

Article

Automatic Age and Gender Recognition Using Ensemble Learning

Ergün Yücesoy

Vocational School of Technical Sciences, Ordu University, Ordu 52200, Turkey; yucesoye@odu.edu.tr;
Tel.: +90-452-2354865-4145

Abstract: The use of speech-based recognition technologies in human–computer interactions is increasing daily. Age and gender recognition, one of these technologies, is a popular research topic used directly or indirectly in many applications. In this research, a new age and gender recognition approach based on the ensemble of different machine learning algorithms is proposed. In the study, five different classifiers, namely KNN, SVM, LR, RF, and E-TREE, are used as base-level classifiers and the majority voting and stacking methods are used to create the ensemble models. First, using MFCC features, five base-level classifiers are created and the performance of each model is evaluated. Then, starting from the one with the highest performance, these classifiers are combined and ensemble models are created. In the study, eight different ensemble models are created and the performances of each are examined separately. The experiments conducted with the Turkish subsection of the Mozilla Common Voice dataset show that the ensemble models increase the recognition accuracy, and the highest accuracy of 97.41% is achieved with the ensemble model created by stacking five classifiers (SVM, E-TREE, RF, KNN, and LR). According to this result, the proposed ensemble model achieves superior accuracy compared to similar studies in recognizing age and gender from speech signals.

Keywords: ensemble learning; hyperparameter tuning; majority voting; speaker age and gender recognition; stacking



Citation: Yücesoy, E. Automatic Age and Gender Recognition Using Ensemble Learning. *Appl. Sci.* **2024**, *14*, 6868. <https://doi.org/10.3390/app14166868>

Academic Editors: Douglas O'Shaughnessy and Jie Zhang

Received: 21 May 2024

Revised: 28 July 2024

Accepted: 30 July 2024

Published: 6 August 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Speech production is a complex process that begins in the brain and requires the coordination of the lungs, larynx, vocal cords, tongue, lips, mouth, and facial muscles. It depends on many anatomical features such as the dimensions of the mouth, pharynx, and nasal cavity; the shape and size of the tongue and lips; the position of the teeth; and the elasticity and density of the tissues [1]. These features vary from person to person, and accordingly, the speech produced by each person is different. In this context, a speech signal contains verbal content as well as information about the speaker, such as his/her identity, language, age, gender, accent, and emotional state [2]. Human beings can easily extract this information from speech signals and use it frequently in their communication with each other. The information outside the verbal content of the speech is called paralinguistic information, and this information is extremely important for human–computer interactions. Recently, paralinguistic information extracted from speech has begun to be used in many areas such as banking systems, call centers, intelligent voice assistants, security, and advertising [3,4]. Age and gender are among the most important pieces of paralinguistic information, and this information can be used for different purposes such as determining the appropriate customer representative; selecting music, advertising, or different content; or defining authorizations to access certain resources [5–7]. In addition, age and gender recognition can be used as preprocessing in other speech-based systems such as speaker, speech, or emotion recognition to develop age- and gender-specific models and their performance can be increased by using these models [6].

There are many studies on recognizing the age and gender classes of the speaker from the speech signal. In some of these studies, the age and gender information of the speakers

is considered together, while in others, this information is evaluated separately [7–13]. In gender recognition, it is aimed to determine the speakers as male and female and in age recognition, it is aimed to determine the speaker's exact age in years or the age group such as young, adult, or elderly. In the studies where age and gender are examined together, speakers are classified according to definitions that indicate both their age and gender groups, such as young females. As for age groups, class definitions such as young, adult, and elderly or with age ranges in the tens such as the twenties, thirties, and forties are generally used. On the other hand, some studies specifically focus on child speakers [14,15]. In this research, age and gender classifications were considered together and a classification with a total of twelve classes was carried out, defined by the combination of two classes (male and female) according to gender and seven classes (20s, 30s, 40s, 50s, 60s, 70s, and 80s) according to age. This grouping is more inclusive than the grouping of young, adult, and elderly and is therefore a more challenging task.

The task of recognizing the age and gender of the speaker from the speech signal consists of two critical stages, namely feature extraction and classification. The first of these steps, feature extraction, is the process of extracting measurement values representing the speech and the speaker from the speech signal, and the effectiveness of this process directly affects the recognition rate. There are various methods used in the literature to extract features from speech signals. Mel spectrogram, mel-frequency cepstral coefficients (MFCCs), linear prediction cepstrum coefficients (LPCCs), spectral features, chroma, fundamental frequency (F0), zero-crossing rate (ZCR), and harmonic features are some of the most commonly used speech features [16,17]. After the feature extraction stage, a machine learning model is trained with the determined feature vectors, and the model is evaluated with the test dataset. Artificial neural networks, support vector machines, k-nearest neighbors, convolutional neural networks, logistic regression, naive Bayes, random forest, recurrent neural networks, decision tree, and long short-term memory are some of the widely used machine learning algorithms [10,11,18]. Recently, two approaches have become dominant in machine learning: deep learning and ensemble learning [19]. Deep learning is a branch of machine learning based on a neural network structure inspired by biological models of computation and cognition in the human brain. Deep learning models can predict very complex non-linear relationships by automatically learning distinctive features from raw input data and can solve the most challenging problems [20]. However, they require large amounts of data and computational resources. Also, complex deep-learning models face the risk of overfitting. On the other hand, the ensemble learning approach is based on the idea of creating a stronger model by combining multiple individual models [19]. Additionally, the diversity provided by the base-level models that form the ensemble model also helps reduce the risk of overfitting. Ensemble learning has been successfully applied in various fields, and superior performances have been achieved compared to single models [21]. This research aims to develop a new model based on ensemble learning to recognize age and gender from the speech signal. For this purpose, eight different ensemble models were created by combining five base-level classifiers in different orders, and the optimum ensemble model was determined by evaluating the performances of each. The results obtained in the experiments reveal that the ensemble model proposed in this research provides superior age and gender recognition accuracy compared to the existing methods in the literature.

The main contributions of the study can be summarized as follows:

- This is the first study, to our knowledge, that uses ensemble learning for classification tasks while considering both the speaker's age and gender.
- The classification in the study is based on a more detailed class definition, where speakers are categorized as male and female with 10-year age intervals.
- Even though the proposed ensemble model takes into account a higher number of classes, it demonstrates superior performance compared to state-of-the-art methods.

2. Related Works

Goyal et al. [8] discussed predicting the age and gender classes of the speaker from the speech signal using multiLayer perceptron (MLP) architecture. In the study, features such as MFCCs, pitch, formants, and chroma were extracted from each speech signal and then selected using the principal component analysis and redundant feature elimination techniques. In the tests performed on the Mozilla Speech database, the accuracy of the developed model was measured as 89.58%. In another study, Tursunov et al. [6] proposed a new end-to-end age and gender recognition model based on CNN with multiple attention module (MAM). In the study, the multiple attention module was used to extract salient features from the input. The proposed model was tested with Common Voice and a local Korean speech dataset. In these tests, the proposed model's gender, age, and age-gender classification accuracies on the Common Voice dataset were 96%, 73%, and 76%, while its accuracies on the Korean speech dataset were 97%, 97%, and 90%, respectively. In their study, Kwasny and Hemmerling [7] applied various deep neural network-based embedding architectures such as x-vector and d-vector for the task of age estimation and gender classification. The system with the best performance reached a mean absolute error (MSE) rate of 5.12 for males and 5.29 for females in age estimation and 99.60% accuracy in gender recognition. In another study, Sánchez-Hevia et al. [10] analyzed the performance of different types of deep neural networks to jointly determine the age and gender classes of the speaker from the speech signal. The study examined convolutional neural networks and the networks that use historical information such as recurrent convolutional neural networks and temporal convolutional networks. In experiments conducted with the Mozilla Common Voice dataset, error rates of below 2% in gender classification and below 20% in age classification were achieved. In [13], different prediction models were created to predict the age, gender, and emotion classes from audio clips, and the test accuracies of each model were compared. In the study, the highest gender prediction accuracy was measured as 96.4% with the CatBoost model, and the highest age prediction accuracy was measured as 70.4% with the random forest model. In another study, Almomani et al. [4] proposed a novel system to classify speech based on the gender, age, and accent of the speakers using machine learning based on an adaptive backpropagation and bagging algorithm. In the experiments, the gender classification accuracy of the adaptive backpropagation algorithm was 98% and the accuracy of the bagging algorithm was 98.10%, while the highest accuracy in age classification was 55.39% with the bagging algorithm. In the study of Kone et al. [22], two different models were proposed for age and gender predictions. While a sequential model with five hidden layers was used for gender prediction, PCA and logistic regression were used as a pipeline for age prediction. In experiments conducted on the Common Voice dataset, an accuracy of around 91% for gender and 59% for age was achieved. In another study, Haluška et al. [12] proposed a CNN model for age and gender recognition from speech signals. In the study, variability was added to the training set by masking at random frequencies on the spectrograms, and thus the generalizability of the model was improved. In the study, the gender classification accuracy of the proposed model was reported as 94.99% and the age classification accuracy was reported as 75.24%.

3. Dataset and Methods

3.1. Dataset

The Mozilla Common Voice dataset [23] was used to evaluate the models proposed in this research. It is a multilingual speech corpus consisting of audio recordings of texts taken from public domain sources and voiced by volunteers. An ordinary Internet user can contribute to this corpus as a speaker or a listener after the registration stage. This dataset is collected for speech recognition purposes, but since the age, gender, language, and accent information of the speakers are available in the dataset, it is also used for different purposes. The dataset consists of three subsections. Audio clips that are listened to by at least two listeners and whose audio-text match is approved by the majority vote of the listeners are in the "valid" subsection, while those that are rejected are in the "invalid" subsection.

Audio clips with fewer than two votes or equal valid and invalid votes are in the “other” subsection. All audio clips are stored as mp3 files, and the information about each recording is available in a csv file. In the dataset, speakers are divided into nine groups according to their age. These age groups and age ranges are as follows: teens: “<19”, twenties: “19–29”, thirties: “30–39”, forties: “40–49”, fifties: “50–59”, sixties: “60–69”, seventies: “70–79”, eighties: “80–89”, and nineties: “>89”.

This study utilized the Turkish subset of the Common Voice dataset, version 18.0, which consists of 121 h of verified speech data. However, some recordings in the dataset do not include age or gender information and there are instances where the audio-text match received down-votes. Additionally, certain classes within the dataset have very few or no records, such as “teens”, “seventies male”, and “eighties male”. To address these issues, recordings with missing age or gender information, those down-voted, and classes with insufficient records were removed from the dataset. As a result, 12 classes remained in the dataset, comprising two gender classes and seven age classes. However, the records in the dataset are clustered into certain classes. For example, while there are 670 records in the twenties female class, there are 10,555 records in the twenties male class. To eliminate this imbalance, a balanced dataset containing 8040 conversations was created by randomly selecting an equal number of records from each age and gender classes, and this dataset was used in experimental studies. The class definitions in this dataset are given in Table 1.

Table 1. The class definitions used in the study.

Class ID	Class Label	Class Definition	Age Range	Gender
#1	F20s	Twenties female	20–29	Female
#2	M20s	Twenties male	20–29	Male
#3	F30s	Thirties female	30–39	Female
#4	M30s	Thirties male	30–39	Male
#5	F40s	Forties female	40–49	Female
#6	M40s	Forties male	40–49	Male
#7	F50s	Fifties female	50–59	Female
#8	M50s	Fifties male	50–59	Male
#9	F60s	Sixties female	60–69	Female
#10	M60s	Sixties male	60–69	Male
#11	F70s	Seventies female	70–79	Female
#12	F80s	Eighties female	80–89	Female

3.2. Feature Extraction

Extracting relevant features from the speech signal is one of the most critical steps that affects the performance of recognition systems. As a result of this process, speech signals are converted from their natural waveforms into a more compact parametric representation [7]. Various features can be used to recognize the age and gender of a speaker from the speech signal. However, mel-frequency cepstral coefficients (MFCCs) are the most effective speech feature type and were also used in this study. The feature extraction process of MFCCs includes many steps. The first two steps are framing and windowing. In the framing, the signal is divided into short segments that are assumed to be stationary. At this stage, an overlap area is created between two frames to prevent information loss between adjacent frames. The typical window length is 20–30 ms, and the overlap rate is 50%. In the windowing step, a window function is applied to each frame to reduce the discontinuity at the edges of the frames. In speech processing studies, the Hamming or Hamming window function is generally used. In the next step, after calculating the magnitude spectrum of the signal by applying DFT to each frame, it was converted into a mel spectrum by passing it

through a filter bank based on the mel scale, which represents the frequency perceived by the human ear [24]. Finally, after taking the logarithm of the mel spectrogram, the inverse cosine transformation was applied to obtain MFCCs. Thus, an acoustic vector representing each speech signal was obtained. In this study, Python's "Librosa" package was used with the following parameters to calculate MFCC features: number of MFCCs: 40, sampling rate: 2050, window length: 2048, window function: Hanning, and hop length: 512.

3.3. Ensemble Learning

Ensemble learning is a machine learning approach in which multiple learners, called base learners, are trained and combined to solve the same problem. In this paradigm, each base learner is considered an expert, and the idea that more accurate and robust models can be obtained by correctly combining the individual judgments of different experts constitutes the main hypothesis of this paradigm [25]. In numerous experimental and theoretical studies, it has been shown that the accuracy of ensemble models is generally superior to that of a single model [26]. There are various ensemble techniques depending on the approaches used in training and combining the base learners [27]. In this article, two popular ensemble methods, majority voting, and stacking, were investigated to classify speakers by age and gender.

3.3.1. Majority Voting

In classification problems, the voting method is often used to combine multiple base-level classifiers. Majority voting and weighted voting are two widely used voting methods [27]. In majority voting, the class label of the unknown sample is predicted according to the class with the highest votes among the members forming the ensemble model. This technique is also known as plurality voting and does not require any parameter optimization after training the base-level classifiers. This method is expressed as follows:

$$\text{class}(x) = \arg \max \left(\sum_k f(y_k(x), c_i) \right) \quad (1)$$

where $y_k(x)$ is the prediction of the k 'th model and $f(y, c)$ is the indicator function with the following definition:

$$f(y, c) = \begin{cases} 1, & y = c \\ 0, & y \neq c \end{cases} \quad (2)$$

The weighted majority voting is a trainable version of majority voting. In this method, different weights are assigned to the predictions of the base-level classifiers in the ensemble according to the performance or reliability of each classifier. After multiplying the predictions of each classifier by the weight assigned to it, these values are summed to calculate the weighted votes for each class. The final prediction is made according to this value, and the class with the highest weighted vote is determined as the class of the input data. In weighted voting, the selection of the weights of the base-level classifiers is a critical issue. Various approaches based on artificial intelligence such as genetic algorithms, particle swarm optimization, and fuzzy sets have been proposed on this issue [28].

3.3.2. Stacking

Stacking is an ensemble learning approach in which the results of the base classifiers forming the ensemble are combined with a meta-classifier. This approach consists of two stages called Level 0 and Level 1, and its schematic diagram is as shown in Figure 1. In the first stage, a set of base-level classifiers is generated from the initial training dataset. Then, the predictions of these classifiers are stacked and the predictions are applied to the meta-classifier to produce the final predictions of the ensemble model. The training set of the meta-classifier is created according to the leave-one-out or cross-validation procedure. Assuming that there are L base-level classifiers and a training set consisting of m samples, in the leave-one-out procedure, one sample from the training set, s_i , is reserved for testing and the base-level classifiers are trained with the remaining $m - 1$ samples. Then, the

predictions of the base-level classifiers are generated for the sample, s_i . This process is repeated for all samples, and the training set of the meta-classifier is created by stacking the correct classes of the samples with the predictions of the L base-level classifiers. If the leave-one-out method is used to train base-level classifiers, L classifier must be trained m times. This process will be quite time-consuming for a large dataset. In this case, the k -fold cross-validation method was used instead of leave-one-out. The training dataset was divided into k parts; one of these parts was used for testing the classifier, and the other $k - 1$ part was used for training the classifier. Thus, the L base-level classifier was trained k times instead of m . Since k will be considerably smaller than m , the use of this method in training base-level classifiers will significantly reduce the computational cost. In this research, the five-fold CV approach was used to generate the training set of the meta-classifier. Any machine learning method can be used as a meta-classifier, but experimental studies have shown that the use of complex models causes overfitting problems [29]. For this reason, simple models such as linear regression are preferred as meta-classifiers [30].

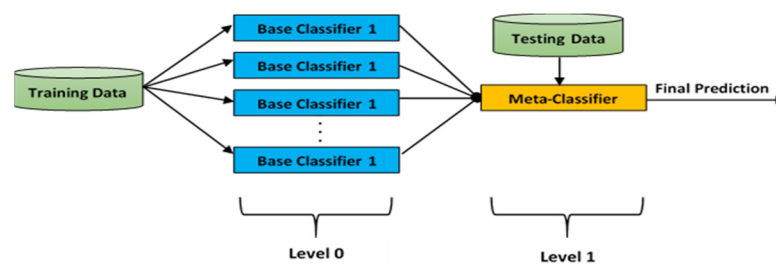


Figure 1. The schematic diagram of stacking ensemble learning.

3.3.3. Base-Level Classifiers

In classification based on ensemble learning, it is very important to determine the base-level classifiers that form the ensemble model. The basic idea in determining base-level classifiers is to provide rich diversity to the ensemble by using various types of machine learning algorithms. Therefore, it is recommended to include in the ensemble various models that make different assumptions in solving predictive modeling tasks, such as linear models, decision trees, support vector machines, and KNN. To provide the desired diversity in this study, a classifier pool consisting of KNN, SVM, RF, LR, and extra trees was selected as the base-level classifier. A brief description of these classifiers is given in the following subsection.

K-Nearest Neighbor (KNN)

KNN is one of the most commonly used supervised machine learning methods. The KNN method, which stands out with its ease of use and flexibility, has no training phase and is therefore also known as lazy learning [31]. The letter K in the name of the algorithm represents the number of neighbors included in the classification process and is one of the most important parameters affecting classification accuracy. In the first stage of the algorithm, to determine the class to which a test sample belongs, the distances between the test sample and all samples in the training dataset are calculated. Then, the K nearest neighbors are selected according to the calculated distances, and the class label of the test sample is determined according to the majority vote among the K nearest neighbors. There are different distance metrics such as Euclidean, Manhattan, and Minkowski used to determine the nearest neighbors. However, no value gives the best results in all datasets for both the distance metric and the number of nearest neighbors. For this reason, the optimum values of these parameters should be determined through a process called hyperparameter tuning. There are different methods used for hyperparameter tuning, such as manual tuning, random search, and grid search [32]. Of these, the grid search requires less experience and computational load and is therefore more popular. In the grid search, the aim is to determine the values that give the best results by scanning

hyperparameter values within a certain range. The grid search method was used for the parameter optimization of all classifiers developed in this study.

Support Vector Machine (SVM)

SVM is one of the most powerful supervised machine learning methods used in both classification and regression problems. However, it is used more frequently for classification purposes. SVMs are mainly developed for binary classification, but they can also be extended to multi-class classification tasks using various methods [33]. The purpose of the SVM algorithm is to find the best possible decision boundary that separates data points belonging to different classes into their classes. While this boundary is a line in a two-dimensional feature space, it is called a hyperplane in high-dimensional space. There may be multiple hyperplanes that separate data points into their classes. The best of these is the hyperplane with the maximum margin; that is, the distance between classes is maximum, and SVM tries to determine it. The SVM algorithm was originally proposed to create linear classifiers [34]. However, with the use of an approach called the kernel trick, SVMs have been enabled to perform non-linear classification effectively. The kernel trick is a simple method based on projecting data that cannot be classified as linear into a higher dimensional space where they can be classified as linear [35]. There are various kernel functions used for this projection, such as polynomial kernel, Gaussian kernel, and sigmoid kernel. The kernel function is highly dependent on the dataset and the task, so it is usually determined as a result of an optimization process. In this research, the grid search was used to determine the kernel function.

Logistic Regression (LR)

LR is a supervised machine learning algorithm mainly used for classification tasks. LR aims to estimate the probability that a test sample belongs to a particular class. To do this, first a linear combination of the features of the samples is taken and then a non-linear sigmoidal function is applied to them [36]. Logistic regression and linear regression can sometimes be confused since they both have the term regression in their names. However, they are quite different from each other. Linear regression is a regression algorithm, and the output is continuous values. On the other hand, logistic regression is a classification algorithm, and its output is the probability that a sample belongs to a particular class. LR was originally developed for binary classification, but it can be straightforwardly extended to multiclass classification (it is called multinomial logistic regression).

Random Forest (RF)

RF is a decision tree-based ensemble learning approach widely used in both classification and regression problems. In simple terms, random forest can be described as a large group of decision trees (DTs) that work together to produce the best possible conclusions. Each of the individual trees in the RF is constructed using a random subset of the training dataset. This subset is created by randomly selecting samples from the training data with replacement, and this process is called bootstrapping. A similar randomness in the selection of bootstrap samples is also used in feature selection. At each node, a randomly selected subset of features is used instead of all features. These two randomness in the training phase help the RF model make better predictions by reducing the correlation between individual decision trees [31]. To classify a new sample, the input vector of the sample is applied to each DT in the forest, and each DT produces an output by evaluating a different part of the input vector. These results are then combined according to majority voting to determine the final prediction. Various parameters affect the accuracy of the RF classifier. Of these parameters, the number of decision trees in the forest and the number of features considered for splitting at each leaf node are the most critical [37]. In this research, these two parameters were optimized with the grid search method.

Extra Tree (E-TREE)

E-TREE [38] is a modified version of the random forest classifier. Like random forest, it builds a large number of decision trees and splits nodes using a random subset of features. The main difference between E-TREE and RF is the changes in the strategies used to ensure variability of the samples and covariates. The first difference arises in the selection of the training set of each decision tree. While random forest uses randomly selected bootstrap samples, extra trees uses the entire training set. Another difference between the two methods is related to the selection of features used in each decision tree. In RF, a random subset of features is selected for each tree, and the best feature for each node split is selected based on some mathematical criterion (typically the Gini index). On the other hand, E-TREE is more aggressive and randomly selects a threshold value for each feature to split the node. Thus, E-TREE becomes more random than RF because it eliminates the bias caused by the selection of the best feature.

3.3.4. Hyperparameter Optimization

ML algorithms have a set of hyperparameters that control the learning process. The process of determining the optimum values of these parameters is known as hyperparameter optimization and is one of the most important steps affecting model performance. There are various methods for hyperparameter optimization [39]. In this research, grid search cross-validation (GridSearchCV) was used in parameter optimization of base-level classifiers. GridSearchCV is one of the most commonly used methods for optimizing the hyperparameters of a machine learning model. In this method, grid search and cross-validation (CV) techniques were used together to find the best parameter for the model. The grid search method uses a brute-force search technique and systematically explores all possible combinations of hyperparameter values in a given range. CV is a data-splitting method used to evaluate the generalization performance of the model. In CV, the available dataset is divided into multiple folds, and one of these folds is used for validation, while the other folds are used for training. The process is repeated each time using a different fold for the validation set. Finally, the performance of the model is calculated by averaging the results from each validation step. In the GridSearchCV method, the CV step is repeated for all parameter combinations examined by the grid search, and the optimum parameters are determined according to the classification accuracies. In this study, a ten-fold CV was used for hyperparameter optimization. The investigated parameter ranges and determined optimum values for each classifier are given in Table 2.

Table 2. Hyperparameters for the base-level classifiers.

Classifier	Range of Hyperparameters	Selected Values
SVM	C: [1, 10, 100, 1000] gamma: [0.1, 0.01, 0.001, 0.0001] kernel: ['linear', 'rbf', 'poly']	1000 0.0001 rbf
KNN	n_neighbors: [1, 2, ..., 26] metric: ['Euclidean', 'Manhattan', 'Minkowski']	1 Manhattan
RF	max_features: ['sqrt', 'log2'] n_estimators: [100, 300, 500, 700]	Log2 700
E-TREE	max_features: [3, 5, 7, 9, 11] n_estimators: [100, 200, ..., 1000]	9 900
LR	C: [0.001, 0.01, 0.1, 1, 10, 100, 1000] Penalty: ['l1', 'l2'] solver: ['lbfgs', 'liblinear', 'newton-cg']	0.01 l2 newton-cg

4. Experimental Evaluation

4.1. Experimental Setup

The models proposed in this study were implemented in Python using machine learning libraries such as pandas, numpy, librosa, and sklearn. Preprocessing steps, such as dataset preparation, splitting it into training and test sets, and feature extraction, were carried out on a Windows 10 Laptop PC with the following specifications: Intel Core i5 4th generation, 8 GB RAM, and Intel HD Graphics 4600 graphics card. On the other hand, the training and evaluation of the models were performed in the Jupyter Notebook on the Google Colab platform. In the development of all models, 8040 randomly selected speech recordings from the Turkish subsection of the MCV dataset were used. This dataset was divided into training and test parts in a ratio of 75:25. First, hyperparameter optimization was performed with a ten-fold CV using the training dataset. Then, the models were trained using the same dataset with the determined optimum parameters. Finally, the performance of the models on the test dataset was evaluated.

4.2. Performance Measures

In the study, the following metrics were used to evaluate the performance of the proposed models: accuracy, precision, recall, and F1 score. These metrics were calculated based on the relationship between the true class of the data and the model's prediction. There were four different situations between the true and predicted class. These were *TP* (true-positive), *TN* (true-negative), *FP* (false-positive), and *FN* (false-negative). *TP* represents positive samples that were correctly predicted as positive, while *TN* represents negative samples that were correctly predicted as negative. On the other hand, *FP* represents negative samples that were incorrectly predicted as positive, while *FN* represents positive samples that were incorrectly predicted as negative. With these definitions, accuracy, precision, recall, and *F1 score* metrics were defined by the following equations:

$$Accuracy = \frac{TN + TP}{TP + FP + FN + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

4.3. Experimental Results

In this section, the experimental results of five base-level classifiers and eight ensemble models created with these classifiers on the Turkish subsection of the MCV dataset are presented. First, the experimental results of the base-level classifiers and then the ensemble models are given separately.

4.3.1. Performances of Base-Level Classifiers

In this study, five different classifiers, namely KNN, SVM, RF, E-TREE, and LR, were examined as base-level classifiers. First, parameter optimization was performed for each of these classifiers with the cross-validation method. Then, the models were created with the determined optimum parameters and training of each was carried out. Finally, the process was completed by testing the models with a dataset that was not used in the training phase. The same dataset was used in the training and testing phases of all models. This dataset was selected from the Turkish subsection of the MCV dataset according to the criteria specified in Section 3.1, and 75% of the dataset was used for model training and the remaining 25% for model testing. The accuracy, precision, recall, and F1 scores obtained in

these experiments are given in Table 3. The best results for each metric are highlighted in bold in the table.

Table 3. Age and gender classification performances of base-level classifiers.

Base-Level Classifier	Accuracy	Precision	Recall	F1 Score
KNN	91.94	92.17	91.94	91.96
SVM	96.37	96.39	96.37	96.35
RF	95.57	95.71	95.57	95.54
E-TREE	96.02	96.22	96.02	96.00
LR	91.59	91.57	91.59	91.53

As seen in Table 3, SVM is the best classifier in terms of all evaluation metrics among the base-level classifiers. The SVM classifier reached 96.37% accuracy by correctly classifying 1937 of 2010 speeches in the test dataset, followed by E-TREE with 96.02% accuracy, RF with 95.57% accuracy, and KNN classifiers with 91.94% accuracy. Among the base-level classifiers, the classifier with the lowest accuracy is LR with 91.59%. The confusion matrix of the SVM classifier is given in Table 4. The confusion matrix represents the classifier's predictions in detail and is widely used in performance evaluation. From Table 4, it can be seen that the M50s and F80s are the classes with the highest accuracy, with 100% and 98.80%, respectively. On the other hand, F20s and F50s are the classes classified with the lowest accuracy, with 86.34% and 94.35%, respectively.

Table 4. Confusion matrix of the SVM classifier.

	F80s	F50s	M50s	F40s	M40s	F70s	F60s	M60s	F30s	M30s	F20s	M20s
F80s	98.80	0	0	0	0.60	0	0	0	0	0.60	0	0
F50s	0	94.35	0.56	0	0	3.39	0	0	0	0	1.69	0
M50s	0	0	100	0	0	0	0	0	0	0	0	0
F40s	0	0	0	95.48	0	0.65	0	0.65	0	0.65	2.58	0
M40s	0	0	0.66	0.66	94.70	1.32	0	0.66	0	1.32	0	0.66
F70s	0	1.27	0	0	0	98.73	0	0	0	0	0	0
F60s	0.61	0	0	0.61	0	0	97.56	0	0	0	1.22	0
M60s	0.65	0	0	0	0.65	0	0.65	98.06	0	0	0	0
F30s	0	0	0	1.10	0	0	0	0	97.79	0	1.10	0
M30s	1.19	0	0	0	2.38	0	0	0	0	96.43	0	0
F20s	0	3.83	0	0.55	1.64	3.28	0	0.55	3.83	0	86.34	0
M20s	0	0	0	0	1.20	0	0	0	0	0	0	98.80

4.3.2. Performances of Ensemble Classifiers

The ensemble classifiers developed in this study are based on two approaches: majority voting and stacking. While the majority voting classifier combines the predictions of the base-level classifiers according to majority voting, the stacking classifier uses the predictions of the base-level classifiers to train the meta-classifier as shown in Figure 1. In this study, LR was used as a meta-classifier of the stacking ensembles. First, two ensemble models were created by stacking and voting the two base-level classifiers with the highest individual performance. Then, the ensemble models were expanded by increasing the number of the base-level classifiers to three, four, and five, respectively, and the performances of each were evaluated separately. In this process, the base-level classifiers were incorporated into the ensemble models according to their performance orders. The performance metrics of ensemble classifiers developed by majority voting are given in Table 5, and those of stacking are given in Table 6. In both tables, the best results for each metric are highlighted in bold.

Table 5. Performance metrics of the majority voting-based ensemble models.

Base-Level Classifiers	Accuracy	Precision	Recall	F1 Score
SVM + E-TREE	96.27	96.44	96.27	96.21
SVM + E-TREE + RF	96.27	96.43	96.27	96.24
SVM + E-TREE + RF + KNN	96.72	96.84	96.72	96.68
SVM + E-TREE + RF + KNN + LR	96.52	96.64	96.52	96.49

Table 6. Performance metrics of the stacking-based ensemble models.

Base-Level Classifiers	Accuracy	Precision	Recall	F1 Score
SVM + E-TREE	96.87	96.90	96.87	96.86
SVM + E-TREE + RF	97.16	97.20	97.16	97.17
SVM + E-TREE + RF + KNN	96.87	96.90	96.87	96.86
SVM + E-TREE + RF + KNN + LR	97.41	97.44	97.41	97.41

From Tables 5 and 6, it can be seen that stacking-based ensemble provides superior performance compared to majority voting in terms of all evaluated metrics. Among all the models developed in the study, the ensemble model, created by stacking five base-level classifiers (SVM, E-TREE, RF, KNN, and LR), stands out as the model with the highest accuracy. This model achieves 97.41% accuracy by correctly classifying 1958 of 2010 speeches in the test dataset, and its confusion matrix is given in Table 7. According to this matrix, F80s and M50s are the two classes classified with the highest accuracy with 100% accuracy. On the other hand, F20s with 92.35% and F50s with 92.66% are the two classes with the lowest classification accuracy. Among the models developed based on majority voting, the highest accuracy is 96.72%, and this result is obtained by evaluating the results of the SVM, E-TREE, RF, and KNN classifiers according to majority voting.

Table 7. Confusion matrix of the model generated by stacking E-TREE, KNN, SVM and RF.

	F80s	F50s	M50s	F40s	M40s	F70s	F60s	M60s	F30s	M30s	F20s	M20s
F80s	100	0	0	0	0	0	0	0	0	0	0	0
F50s	0	92.66	0	0	0	3.95	0	0	1.13	0	2.26	0
M50s	0	0	100	0	0	0	0	0	0	0	0	0
F40s	0	0	0	98.06	0	0.65	0	0.65	0	0	0.65	0
M40s	0	0	0	0.66	94.70	0.66	0	1.32	0	1.32	0	1.32
F70s	0	0.64	0	0	0	99.36	0	0	0	0	0	0
F60s	0	0	0	0.61	0	0	98.17	0	0	0	1.22	0
M60s	0.65	0	0	0	0.65	0	0	98.71	0	0	0	0
F30s	0	0	0	1.10	0	0	0	0	97.24	0	1.66	0
M30s	0.60	0	0	0	0.60	0	0	0	0	98.81	0	0
F20s	0.55	2.19	0	0.55	1.09	2.19	0	0	1.09	0	92.35	0
M20s	0	0	0	0	0	0	0	0.60	0	0	0	99.40

5. Discussion

This article used a classification approach based on ensemble learning to recognize the speaker's age and gender classes. First, five models that classify speakers according to their age and gender were created using five different machine learning methods, and the performance of each was analyzed. When the results in Table 3, where the performance measurements obtained for these models are given, are examined, SVM, E-TREE, and RF classifiers offer over 95% accuracy, while the accuracy of KNN and LDA classifiers remains around 91%. This difference may be due to the fact that KNN is a type of instance-based learning and does not create a general internal model, while LDA classifies based on a linear combination of features. Then, eight ensemble models were developed in which the results of these classifiers were combined with majority voting and stacking approaches, and the performance of these models was evaluated. In tests performed on the same dataset and in

the same experimental setup, it is seen that both ensemble approaches provide an increase in age and gender classification performance. Additionally, the experimental results reveal that models developed with the stacking approach provide superior results compared to those developed with majority voting. This shows that the stacking approach takes better advantage of the complementary aspects of the classifiers that make up the ensemble. In the study, the ensemble model created by stacking five base-level classifiers is the model with the highest accuracy with an accuracy rate of 97.41%.

The study presents a comparison of the performance of the model based on the ensemble approach with state-of-the-art methods in Table 8. Upon reviewing the table, it becomes apparent that there are differences in the class definitions used in the various studies. Some studies classify speakers solely based on their gender, while others classify them based on both age and gender. The results given in the table indicate that gender classification accuracy is higher than age classification accuracy. This suggests that classifying age is a more challenging task compared to gender classification.

Table 8. Comparison of the proposed model with state-of-the-art methods.

Ref.	Dataset	Classifier	Number of Classes {Class Labels}	Accuracy
[4]	Voice gender and Common Voice	Backpropagation and bagging	2 {male, female} 4 {20s, 40s, 50s, 60s}	98.10% 55.39%
[6]	Korean Speech Recognition	CNN with MAM	2 {male, female} 6 {teens, 20s, 30s, 40s, 50s, 60s} 12 {gender x age categories}	97% 97% 90%
[7]	TIMIT dataset	Deep Neural Networks	2 {male, female}	96.8% to 99.6%
[8]	Common Voice	MLP	10 {male, female, teens, 20s, 30s, 40s, 50s, 60s, 70s, 80s}	89.58%
[10]	Common Voice	Deep Neural Networks	2 {male, female} 6 {YM, YF, AM, AF, SM, SF}	above 98% above 80%
[12]	Common Voice and Samromur	CNN	2 {male, female} 3 {0–15, 20–39, 40+}	94.99% 75.25%
[13]	Common Voice	ML algorithms	2 {male, female} 3 {young, matured, old}	96.4% 70.4%
[22]	Common Voice	RobustScalar, PCA, and LR	2 {male, female} 8 {teens, 20s, 30s, 40s, 50s, 60s, 70s, 80s}	91% 59%
[40]	Common Voice	1D and 2D CNN	10 {M20s, F20s, M30s, F30s, M40s, F40s, M50s, F50s, M60s, F60s}	94.40%
This study	Common Voice	Ensemble model	12 {M20s, F20s, M30s, F30s, M40s, F40s, M50s, F50s, M60s, F60s, F70s, F80s}	97.41%

Among the compared studies, there are very few studies in which age and gender classifications are made together. In one of these studies [6], a CNN model with a multiple attention module (MAM) is proposed to classify speakers according to their age and gender. In that study, it is stated that with the proposed model, the speeches in the Korean Speech Recognition dataset were classified into 12 age and gender classes with 90% accuracy. In another study [10], a six-class classification is made according to age and gender, and in that study, it is reported that over 80% accuracy was achieved on the Common Voice dataset with deep neural networks of different types. The class definition used in my previous study [40] is very similar to the class definition in this study. In that study, a 10-class classification defined according to the speaker's age and gender is performed based on 1D and 2D CNN models. The accuracy of the proposed 2D CNN model on the Turkish subsection of the Common Voice dataset is measured as 94.40%. On the other hand, in this study, the number of age and gender classes, which was 10 in the previous study, is increased to 12 with the addition of the seventies female (F70s) and eighties female (F80s)

classes. In addition, the dataset used in training the models has also been expanded. In this study, although the number of classes has increased, there is no decrease in classification accuracy and even a 3% increase is achieved. This increase can be considered the result of three factors. The first of these factors may be the use of complementary aspects of different classifiers in the ensemble learning approach. The second factor may be the increase in the amount of data used in the training and testing phases of the models. And the last reason for the increase in accuracy may be the new classes added to the dataset (seventies female and eighties female). In the experiments, it is seen that these two classes are the classes classified with the highest accuracy. The high individual classification accuracy of these classes must have increased the accuracy of the entire model by a certain amount.

6. Limitations

This study has certain limitations, which are primarily related to the dataset used. The dataset only consists of Turkish speech data that have been evaluated as valid by volunteer listeners. Additionally, any speech data in the dataset that lacked age or gender information were excluded from the study. In order to ensure the generalizability of the study's results, it is essential to utilize more comprehensive datasets that include data recorded in various languages and recording environments.

7. Conclusions and Future Works

In this study, an ensemble learning based approach is proposed for age and gender classifications using MFCC features extracted from speech signals. For this purpose, five different machine learning algorithms, namely SVM, E-TREE, KNN, RF, and LR, and two different ensembling methods, namely stacking and majority voting, were examined. In the study, a total of 13 models were created, 5 with base-level classifiers and 8 with the ensemble of these classifiers, and performance measurements of each were performed. All classifiers in the study were developed with a dataset consisting of randomly selected speeches from the Turkish subsection of the MCV dataset. In the experiments, it was observed that SVM provides the highest age and gender classification accuracy among the five base-level classifiers. Following SVM, E-TREE shows the next highest accuracy, followed by the RF, KNN, and LR classifiers. After the individual performance evaluations of the base-level classifiers were performed, the performance evaluations of the ensemble models created by majority voting and stacking of these classifiers were carried out. The experimental results obtained in the study show that the stacking-based ensemble models provide superior performance than those based on majority voting. The model created by stacking the predictions of five base-level classifiers with the LR meta-classifier stands out as the model with the highest accuracy. The accuracy of this model on the test dataset is 97.41%, while the highest accuracy achieved by majority voting is 96.72%. The experimental results show that classification models created with both ensemble approaches provide a significant performance increase in age and gender classifications.

There are various potential areas for future research to explore. One main area is to assess the proposed models using different datasets, including child speech, and to experiment with various modern deep learning architectures like transformer architectures. Furthermore, broadening the diversity of the ensemble by incorporating different deep-learning architectures could be another avenue for future research. Lastly, for age and gender recognition, exploring multi-modal models utilizing features extracted from various data sources, such as facial images, body movements, and speech utterances, could be another direction for future work.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Common Voice Corpus 13.0 Turkish database is available via <https://commonvoice.mozilla.org/tr/datasets> (accessed on 15 July 2024).

Conflicts of Interest: The author declares no conflicts of interest.

References

- Mathur, S.; Vyas, J. Acoustic analysis for comparison and identification of normal and disguised speech of individuals. *J. Forensic Sci. Criminol.* **2016**, *4*, 403. [\[CrossRef\]](#)
- Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A.; Neffat, O.S. Speech emotion recognition through hybrid features and Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 4750. [\[CrossRef\]](#)
- Shchetinin, E.Y.; Sevastianov, L. Improving the Learning Power of Artificial Intelligence Using Multimodal Deep Learning. *EPJ Web Conf.* **2021**, *248*, 01017. [\[CrossRef\]](#)
- Almomani, A.; Alweshah, M.; Alomoush, W.; Alauthman, M.; Jabai, A.; Abbass, A.; Gupta, B.B. Age and Gender Classification Using Backpropagation and Bagging Algorithms. *Computers. Mater. Contin.* **2023**, *74*, 3045–3062. [\[CrossRef\]](#)
- Přibíl, J.; Přibílová, A.; Matoušek, J. GMM-based speaker age and gender classification in Czech and Slovak. *J. Electr. Eng.* **2017**, *68*, 3–12. [\[CrossRef\]](#)
- Tursunov, A.; Khan, M.; 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* **2021**, *21*, 5892. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kwasny, D.; Hemmerling, D. Gender and age estimation methods based on speech using deep neural networks. *Sensors* **2021**, *21*, 4785. [\[CrossRef\]](#) [\[PubMed\]](#)
- Goyal, S.; Patage, V.V.; Tiwari, S. Gender and age group predictions from speech features using multi-layer perceptron model. In Proceedings of the 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 10–13 December 2020. [\[CrossRef\]](#)
- Kalluri, S.B.; Vijayaseenan, D.; Ganapathy, S. Automatic speaker profiling from short duration speech data. *Speech Commun.* **2020**, *121*, 16–28. [\[CrossRef\]](#)
- 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. *Multimed. Tools Appl.* **2022**, *81*, 3535–3552. [\[CrossRef\]](#)
- Hızlısoy, S.; Çolakoğlu, E.; Arslan, R.S. Speech-to-Gender Recognition Based on Machine Learning Algorithms. *Int. J. Appl. Math. Electron. Comput.* **2022**, *10*, 84–92. [\[CrossRef\]](#)
- Haluška, R.; Popovič, M.; Pleva, M.; Frohman, M. Detection of Gender and Age Category from Speech. In Proceedings of the 2023 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia, 21–22 September 2023. [\[CrossRef\]](#)
- Zaman, S.R.; Sadekeen, D.; Alfaz, M.A.; Shahriyar, R. One source to detect them all: Gender, age, and emotion detection from voice. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021. [\[CrossRef\]](#)
- Safavi, S.; Russell, M.; Jančovič, P. Automatic speaker, age-group and gender identification from children's speech. *Comput. Speech Lang.* **2018**, *50*, 141–156. [\[CrossRef\]](#)
- Kaya, H.; Salah, A.A.; Karpov, A.; Frolova, O.; Grigorev, A.; Lyakso, E. Emotion, age, and gender classification in children's speech by humans and machines. *Comput. Speech Lang.* **2017**, *46*, 268–283. [\[CrossRef\]](#)
- Byun, S.W.; Lee, S.P. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Appl. Sci.* **2021**, *11*, 1890. [\[CrossRef\]](#)
- Sharma, G.; Umapathy, K.; Krishnan, S. Trends in audio signal feature extraction methods. *Appl. Acoust.* **2020**, *158*, 107020. [\[CrossRef\]](#)
- Nitisara, G.R.; Suyanto, S.; Ramadhani, K.N. Speech age-gender classification using long short-term memory. In Proceedings of the 2020 3rd International Conference on Information and Communications Technology (ICOIAC), Yogyakarta, Indonesia, 24–25 November 2020. [\[CrossRef\]](#)
- Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud. Univ. -Comput. Inf. Sci.* **2023**, *35*, 757–774. [\[CrossRef\]](#)
- Kibrete, F.; Trzepieciński, T.; Gebremedhen, H.S.; Woldemichael, D.E. Artificial intelligence in predicting mechanical properties of composite materials. *J. Compos. Sci.* **2023**, *7*, 364. [\[CrossRef\]](#)
- Alotaibi, Y.; Ilyas, M. Ensemble-Learning Framework for Intrusion Detection to Enhance Internet of Things' Devices Security. *Sensors* **2023**, *23*, 5568. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kone, V.S.; Anagal, A.; Anegundi, S.; Jadhav, P.; Kulkarni, U.; Meena, S.M. Voice-based Gender and Age Recognition System. In Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 5–6 May 2023. [\[CrossRef\]](#)
- Mozilla Common Voice. Available online: <https://commonvoice.mozilla.org/tr/datasets> (accessed on 3 June 2022).
- Stevens, S.S.; Volkman, J.; Newman, E.B. A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.* **1937**, *8*, 185–190. [\[CrossRef\]](#)
- Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [\[CrossRef\]](#)

26. Brown, G. Ensemble Learning. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: Boston, MA, USA, 2011. [\[CrossRef\]](#)
27. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39. [\[CrossRef\]](#)
28. Dogan, A.; Birant, D. A weighted majority voting ensemble approach for classification. In Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, 11–15 September 2019. [\[CrossRef\]](#)
29. Li, Y.; Chen, W. A comparative performance assessment of ensemble learning for credit scoring. *Mathematics* **2020**, *8*, 1756. [\[CrossRef\]](#)
30. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2005.
31. Aljero, M.K.A.; Dimililer, N. A novel stacked ensemble for hate speech recognition. *Appl. Sci.* **2021**, *11*, 11684. [\[CrossRef\]](#)
32. Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **2020**, *415*, 295–316. [\[CrossRef\]](#)
33. Platt, J.; Cristianini, N.; Shawe-Taylor, J. Large margin DAGs for multiclass classification. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 547–553.
34. Vapnik, V.N. Pattern recognition using generalized portrait method. *Autom. Remote Control* **1963**, *24*, 774–780.
35. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh PA, USA, 27–29 July 1992. [\[CrossRef\]](#)
36. Bartosik, A.; Whittingham, H. Evaluating safety and toxicity. In *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*; Academic Press: Cambridge, MA, USA, 2021; pp. 119–137. [\[CrossRef\]](#)
37. Ahmed, S.; Hossain, M.A.; Bhuiyan, M.M.I.; Ray, S.K. A comparative study of machine learning algorithms to predict road accident severity. In Proceedings of the 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), London, UK, 20–22 December 2021. [\[CrossRef\]](#)
38. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [\[CrossRef\]](#)
39. Elgeldawi, E.; Sayed, A.; Galal, A.R.; Zaki, A.M. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics* **2021**, *8*, 79. [\[CrossRef\]](#)
40. Yücesoy, E. Speaker age and gender recognition using 1D and 2D convolutional neural networks. *Neural Comput. Appl.* **2024**, *36*, 3065–3075. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.