Received 29 June 2023, accepted 19 July 2023, date of publication 25 July 2023, date of current version 1 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3298697

RESEARCH ARTICLE

Fuzzy-Based Ensemble Feature Selection for Automated Estimation of Speaker Height and Age Using Vocal Characteristics

UMNIAH HAMEED JAID[1,2] **AND ALIA KARIM ABDULHASSAN**[2]
[1]Department of Computer Science, College of Science, University of Baghdad, Baghdad 10071, Iraq
[2]Department of Computer Science, University of Technology, Baghdad 10066, Iraq

Corresponding author: Umniah Hameed Jaid (umniah.h@sc.uobaghdad.edu.iq)

ABSTRACT Estimating speaker attributes from vocal characteristics is an important research area with applications in forensic science, biometric identification, and human–computer interaction. The accurate estimation of these attributes requires the effective extraction of relevant audio features from the audio signal. This work proposes a new approach for automatic speaker height and age estimation using fuzzy-based ensemble feature selection with speech parameters. This approach derives the initial feature importance from different feature selection (FS) methods. The obtained feature importance values are then sorted into a matrix, which is converted to ranks and passed to the fuzzy c-means (FCM) algorithm to produce the final feature ranking and identify the most distinctive features for estimation. The proposed approach can combine the results of multiple feature importance methods into an ensemble approach or choose the best features based on the feature importance from a single method without requiring a predefined number of top features. Several experiments were performed to evaluate the proposed approach on acoustic features obtained using the OpenSMILE toolkit from the TIMIT dataset. The results show that the proposed approach can effectively select the most informative features, and it outperforms similar studies on the same dataset, with promising results of 5.4, 4.71 mean absolute error (MAE) in height estimation and 5.38, 5.24 MAE for age estimation for males and females, respectively.

INDEX TERMS Age estimation, ensemble feature selection, fuzzy C-Means, height estimation, speaker profiling.

I. INTRODUCTION

Speech is a fundamental mode of communication among humans, enabling the expression of thoughts, ideas, information, and emotions. However, speech signals can convey much more than just spoken words. By listening to individuals' voices, people can easily discern their gender, age, emotional state, and physical condition. Applications estimating a speaker's physical attributes from speech have broad implications in areas such as surveillance, forensics, healthcare, and human–robot interaction. Voice-based biometrics offer advantages in terms of non-intrusiveness, cost-effectiveness, ease of deployment, and user acceptance.

Automated speaker profiling (ASP) can significantly support healthcare and assisted living environments by identifying elderly users and offering them specialized assistance, such as fall detection, medication reminders, or physical therapy guidance. Furthermore, ASP can enhance commercial applications by estimating the age of customers and providing tailored product recommendations, promotional offers, or advertisements. In the forensic domain, ASP can be employed to gather information about criminal suspects from phone calls or other audio recordings. Additionally, ASP can contribute to developing assistive technologies, enabling personalized interactions for users with disabilities or specific communication needs.

Numerous studies have established a connection between an individual's physical attributes and their voice. One such

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

pioneering study carried out by Lass et al. [1] demonstrated a strong correlation between a person's voice and their physical size. Subsequent research indicated that listeners could estimate the relative size of speakers (i.e., height and weight) based on their voices [2], [3]. Moreover, a study employing magnetic resonance imaging (MRI) on 129 participants verified the correlations between a person's height, weight, and vocal tract length (VTL) [4]. Certain vocal characteristics, such as speech rate, can provide insights into a speaker's age; for instance, younger speakers tend to speak faster than older speakers [5], and the fundamental frequency (F0) generally declines with age, particularly among female speakers [6]. In addition, F0 plays a crucial role in gender detection because male speakers typically have lower-frequency voices than female speakers.

The research on extracting paralinguistic content has grown rapidly, with speaker-profiling tasks typically involving extracting key features from raw speech signals and applying machine learning models for predictions. Various features, such as Fundamental Frequency (F0),(Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and formants, have been used for speaker-profiling tasks, such as height estimation, age estimation, and gender detection. Some studies have combined spectral features with temporal and prosodic features, while others have employed statistical approaches using short-term features as supervectors. Deep learning methods have also been used for feature extraction and speaker representation.

However, finding an optimal feature set for representing multiple physical attributes remains a challenge. FS methods, such as CatBoost [7], Relief-based algorithms [8], and dimensionality reduction methods like PCA [9] and LDA [10], have been applied to different feature sets in the speaker-profiling literature. These methods aim to identify the most relevant features for various speaker-profiling tasks, including height estimation, age estimation, and gender detection, to improve the accuracy of classification models.

The effectiveness of machine learning (ML) algorithms is directly tied to the quality of features extracted from the data. ML algorithms may underperform when working with datasets containing numerous features, particularly when the feature set includes significant redundancy and irrelevance [11]. To tackle this issue, researchers employ feature selection (FS) methods to identify and select only the most relevant features representing the essential properties of speech data.

FS algorithms can be broadly categorized into three types: filter methods, wrapper methods, and embedded methods [12]. Filter methods select crucial features based on data properties but do so without considering estimators in the selection process. This approach is often used to preprocess data before applying a learning algorithm [13]. Wrapper methods, on the other hand, use a learning algorithm to assess feature importance. The selection of features is based on their ability to enhance the estimator's performance, and the process is repeated until an optimal subset of features has been identified. Wrapper methods require more computational resources than filters [12]. Embedded methods, as the name suggests, combine filter and wrapper methods. They employ both intrinsic data properties and a learning algorithm to identify key features. Embedded methods require more computational resources than filters but less than wrappers [14].

Although FS methods are effective, they also have limitations because no single feature selection method can guarantee the best feature subset for all datasets. To address these limitations, a new research direction has emerged, known as ensemble FS. This approach integrates multiple FS methods to improve the accuracy of the selected features [15].

The current work proposes a novel fuzzy-based ensemble feature selection method to leverage the capabilities of different types of FS methods and select the most discriminative features for speaker profiling, including age and height estimation, using the FCM algorithm. By employing fuzzy logic, different degrees of membership are assigned to each feature based on its rank, which allows for handling uncertainty in the FS process. This approach can enhance the overall performance of speaker-profiling models by selecting the most relevant features while reducing the impact of irrelevant or redundant features. Furthermore, the proposed fuzzy-based ensemble feature selection method determines the most important features for speaker profiling by applying fuzzy logic to the feature importance of individual FS methods. This approach can address the limitations of individual FS methods and improve the accuracy of speaker-profiling models. Although the proposed fuzzy-based ensemble feature selection method attempts to address some of the limitations of individual FS methods, it is not without its own limitations. One limitation is the need to try different combinations of FS methods in the ensemble to arrive at the best results, which can be time-consuming and computationally expensive.

The remainder of the current paper is structured as follows: The next section describes the methodology followed, including the dataset used and the main stages of the proposed method. Section III presents detailed experimental results and evaluation measures used, Section IV discusses the results, and finally, section V concludes the paper.

II. METHODOLOGY

The present work proposes an improved approach for predicting a speaker's height and age using voice signals, building on the concept of ensemble FS techniques [16]. These techniques combine the strengths of different FS methods to provide more robust and accurate results. The proposed method employs FCM clustering to integrate multiple feature-ranking methods, hence selecting the most relevant features for the task.

The proposed methodology comprises three stages. First, data are collected from the TIMIT dataset, followed by feature extraction and preprocessing to ensure the relevance and meaningfulness of the extracted features. Second, the proposed FS algorithm is applied to the extracted features, aiming to identify an optimal subset. Finally, height and age estimation is performed using support vector regression (SVR). A comprehensive illustration of the entire process is presented in Fig. 1. The subsequent sections provide a detailed explanation of each stage and outline the main components of the proposed method.

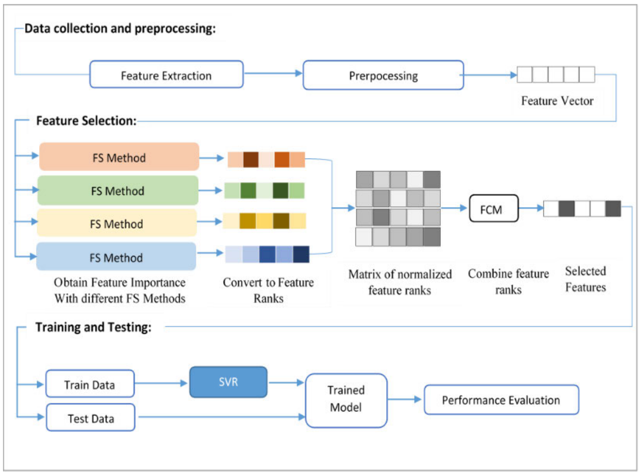


FIGURE 1. Illustration of the proposed method for ensemble FS.

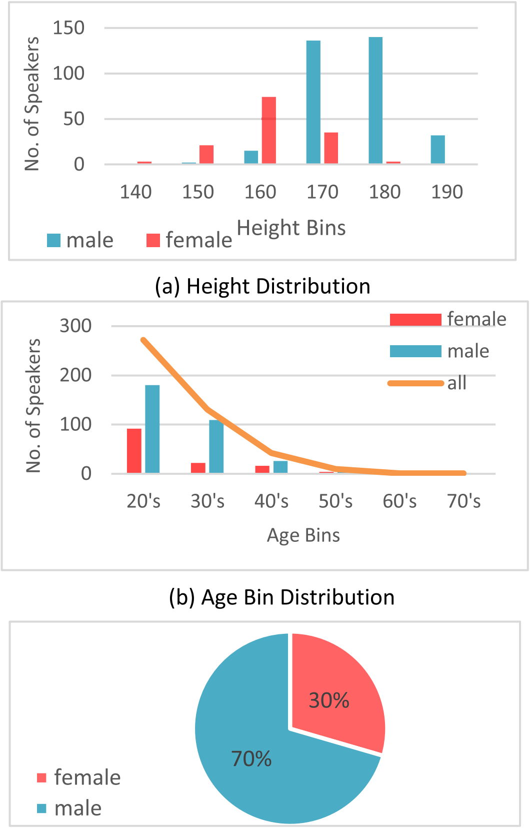
A. DATASET

For speaker profiling, the TIMIT dataset was employed [16], which is an automatic speech recognition (ASR) dataset that contains participant metadata, including height, age, education level, ethnicity, and regional dialect. Each participant contributed 10 recordings of speech transcripts, resulting in 6,300 utterances divided into non overlapping training and test sets. The training set consisted of 326 male speakers and 136 female speakers (i.e., a total of 462 speakers), and the test set consisted of 168 speakers (112 male and 56 female). For this work, the standard TIMIT train/test split was used.

Fig. 2a shows the height distributions of male and female speakers, Fig. 2b shows their age distributions, and Fig. 2c shows their gender distributions. The original split of the data caused imbalances among the three categories of height, age, and gender because the dataset was originally designed for speech recognition. These imbalances affected the performance and predicted the results of the model, as reported by several studies [17], [18]. However, to maintain comparability with other studies, the original train/test split was maintained, and the data in the training set was further split into 80% training and 20% validation sets.

B. FEATURE EXTRACTION

In the present study, the OpenSMILE toolkit was employed to extract a comprehensive set of spectral, prosodic, voice quality, and articulatory features. The (extended) Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set [28] was selected for feature extraction because it encompasses a broad range of acoustic and prosodic features, offers a consistent baseline for evaluation, and mitigates discrepancies because of varying parameter sets. A study by [28] proposed two versions of eGeMPAS: minimalistic and extended versions. The minimalistic set comprises 18 low-level descriptors (LLDs) classified into three categories:



(a) Height Distribution

(b) Age Bin Distribution

(c) Gender Distribution

FIGURE 2. The distributions for male and female speakers in the TIMIT dataset across different height and age bins.

Frequency-related parameters, such as F0, Jitter, formants 1, 2, and 3, frequency, and first-formant bandwidth.

Energy/amplitude-related parameters, including shimmer, loudness, and HNR.

Spectral parameters like alpha ratio, Hammarberg index, spectral slope, the relative energy of formants 1, 2, and 3, harmonic differences H1–H2, and harmonic differences H1–A3.

By calculating the arithmetic means and standard deviations (stddevs) for all 18 LLDs, 36 parameters were obtained. An additional eight functionals were also applied to loudness and pitch at the 20th, 50th, and 80th percentiles, as well as in the range of the 20th to 80th percentiles, together with the means and standard deviations (stddevs) of the slopes of rising and falling signal parts, resulting in 52 parameters. Moreover, the arithmetic means of the alpha ratios, Hammarberg indices, and spectral slopes produced 56 parameters. Finally, six temporal features were included: the rate of loudness peaks, the mean lengths and stddevs of continuously voiced regions where F0 was nonzero, the mean lengths and stddevs of unvoiced regions, and the number of continuous voiced regions per second, yielding a total of 88 acoustic features.

C. DATA PREPROCESSING

To reduce the variance in feature expressions and remove the learning model's bias toward specific values, the quantile normalization technique was employed for feature transformation, which forces different samples with distinct statistical distributions to conform to the same target distribution.

This technique operates by initially organizing all feature values in ascending order and subsequently assigning ranks to the sorted values. The mean of each rank is then calculated, guaranteeing that the resulting dataset maintains the same distribution of values across all features while preserving the original relationships between samples.

In addition to the quantile normalization technique applied to the feature values, a square transformation is utilized for preprocessing the target labels. The purpose of this transformation is to stabilize the variance and improve the normality of the distribution of the target labels, thereby enhancing the performance of the learning model.

The square transformation is applied by taking the square of each target label value. This transformation particularly benefits the models in cases where the target labels exhibit a skewed distribution. By applying the square transformation, the distribution of target labels becomes more symmetric and less skewed, which helps the learning model identify patterns and relationships in the data more efficiently. Consequently, this leads to more accurate and reliable predictions for speaker-profiling tasks, such as height and age estimation.

D. FEATURE SELECTION METHODS

The present work employed an ensemble FS process that incorporates various methods for measuring feature importance. These methods were categorized into three main groups: filter methods, wrapper methods, and embedded methods. Three filter methods, two wrapper methods, and one embedded method were used for a total of six methods to evaluate feature importance.

The filter methods assessed the relevance of features independently of any predictive model, often relying on statistical measures to determine their importance. The three filter methods used were Pearson's correlation feature importance (PCFI), Relief, and Fisher score.

PCFI is based on Pearson's correlation coefficient, a statistical measure quantifying the linear relationship between two variables. Features with high positive or negative correlations with the target variable are deemed more important for prediction because they contain more information about the target.

ReliefF is another filter method that extends the original Relief algorithm to handle incomplete and multi-class datasets. Relief evaluates the importance of a feature based on its ability to differentiate between instances that are close to each other in the feature space [22]. Higher Relief scores indicate more important features.

The Analysis of Variance (ANOVA) is a third filter method that aims to identify significant differences between groups while minimizing variations within each group [19]. ANOVA uses an F-value to achieve this. Higher F-values suggest features that are more effective in distinguishing between groups. The F-value is determined by the ratio of the between-group variability to the within-group variability. A high F-value indicates a feature that contributes significantly to the differentiation between groups.

In the estimation process of height and age, the target variables have been transformed and are considered as categorical inputs. This approach ensures a nuanced understanding of the data by treating these typically continuous variables - age and height - as distinct categories, instead of a continuum.

Wrapper methods evaluate feature importance based on the performance of a predictive model. The present study used two wrapper methods: single feature performance (SFP) and permutation feature importance (PFI). SFP measures feature importance by training a model with each feature individually and determining the model's performance score. The higher the performance of the model on a feature, the more indicative it is of the feature's discriminative power. PFI, as described in [20], measures feature importance by the decrease in a model performance score caused by randomly shuffling the values of a single feature. The drop in the model performance score reflects how much the model relies on that feature.

Finally, the present study has used an embedded method, random forest feature importance (RFFI) [21], which identifies the most important features for making predictions using an ensemble of decision trees. The feature importance is computed for each tree in the forest and averaged over all trees to obtain a more stable estimate of feature importance [22].

E. FCM

Fuzzy c-means (FCM) is a clustering algorithm that is an extension of the traditional K-means algorithm that allows data points to belong to multiple clusters to varying degrees. This is achieved by assigning each data point a membership value for each cluster that represents the degree to which the data point belongs to that cluster. These membership values are often referred to as "fuzzy" membership values. FCM has several advantages over traditional clustering, including greater flexibility in handling data that may belong to multiple clusters and the ability to model clusters that have overlapping boundaries.

Given a dataset of N feature and M feature ranks [x_1, $x_2, \ldots, x_N$], the FCM algorithm works as follows:

1. Initialize the algorithm with the number of clusters (c), where the clusters indicate selected and not selected features, and a fuzziness parameter (m).
2. Initialize an MxC fuzzy membership matrix $U = [u_{ij}]$ randomly where u_{ij} is the degree of membership of feature i in cluster j, subject to the constraints that

each element u_{ij} is between 0 and 1, with higher values indicating stronger membership, and that the sum of membership values for each feature is 1.

3. Initialize the centroid matrix V, which is a CxN matrix. The values in the centroid matrix represent the centroids of each cluster.
4. Calculate the cluster centroids as the weighted mean of the feature ranks, where the weights are the membership values u_{ij}. That is, for each cluster j, calculate the following:

$$v_j = \sum_{i=1}^{M} u_{ij} x_i \Big/ \sum_{i=1}^{M} u_{ij} \tag{1}$$

5. Update the membership values for each feature by calculating its degree of belongingness to each cluster based on its distance from each cluster center. The new membership value for feature i in cluster j is given by the following:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} d(x_i, v_k)/d(x_i, v_j)^{2/m} - 1} \tag{2}$$

where $d(x_i, v_k)$ is the Euclidean distance between data point i and cluster centroid vk, and m is a weighting exponent that controls the degree of fuzziness.

6. Repeat steps 4–5 until the membership values converge or a maximum number of iterations is reached.
7. Assign each feature to the cluster with the highest membership value.

III. EXPERIMENTS

For height and age estimation experiments, the TIMIT dataset is used with the standard train and test split. Because the TIMIT dataset contains 10 utterances for each speaker, the effect of prediction is evaluated at the speaker level rather than the utterance level, which can be done by aggregating the results of each speaker over their utterances and obtaining the mean value of the prediction. The evaluation metrics used are MAE and root mean squared error (RMSE), as follows:

$$MAE = \frac{\sum |y_i - x_i|}{n} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum (y_i - x_i)^2}{n}} \tag{4}$$

where y_i is the predicted value, x_i is the target value, and n is the number of observations.

For the FS algorithms, python's sklearn library [23] is used for RFFI, PCFI, and ANOVA. The feature_engine library described in [24] is used for PFI and SFP methods. For reproducibility, default parameters were used in all the algorithms, and a random seed of 42 was used in algorithms that require randomization.

A. PERFORMANCE OF INDIVIDUAL FS METHODS

The first set of experiments are done with individual FS methods. Tables 1 and 2 show the MAEs obtained with the Top 20, 50, and 75 features of each FS method for age and height estimation, respectively. Additionally, these results are compared with those obtained by applying fuzzy-ensemble to individual FS methods.

TABLE 1. Comparison of MAE in age estimation using different feature selection methods and proposed method.

	Pearson Correlation (PCFI)			ANOVA		
	All	Female	Male	All	Female	Male
Top 20	5.40	5.56	5.31	5.50	5.65	5.40
Top 50	5.34	5.32	5.35	5.56	5.59	5.53
Top 75	5.48	5.47	5.48	5.49	5.52	5.47
Proposed	5.32	5.23	5.36	5.49	5.73	5.37
	Random Forests Feature Importance (RFFI)			Single Feature Performance (SFP)		
	All	Female	Male	All	Female	Male
Top 20	5.48	5.84	5.30	5.41	5.61	5.30
Top 50	5.49	5.85	5.30	5.48	5.57	5.44
Top 75	5.47	5.81	5.30	5.45	5.60	5.37
Proposed	5.43	5.58	5.35	5.39	5.35	5.41
	Permutation Feature Importance (PFI)			Relief		
	All	Female	Male	All	Female	Male
Top 20	5.51	5.89	5.31	5.49	5.63	5.41
Top 50	5.43	5.75	5.27	5.36	5.52	5.28
Top 75	5.47	5.82	5.30	5.46	5.39	5.50
Proposed	5.44	5.64	5.34	5.42	5.53	5.36
No Feature selection				5. 50	5.49	5.51

TABLE 2. Comparison of MAE in height estimation using different feature selection methods and proposed method.

	Correlation			ANOVA		
	All	Female	Male	All	Female	Male
Top 20	5.28	5.03	5.41	5.25	4.96	5.4
Top 50	5.24	5.0	5.32	5.33	5.12	5.43
Top 75	5.24	5.18	5.27	5.34	5.2	5.41
Proposed	5.21	5.08	5.28	5.21	5.01	5.32
	Random Forests Feature Importance (RFFI)			Single Feature Performance (SFP)		
	All	Female	Male	All	Female	Male
Top 20	5.35	5.16	5.44	5.31	5.15	5.39
Top 50	5.3	5.26	5.31	5.28	5.08	5.38
Top 75	5.3	5.23	5.33	5.28	5.21	5.32
Proposed	5.26	5.04	5.37	5.2	4.91	5.34
	Permutation Feature Importance (PFI)			Relief		
	All	Female	Male	All	Female	Male
Top 20	5.33	5.14	5.42	5.31	5.27	5.33
Top 50	5.34	5.24	5.4	5.61	5.57	5.63
Top 75	5.31	5.27	5.33	5.32	5.33	5.32
Proposed	5.24	4.91	5.40	5.15	4.84	5.31
No Feature selection				5.28	5.25	5.3

Upon examining the overall MAE for age estimation, performance varies across different groups and FSs. Although some top N features enhance the outcomes, the improvement is not consistent. As for the proposed method, except for PFI and Relief, using the proposed method yields better results compared with those obtained without FS and compared with using the top N features. Although the improvements are quite small, the results demonstrate the ability of the proposed method to automatically select the number of features without setting a specific number while achieving comparable performance.

Similarly, for height estimation, the proposed approach consistently results in lower or very close MAE values for all gender groups (All, Female, Male) compared with the Top 20, Top 50, and Top 75 FSs. However, all the FS methods, including the proposed method, improved the performance of female height estimation while leading to worse performance for male height estimation compared with the estimation without feature selection. This indicates a bias toward the minority group of female speakers.

B. ENSEMBLE OF DIFFERENT COMBINATIONS OF FS METHODS

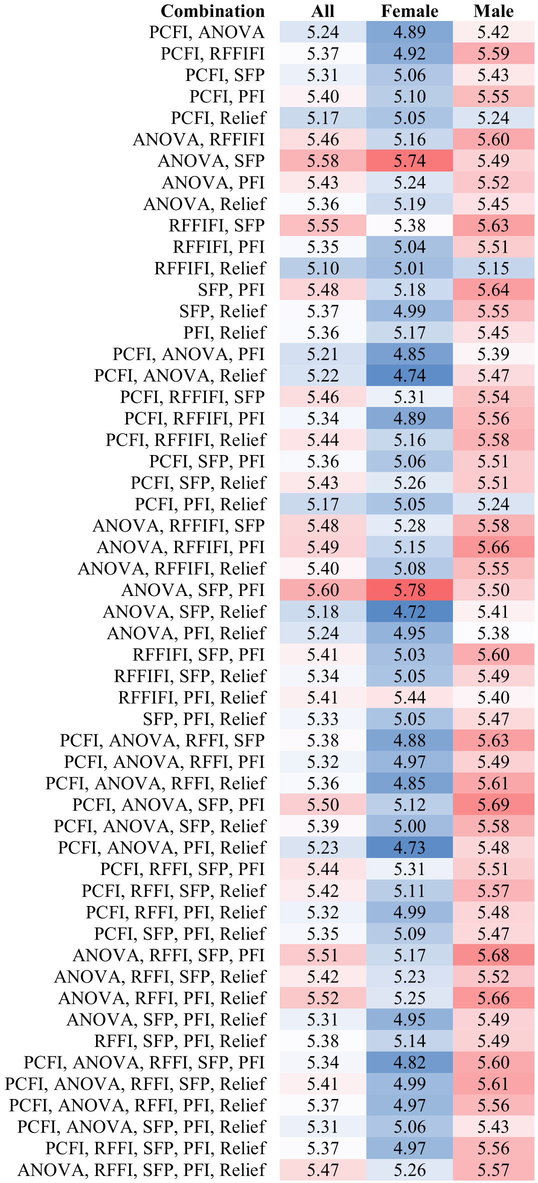
The second group of experiments explored combinations of two, three, four, five, and six FS methods for ensemble feature selection. There were 16 different combinations of two FS methods, 20 combinations of three FS methods, 15 combinations of four FS methods, and six combinations of five FS methods. The results obtained using these combinations with the proposed method for height and age estimation are illustrated in Fig. 3 and 4.

In the analysis of age estimation data, the combinations involving 'PCFI' and 'ANOVA' methods for males and 'PCFI' and 'PFI' methods for females emerged as optimal, yielding the lowest MAE. These methods also featured prominently in the top 10 combinations with the lowest MAE, underscoring their consistent and substantial impact on model performance. The unique presence of 'Relief' in the best-performing combinations suggests a potentially positive contribution to the accuracy of age estimation.

However, when these top-performing methods were paired with others without the mutual presence ('PCFI, RFFI, PFI' for females and 'ANOVA, RFFI, PFI' for males), they resulted in the highest MAE values. This indicates that the efficacy of these methods might not be as potent when used individually or in certain combinations, shedding light on the nuanced and possibly synergistic dynamics at play in age estimation performance.

In height estimation, the combination of PCFI and ANOVA methods demonstrated the lowest MAE. Furthermore, the ANOVA and Relief methods appeared frequently in the top 5 combinations with the lowest MAE, suggesting their significant and positive impact on height estimation. Specifically, the Relief method consistently improved the model's performance when combined with other methods but didn't feature in the worst combinations, indicating its robustness and reliability.

Conversely, the highest MAE values were seen when the Relief method was not included in the combinations, emphasizing its contribution to the model's performance. Similar to age estimation, these findings show the potential synergistic effects of combining these feature selection methods.



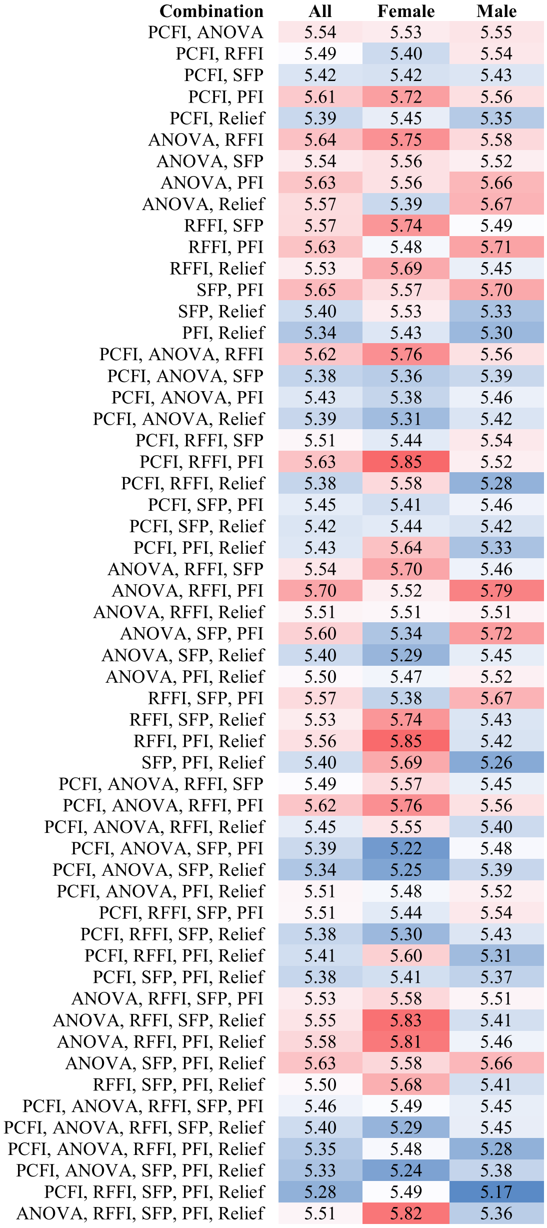
Combination	All	Female	Male
PCFI, ANOVA	5.24	4.89	5.42
PCFI, RFFIFI	5.37	4.92	5.59
PCFI, SFP	5.31	5.06	5.43
PCFI, PFI	5.40	5.10	5.55
PCFI, Relief	5.17	5.05	5.24
ANOVA, RFFIFI	5.46	5.16	5.60
ANOVA, SFP	5.58	5.74	5.49
ANOVA, PFI	5.43	5.24	5.52
ANOVA, Relief	5.36	5.19	5.45
RFFIFI, SFP	5.55	5.38	5.63
RFFIFI, PFI	5.35	5.04	5.51
RFFIFI, Relief	5.10	5.01	5.15
SFP, PFI	5.48	5.18	5.64
SFP, Relief	5.37	4.99	5.55
PFI, Relief	5.36	5.17	5.45
PCFI, ANOVA, PFI	5.21	4.85	5.39
PCFI, ANOVA, Relief	5.22	4.74	5.47
PCFI, RFFIFI, SFP	5.46	5.31	5.54
PCFI, RFFIFI, PFI	5.34	4.89	5.56
PCFI, RFFIFI, Relief	5.44	5.16	5.58
PCFI, SFP, PFI	5.36	5.06	5.51
PCFI, SFP, Relief	5.43	5.26	5.51
PCFI, PFI, Relief	5.17	5.05	5.24
ANOVA, RFFIFI, SFP	5.48	5.28	5.58
ANOVA, RFFIFI, PFI	5.49	5.15	5.66
ANOVA, RFFIFI, Relief	5.40	5.08	5.55
ANOVA, SFP, PFI	5.60	5.78	5.50
ANOVA, SFP, Relief	5.18	4.72	5.41
ANOVA, PFI, Relief	5.24	4.95	5.38
RFFIFI, SFP, PFI	5.41	5.03	5.60
RFFIFI, SFP, Relief	5.34	5.05	5.49
RFFIFI, PFI, Relief	5.41	5.44	5.40
SFP, PFI, Relief	5.33	5.05	5.47
PCFI, ANOVA, RFFI, SFP	5.38	4.88	5.63
PCFI, ANOVA, RFFI, PFI	5.32	4.97	5.49
PCFI, ANOVA, RFFI, Relief	5.36	4.85	5.61
PCFI, ANOVA, SFP, PFI	5.50	5.12	5.69
PCFI, ANOVA, SFP, Relief	5.39	5.00	5.58
PCFI, ANOVA, PFI, Relief	5.23	4.73	5.48
PCFI, RFFI, SFP, PFI	5.44	5.31	5.51
PCFI, RFFI, SFP, Relief	5.42	5.11	5.57
PCFI, RFFI, PFI, Relief	5.32	4.99	5.48
PCFI, SFP, PFI, Relief	5.35	5.09	5.47
ANOVA, RFFI, SFP, PFI	5.51	5.17	5.68
ANOVA, RFFI, SFP, Relief	5.42	5.23	5.52
ANOVA, RFFI, PFI, Relief	5.52	5.25	5.66
ANOVA, SFP, PFI, Relief	5.31	4.95	5.49
RFFI, SFP, PFI, Relief	5.38	5.14	5.49
PCFI, ANOVA, RFFI, SFP, PFI	5.34	4.82	5.60
PCFI, ANOVA, RFFI, SFP, Relief	5.41	4.99	5.61
PCFI, ANOVA, RFFI, PFI, Relief	5.37	4.97	5.56
PCFI, ANOVA, SFP, PFI, Relief	5.31	5.06	5.43
PCFI, RFFI, SFP, PFI, Relief	5.37	4.97	5.56
ANOVA, RFFI, SFP, PFI, Relief	5.47	5.26	5.57

FIGURE 3. Heatmap of MAE for height estimation performance using different combinations of feature selection methods.

It is important to note that these results may be specific to the dataset and the problem under investigation. Consequently, it is recommended that various combinations of FS methods be explored to identify the optimal solution for a given problem. However, these findings emphasize the benefits of method combinations in enhancing the model's performance rather than individual feature selection techniques.

IV. DISCUSSION

The primary objective of this study is to examine the efficacy of employing fuzzy-ensemble feature selection methods for improving the accuracy and efficiency of height and age prediction from speech data.



Combination	All	Female	Male
PCFI, ANOVA	5.54	5.53	5.55
PCFI, RFFI	5.49	5.40	5.54
PCFI, SFP	5.42	5.42	5.43
PCFI, PFI	5.61	5.72	5.56
PCFI, Relief	5.39	5.45	5.35
ANOVA, RFFI	5.64	5.75	5.58
ANOVA, SFP	5.54	5.56	5.52
ANOVA, PFI	5.63	5.56	5.66
ANOVA, Relief	5.57	5.39	5.67
RFFI, SFP	5.57	5.74	5.49
RFFI, PFI	5.63	5.48	5.71
RFFI, Relief	5.53	5.69	5.45
SFP, PFI	5.65	5.57	5.70
SFP, Relief	5.40	5.53	5.33
PFI, Relief	5.34	5.43	5.30
PCFI, ANOVA, RFFI	5.62	5.76	5.56
PCFI, ANOVA, SFP	5.38	5.36	5.39
PCFI, ANOVA, PFI	5.43	5.38	5.46
PCFI, ANOVA, Relief	5.39	5.31	5.42
PCFI, RFFI, SFP	5.51	5.44	5.54
PCFI, RFFI, PFI	5.63	5.85	5.52
PCFI, RFFI, Relief	5.38	5.58	5.28
PCFI, SFP, PFI	5.45	5.41	5.46
PCFI, SFP, Relief	5.42	5.44	5.42
PCFI, PFI, Relief	5.43	5.64	5.33
ANOVA, RFFI, SFP	5.54	5.70	5.46
ANOVA, RFFI, PFI	5.70	5.52	5.79
ANOVA, RFFI, Relief	5.51	5.51	5.51
ANOVA, SFP, PFI	5.60	5.34	5.72
ANOVA, SFP, Relief	5.40	5.29	5.45
ANOVA, PFI, Relief	5.50	5.47	5.52
RFFI, SFP, PFI	5.57	5.38	5.67
RFFI, SFP, Relief	5.53	5.74	5.43
RFFI, PFI, Relief	5.56	5.85	5.42
SFP, PFI, Relief	5.40	5.69	5.26
PCFI, ANOVA, RFFI, SFP	5.49	5.57	5.45
PCFI, ANOVA, RFFI, PFI	5.62	5.76	5.56
PCFI, ANOVA, RFFI, Relief	5.45	5.55	5.40
PCFI, ANOVA, SFP, PFI	5.39	5.22	5.48
PCFI, ANOVA, SFP, Relief	5.34	5.25	5.39
PCFI, ANOVA, PFI, Relief	5.51	5.48	5.52
PCFI, RFFI, SFP, PFI	5.51	5.44	5.54
PCFI, RFFI, SFP, Relief	5.38	5.30	5.43
PCFI, RFFI, PFI, Relief	5.41	5.60	5.31
PCFI, SFP, PFI, Relief	5.38	5.41	5.37
ANOVA, RFFI, SFP, PFI	5.53	5.58	5.51
ANOVA, RFFI, SFP, Relief	5.55	5.83	5.41
ANOVA, RFFI, PFI, Relief	5.58	5.81	5.46
ANOVA, SFP, PFI, Relief	5.63	5.58	5.66
RFFI, SFP, PFI, Relief	5.50	5.68	5.41
PCFI, ANOVA, RFFI, SFP, PFI	5.46	5.49	5.45
PCFI, ANOVA, RFFI, SFP, Relief	5.40	5.29	5.45
PCFI, ANOVA, RFFI, PFI, Relief	5.35	5.48	5.28
PCFI, ANOVA, SFP, PFI, Relief	5.33	5.24	5.38
PCFI, RFFI, SFP, PFI, Relief	5.28	5.49	5.17
ANOVA, RFFI, SFP, PFI, Relief	5.51	5.82	5.36

FIGURE 4. Heatmap of MAE for age estimation performance using different combinations of feature selection methods.

The results presented in Table 1 and Table 2 show that employing FS provided improvements over using the full feature set in terms of MAE and a number of features. Moreover, employing the proposed method on individual FS methods considerably improved the results in some FS methods without the need to set a prior number of features to select. As illustrated in Figs. 5 and 6, applying the fuzzy-ensemble method to individual FS methods and their combinations led to notable improvements. The algorithm automatically selected the optimal number of features, resulting in enhanced accuracies. This demonstrates that the proposed method can contribute to performance improvements in both accuracy and efficiency.

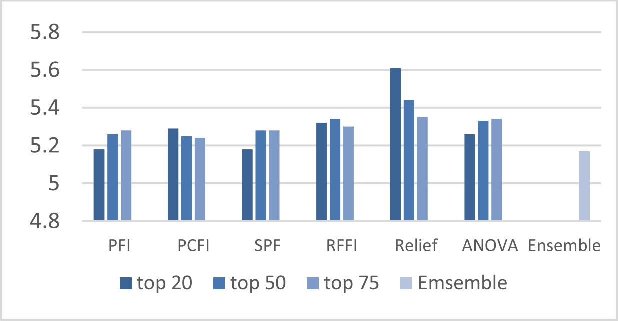


FIGURE 5. Comparison of MAE in age estimation using Top 20, 50, and 75 features across different feature selection methods and the proposed method.

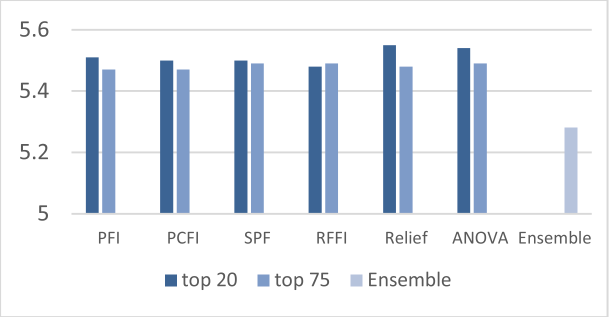


FIGURE 6. Comparison of MAE in Height estimation using Top 20, 50, and 75 features across different feature selection methods and the proposed method.

In Addition, we conduct an in-depth analysis of the performance of the proposed method across various subgroups of height and age. The testing data is categorized into different subgroups based on height and age, allowing us to examine the specific performance characteristics within each subgroup. Figures 7 presents an overview of the subgroups for both male and female speakers, displaying the height estimation performance and the number of speakers in each subgroup within the testing split. Notably, we observe that significant errors arise for speakers in subgroups corresponding to a low number of representative samples. This discrepancy in performance could potentially be attributed to the limited amount of training data available for these particular subgroups. As can be seen in Figure 2 that illustrates gender-specific histograms of speaker heights for both the training and testing datasets, revealing a notable mismatch between the height distributions. These mismatches in the training and testing height histograms might contribute to the occurrence of substantial errors for extreme height values.

Similarly, we evaluate the performance of the proposed approach in age estimation by dividing the data into distinct age subgroups, depicted in Figures 8.a and 8.b. Notably, we find that the MAE is particularly high for three specific age groups (ranging from 45 to 75 years) in both genders. This observation is further substantiated by the scarcity of training speakers within these age groups. Consequently, the MAE error within these three groups exceeds 22 years, significantly influencing the overall MAE performance.

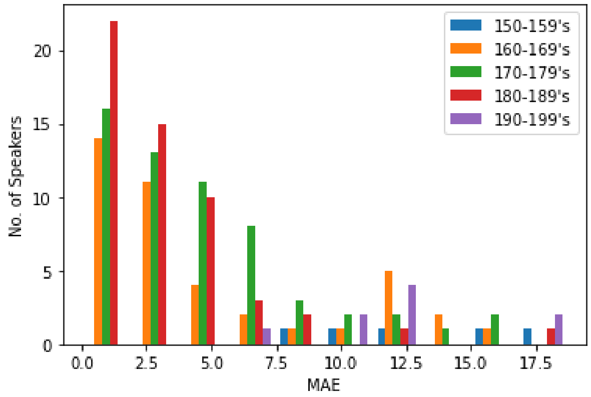


FIGURE 7. Histogram of MAE for height estimation across different height subgroups of male and female speakers using the proposed method.

After conducting our subgroup evaluations, we also visually inspected the performance of our model using regression plots, for representing the correlation between predicted and actual values of age and height. Figures 9 and 10 show the regression plots for height and age prediction, respectively. In addition to MAE and RMSE, the standard deviation of error was calculated for each prediction to further understand the variability in the prediction errors.

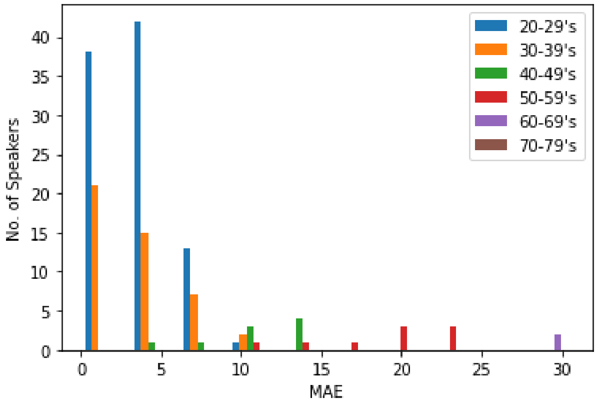


FIGURE 8. Histogram of MAE for age estimation across different age subgroups of male and female speakers using the proposed method.

In Figure 9, each point represents a test sample, with the x-axis indicating the actual age and the y-axis showing the predicted heights. Similarly, Figure 10 illustrates the relationship between the actual and predicted age. A perfect prediction would result in a data point falling along the 45-degree line, where the actual value equals the predicted value.

The plots show that our model performs quite accurately for the majority of data points, as they cluster around the 45-degree line. However, there are some outliers, particularly for extreme age and height values, which aligns with our findings from the subgroup analysis. The deviation of these points from the ideal line indicates the areas where our model's performance could be further improved, especially for these extreme cases. The standard deviation of error was found to be 7.81 for age prediction and 6.99 for height prediction

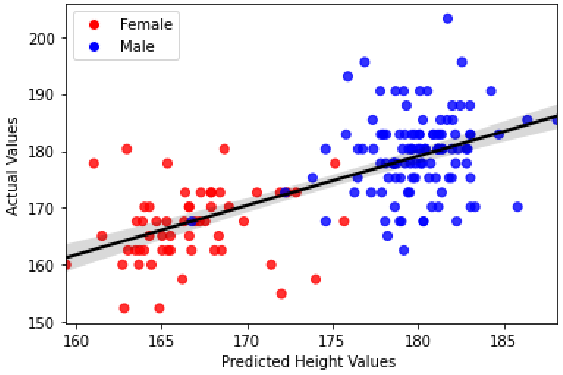


FIGURE 9. Regression plot for height estimation of male and female speakers using the proposed method.

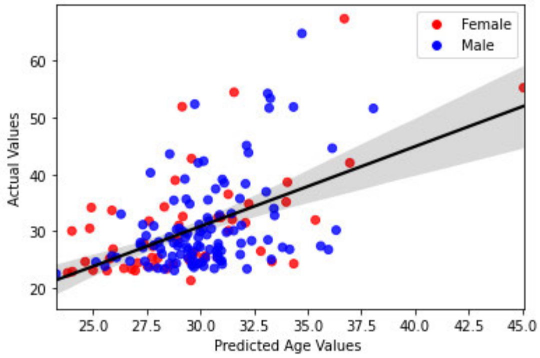


FIGURE 10. Regression plot for age estimation of male and female speakers using the proposed method.

Additionally, a comparison between the best-performing method and several previous studies on predicting speakers' height and age using the same dataset, training procedure, and evaluation metrics shows that the proposed method outperformed these studies, as shown in Tables 3 and 4.

The Fuzzy-Ensemble Feature Selection method we've proposed, originally developed with speaker recognition in mind, has the potential to influence a broad range of fields. Its demonstrated ability to efficiently select optimal feature sets from large, complex datasets makes it a valuable tool not only for speech analysis, but also in domains as diverse as medical diagnostics, financial data analysis, climate modeling, and image processing. Any field that deals with high-dimensional data could benefit from this approach. In short, our method has the potential to provide substantial contributions to feature selection processes across numerous disciplines, paving the way for improved model accuracy and efficiency.

An essential aspect to address is the scalability of our Fuzzy-Ensemble Feature Selection approach. While the proposed methodology itself is highly efficient, it is pertinent to note that the time complexity predominantly depends on

TABLE 3. Comparison of performance in height estimation between previous studies and the proposed method.

Study	MAE		RMSE	
	Male	Female	Male	Female
Kaushik. Et. al [25]	5.24	5.09	6.92	6.34
Singh et. al [26]	**5.0**	5.0	6.7	**6.1**
Kalluri. Et. al [17]	5.2	4.8	6.8	**6.1**
Williams [27]	5.37	5.49	-	-
Mporas [8]	5.30	5.1	6.8	6.3
Gupta. Et. al [28]	5.58	5.07	7.3	6.43
Proposed	5.4	**4.71**	**6.61**	6.24

TABLE 4. Comparison of performance in age estimation between previous studies and proposed method.

Study	MAE		RMSE	
	Male	Female	Male	Female
Singh et. al[26]	5.5	6.5	7.8	8.9
Gupta. Et.al[28]	**3.9**	**4.48**	**5.54**	**6.49**
Kwasny [29]	5.12	5.29	7.42	8.63
Kalluri [17]	5.2	5.6	8.1	8.7
Kaushik [25]	5.62	6.08	7.85	8.75
Zazo et.al. [30]	6.97	7.79	-	-
Proposed	5.38	5.24	8.0	8.36

the nature of feature selection methods employed within the ensemble. Certain feature selection methods could be computationally demanding, leading to increased time complexity when applied to larger datasets or real-time applications. Thus, striking a balance between improved prediction accuracy and computational efficiency emerges as a crucial factor for consideration when planning the expansion or application of this approach in diverse contexts.

A. SELECTED FEATURES

Upon examining the selected features for height and age estimation, the following observations were made:

For height estimation, the selected features primarily focused on fundamental frequency (F0) and formant frequency (F1, F2, F3) features. Additionally, loudness features, such as mean, standard deviation, percentiles, and slope-related attributes, were included. Spectral features, such as spectral flux, were also considered. Regarding MFCC features, the algorithm selected only the first four coefficients and their respective standard deviations. Other chosen features encompassed jitter and shimmer attributes, HNR, the first three formants, and their standard deviations, as well as voicing and voice quality characteristics.

In contrast, the features selected for age estimation encompassed a wider array of acoustic properties, capturing more aspects of speech signals. These included pitch, loudness, spectral characteristics, voice quality, and voicing-related features. Specific attributes consisted of the mean and standard deviation of the first MFCC, along with the third and fourth MFCCs in the voiced section. Additional formant features, such as the standard deviation of formant frequencies (F2 and F3), bandwidths, and amplitudes relative to F0, were also considered. Voicing-related features such as loudness peaks per second, voiced segments per second, and mean voiced segment length with its standard deviation were included as well. Moreover, the equivalent sound level representing the overall loudness of the speech signal was considered.

Further to our analysis, we carried out an additional experiment to identify the most recurrent features across most FS methods for both height and age estimations. By conducting this, we aimed to uncover the potential universal attributes that might be critical in predicting these parameters across diverse FS methods.

Out of the 88 features in the feature set, the features that were selected by most FS methods for both tasks of height and age prediction include F1bandwidth_sma3nz_amean, slopeV0-500_sma3nz_amean,F0semitoneFrom27.5Hz_sma3nz_percentile20.0, F0semito\eFrom27.5Hz_sma3nz_percentile80.0, logRelF0-H1-A3_sma3nz_amean,

These features consistently appeared as top-ranking features across the majority of the FS methods for height and age estimation.

To gain a better understanding of the relationship between each of these common features and the target parameters (height and age), we plotted their correlations. Figure 11 depicts some of the results for the training portion of the TIMIT dataset. Each dot in the plots represents an individual data sample, and the degree of correlation is represented by the trend line.

From these figures, it is observed that negative correlations are observed throughout, indicating a general trend of decreasing voice parameters with increasing age or height. In terms of age, the voice slope exhibits a weak negative correlation for males (-0.19) and a somewhat stronger correlation for females (-0.36). A similar pattern emerges with height, albeit weaker correlations are seen. Formant (F1) bandwidth demonstrates a slight negative correlation with age and height for both genders, with the strongest association observed in males' height (-0.17). The 80th percentile of fundamental frequency (F0), typically associated with higher frequency voice components, also shows weak negative correlations with age and height, but with a more pronounced relationship between females' age and F0 (-0.31). Interestingly, a stronger negative correlation was observed in the 20th percentile F0, representative of lower frequency voice components, and females' age (-0.38). These statistical relationships suggest trends in how age and height may influence voice parameters.

These results provide valuable insights for feature prioritization and selection in future iterations of our model, potentially leading to further improvements in prediction accuracy.

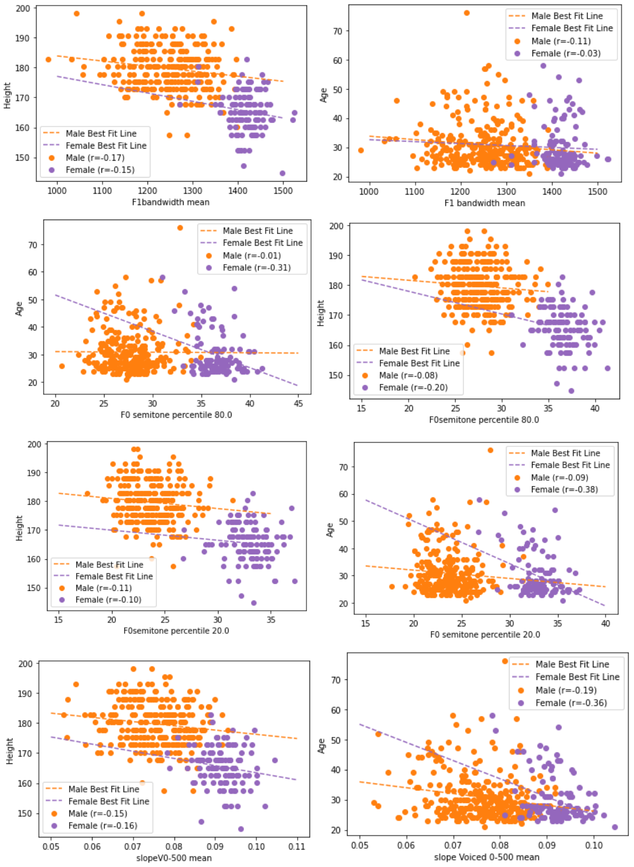


FIGURE 11. **Scatter plot of the estimates of the most occurring features in FS methods with the speaker height and age for the TIMIT training set. Value in the brackets shows the correlation (r) between the features and corresponding physical parameters for male and female speakers. The best-fit line is also shown for both male and female speakers separately.**

V. CONCLUSION

In this research, we have proposed and evaluated a novel ensemble FS method using FCM clustering to improve the estimation of height and age from voice data. This method leverages multiple filter-based, wrapper-based, and embedded FS algorithms, automatically selecting an optimal number of features based on their importance across different FS methods.

The results demonstrated that the proposed method enhanced the discriminative power of the system, yielding notable improvements in MAE, RMSE, and the number of features, compared to using a full feature set. This was further illustrated by the improvements observed in both individual and combined FS methods when the fuzzy-ensemble method was applied.

The proposed approach showcased its potential by outperforming various previous studies, making it a promising tool for feature selection in voice-based applications such as height and age estimation. Collectively, findings from both age and height estimation analyses highlight the significance of using combinations of feature selection rather than individual feature selection techniques.

Despite the effectiveness of the proposed method, certain limitations were identified, including the need to experiment with various FS method combinations to arrive at the best combination for the task. Furthermore, although the objective of this study is to illuminate the potential for estimating a variety of speaker traits using ensemble feature selection, it's important to acknowledge that the dataset employed in this research doesn't sufficiently encapsulate the vast range of variables influencing human voice. To execute speaker profiling in real-world situations, a more representative dataset is necessary. In future studies, these challenges could potentially be addressed by integrating our ensemble FS method with advanced optimization approaches such as genetic algorithms and automating the process of selecting.

REFERENCES

[1] N. J. Lass and W. S. Brown, "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," *J. Acoust. Soc. Amer.*, vol. 63, no. 4, pp. 1218–1220, Apr. 1978.

[2] D. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *J. Acoust. Soc. Amer.*, vol. 117, no. 1, pp. 305–318, Jan. 2005, doi: 10.1121/1.1828637.

[3] W. A. van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," *Lang. Speech*, vol. 38, no. 3, pp. 267–287, Jul. 1995.

[4] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 106, no. 3 Pt 1, pp. 1511–1522, Sep. 1999, doi: 10.1121/1.427148.

[5] C. Müller, "Automatic recognition of speakers^2 age and gender on the basis of empirical studies," in *Proc. 9th Int. Conf. Spoken Lang. Process. Interspeech*, Sep. 2006, pp. 2118–2121.

[6] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *J. Voice*, vol. 32, no. 5, pp. 644.e1–644.e9, Sep. 2018.

[7] A. Badr and A. Abdul-Hassan, "CatBoost machine learning based feature selection for age and gender recognition in short speech utterances," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 3, pp. 150–159, Jun. 2021, doi: 10.22266/ijies2021.0630.14.

[8] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio features selection for automatic height estimation from speech," in *Proc. Hellenic Conf. Artif. Intell.* Cham, Switzerland: Springer, 2010, pp. 81–90.

[9] K. S. Babu and D. Vijayasenan, "Robust features for automatic estimation of physical parameters from speech," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2017, pp. 1515–1519.

[10] A. A. Badr and A. K. Abdul-Hassan, "Estimating age in short utterances based on multi-class classification approach," *Comput., Mater. Continua*, vol. 68, no. 2, pp. 1713–1729, 2021, doi: 10.32604/cmc.2021.016732.

[11] J. Gonzalez-Lopez, S. Ventura, and A. Cano, "Distributed multi-label feature selection using individual mutual information measures," *Knowl.-Based Syst.*, vol. 188, Jan. 2020, Art. no. 105052.

[12] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, Mar. 2019.

[13] E. K. Naka, V. G. Guliashki, and G. I. Marinova, "A comparative analysis of different feature selection methods on a Parkinson data," in *Proc. 15th Int. Conf. Adv. Technol., Syst. Services Telecommun. (TELSIKS)*, Oct. 2021, pp. 366–371.

[14] Z. Soumaya, B. D. Taoufiq, N. Benayad, K. Yunus, and A. Abdelkrim, "The detection of Parkinson disease using the genetic algorithm and SVM classifier," *Appl. Acoust.*, vol. 171, Jan. 2021, Art. no. 107528.

[15] C. Chen, Y. Tsai, F. Chang, and W. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Syst.*, vol. 37, no. 5, Oct. 2020, Art. no. e12553.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1," NASA STI/Recon, Washington, DC, USA, Tech. Rep. 93-27403, Feb. 1993.

[17] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Commun.*, vol. 121, pp. 16–28, Aug. 2020, doi: 10.1016/j.specom.2020.03.008.

[18] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6580–6584.

[19] H. A. Abdulmohsin, B. Al-Khateeb, and S. S. Hasan, "Speech gender recognition using a multilayer feature extraction method," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*. Cham, Switzerland: Springer 2022, pp. 113–122.

[20] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[21] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML)*. Antwerp, Belgium: Springer 2008, pp. 313–325.

[22] M. Alduailij, Q. W. Khan, M. Tahir, M. Sardaraz, M. Alduailij, and F. Malik, "Machine-learning-based DDoS attack detection using mutual information and random forest feature importance method," *Symmetry*, vol. 14, no. 6, p. 1095, May 2022.

[23] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[24] S. Galli, "Feature-engine: A Python package for feature," *J. Open Source Softw.*, vol. 6, no. 65, p. 3642, 2021.

[25] M. Kaushik, V. Tung Pham, and E. Siong Chng, "End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN," 2021, *arXiv:2101.05056*.

[26] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *Proc. 4th Int. Conf. Biometrics Forensics (IWBF)*, 2016, pp. 1–6.

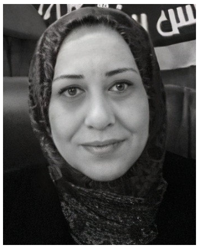
[27] J. H. Hansen, K. Williams, and H. Boril, "Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models," *J. Acoust. Soc. Amer.*, vol. 138, no. 2, pp. 1052–1067, Aug. 2015, doi: 10.1121/1.4927554.

[28] T. Gupta, D.-T. Truong, T. T. Anh, and C. E. Siong, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," 2022, *arXiv:2203.11774*.

[29] D. Kwasny and D. Hemmerling, "Joint gender and age estimation based on speech signals using x-vectors and transfer learning," 2020, *arXiv:2012.01551*.

[30] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018, doi: 10.1109/access.2018.2816163.

UMNIAH HAMEED JAID received the B.Sc. degree in computer science from the University of Baghdad, Baghdad, Iraq, in 2007, and the M.Sc. degree in computer science from the University of Glasgow, Glasgow, U.K., in 2012. She is currently pursuing the Ph.D. degree with the University of Technology, Baghdad. She has been a Lecturer with the Department of Computer Science, University of Baghdad, since 2008. Her research interests include speech processing, machine learning and pattern recognition, and artificial intelligence.



ALIA KARIM ABDULHASSAN received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Technology, Baghdad, Iraq, in 1993, 1999, and 2004, respectively. She has been a Professor with the University of Technology, since 2019, where she is currently the Dean of the Computer Science Department. She has been supervising more than 26 M.Sc. and Ph.D. thesis in computer science, since 2007. She has authored and coauthored more than 100 papers in international conferences and journals. Her current research interests include soft computing, green computing, artificial intelligence, data mining, software engineering, electronic management, and computer security.