol. 10, 2022, pp. 122136–122158, doi: 10.1109/ACCESS.2022.3223444.

[16] Jia, W.—Sun, M.—Lian, J.—Hou, S.: Feature Dimensionality Reduction: A Review. Complex & Intelligent Systems, Vol. 8, 2022, No. 3, pp. 2663–2693, doi: 10.1007/s40747-021-00637-x.

[17] Zheng, J.—Yang, Z.—Ge, Z.: Deep Residual Principal Component Analysis as Feature Engineering for Industrial Data Analytics. IEEE Transactions on Instrumentation and Measurement, Vol. 73, 2024, pp. 1–10, doi: 10.1109/TIM.2024.3420267.

[18] Chien, J. T.—Chen, B. C.: A New Independent Component Analysis for Speech Recognition and Separation. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, 2006, No. 4, pp. 1245–1254, doi: 10.1109/TSA.2005.858061.

[19] Deng, M.—Chen, J.—Wu, Y.—Ma, S.—Li, H.—Yang, Z.—Shen, Y.: Using Voice Recognition to Measure Trust During Interactions with Automated Vehicles. Applied Ergonomics, Vol. 116, 2024, Art. No. 104184, doi: 10.1016/j.apergo.2023.104184.

[20] Kaczmarczyk, K.—Miałkowska, K.: Backtesting Comparison of Machine Learning Algorithms with Different Random Seed. Procedia Computer Science, Vol. 207, 2022, pp. 1901–1910, doi: 10.1016/j.procs.2022.09.248.

[21] Dutta, S.—Arunachalam, A.—Misailovic, S.: To Seed or Not to Seed? An Empirical Analysis of Usage of Seeds for Testing in Machine Learning Projects. 2022 IEEE Conference on Software Testing, Verification and Validation (ICST), 2022, pp. 151–161, doi: 10.1109/ICST53961.2022.00026.

[22] Vadovský, M.—Paralič, J.: Parkinson's Disease Patients Classification Based on the Speech Signals. 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), 2017, pp. 000321–000326, doi:

10.1109/SAMI.2017.7880326.

[23] Wong, T. T.—Yeh, P. Y.: Reliable Accuracy Estimates from k-Fold Cross Validation. IEEE Transactions on Knowledge and Data Engineering, Vol. 32, 2020, No. 8, pp. 1586–1594, doi: 10.1109/TKDE.2019.2912815.

[24] Sikder, J.—Datta, N.—Tripura, S.—Das, U. K.: Emotion, Age and Gender Recognition Using SURF, BRISK, M-SVM and Modified CNN. 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), IEEE, 2022, pp. 1–6, doi: 10.1109/ICECET55527.2022.9872771.

[25] Kwasny, D.—Hemmerling, D.: Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks. Sensors, Vol. 21, 2021, No. 14, Art. No. 4785, doi: 10.3390/s21144785.

[26] Maseri, M.—Mamat, M.: Performance Analysis of Implemented MFCC and HMM-Based Speech Recognition System. 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2020, pp. 1–5, doi: 10.1109/IICAIET49801.2020.9257823.

[27] Jain, K.—Chaturvedi, A.—Dua, J.—Bhukya, R. K.: Investigation Using MLP-SVM-PCA Classifiers on Speech Emotion Recognition. 2022 IEEE 9th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2022, pp. 1–6, doi: 10.1109/UPCON56432.2022.9986457.

[28] Murugappan, M.—Baharuddin, N. Q. I.—Jerritta, S.: DWT and MFCC Based Human Emotional Speech Classification Using LDA. 2012 International Conference on Biomedical Engineering (ICoBE), 2012, pp. 203–206, doi: 10.1109/ICoBE.2012.6179005.

[29] Zaman, S. R.—Sadekeen, D.—Alfaz, M. A.—Shahriyar, R.: One Source to Detect Them All: Gender, Age, and Emotion Detection from Voice. 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 2021, pp. 338–343, doi: 10.1109/COMPSAC51774.2021.00055.

[30] Fulop, S. A.: Speech Spectrum Analysis. Springer, 2011, doi: 10.1007/978-3-642-17478-0.

[31] Ardila, R.—Branson, M.—Davis, K.—Kohler, M.—Meyer, J.—Henretty, M.—Morais, R.—Saunders, L.—Tyers, F.—Weber, G.: Common Voice: A Massively-Multilingual Speech Corpus. In: Calzolari, N., Béchet, F., Blache, P. et al. (Eds.): Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020). European Language Resources Association, 2020, pp. 4218–4222, https://aclanthology.org/2020.lrec-1.520/.

[32] Shafran, I.—Riley, M.—Mohri, M.: Voice Signatures. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 31–36, doi: 10.1109/ASRU.2003.1318399.

[33] Přibil, J.—Přibilová, A.—Matoušek, J.: GMM-Based Speaker Gender and Age Classification After Voice Conversion. 2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016, pp. 1–5, doi: 10.1109/SPLIM.2016.7528391.

[34] Bocklet, T.—Maier, A.—Bauer, J. G.—Burkhardt, F.—Noth, E.: Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and

Support Vector Machines. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 1605–1608, doi: 10.1109/ICASSP.2008.4517932.

[35] Kennedy, J.—Lemaignan, S.—Montassier, C.—Lavalade, P.—Irfan, B.—Papadopoulos, F.—Senft, E.—Belpaeme, T.: Children Speech Recording (English, Spontaneous Speech + Pre-Defined Sentences). Zenodo, 2016, doi: 10.5281/zenodo.200495.

[36] Iloanusi, O.—Ejiogu, U.—Okoye, I. E.—Ezika, I.—Ezichi, S.—Osuagwu, C.—Ejiogu, E.: Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca. 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), 2019, pp. 175–179, doi: 10.1109/ISCMI47871.2019.9004306.

[37] Common Voice. Mozilla, 2021, https://commonvoice.mozilla.org/ (Accessed: 15 September 2021).

[38] Uzun Ozsahin, D.—Emegano, D. I.—Hassan, A.—Aldakhil, M.—Banat, A. M.—Duwa, B. B.—Ozsahin, I.: Chapter Fifteen – A Speech Recognition System Using Technologies of Audio Signal Processing. In: Uzun Ozsahin, D., Ozsahin, I. (Eds.): Practical Design and Applications of Medical Devices. Elsevier, 2024, pp. 203–216, doi: 10.1016/b978-0-443-14133-1.00001-x.

[39] Yuan, X.—Wang, Y.—Wang, C.—Ye, L.—Wang, K.—Wang, Y.—Yang, C.—Gui, W.—Shen, F.: Variable Correlation Analysis-Based Convolutional Neural Network for Far Topological Feature Extraction and Industrial Predictive Modeling. IEEE Transactions on Instrumentation and Measurement, Vol. 73, 2024, pp. 1–10, doi: 10.1109/TIM.2024.3373085.

[40] Argouarc'h, E.—Desbouvries, F.: Binary Classification Based Monte Carlo Simulation. IEEE Signal Processing Letters, Vol. 31, 2024, pp. 1449–1453, doi: 10.1109/LSP.2024.3396403.

[41] Mienye, I. D.—Jere, N.: A Survey of Decision Trees: Concepts, Algorithms, and Applications. IEEE Access, Vol. 12, 2024, pp. 86716–86727, doi: 10.1109/ACCESS.2024.3416838.

[42] Ige, A. O.—Sibiya, M.: State-of-the-Art in 1D Convolutional Neural Networks: A Survey. IEEE Access, Vol. 12, 2024, pp. 144082–144105, doi: 10.1109/ACCESS.2024.3433513.

[43] Rasel, M. A.—Obaidellah, U. H.—Kareem, S. A.: Convolutional Neural Network-Based Skin Lesion Classification with Variable Nonlinear Activation Functions. IEEE Access, Vol. 10, 2022, pp. 83398–83414, doi: 10.1109/ACCESS.2022.3196911.

[44] Mathur, A.—Foody, G. M.: Multiclass and Binary SVM Classification: Implications for Training and Classification Users. IEEE Geoscience and Remote Sensing Letters, Vol. 5, 2008, No. 2, pp. 241–245, doi: 10.1109/LGRS.2008.915597.

[45] Freire, P.—Srivallapanondh, S.—Spinnler, B.—Napoli, A.—Costa, N.—Prilepsky, J. E.—Turitsyn, S. K.: Computational Complexity Optimization of Neural Network-Based Equalizers in Digital Signal Processing: A Comprehensive Approach. Journal of Lightwave Technology, Vol. 42, 2024, No. 12, pp. 4177–4201, doi: 10.1109/JLT.2024.3386886.

[46] Lemm, S.—Curio, G.—Hlushchuk, Y.—Muller, K. R.: Enhancing the Signal-to-Noise Ratio of ICA-Based Extracted ERPs. IEEE Transactions on Biomedical Engineering, Vol. 53, 2006, No. 4, pp. 601–607, doi: 10.1109/TBME.2006.870258.

[47] Li, Z.—Kamnitsas, K.—Glocker, B.: Analyzing Overfitting Under Class Imbalance in Neural Networks for Image Segmentation. IEEE Transactions on Medical Imaging, Vol. 40, 2021, No. 3, pp. 1065–1077, doi: 10.1109/TMI.2020.3046692.

[48] Qawaqneh, Z.—Mallouh, A. A.—Barkana, B. D.: Deep Neural Network Framework and Transformed MFCCs for Speaker's Age and Gender Classification. Knowledge-Based Systems, Vol. 115, 2017, pp. 5–14, doi: 10.1016/j.knosys.2016.10.008.

[49] Büyük, O.—Arslan, M. L.: Combination of Long-Term and Short-Term Features for Age Identification from Voice. Advances in Electrical and Computer Engineering, Vol. 18, 2018, No. 2, pp. 101–108, doi: 10.4316/aece.2018.02013.

[50] Tursunov, A.—Mustaqeem—Choeh, J. Y.—Kwon, S.: Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module Through Speech Spectrograms. Sensors, Vol. 21, 2021, No. 17, Art. No. 5892, doi: 10.3390/s21175892.

[51] Vlaj, D.—Zgank, A.: Acoustic Gender and Age Classification as an Aid to Human-Computer Interaction in a Smart Home Environment. Mathematics, Vol. 11, 2022, No. 1, Art. No. 169, doi: 10.3390/math11010169.

**Laxmi Kantham Durgam** received his B.Tech. degree in 2015 from the Department of Electronics and Communication Engineering at the Jawaharlal Nehru Technological University (JNTU), Hyderabad, India. In 2020, the JNTUH university awarded him a Master's degree in digital systems and computer electronics. Currently, he is pursuing his Ph.D. degree in speech signal processing with the Electronics and Communication Engineering Department at the National Institute of Technology (NIT) Warangal, India. At the Technical University of Košice, Slovakia, he is presently a visiting researcher (NSP SAIA) at the Department of Electronics and Multimedia Communications within the Faculty of Electrical Engineering and Informatics. His areas of interest in the study are deep learning and its applications, computer vision, machine learning, optimization, human-machine interaction, image and video processing, speech processing, speech recognition, and speaker diarization.

**Ravi Kumar Jatoth** graduated in 2003 with a B.Eng. in electronics and communications engineering from the Osmania University in Hyderabad, India. In 2005, he earned his Master's degree in instrumentation and control systems from the Jawaharlal Nehru Technological University (JNTU) in Hyderabad. In 2014, he earned his Ph.D. from the National Institute of Technology (NIT) Warangal in Warangal, India. He is presently Professor at NIT Warangal's Department of Electronics and Communication Engineering. He founded the Laboratory of Digital Signal Processing at NITW. His research includes digital signal processing, artificial intelligence, computer vision, process controller design, tracking algorithms, signal processing, and graph and image processing. He has received over 1 584 citations for more than 100 publications published in journals and conference proceedings.

**Daniel Hládek** received his M.Sc. degree in artificial intelligence at the Technical University of Košice, Slovakia, in 2006 and his Ph.D. degree in artificial intelligence in 2009. From 2009 to 2015, he was Research Assistant in the Laboratory of Speech Communication Technologies. Since 2015, he has been Assistant Professor at the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia. His research interests include natural language processing, natural language understanding, natural language generation, text mining, statistical language modeling, question answering, automatic spelling correction, spontaneous speech recognition, and human–computer interaction. He also investigates questions related to the detection of hate speech and offensive language.

**Stanislav Ondáš** graduated from the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice, Slovakia, in 2004. His Ph.D. degree he received at the same department in the field of telecommunications in 2008. He is Associate Professor currently working in the Laboratory of Speech and Mobile Technologies in the same department. His interests include children's audiometry, speech therapy, spoken dialogue systems, dialog management and human-robot interaction. Additionally, he was involved in several COST activities and national and international initiatives and projects. To date, he has received over 380 citations for more than 70 publications published in journals and conference proceedings.


**Matúš Pleva** graduated with his Ph.D. in telecommunications from the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice, Slovakia (2010). He works as Associate Professor in the field of informatics and is Head of the Department. His research interests are acoustic modeling, acoustic event detection, speaker recognition, speech processing, human–machine interaction, embedded systems and parallel computing, security and biometrics, computer networking, IoT, etc. He just successfully finished the "Deep Learning for Advanced Speech Enabled Applications" bilateral project with the National Taipei University of Technology and the "Content Innovation and Lecture Textbooks for Biometric Safety Systems" project as principal investigator. He leads an intense bilateral collaboration between TUKE and CAVS, MSU, US in the field of robotics and HCI. He also participated in more than 50 national and international projects and COST actions. He has published over 150 technical papers in journals and conference proceedings with over 1 400 citations to date.

**Jozef Juhár** graduated from the Technical University of Košice, Slovakia, in 1980 with his degree in radioelectronics. He founded the Laboratory of Speech Communication Technologies at the Technical University of Košice, Slovakia, where he currently holds the position of Full Professor in the Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics. He has authored and co-authored over 400 scientific papers. His areas of interest in research are digital speech and audio analysis and synthesis and natural language processing for intelligent applications, and the study and development of spoken dialogue systems, human–computer interfaces, and speech recognition systems.