Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4444388, 13 pages
https://doi.org/10.1155/2022/4444388

Research Article

Speaker Gender Recognition Based on Deep Neural Networks and ResNet50

Abeer Ali Alnuaim,[1] **Mohammed Zakariah**,[2] **Chitra Shashidhar**,[3] **Wesam Atef Hatamleh**,[4] **Hussam Tarazi**,[5] **Prashant Kumar Shukla**,[6] **and Rajnish Ratna**[7]

[1]*Department of Computer Science and Engineering, College of Applied Studies and Community Services, King Saud University, P.O. Box 22459, Riyadh 11495, Saudi Arabia*
[2]*College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia*
[3]*Department of Commerce and Management, Seshadripuram College, Seshadripuram, Bengaluru 20, India*
[4]*Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia*
[5]*Department of Computer Science and Informatics, School of Engineering and Computer Science, Oakland University, Rochester Hills, MI, 318 Meadow Brook Rd, Rochester, MI 48309, USA*
[6]*-Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522502 Andhra Pradesh, India*
[7]*Gedu College of Business Studies, Royal University of Bhutan, Bhutan*

Correspondence should be addressed to Rajnish Ratna; rajnish.gcbs@rub.edu.bt

Received 22 February 2022; Revised 9 March 2022; Accepted 11 March 2022; Published 24 March 2022

Academic Editor: Mohammad Farukh Hashmi

Several speaker recognition algorithms failed to get the best results because of the wildly varying datasets and feature sets for classification. Gender information helps reduce this effort since categorizing the classes based on gender may help lessen the impact of gender variability on the retrieved features. This study attempted to construct a perfect classification model for language-independent gender identification utilizing the Common Voice dataset (Mozilla). Most previous studies are doing manual extracting characteristics and feeding them into a machine learning model for categorization. Deep neural networks (DNN) were the most effective strategy in our research. Nonetheless, the main goal was to take advantage of the wealth of information included in voice data without requiring significant manual intervention. We trained the deep learning network to choose essential information from speech spectrograms for the classification layer, performing gender detection. The pretrained ResNet 50 fine-tuned gender data successfully achieved an accuracy of 98.57% better than the traditional ML approaches and the previous works reported with the same dataset. Furthermore, the model performs well on additional datasets, demonstrating the approach's generalization capacity.

1. Introduction

Neural networks are state-of-the-art in various classification tasks, including video and audio segmentation. Determining the speaker's identity from an audio clip of their speech is a classification problem of this sort. Nevertheless, the quantity of data necessary to produce acceptable results is one of the most critical trade-offs in neural network training. Gender is a component that results in an average physiological difference. It may increase the identification system's accuracy because males and females have diverse emotional expressions and voice processes. Integrating gender data in the development and testing processes makes the data more trustworthy. The neural network obtains another element

for identifying the task-specific voice qualities for the two genders [1]. Gender categorization by the audio signal is crucial for various applications, involving targeted answers and advertising by voice assistants, population statistics via age group analysis, and automated profiling of an individual using speech data to help in a criminal probe. Additionally, for a model with a gender-specific search space, even a small quantity of data will significantly contribute to various audio systems, such as automated voice recognition, speaker identification, and content-based multimedia indexing.

A collection of features is utilized to determine the gender of a sound. The Mel-scaled power spectrogram (Mel), Mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), spectral contrast (Contrast), and tonal centroid are among the most often used features for speech gender detection (Tonnetz). Machine learning (ML) approaches are used to build a high-quality system for distinguishing voice gender using the retrieved attributes. Each classification approach, in specific, generates a collection of hypothesis models and chooses the most optimum one. This model identifies the unknown voice label by acquiring the audio attributes and classifying the voice gender.

As the feature extraction stage has progressed to the point that many academics now see it as feature engineering, intending to develop robust feature vectors that accurately characterize structures in methods relevant to the job at hand. The primary goal of feature engineering is to create features that cluster patterns belonging to the same class together in the feature space while keeping them as far apart as possible from other categories. But autonomous representation learning has attracted increased interest to study deep learning methodologies more readily and extensively. The classification scheme is constructed with deep learning so that the encoder acquires the optimal attributes for describing patterns throughout the training phase. Additionally, because of specialized deep architectures, such as CNN, the input structures are sometimes depicted as a picture. CNN is a specialized architecture for handling the image classification task, among other things. This has prompted academics working with CNNs to create ways for transforming an audio input to a time-frequency image. Hence, another approach tackles the gender categorization issue arose by using the auditory spectrograms as inputs to our system. It sounds spectrograms are visual representations of audio. Spectrograms are very comprehensive and precise representations of speech that have been extensively employed in auditory categorization applications [2–4]. Deep neural networks (DNNs) trained on extracted features are very effective in removing data and have been successfully used in applications such as speech recognition [3, 5] and picture identification [5–7]. CNN's can effectively leverage the invariance inherent in spectrograms for convolutional and pooling operations [8].

The purpose of this article is to analyze deep learning techniques to typical machine learning models trained on handcrafted features to determine if deeply learned attributes are adequate for gender categorization tasks. Deep neural networks (DNNs) and convolutional neural networks

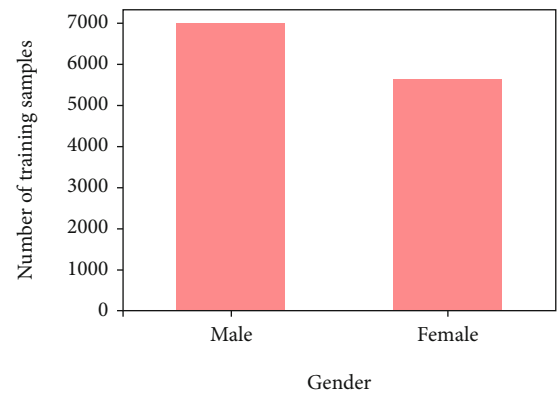

Figure 1: Statistics of voice samples in different genders.

are the most accurate classifiers and feature extractors for speech gender detection, according to experimental data (CNN) classification. When used for audio categorization, [9] demonstrated the performance of prominent CNN architectures like AlexNet, VGG, Inception, and ResNet. Their method required decomposing the audio time series using a short-time Fourier transform to generate a spectrogram utilized to enter the CNN. The issue with many of these models is that they are huge and have many learnable parameters. Therefore, we concentrated on a database with just a few thousand samples since it seems improbable that these vast networks could be trained with our sparse data. Transfer learning is one way to circumvent this. This is accomplished by employing a pretrained network, freezing the values of most levels, and retraining just the last few layers using our audio data for training. This is one of the directions we took with this job. The performance achieved by fine-tuning various pretrained CNNs (ResNet 34 and ResNet50 on ImageNet) is optimal for our gender audio categorization issues.

The following are the significant contributions made to the community:

(1) Presented the study of impacts of a set of handcrafted voice qualities as possible appropriate features for gender classification algorithms

(2) The work contributes to our understanding of the extent to which picking voice signal features aided in the development of machine learning models

(3) Compared the performance of various classical machine learning models and DNN trained on handcrafted voice features

(4) Performance study of fine-tuning the pretrained ResNet34 and ResNet50 on audio spectrograms. The purpose of this study was to determine whether spectrograms give sufficient detail for accurate gender audio classification

(5) Presented the research on the effectiveness of speech classifiers on various corpus datasets

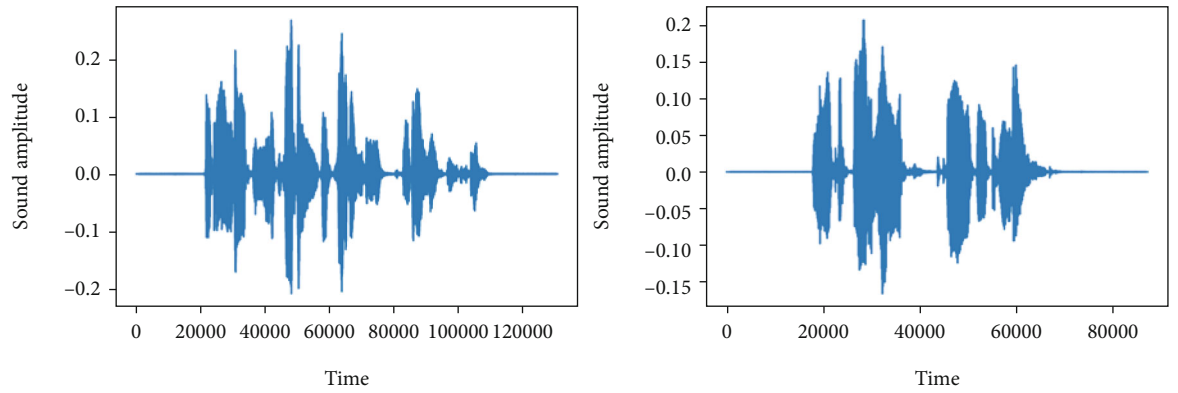

Figure 2: Waveform for the female voice.

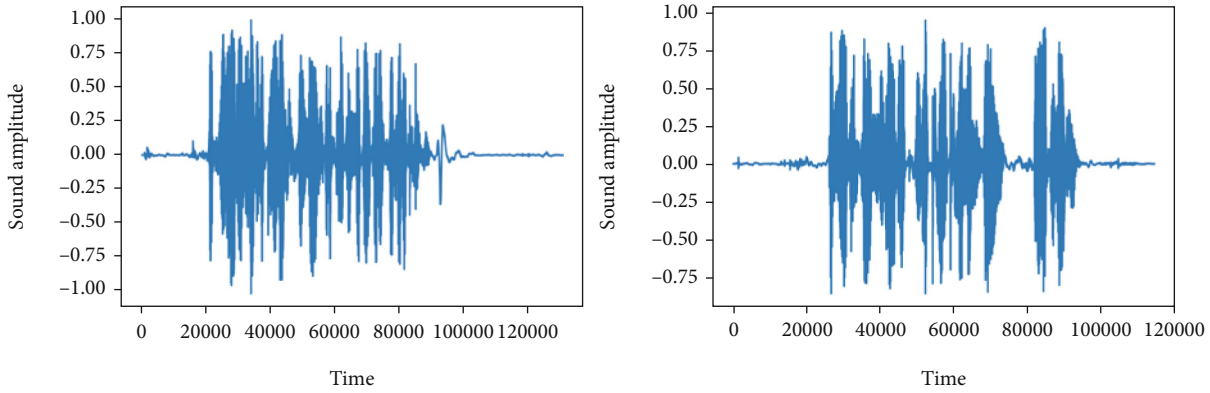

Figure 3: Waveform for the male voice.

The rest of the paper is organized as follows: Section 2 describes the previous works in the domain; Section 3 details about the methodology, which consists of dataset, preprocessing techniques, feature extraction, model architecture, and evaluation metrics; Section 4 describes the results and Section 5 with discussion; and Section 6 with conclusions followed by a list of relevant references.

2. Literature Review

Numerous studies have been undertaken to determine the effectiveness of voice classifiers to increase the precision of programs being used. [10] recognized the speaker's gender on the TIDIGITS database with an accuracy rate of 98.65% using two-level classifiers (pitch frequency and GMM classifier). [11] analyzed voices from the Ivie corpus using four classifiers: GMM, multilayer perceptron (MLP), vector quantization (VQ), and learning vector quantization (LVQ). They had a reliability percentage of 96.4%. [12] integrated the acoustic voice levels determined by five distinct approaches into a single score level. The findings were attained using the gender dataset with an 81.7% success rate for the gender category. [13] developed a method to recognize speakers cantered on a fusion score of seven subsystems employing the MFCC, PLP, and prosodic feature vectors on three distinct classifiers: GMM, SVM, and GMM-SV-based SVM. The categorization rate of success for gender identification is 90.4% when utilizing the aGender dataset. [14] used two classifiers to a private dataset to determine gender voice: SVM and decision tree (DT) using the MFCC feature. The total accuracy of gender categorization using MFCC-SVM and MFCC-DT was 93.16% and 91.45%, correspondingly. [15] developed a method to increase the MFCC features and then modify the weighting between the DNN tiers. These enhanced MFCC attributes are assessed using DNN and I-Vector classifiers, which achieve an overall accuracy rate of 58.98% and 56.13%, accordingly. [16] examined two arrangement approaches (DNN and SVM) for robust sound classifications utilizing single and combination feature vectors. The findings indicated that the DNN strategy outperformed the noise approach because of its robustness and poor sensitivity to sound.

[17] predicted age and gender using deep neural networks. Lately, raw waveform processing in speech has become a trend in the speech field, and it has been shown that employing raw audio waveforms improves voice recognition effectiveness [18]. When raw audio-based neural networks are used, voice activity identification [19] and speaker verification [20] have also shown considerable performance gains. [19] attempts to decipher and explain how CNNs

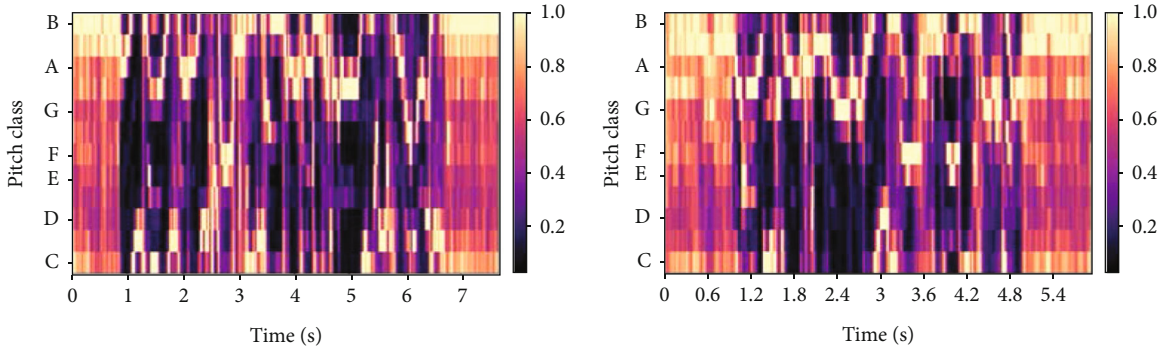

Figure 4: Chromagram for the female voice.

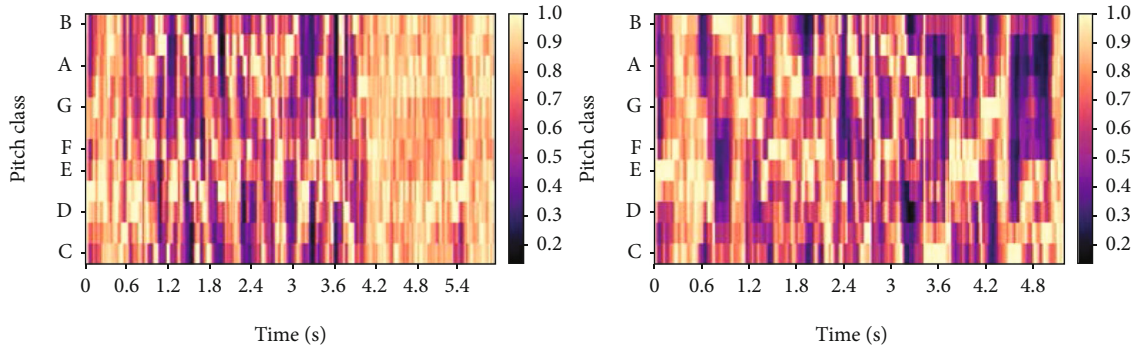

Figure 5: Chromagram for the male voice.

categorize audio information. CNN's are learned using both spectrogram and raw audio inputs, and a layer-wise relevance propagation (LRP) technique is employed to examine how the systems choose features and make choices. They demonstrated the distinct patches on the input signal that strongly connect to each output label. The paper's findings show that spectrogram inputs result in greater accuracy than raw auditory data.

In certain works, they employed gender recognition as a factor. [1] used speech recordings to build an emotion recognition model. The suggested approach includes an R-CNN and a gender information block. The suggested approach improves accuracy by 5.6%, 7.3%, and 1.5% in Mandarin, English, and German, respectively, compared to existing highest accuracy algorithms. [21] also illustrates the importance of gender and linguistic variables in vocal expression classification. They found that higher energy emotions like anger, joy, and surprise were easier to discern in male voices speaking a harsh language like German. Disgust and Fear were easier to discern in female voices in each language. They also found that when analyzing emotion across gender and language, signal amplitude, and energy are critical.

3. Materials and Methods

3.1. Dataset. Familiar Voice is a corpus of speech data [20] read by users on the Common Voice website (http://voice.mozilla.org/) based on text from various public domain sources, including user-submitted blog posts, old books, movies, and other publicly available speech corpora. Its main goal is to develop and test automated speech recognition (ASR) software. There are 8,64,448 MP3 audio files in the data collection. A .tsv file including the filename, sentence, accent, age, gender, locale, upvotes, and downvotes was also included in the dataset. The audio clips were saved using the same name in the associated.tsv file. The.tsv file was filtered by deleting any missing attributes in gender, age, and accents and choosing the rows with column "downvotes" of 0. Furthermore, the gender options were limited to "male" and "female." As a result, the total dataset is limited to 3,94,818 rows with labels for and age (from diverse geographic areas, each with its own linguistic dialects and accents.

We only took a subset of the created dataset for the current work. The number of audio files in the female and male categories was nearly equal to avoid bias (Figure 1). We separated the .tsv file into two columns for further processing: filename and gender. There were 6995 male audio files and 5662 female audio files in the filtered audio collection. The WAV format converted all audio files for frequency spectrum analysis [22]. Figures 2 and 3 show the time-domain representation of the female and male audio waveforms, respectively. The chromagram of the male and female audio categories is depicted in Figures 4 and 5, which shows how the pitches for twelve different pitch classes change between the gender groups.

4. Feature Extraction

4.1. Fourier Transforms. Audio signals can be complicated combinations of several sound components. First, it is

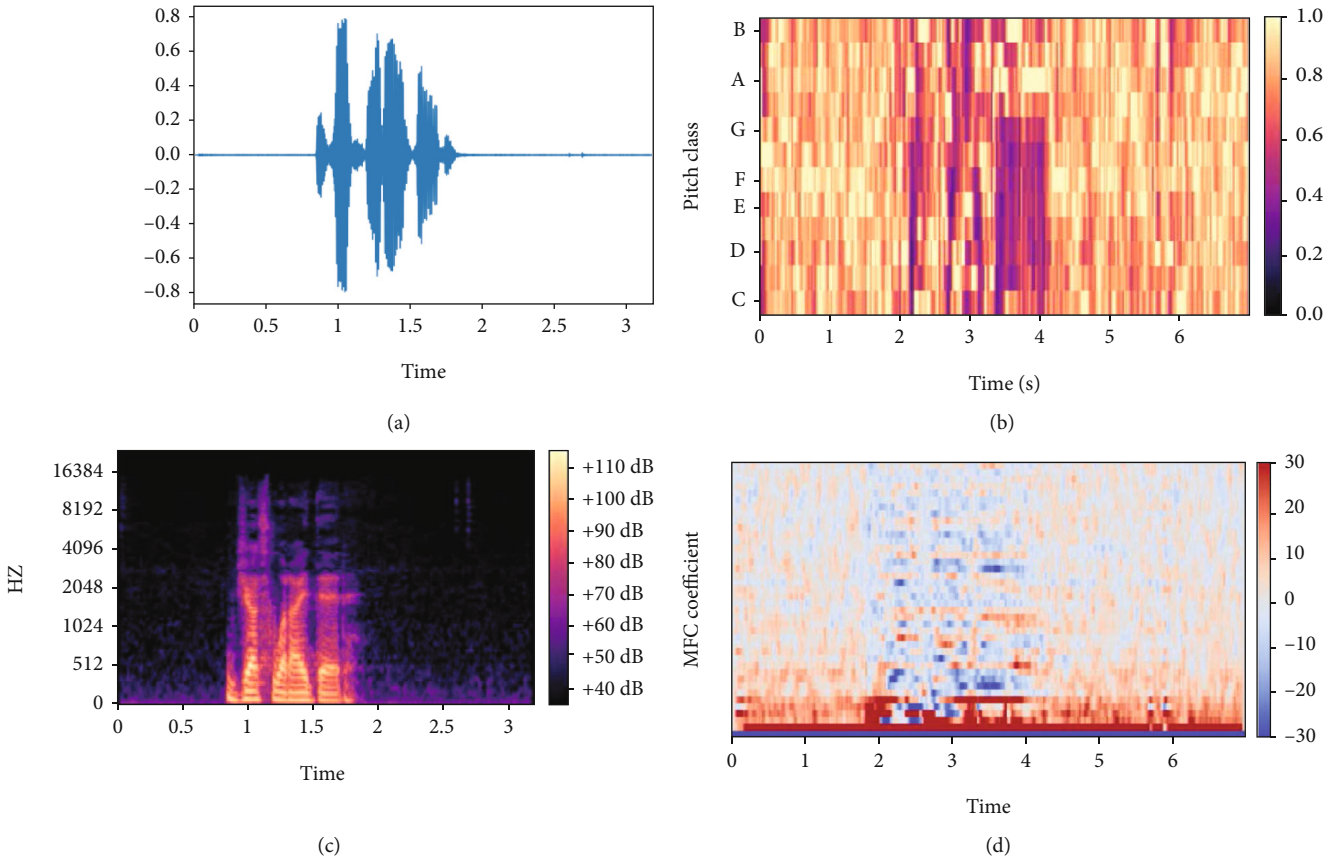

Figure 6: (a) Original time domain signal. (b) Chromogram. (c) Mel spectrogram. (d) MFC coefficients.

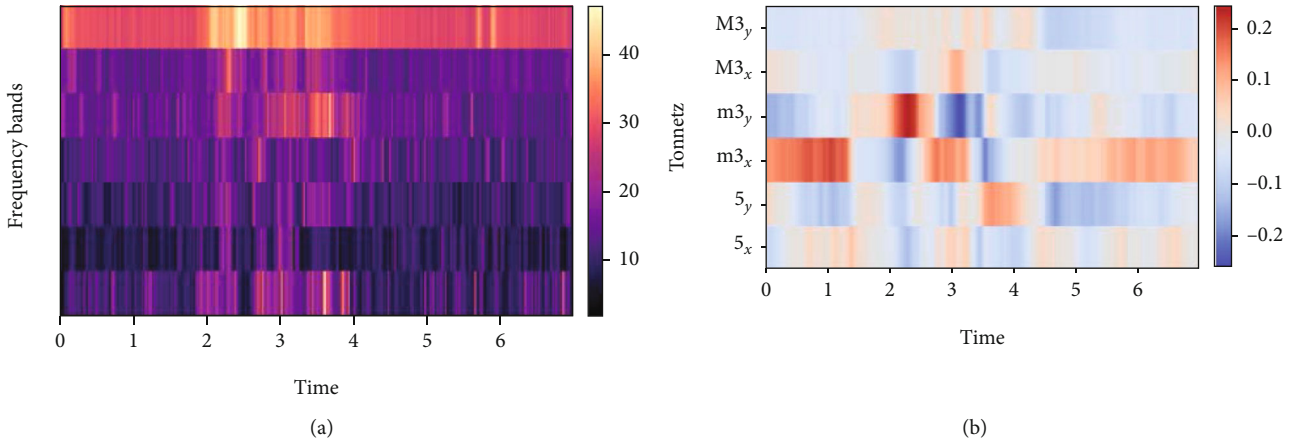

Figure 7: (a) Spectral contrast. (b) Tonal centroids (Tonnetz).

Table 1: Hyperparameters of the ML models.

Models	Hyperparameters
K neighbor classifier	$K = 3$
SVC (kernel = 'linear')	$C = 0.025$
Decision tree classifier	Max depth = 5
Random forest classifier	Max depth = 5, n_estimators = 10, max_features = 1

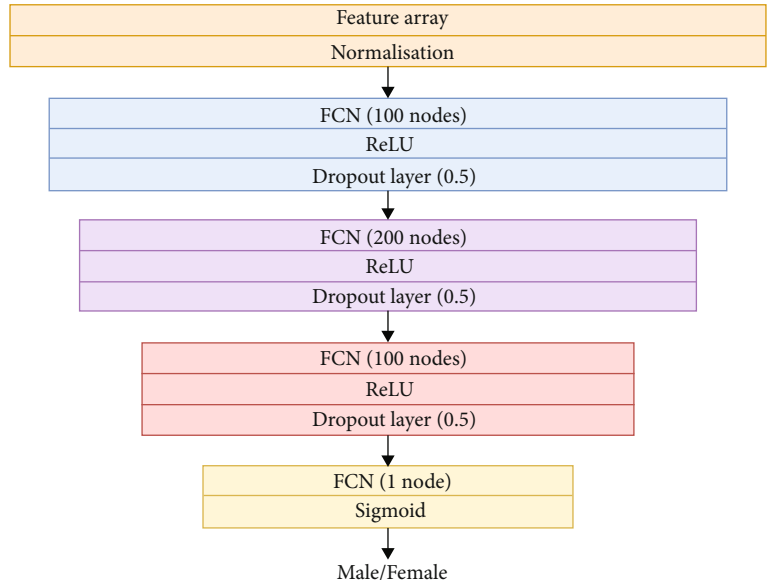

Figure 8: Deep neural network architecture.

Table 2: ResNet 34 and ResNet 50 detailed architecture.

Layer name	Output size	34 layers	50 layers
Conv 1	112×112	7×7, 64, stride 2	
Conv 2.x	56×56	3×3 max pool, stride 2 $[3 \times 3, 64\ 3 \times 3, 64] \times 3$	$[1 \times 1, 64\ 3 \times 3, 64\ 1 \times 1, 256] \times 3$
Conv 3.x	28×28	$[3 \times 3{,}128\ 3 \times 3{,}128] \times 4$	$[1 \times 1{,}128\ 3 \times 3{,}128\ 1 \times 1, 512] \times 4$
Conv 4.x	14×14	$[3 \times 3{,}256\ 3 \times 3{,}256] \times 6$	$[1 \times 1{,}256\ 3 \times 3{,}256\ 1 \times 1, 1024] \times 6$
Conv 5.x	7×7	$[3 \times 3{,}512\ 3 \times 3{,}512] \times 3$	$[1 \times 1{,}512\ 3 \times 3{,}512\ 1 \times 1, 2048] \times 3$
FLOPs	1×1	Average pool, 1000-d fc, softmax 3.6×10^9	3.8×10^9

Table 3: The classification performance on the test set for different models.

Features	Model	Accuracy (%)
MFCC, Mel spectrogram, Chroma STFT, Tonnetz, special contrast	Designed DNN	95.97
	MLP	95.81
	K neighbor classifier	95.10
	Random forest classifier	94.23
	SVC RBF kernel	93.92
	SVC	91.63
	Ada boost classifier	90.13
	Decision tree classifier	88.70
	Quadratic discriminant analysis	77.33
	Gaussian NB	72.27

Table 4: The classification performance of DNN on the test dataset.

Features	Accuracy (%)	$F1$ score (%)	Precision (%)	Recall (%)
MFCC, Mel spectrogram, Chroma STFT, Tonnetz, special contrast	95.97	95.91	96.03	95.82

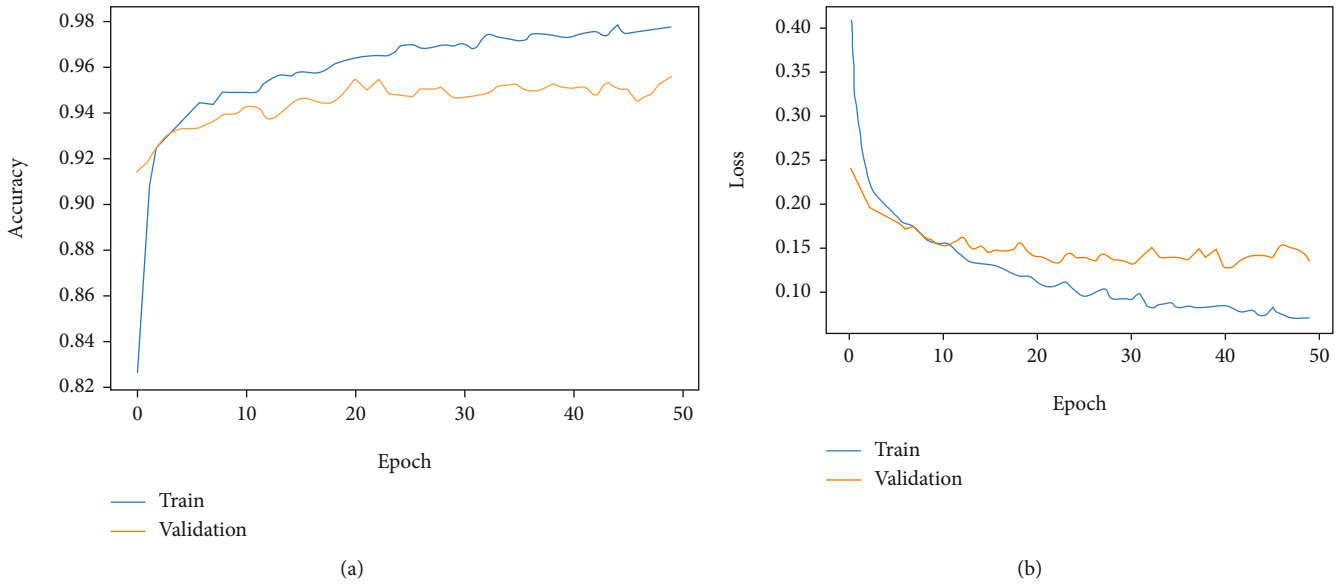

Figure 9: Accuracy and loss graphs for DNN.

Table 5: Comparison of the performance of two ResNet architectures.

Model	Accuracy	*F*1 score	Recall	Precision
ResNet34	97.94	98.18	98.32	98.04
ResNet50	98.57	98.74	99.02	98.47

decomposed into building components that can be processed more efficiently to understand a signal better. When these building blocks are exponential functions, the procedure is known as Fourier analysis. The Fourier transform [23] converts a time-dependent signal to a frequency-dependent role, revealing the original signal's frequency spectrum. The Fourier transform gives the signal's frequencies as well as their magnitudes. The inverse Fourier transform converts the frequency-domain representation of a given signal into the original signal.

4.2. Fast Fourier Transform (FFT). The fast Fourier transform (FFT) is a mathematical procedure frequently used to estimate the discrete Fourier transform of any sequence (DFT). We have a series of amplitudes sampled from a continuous audio stream in this scenario. Using the FFT technique, each frame of those N samples is converted from a time-domain discrete signal to a frequency-domain signal [24]. The FFT is considered an effective computing implementation of the DFT method, which is specified on a set of N samples $\{x_n\}$ as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\left(-\frac{j2\pi kn}{N}\right)} \; k = 0, 1, 2, \cdots, N-1. \tag{1}$$

4.3. Chroma-STFT (Short-Time Fourier Transform). An octave is defined as the distance of 12 pitches in our scale. Tone height and Chroma are two components of a pitch. The tone height is the octave number, and the Chroma is the pitch spelling. A pitch class is the collection of all pitches with the same Chroma. We combine all spectral information for one pitch class with Chroma features into a single coefficient. The Chroma value of audio is essentially a representation of the intensity of the twelve unique pitch classes (semitones or Chroma) used in music analysis. They can be used to distinguish between audio signal's pitch class profiles. Chroma STFT, as illustrated in Figure 6(b), generates a chromagram from a waveform or power spectrogram. It computed Chroma features via a short-term Fourier transformation. STFT encodes information about the pitch and signal structure classification. It depicts the spike with high values (as indicated by the graph's color bar) in low values (dark regions).

4.4. Mel Spectrogram. A spectrogram is a visual representation of a signal's frequency spectrum. The Mel scale [25] is a mathematical representation of how the human ear works, demonstrating that people do not perceive frequencies on a linear scale. Humans are more sensitive to differences at lower frequencies than at higher frequencies. In mathematical terms, the Mel scale is the outcome of a nonlinear transformation of the frequency scale [26]. The term "Mel-frequency scale" refers to a scale that is defined as

$$\text{mel} = 2595 * \log_{10}\left(\frac{(1+f)}{700}\right), \tag{2}$$

where f is the signal values in hertz. mel is the signal values in the Mel scale.

The term "Mel spectrogram" refers to a spectrogram that has been converted to the Mel scale. For example, Mel spectrogram returns a power spectrogram coefficient that has been Mel scaled. Mel spectrogram object is shown in Figure 6(c).

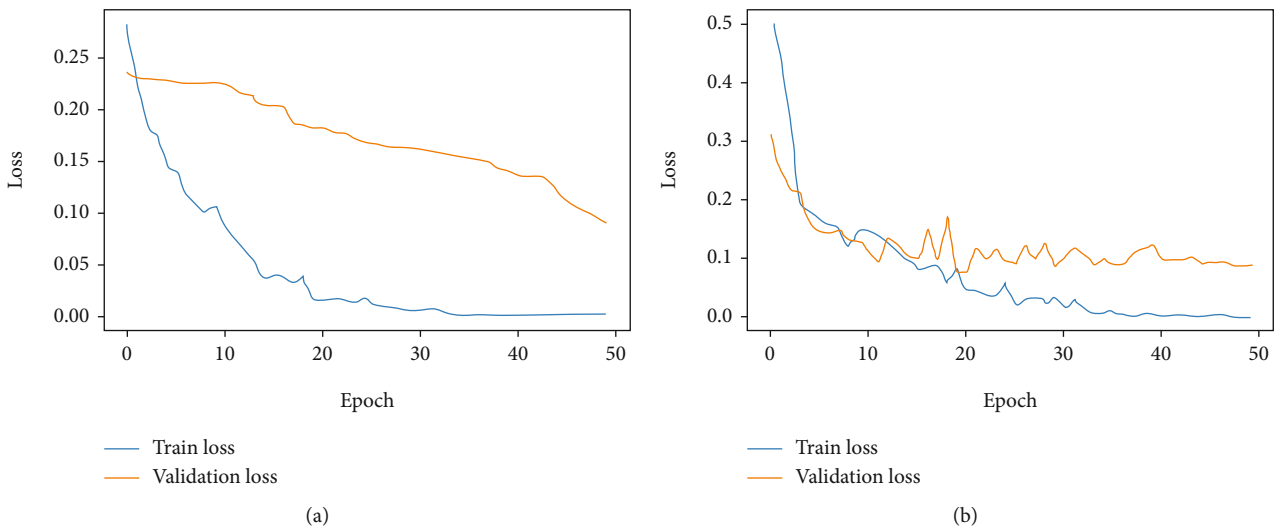

Figure 10: Loss graphs: (a) ResNet34 and (b) ResNet50.

4.5. MFCC. MFCC accurately portrays the vocal tract, a filtered shape of a human voice, and a short-time power spectrum envelope. MFCC is nothing but the coefficients that make up the Mel-frequency cepstrum) as shown in Figure 6(d). Generally, the first 13 coefficients of MFCC (the lower dimensions) are considered features because they reflect the spectral envelope. And the higher dimensions that have been deleted express the spectral subtleties. For distinct phonemes, envelopes are sufficient to convey the distinction, allowing us to distinguish phonemes using MFCC. It is a significant feature in various study fields that utilize audio signals [27].

4.6. Spectral Contrast. The spectral contrast of an audio signal is the energy of frequency at each timestamp. It is not easy to measure energy as most audio files contain a changing frequency. However, spectrum contrast helps to measure energy fluctuation. Spectral contrast is estimated based on [28], where it considers the spectral peak, the spectral valley, and the difference in frequency between each subband. The spectral contrast is depicted in Figure 7(a), which uses the relative spectral distribution rather than the average spectral envelope to represent it. High contrast signals are clear and narrow-band, while low contrast signals are broad-band noise.

4.7. Tonnetz Features. Tonnetz computes tonal centroid information from musical audio streams. It followed the methodology by [10]. Figure 7(b) shows Tonnetz centroids.

4.8. Model Development

4.8.1. Traditional Machine Learning. The traditional machine learning model's hyperparameters are specified in Table 1.

4.8.2. Deep Neural Network (DNN). Five layers form the planned neural network (Figure 8). The retrieved feature vector had a length of 192. Before feeding the whole feature vector into the neural network design, it was normalized. The label vector was binary because our objective was to categorize the audio stream into male or female. The first fully connected layer (FCN) is configured to accept a 192-dimensional feature vector. The nonlinear ReLU function was used as an activation, followed by a dropout layer. Dropout is a strategy for preventing overfitting in deep neural networks [29], and we provided a dropout rate of 0.5 while designing the network. The design of the second and third hidden levels was identical. The output layer is composed of a single node, and the activation function was Sigmoid due to its superior performance in binary classification.

4.9. Residual Network (ResNet). ResNet [7] is an exceptional architecture composed of residual layers. Additionally, ResNet is unique in that it uses global average pooling layers at the network's end rather than the more conventional set of fully connected layers. These architectural advancements result in a model that is eight times more detailed than VGGNet while remaining significantly smaller. This study examines ResNet30 (a 34-layer residual network) and ResNet50 (54-layer residual network) (Table 2). Both CNN's utilized spectrograms created from the audio wave file as input.

4.10. Training. The test and validation datasets together make up 10% of the total data. The DNN architecture is built using the Keras framework, supported by TensorFlow and written in Python. For the ResNet34 and ResNet 50 network training, we used the FAST AI library built on top of PyTorch. All other processing and analysis were performed on the DNN and ResnNets using NumPy, OpenCV, Scikit-learn, and other open-source tools. The training was conducted on a 32 GB NVIDIA Quadro P1000 GPU. The movement began with a 0.001 learning rate. We used the Adam algorithm as an optimizer [30]. The training could last up to 50 epochs with a batch size of 64. However, if the validation loss does not decrease continuously over a long period, early stopping will occur. It is applied to the test dataset to validate the trained model's performance.

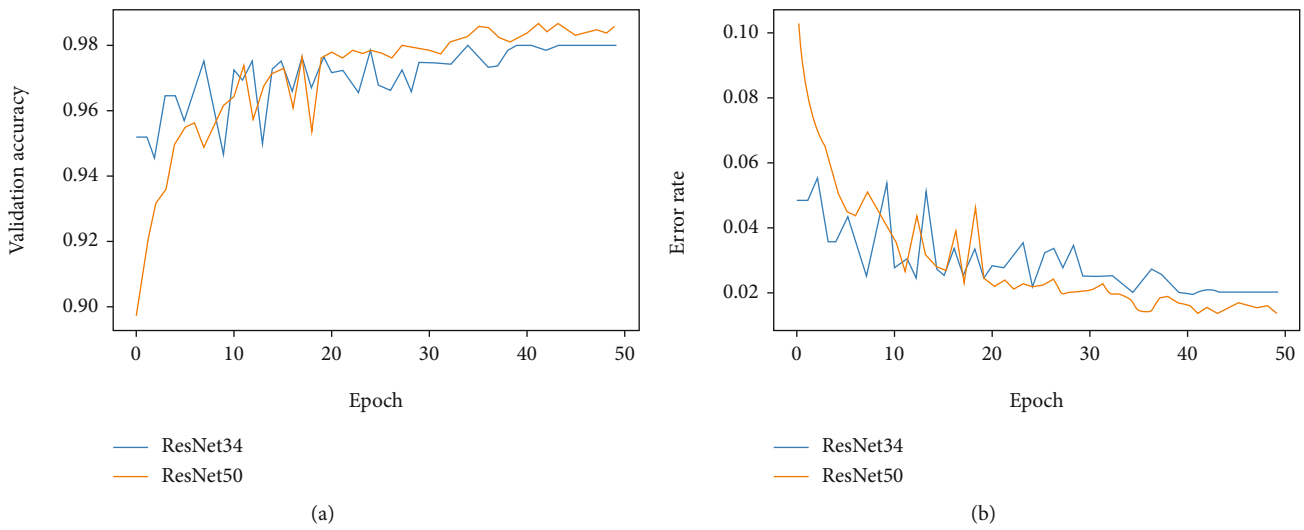

Figure 11: (a) Validation accuracy. (b) Error rate of Resnet34 and ResNet50.

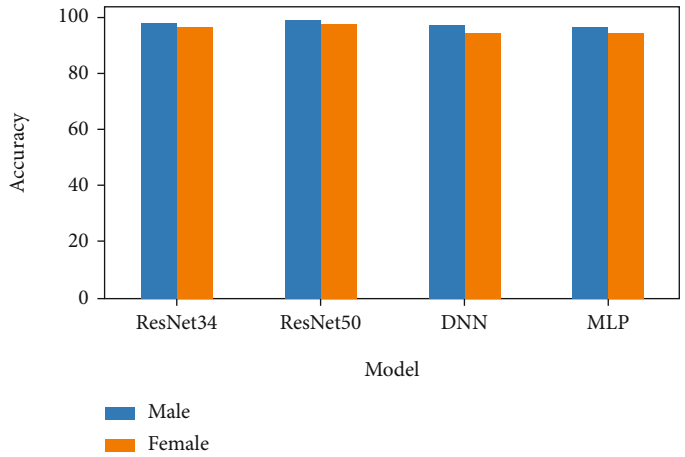

Figure 12: Classification accuracy of female and male classes for different models on the Mozilla dataset.

Table 6: Comparison of the performance of developed ResNet50 on SVD and RAVDESS datasets.

Dataset	Accuracy	*F*1 score	Recall	Precision
SVD	72.74	77.79	95.48	65.63
RAVDESS	91.5	90.89	97.91	84.81

4.11. Performance Evaluation. The scikit-learn handles the performance evaluation part. They employ metrics based on the "Confusion Matrix" to assess the binary classification model's performance. Some of the important values that contribute to the performance evaluation were as follows:

(i) False positive (FP): when data is negative yet the model predicts positive

(ii) False negative (FN): positive data predicted by a model as negative

(iii) True positive (TP): when both the data and the model anticipate positive

These values were evaluated considering the problem as a binary classification problem: 0 (negative) and 1 (positive). Here, we considered the 'male' category as positive and 'female' category as negative.

The model performance metrics like accuracy (Equation (3)), recall (Equation (4)), precision (Equation (5)), and *F*1 score (Equation (6)) were derived from the TP, FP, and FN values.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

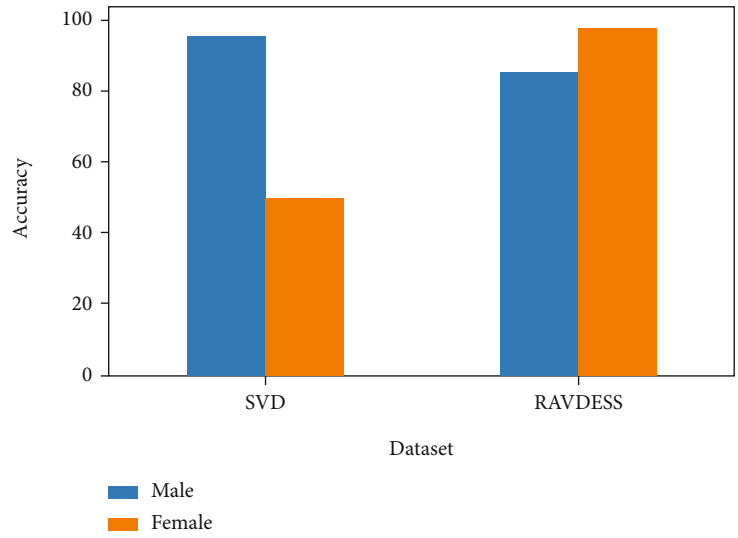

Figure 13: The class wise accuracy of the ResNet 50 on SVD and RAVDESS datasets.

Table 7: Comparison of performance of the proposed model with the other works.

Work	Accuracy (%)	Recall (%)	Precision (%)	$F1$ score (%)
[42]	94.32			
[41]	96.4	96.4	96.4	96.4
[43]	97.8			
Ours	98.57	99.02	98.47	98.74

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

5. Results

The research goal was to design a system that could accurately identify a person's gender using their voice. A set of experiments were conducted in order to determine the most appropriate classifier for gender classification. Choosing a classifier for the gender detection problem in multimedia applications is based on identifying genders in data. Other aspects, like the amount of time required for training and classification, can also impact the decision, especially given a large amount of data to be analyzed. We also aimed at identifying whether all the handcrafted features are necessary to develop a highly accurate model. For this reason, we conducted many sets of contrast tests to see whether deep features recovered by CNN or traditional handmade features performed better on the gender dataset. The amounts of speech samples allotted to male and female classes were purposefully kept equal in the dataset.

5.1. Classification Based on Manually Extracted Features. A total of 12 chromatograms, 128 Mel spectrograms, 40 MFCCs, and 12 special contrast and Tonnetz features were retrieved from the Mozilla Voice dataset. Afterward, we used the retrieved features to train machine learning algorithms. Additionally, we designed a DNN suitable for classifying the genders from the extracted features of audio data. The test accuracy is observed and analyzed for these trained models. The experimental results are summarized in Table 3. The DNN shows the best accuracy out of all the models trained. Table 4 provides a detailed performance analysis of the DNN on the test dataset. The accuracy and loss plots are also shown in Figure 9.

5.2. Classification Using CNN Models. We have carried out another experiment to observe the effectiveness of the pretrained CNN models. The models were trained on spectrogram images other than handcrafted features from the audio dataset. ResNet34 and ResNet50 models are initialized with pretrained weights and fine-tuned on features extracted from the spectrograms of each audio file. The ImageNet pretrained models were trained for 50 epochs. The learning rate was 0.001. These two comparatively complex models are of larger size, and the training time for every epoch is much longer. Table 5 displayed the performance of the two models based on the precision, recall, $F1$ score, and accuracy. The transfer learning performs better than the DNN in terms of all performance metrics. This confirmed our expectation that the lower level features learned by the convolution layers from image data can also be applied to the audio data.

Figure 10 shows the train and validation loss graphs of ResNet34 and ResNet 50 models. The more convergence is visible in the ResNet50. Figure 11 also shows the comparison of validation accuracy and error rate for these two models. The results demonstrate that the pretrained ResNet50 outperforms ResNet34. The effectiveness of each gender identification is shown in Figure 12. Male and female prediction scores are almost similar.

The study on the features suggests that the even MEL spectrogram images contain discriminating information for the gender classification. Even though the ResNet50 model is complex and took more training time, it shows impressive increment in the accuracy compared to other models trained on handcrafted features. Overall the architecture with more

convolutional layers obtained better results than all the handcrafted approaches in the given dataset.

5.3. Performance on Other Datasets. Based on the results of the aforementioned experiments, it is clear that ResNet50 is the most appropriate model for gender categorization on the Mozilla dataset. Although the model should be capable of performing well on test samples drawn from the same parent dataset that was used for training, it should also be capable of performing well on additional datasets gathered in a variety of contexts. For this reason, we conducted performance study on two separate datasets; Saarbruecken Voice Database (SVD) [31] and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [32]. We conducted a performance analysis on these two different datasets in order to better understand the generalization potential of the ResNet50 model that we constructed. The findings of the experiment are depicted in Table 6 and Figure 13.

6. Discussion

Paralinguistic analysis, which includes tasks such as age and gender detection, is a rapidly increasing research topic. Speech is a time-varying signal generated by the vocal tract, and the form of the vocal tracts varies significantly between males and females. Thus, automatic gender identification using voice has a broad range of applications. The age and gender of the speaker may be used by the interactive voice response system (IVR) to refer the speaker to an appropriate consultant [33] or to play background music suited for the speaker's gender/age group [34]. The variance in pitch across genders is relatively large, making it difficult to categorize the voice into different classes (e.g., emotions and pathologies) [1]. Hence, when used in conjunction with speaker identification, precise gender classification considerably lowers uncertainty in such classification. By narrowing the search space to a class of a particular gender, the classification algorithm will become increasingly accurate and reduce the mutual influence of each other. Compared to gender-neutral systems, automatic speech recognition (ASR) systems using gender-specific models get more excellent recognition rates [35].

We chose the Common Voice corpus for multilingual speech research for a variety of reasons, not the least of which being its immense size. We predicted that the results from this dataset would be more applicable to a wide variety of real-world applications, as it comprises a greater variety of languages and the amount of data per language varies from extremely tiny to relatively large [20].

The first set of classification studies concerns the classification of two gender audio classes using extracted features. As [36] shown, we used five frequency domain features to compare the performance of classical machine learning and the newly designed DNN model. The findings indicated that the proposed DNN achieved the highest accuracy (Table 3) and that the model converges to an optimal weight value for the gender classification problem (Figure 9).

In the second set of experiments, we trained CNNs using voice spectrograms since the human ear similarly perceives sounds in terms of varying frequencies across time [37]. Additionally, a two-dimensional representation of the speech signal serves as an appropriate input for CNN models for speech analysis [38]. This experiment is aimed at determining whether the deep learning method is appropriate for gender classification. If CNN effectively identifies the gender classes, we can bypass the bottleneck associated with manual feature extraction. According to Table 3, the pretrained ResNet50 exhibits the highest accuracy, $F1$ score, recall, and precision. Additionally, the study implies that pretrained knowledge is crucial in the network's first stages. This is because pretrained models treat spectrograms identically to images. Also, the network's fully connected layers suffer the most change, as they are task-specific [39].

When standard speech features are compared to the deep feature extracted in this study, the latter performs nearly 2% better at gender recognition than the former. As a result, the deep feature extracted in this study is more appropriate for gender representation than the manually extracted features. Additionally, when the number of layers is large, the recognition accuracy is high (Table 3). However, as the number of layers increases, the amount of calculation increases slightly and the duration of classification recognition increases [40]. As a result, we need a trade-off between accuracy and model size. Because the primary objective of this project is accuracy, ResNet50 clearly outperforms other custom and traditional models. However, if we are training models or implementing them on resource-constrained platforms with a tolerance for accuracy loss, our model choices may be different.

6.1. Comparative Analysis. There were numerous attempts on the Mozilla Common Voice dataset to classify gender. The audio samples were transformed into 20 statistical features by [41]. They trained some machine learning models and discovered that CatBoost outperforms all other predictive models with a test accuracy of 96.4%. [42] used a neural network to test the efficiency of the MFCC and Mel spectrograms. They discovered that using a combination of MFCC and Mel spectrograms, the proposed model yielded 94.32%. Following up on their previous work, the same authors [43] propose a 1-D convolutional neural network (CNN) model to recognize gender using several features taken from speech signals such as MFCC, Mel spectrogram, and Chroma. By merging the MFCC, Mel, and Chroma feature sets, the 1-D CNN model achieved a higher accuracy of 97.8%. However, our proposed model surpasses these earlier investigations (Table 7).

7. Conclusions

Recognizing the gender of the human Voice has been regarded as a difficult task due to its importance in various applications. For this study, we used Mozilla's "Common Voice" database, an open-source, multilanguage collection of voices with information about the speaker's gender. This article examines gender detection from Voice using various

machine learning methods. The study discovered that fine-tuning simple pretrained ImageNet models trained on audio spectrograms result in state-of-the-art performance on the MOZILLA dataset, as well as acceptable performance on the SVD and RAVDESS datasets. We see that pretrained models retain much of their past knowledge, particularly in the early layers during fine-tuning. Only the network's intermediate layers are significantly altered to adapt the model to the audio categorization challenge. Additionally, we discovered that CNN models learnt deep features from energy distributions in spectrograms outperformed handmade feature extraction methods. Gender discrimination results are also rather good, with an accuracy of 98.57%, which is close to the best documented in the literature.

Data Availability

The dataset used in this study is available at https://commonvoice.mozilla.org/en/datasets

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The authors extend their appreciation to the Researchers supporting project number (RSP-2021/314) King Saud University, Riyadh, Saudi Arabia.

References

[1] T.-W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.

[2] P. Rao, "Audio signal processing," in *Speech, Audio, Image and Biomedical Signal Processing Using Neural Networks*, pp. 169–189, Springer, 2008.

[3] A. Hannun, C. Case, J. Casper et al., "Deep speech: Scaling up end-to-end speech recognition," 2014, https://arxiv.org/abs/1412.5567.

[4] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2017.

[5] D. Amodei, S. Ananthanarayanan, R. Anubhai et al., "Deep speech 2: end-to-end speech recognition in english and mandarin," in *International conference on machine learning*, pp. 173–182, New York City, NY, USA, 2016.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[7] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, vol. 6, pp. 770–778, 2016.

[8] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2744–2748, Kos, Greece, 2017.

[9] S. Hershey, S. Chaudhuri, D. P. Ellis et al., "CNN architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, New Orleans, LA, USA, 2017.

[10] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, 225 pages, 2012.

[11] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal based gender identification system using four classifiers," in *2012 International conference on multimedia computing and systems*, pp. 184–187, Tangiers, Morocco, 2012.

[12] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[13] E. Yücesoy and V. V. Nabiyev, "A new approach with score-level fusion for the classification of a speaker age and gender," *Computers and Electrical Engineering*, vol. 53, pp. 29–39, 2016.

[14] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 3, no. 12, 2012.

[15] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5–14, 2017.

[16] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information science*, vol. 396, pp. 24–32, 2017.

[17] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in *2016 International Conference on Signal Processing and Communication (ICSC)*, pp. 244–249, Noida, India, 2016.

[18] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15–32, 2019.

[19] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," 2018, https://arxiv.org/abs/1807.03418.

[20] R. Ardila, M. Branson, K. Davis et al., "Common Voice: a massively-multilingual speech corpus," 2019, https://arxiv.org/abs/1912.06670.

[21] Z. Dair, R. Donovan, and R. O'Reilly, "Linguistic and gender variation in speech emotion recognition using spectral features," 2021, https://arxiv.org/abs/2112.09596.

[22] S. A. Fulop, *Speech Spectrum Analysis*, Springer Science & Business Media, 2011.

[23] D. Gabor, "Theory of communication. Part 1: the analysis of information," *The journal of the Institution of Electrical Engineers. Radio and communication engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[24] K.-I. Kanatani, *Group-Theoretical Methods in Image Understanding*, vol. 20, Springer Science & Business Media, 2012.

[25] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: a revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[26] T. Qiao, S. Zhang, Z. Zhang, S. Cao, and S. Xu, "Sub-spectrogram segmentation for environmental sound classification via convolutional recurrent neural network and score level

fusion," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 318–323, Nanjing, China, 2019.

[27] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics*, vol. 28, no. 4, pp. 357–366, 1980.

[28] W. H. Abdulla, N. K. Kasabov, and D.-N. Zealand, "Improving speech recognition performance through gender separation," *Changes*, vol. 9, p. 10, 2001.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[30] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: a preliminary study," *2017 international conference and workshop on bioinspired intelligence (IWOBI)*, 2017, pp. 1–4, Funchal, Portugal, 2017.

[31] W. J. Barry and M. Pützer, *Saarbrücken Voice Database*, Institute of Phonetics, Univ. of Saarland, 2022, http://www.stimmdatenbank.coli.uni-saarland.de/.

[32] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, article e0196391, 2018.

[33] 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.

[34] D. Mahmoodi, H. Marvi, M. Taghizadeh, A. Soleimani, F. Razzazi, and M. Mahmoodi, "Age estimation based on speech features and support vector machine," in *2011 3rd Computer Science and Electronic Engineering Conference (CEEC)*, pp. 60–64, University of Essex, UK, 2011.

[35] M. Abdollahi, E. Valavi, and H. A. Noubari, "Voice-based gender identification via multiresolution frame classification of spectro-temporal maps," in *2009 International Joint Conference on Neural Networks*, pp. 1–4, Atlanta, GA, USA, 2009.

[36] R. S. Alkhawaldeh, "DGR: gender recognition of human speech using one-dimensional conventional neural network," *Scientific Programming*, vol. 2019, Article ID 7213717, 12 pages, 2019.

[37] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," *Babel*, vol. 37, no. 4k, p. 10k, 2020.

[38] A. Tursunov, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, 2021.

[39] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN models for audio classification," 2020, https://arxiv.org/abs/2007.11154.

[40] F. Li, M. Liu, Y. Zhao et al., "Feature extraction and classification of heart sound using 1D convolutional neural networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, 11 pages, 2019.

[41] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 338–343, Madrid, Spain, 2021.

[42] K. Chachadi and S. R. Nirmala, "Voice-based gender recognition using neural network," in *Information and Communication Technology for Competitive Strategies (ICTCS 2022)*, pp. 741–749, Springer, 2022.

[43] K. Chachadi and S. R. Nirmala, "Gender recognition from speech signal using 1-D CNN," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT*, pp. 349–360, Hyderabad, India, 2022.