Age And Gender Prediction From Human Voice For Customized Ads In E-Commerce

Muhammed Shameem P
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
shameem@meaec.edu.in

Muhammed Faheem
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
19ncs06@meaec.edu.in

Muhammed Sahad V V
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
19gcs42@meaec.edu.in

Nafla Ashraf
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
19mcs33@meaec.edu.in

Shad Shaharyar
Computer Science and Engineering
MEA Engineering College
Perinthalmanna, India
19ncs04@meaec.edu.in

***Abstract*—In e-commerce, personalization is key to delivering a satisfying customer experience. Age and gender prediction from human voice can play an important role in achieving this by enabling e-commerce companies to deliver more targeted advertisements to their customers. This technology works by using machine learning models to analyse audio recordings of a person's voice and predict their age and gender. The method of predicting age and gender from the human voice, including data collection, model training, and performance evaluation, is covered in detail in this work. We also explore the potential applications of this technology in e-commerce, including personalized product recommendations and targeted advertising. Our findings suggest that age and gender prediction from human voice have the potential to significantly enhance the customer experience in e-commerce and drive business growth. Machine learning models are the basic foundation for the classification employed in this research. The age and gender categories are determined using the Convolutional Neural Network technique. On the audio data that is acquired from Kaggle, the models are methodically trained. Only the audio recordings with the corresponding labels were included in the dataset. Before being fed into the machine learning models, the noisy data from the dataset has been removed. Mel Frequency Cepstral Coefficients (MFCC) are a tool we use in our project to extract features.**

***Index Terms*—E-commerce, Personalization, Customer experience, Convolutional Neural Network, Mel Frequency Cepstral Coefficient**

I. Introduction

The e-commerce industry has been rapidly growing over the past decade, and with this growth comes a greater focus on providing customers with a personalized and satisfying experience. A crucial aspect of personalization is understanding the demographic characteristics of customers, such as their age and gender. This information can then be used to tailor the customer experience, including product recommendations and advertising, to better meet the needs and preferences of each individual. Age and gender prediction from human voice is a novel technology that uses machine learning models to analyze audio recordings of a person's voice and predict their age and gender. The process involves collecting a dataset of audio recordings, preprocessing the data, training a machine learning model, and evaluating its performance. After the model has been trained, it may be used to forecast the gender and age of fresh audio recordings, enabling e-commerce businesses to offer their clients more specialised adverts and product recommendations. In this paper, we present a comprehensive overview of age and gender prediction from human speech, including data collection, model training, and performance evaluation. Also explore the potential applications of this technology in e-commerce, including personalized product recommendations and targeted advertising. Age and gender prediction from human speech is not only beneficial for e-commerce businesses but can also be used in various other fields, such as healthcare, security, and law enforcement. Our goal is to demonstrate the value of age and gender prediction from human voice in enhancing the customer experience in e-commerce and driving business growth. We believe that this technology has the potential to revolutionize the way e-commerce companies interact with their customers and deliver personalized experiences.

II. Related Works

A. Voice based Age, Accent and Gender Recognition

The methodology in the paper [1] involves several steps to accurately recognize the age, accent, and gender of a speaker using their voice. The first step involves recording audio samples from a diverse group of speakers. The recorded audio signals were then pre-processed to remove noise and improve the quality of the signals. This involved techniques such as filtering, resampling and normalization. After preprocessing, the authors used Mel Frequency Cepstral Coefficients (MFCCs) to extract features from the audio signals, which were then used to train a deep neural network (DNN) to classify the speaker's age, accent, and gender [1]. The proposed method's performance was evaluated using metrics such as accuracy, precision,

recall, and F1-score, and it was compared to other methods in the literature to demonstrate its effectiveness. Overall, the methodology focuses on using a 2 DNN to classify speaker characteristics based on audio signals, and evaluating the performance of the proposed method. The proposed method may not generalize well to new speaker populations or unseen accents and may require additional training data to improve performance [1].

B. Age group classification and gender recognition from speech with temporal convolutional neural networks

The methodology in the paper [2] involves using a temporal convolutional neural network (TCN) to recognize the age group and gender of a speaker from their speech. The authors collected speech samples from a diverse group of speakers and labeled the data with their age group and gender. The speech signals were transformed into log Mel spectrograms, which were then used as input features for the TCN. The authors designed a TCN architecture with multiple convolutional and pooling layers to process the log Mel spectrograms and produce a prediction for the age group and gender of the speaker. The TCN was trained on the collected speech data and the extracted features using cross-entropy loss as the loss function [2]. The performance of the proposed method was evaluated using metrics such as accuracy and confusion matrices, and the results were compared with those of other methods in the literature. The recognition accuracy may be influenced by the context in which the speech was recorded, such as background noise or the presence of other speakers.

C. Gender Recognition by Voice Using an Improved Self-Labeled Algorithm

The paper [3] describes a method for recognizing gender by a voice that employs an improved self-labeled algorithm. The paper's methodology involves using a self-labeled algorithm for gender recognition by voice. This means that the algorithm is trained on a dataset of speech samples in which the gender of the speaker is known, but the algorithm is also in charge of labeling the gender of each sample. An optimized feature extraction process and a decision-making mechanism based on a combination of different classifiers are used to improve the algorithm. The algorithm's performance is evaluated and compared to other existing methods for gender recognition by voice using a publicly available dataset of speech samples. Gender recognition has limitations because it is inherently difficult and prone to errors, especially when the gender of the speaker is ambiguous or non-binary.

D. Gender and Age Estimation Methods Based on Speech Using Deep Neural Network

The paper [4] proposes a methodology for predicting the gender and age of a speaker based on speech using deep neural networks (DNNs). The proposed methodology involves collecting a dataset of speech recordings, preprocessing the data, training a DNN model, and evaluating its performance.

The paper uses a combination of long short-term memory (LSTM) and convolutional neural network (CNN) architectures to extract features from speech recordings and predict the gender and age of the speaker. The proposed methodology was tested on the publicly available VoxCeleb dataset, and the results showed that the DNN model achieved high accuracy in predicting gender and age.

However, there are several limitations to this paper. Firstly, the VoxCeleb dataset used in the paper only includes recordings of celebrities, which may not be representative of the general population. Secondly, the dataset includes only speech recordings in English, which may limit the applicability of the model to predict the gender and age of speakers in other languages.

E. Human Age Estimation Using Deep Learning from Gait Data

The paper [5] presents a method for estimating the age of individuals based on their gait data. The methodology involves collecting gait data, pre-processing it to remove noise and outliers, and feeding the data into a deep learning model, specifically a Convolutional Neural Network (CNN). The model is trained on a large gait dataset, and its performance is measured with mean absolute error (MAE) and root mean squared error (RMSE). With low MAE and RMSE values, the results show that the proposed deep learning approach is effective in estimating age from gait data. The authors aim to demonstrate the usefulness of deep learning in human age estimation, which has applications in fields such as biometrics and human- computer interaction. The method may not be applicable to all populations, as gait patterns and aging-related characteristics can vary widely across different groups.

F. Speech Enhancement Method Based On LSTM Neural Network for Speech Recognition

The methodology involves using a Long Short-Term Memory (LSTM) neural network to remove noise and other distortions from speech signals, in order to improve speech recognition performance [6]. The authors pre-process the speech signals to extract features, which are then input into the LSTM network. The network is trained on a large speech dataset with both clean and noisy speech signals, in order to learn how to enhance speech signals by removing noise and other distortions. The performance of the method is evaluated The LSTM network used in the method may be complex and computationally expensive, which may limit its practical implementation in real-world applications.using various speech recognition metrics, including Word Error Rate (WER), Frame Error Rate (FER), and Signal to Noise Ratio (SNR) [6]. The results show that the LSTM-based speech enhancement method outperforms traditional speech enhancement methods, achieving higher speech recognition accuracy.

G. Age and Gender Classification using Con- volutional Neural Networks

The methodology in the paper [7] involves using Convolutional Neural Networks (CNNs) for classifying individuals

into different age and gender groups based on their facial images. The authors first pre-process the facial images to extract features and then input these features into the CNN. CNN is trained on a large dataset of facial images labeled with age and gender information. The authors evaluate the performance of the method using various classification metrics, including accuracy, precision, recall, and F1-score, and compare it with other state-of-the-art methods. The results show that the CNN-based method outperforms other methods, 3 achieving higher accuracy and better performance on age and gender classification [7]. The accuracy and generalizability of age and gender classification based on facial photographs may be constrained because of variables including lighting, facial expressions, and individual variances.

H. Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Network

The paper [8] proposes a methodology for predicting the age of a speaker based on short speech utterances using a long short-term memory (LSTM) recurrent neural network. The proposed methodology involves collecting a dataset of short speech utterances, preprocessing the data, training an LSTM model, and evaluating its performance.

The paper uses a combination of feature extraction techniques and an LSTM model to predict the age of the speaker from short speech utterances. The proposed methodology was tested on a dataset of Mandarin Chinese speech, and the results showed that the LSTM model achieved high accuracy in predicting age.

However, there are several limitations to this paper. Firstly, the dataset used in the paper only includes speech recordings in Mandarin Chinese, which may limit the applicability of the model to predict the age of speakers in other languages. Secondly, the dataset includes only a limited number of speakers, which may not be representative of the general population.

I. Age Prediction using Image Dataset using Machine Learning

The paper [9] proposes a methodology for predicting age from facial images using machine learning techniques. The proposed methodology involves collecting a dataset of facial images, preprocessing the data, training a machine learning model, and evaluating its performance.

The paper uses a convolutional neural network (CNN) architecture to extract features from facial images and predict the age of the person in the image. The proposed methodology was tested on the publicly available Adience benchmark dataset, and the results showed that the CNN model achieved high accuracy in predicting age. The Adience dataset used in the paper only consists of a limited number of images, which may not be representative of the general population.

J. I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction

The paper [10] presents a speaker recognition method that utilizes the i-vector technique for feature extraction and dimensionality reduction. The i-vector technique involves projecting speaker-specific information onto a low-dimensional subspace, allowing for better speaker representation and discrimination. The paper proposes a method to improve the i-vector extraction process by incorporating a dimensionality reduction technique. The method is evaluated on a speaker recognition task and the results show improved speaker recognition accuracy compared to conventional i-vector extraction methods.

III. METHODOLOGY

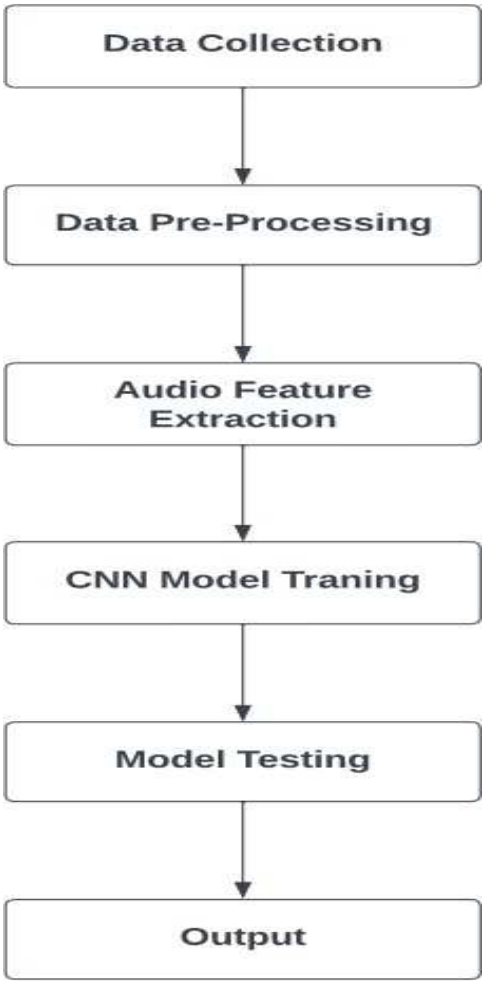

Fig. 1. Architecture diagram

A. Data Collection

In age and gender prediction from human voice, data collection is an essential step to obtain a representative and diverse dataset that can be used to train and evaluate machine learning models. The dataset used in this case is from Mozilla Common Voice, which is a crowdsourced dataset of speech recordings in multiple languages, collected by Mozilla. The Common Voice dataset is collected through a web interface that allows

individuals to record and upload their voice samples. The dataset includes information such as age, gender, accent, and other metadata associated with each recording. The Common Voice dataset is open-source, which means that it can be freely used for research and other non-commercial purposes.

The Common Voice dataset on Kaggle includes over 7,000 hours of speech data contributed by over 42,000 individuals, speaking in 60 languages. The dataset contains a variety of speech styles, accents, and dialects, making it a valuable resource for speech recognition and natural language processing research.

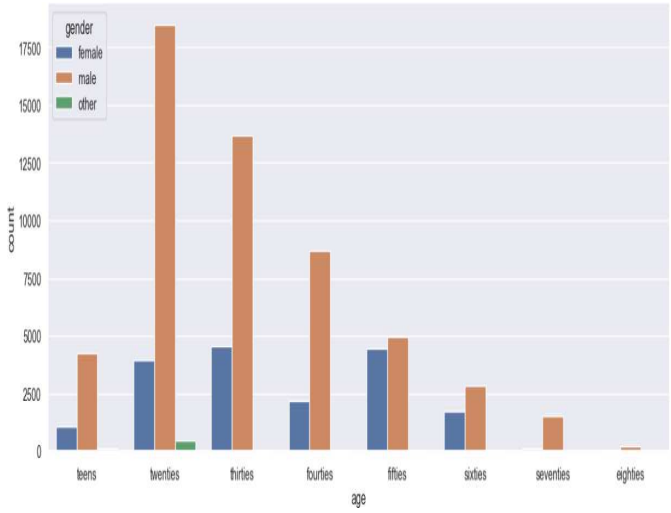

Fig. 2. Graphical representation of dataset

B. Data Preprocessing

1) Data Preparation: Data preprocessing is a crucial step in age and gender prediction from human voice as it helps to prepare the data for machine learning algorithms. The goal of data preprocessing is to transform the raw audio data into a format that can be fed into machine learning models for training and prediction. In data preparation, unknown values are removed from the dataset and converted into a new dataset with only known values.

2) Label Encoding: Teens, twenties, thirties, forties, fifties, sixties, seventies, and eighties are the age categories in the dataset. Male, female, and other are the three gender classifications. The process of transforming the categorical variables listed above into numeric values is done here.

C. Audio Feature Extraction

Audio feature extraction involves extracting meaningful features from the audio signal that can be used to predict the age and gender of the speaker. The proposed system's audio features includes:

1) Mel-frequency cepstral coefficients (MFCCs): The MFCC coefficients capture the spectral envelope of the audio signal, which is related to the shape of the vocal tract and can be used to distinguish between different speakers, genders, and age groups. Males and females tend to have different formant frequencies, which are related to the resonant frequencies of the vocal tract, and these differences can be captured by the MFCC coefficients. Similarly, as a person ages, there may be changes in the shape of the vocal tract that can also be captured by the MFCC coefficients.

Therefore, by extracting MFCC features from the audio signals in the dataset, we can use these features as input to machine learning algorithms to predict the gender and age of the speaker.

2) Spectral centroid: The spectral centroid represents the frequency around which most of the spectral energy is concentrated and can be used as a feature to capture differences in the vocal characteristics of males and females, as well as changes in vocal characteristics as a person ages.

Mean fundamental frequencies in male and female typically differ (or pitch), which can be captured by the spectral centroid feature. Males generally have a lower pitch than females, resulting in a lower spectral centroid value for male voices compared to female voices. Similarly, as a person ages, there may be changes in the pitch of their voice due to changes in the length and tension of their vocal cords, which can also be captured by the spectral centroid feature.

3) Spectral bandwidth: The spectral bandwidth represents the width of the frequency band around the spectral centroid that contains a certain percentage of the spectral energy and can be used as a feature to capture differences in the vocal characteristics of males and females, as well as changes in vocal characteristics as a person ages.

Males and females tend to have different spectral bandwidths due to differences in the shape and size of their vocal tracts. Females generally have a narrower bandwidth than males, resulting in a higher spectral bandwidth value for female voices compared to male voices. Similarly, as a person ages, there may be changes in the shape and size of their vocal tract that can also be captured by the spectral bandwidth feature.

4) Spectral rolloff: The spectral rolloff represents the point in the frequency spectrum where a certain percentage of the total spectral energy is contained in the lower frequencies and can be used as a feature to capture differences in the vocal characteristics of males and females, as well as changes in vocal characteristics as a person ages.

Spectral rolloff values between males and females typically differ due to differences in the shape and size of their vocal tracts. Females generally have a higher spectral rolloff value than males, indicating that most of the spectral energy is concentrated in the higher frequencies. Similarly, as a person ages, there may be changes in the shape and size of their vocal tract that can also be captured by the spectral rolloff feature.

5) Glottal features: Glottal features can be used to capture the characteristics of the glottal source signal, which is the signal generated by the vocal folds during speech production. The glottal source signal contains information about the shape and tension of the vocal folds and can be used as a feature to capture differences in the vocal characteristics of males and

females, as well as changes in vocal characteristics as a person ages.

Men and women often have different mean fundamental frequencies, or pitch, which can be captured by glottal features. Males generally have a lower pitch than females, resulting in differences in the shape and tension of their vocal folds, which can be captured by glottal features. Similarly, as a person ages, there may be changes in the tension and elasticity of the vocal folds, which can also be captured by glottal features.

6) Rhythm features: Rhythm features can be used to capture the temporal patterns and characteristics of speech production. Rhythm features can be used to capture differences in the vocal characteristics of males and females, as well as changes in vocal characteristics as a person ages.

Males and females have different speaking rates and patterns of stress and intonation, which can be captured by rhythm features. Males generally have a slower speaking rate than females and tend to use more monotonous intonation patterns, which can be captured by rhythm features. Similarly, as a person ages, there may be changes in their speaking rate and patterns of stress and intonation, which can also be captured by rhythm features.

D. CNN Model Training

The extracted audio features, such as MFCCs, spectral centroid, spectral bandwidth, spectral rolloff, glottal features, and rhythm features, are combined into a single input feature vector that is fed into the CNN model. The model then learns to identify patterns in the combination of features that are indicative of age and gender.

The CNN model used for this task may consist of multiple convolutional and pooling layers, followed by one or more fully connected layers. The input feature vector is fed into the first convolutional layer, which convolves the filters over the input feature vector to produce feature maps. The pooling layers then perform downsampling on the feature maps to reduce their dimensionality while retaining the most important features.

The fully connected layers take the high-level features learned by the previous layers and use them to predict the age and gender of the speaker. The model is trained using an optimization algorithm such as stochastic gradient descent (SGD) to minimize the difference between the predicted and actual labels of the training data.

The training process usually involves splitting the labeled dataset into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate the model's performance and prevent overfitting. The model is trained for 100 epochs.

E. Model Testing

After training, a testing dataset is used to evaluate the CNN model's ability to predict gender and age from voice. The test dataset is a separate set of audio recordings that the model has not seen before during training.

During the testing phase, the CNN model takes the extracted audio features as input and makes age and gender predictions for each audio recording in the test dataset. The model's predictions are then compared to the actual age and gender labels of the test dataset to evaluate the model's accuracy.

F. Output

The output of the CNN model for age and gender prediction from voice is a set of predicted age and gender labels for each input audio recording. Specifically, the model outputs a probability distribution over the possible age and gender labels for each audio recording.

For age prediction, the model may output a probability distribution over a category of ages, such as teens, twenties, thirties, forties, fifties, sixties, seventies, and eighties. The predicted age label for each audio recording is typically the age range with the highest probability.

For gender prediction, the model may output a probability distribution over three possible labels, male, female, and others. The predicted gender label for each audio recording is typically the label with the highest probability.

IV. Result

With 100 epochs, the model showed an accuracy of around 70%. There are 73,767 audio samples in the collection. 20% of the dataset is for testing, while the remaining 80% is for training. Fig. 3 displays the model's training and testing accuracy. The accuracy score is shown on the y-axis, while the number of epochs is shown on the x-axis. Fig. 4 depicts the model's training loss and validation loss. The y-axis displays the amount of loss, while the x-axis displays the number of epochs. The confusion matrix for the trained model is shown in Fig. 5. It is employed to assess how well the trained machine learning model is performing. For each class in the prediction, the number of accurate predictions is displayed.

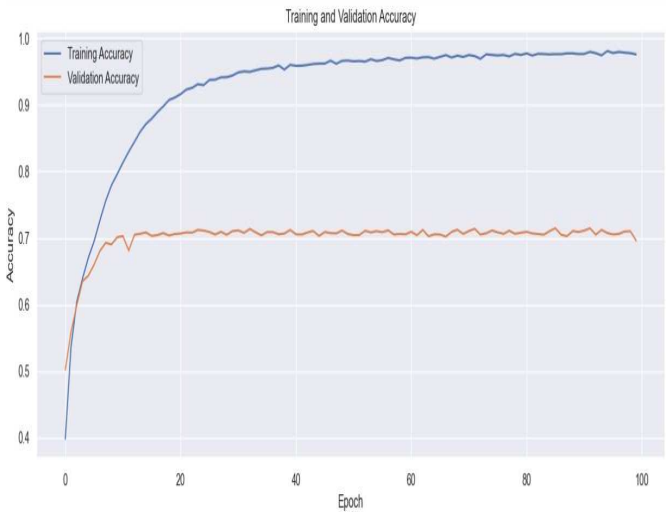

Fig. 3. Training loss and Validation loss

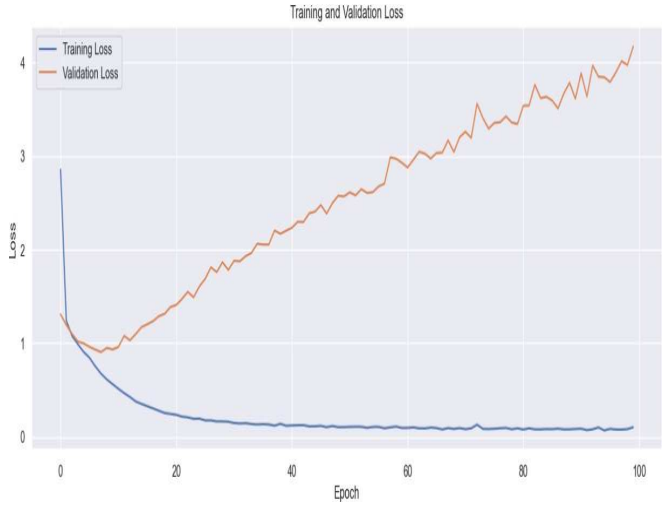

Fig. 4. Training loss and Validation loss

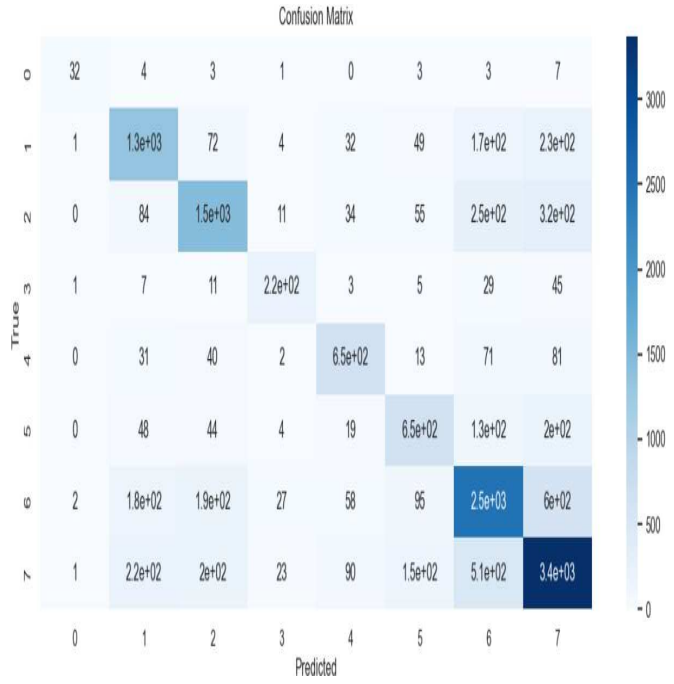

Fig. 5. Confusion Matrix

V. Conclusion

In conclusion, the prediction of age and gender from human voice has the potential to improve the relevance and effectiveness of e-commerce ads. By using machine learning models trained on audio signals, it is possible to accurately predict the age and gender of speakers and tailor ads to their demographic characteristics. This approach can result in a better user experience, increased engagement with ads, and higher conversion rates for e- commerce companies. However, it is important to consider the ethical implications of using personal information such as age and gender in e-commerce ads. Users' privacy must be protected, and data must be collected and used in accordance with applicable laws and regulations. Additionally, the accuracy of the predictions must be rigorously evaluated and validated to avoid any negative consequences such as discrimination or biased outcomes. Overall, age and gender prediction from human voice is a promising area of research and has the potential to bring significant benefits to the e-commerce industry. With careful consideration of the technical and ethical implications, this technology has the potential to enhance the user experience and drive business results.

Acknowledgment

We would like to express our gratitude to the research participants who generously shared their voice samples and demographic information. We also thank our colleagues and mentors for their guidance and insights, and our families and friends for their unwavering support. We hope that our findings on age and gender prediction from voice for customized ads in e-commerce will contribute to the field of personalized advertising.

References

[1] Deepa Angadi; Manoj K R; Nagendra N S; Nithin Kumar B, "Voice based Age, Accent and Gender Recognition", July 2021, Computer Science and Engineering, MVJ College of Engineering, Bengaluru.
[2] Hector A. SanchezHevia; Roberto Gil-Pita; Manuel Utrilla-Manso; Manuel Rosa- Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks", 22 September 2021.
[3] Ioannis E. Livieris, Emmanuel Pintelas, Panagiotis Pintelas, "Gender Recognition by Voice Using an Improved Self-Labeled Algorithm", 5 March 2019
[4] Damian Kwasny, Daria Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks" ,13 July 2021, Department of Measurement and Electronics, AGH University of Science and Technology.
[5] Refat Khan Pathan; Mohammad Amaz Uddin; Nazmun Nahar; Ferdous Ara; Mohammad Shahadat Hossain; Karl Andersson, "Human Age Estimation Using Deep Learning from Gait Data", 2021.
[6] Ming Liu, Yujun Wang, Jin Wang, Jing Wang, Xiang Xie, "Speech Enhancement Method Based On LSTM Neural Network for Speech Recognition", 08 October 2019.
[7] Gil Levi ; Tal Hassner, "Age and Gender Classification using Convolutional Neural Networks ", 2021, Department of Mathematics and Computer Science, The Open University of Israel.
[8] Ruben Zazo ; Phani Sankar Nidadavolu; Nanxin Chen; Joaquin Gonzalezrodriguez; Najim Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks", 15 march 2018
[9] Ishita Verma; Urvi Marhatta; Sachin Sharma; Vijay Kumar, "Age Prediction using Image Dataset using Machine Learning", 23 July 2020, International Journal of Innova- tive Technology and Exploring Engineering (IJITEE).
[10] Noor Salwani Ibrahima; Dzati Athiar Ramlia, "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction", 2018, 22nd International Conference on Knowledge-Based and Intelligent Information Engineering Systems.
[11] Abraham Woubie Zewoudie, Jordi Luque, Javier Hernando. The use of long-term features for GMM- and i-vector-based speaker diarization systems. EURASIP Journal on Audio, Speech, and Music Processing, 2018.
[12] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling tech- niques of speakers," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), vol. 1. May 2002, pp. I-137–I-140.