Intechno Journal :
Information Technology Journal
Vol. 8 No. 1 July 2026
Homepage: https://jurnal.amikom.ac.id/index.php/intechno
INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL
e-ISSN: 2655-1438
p-ISSN: 2655-1632

Robust Deep Learning Approach for Automatic Age and Gender Recognition Based on Voice

M. Nasyid Yunitian Rizal[1], Theopilus Bayu Sasongko[2], Arifiyanto Hadinegoro[3], Kumara Ari Yuana[4], Wahid Miftahul Ashari[5]

[1,2,3,4]Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
[5]Computer Engineering, Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received: 17 June 2026
Revised: 2 July 2026
Accepted: 10 July 2026
Published: 31 July 2026

Keywords:

Age and Gender Recognition
Deep Learning
Bi-LSTM
CNN
Transformer

ABSTRACT

Voice-based age and gender recognition plays an important role in biometric authentication, personalized human–computer interaction, and digital forensic applications. However, existing single deep learning architectures often struggle to simultaneously capture local acoustic patterns, temporal dependencies, and long-range contextual information from speech signals. This study proposes a hybrid CNN–BiLSTM–Transformer framework to improve the accuracy and robustness of multi-class age and gender classification. The proposed approach employs Mel-spectrogram representations generated from the Mozilla Common Voice dataset, followed by audio standardization and feature extraction. A Convolutional Neural Network (CNN) enhanced with Squeeze-and-Excitation blocks extracts discriminative spectral features, a Bidirectional Long Short-Term Memory (Bi-LSTM) network models bidirectional temporal dependencies, and a Transformer Encoder captures global contextual relationships through a multi-head self-attention mechanism. The model was evaluated on 40,392 speech samples across 12 age–gender categories. Experimental results achieved an overall classification accuracy of 91%, outperforming standalone CNN and Bi-LSTM models, which obtained accuracies of 84% and 74%, respectively. In addition, the proposed model demonstrated balanced performance with macro-average precision, recall, and F1-scores of 0.91, 0.92, and 0.91. The novelty of this research lies in the integration of complementary spatial, temporal, and global attention mechanisms within a unified architecture for large-scale multi-class voice-based demographic classification, providing an effective and scalable solution for intelligent biometric and speech analysis systems.

Corresponding Author:

Theopilus Bayu Sasongko
Email: theopilus.27@amikom.ac.id

1. INTRODUCTION

Rapid advancements in information technology have significantly transformed the way humans interact with machines. Modern intelligent systems are expected not only to execute commands but also to understand user characteristics in order to provide more personalized and adaptive services. Among these characteristics, the automatic recognition of age and gender has become a fundamental capability [1]. By identifying user demographics, systems can dynamically adjust communication styles, content delivery, and service strategies, creating new opportunities to enhance user interaction across various industries. Ultimately, these technologies aim to establish more natural and intelligent human–machine interactions, thereby improving user satisfaction

and long-term engagement. The practical applications of age and gender recognition are extensive and have been adopted in a wide range of real-world scenarios, from content personalization to access control systems. In the e-commerce sector, for instance, voice assistants can analyze users' speech to estimate demographic attributes and provide product recommendations or targeted advertisements tailored to specific age and gender groups, an approach that has been shown to improve user engagement and purchasing decisions [2].

Another important application is in automotive security systems, where age detection can be used to restrict unauthorized access. For example, the system can distinguish between adult and child voices to prevent children from starting a vehicle without supervision, thereby enhancing safety and reducing potential risks [3]. Previous studies have demonstrated the effectiveness of deep learning architectures for speech-based demographic classification. An early study investigated Recurrent Neural Network (RNN) architectures, specifically comparing Simple RNN and Long Short-Term Memory (LSTM) for gender classification. The results showed that LSTM outperformed conventional sequential models in capturing speech patterns. Importantly, the authors suggested that combining multiple architectures could further improve feature representation, providing a strong motivation for the development of hybrid models [4].

Another study employed a pure Convolutional Neural Network (CNN) by treating speech signals as spectrogram images. CNN achieved promising performance in extracting spatial features for age and gender classification. However, its main limitation lies in modeling speech as static visual patterns, making it less effective at capturing long-term temporal dependencies. Consequently, CNN serves as an effective feature extractor but requires complementary temporal modeling techniques [5]. Transformer-based approaches have also demonstrated competitive performance, particularly in gender prediction, due to their ability to capture global contextual information across entire utterances. Nevertheless, the use of average pooling in many Transformer architectures may lead to the loss of fine-grained local sequential information. This limitation motivates the integration of Bi-Directional Long Short-Term Memory (Bi-LSTM) to preserve local temporal dependencies before global attention modeling [6].

A closely related study proposed a CNN-BiLSTM-Transformer architecture for medical speech classification. The experimental results confirmed that the combination of CNN for spatial feature extraction, Bi-LSTM for bidirectional temporal modeling, and Transformer for global attention achieved superior and more stable performance than individual architectures. This framework provides the primary architectural reference for the present study, which extends its application to age and gender classification using the Common Voice dataset [7]. The superiority of hybrid architectures has been further validated by comparative studies in speech emotion recognition. CNN-LSTM models consistently outperformed standalone CNN and Transformer models, indicating that combining complementary learning mechanisms is more effective than relying on a single architecture [8]. Similarly, another study directly compared CNN-LSTM and pure CNN models for gender and age classification reported significant performance improvements with the hybrid approach. This study using multi corpus benchmarking [9].

Another relevant work introduced a parallel CNN-Transformer framework, where CNN extracted local features and Transformer modeled global representations, achieving promising results on the Common Voice dataset. Although effective, the parallel design did not explicitly address sequential temporal dependencies. To overcome this limitation, the present study incorporates Bi-LSTM as an intermediate temporal modeling component, enabling comprehensive local, sequential, and global feature learning [10]. Finally, an ensemble learning approach achieved high accuracy on the Common Voice dataset. Despite its strong performance, the study was limited to a single-language dataset (Turkish) and excluded adolescent age groups, which present unique challenges due to voice changes during puberty. Moreover, ensemble models generally incur higher computational costs, reducing their suitability for real-time applications. In contrast, the present study proposes a computationally efficient single hybrid CNN-BiLSTM-Transformer architecture designed to handle multilingual and multi-accent speech data while incorporating adolescent age categories, thereby addressing broader demographic classification challenges [11]. To achieve accurate voice-based age and gender classification, numerous studies have explored a variety of deep learning approaches. However, conventional sequential models such as Recurrent Neural Networks (RNNs) suffer from the vanishing gradient problem, limiting their ability to capture long-term temporal dependencies and resulting in degraded performance for longer speech recordings [12]. Furthermore, Mel-spectrograms exhibit two-dimensional time-frequency representations with rich spatial patterns, making RNN-based architectures less effective at directly extracting local spectral features [13].

Convolutional Neural Networks (CNNs) have demonstrated superior capability in learning local spatial representations from spectrogram images; nevertheless, their limited temporal modeling capacity restricts their

effectiveness in capturing long-range sequential dependencies when used independently[14]. Meanwhile, Transformer architectures have shown remarkable success in modeling global contextual relationships through self-attention mechanisms, although recent studies suggest that their performance can be further enhanced by integrating sequential models such as Bidirectional Long Short-Term Memory (Bi-LSTM) networks to capture fine-grained local temporal dynamics [15].

Motivated by these limitations, the main contribution proposes a hybrid CNN–BiLSTM–Transformer architecture for voice-based age and gender classification are summarized as follows:

1. A novel hybrid CNN–BiLSTM–Transformer architecture is proposed for simultaneous voice-based age and gender classification, enabling complementary learning of local spectral features, bidirectional temporal dependencies, and long-range contextual information within a unified end-to-end framework.
2. A modified CNN feature extractor incorporating Squeeze-and-Excitation (SE) blocks and GELU activation is introduced to enhance discriminative spectral feature learning by adaptively emphasizing informative channels while suppressing irrelevant responses.
3. A comprehensive experimental evaluation iss conducted on the Mozilla Common Voice dataset consisting of 40,392 speech samples distributed across 12 age–gender classes, demonstrating the effectiveness of the proposed architecture over individual CNN and Bi-LSTM baseline models.

By integrating complementary spatial, temporal, and global contextual representations, the proposed hybrid architecture addresses the inherent complexity and variability of human speech signals more effectively than single-model approaches[16]. Consequently, the proposed model is expected to provide a more comprehensive representation of vocal characteristics, leading to significant improvements in age and gender classification performance. The rest of this study will then review related works, Section 3 describes the proposed approach method, Section 4 presents the empirical results and discussion, the Last section describes the conclusions and future work.

2. METHOD

This study develops a Hybrid CNN–BiLSTM–Transformer model for voice-based age and gender classification. CNN is utilized for feature extraction from speech representations, BiLSTM captures contextual temporal information in both forward and backward directions, and the Transformer enhances the model by focusing on the most relevant speech characteristics. The experiments were conducted using the Mozilla Common Voice dataset, a publicly available speech corpus obtained from Kaggle.

2.1 Data Collection

This study employed the publicly available Common Voice dataset, sourced from the Mozilla Common Voice project[17] and accessed through Kaggle. The dataset consists of speech recordings contributed by speakers from diverse demographic backgrounds and validated through a crowdsourcing process. Audio files are provided in MP3 format, while the corresponding metadata are stored in CSV (Comma-Separated Values) files. The Common Voice dataset is divided into three subsets: training, development, and testing. This research specifically utilizes the demographic attributes of age and gender as target labels. Age labels are categorized into predefined age groups, such as teens, twenties, thirties, and subsequent ranges. The dataset was selected due to its substantial diversity in speaker characteristics, including variations in accent, speaking style, and natural recording environments. Such diversity enhances the robustness and generalizability of the proposed model, enabling it to effectively recognize a wide range of voice characteristics rather than being limited to a specific speaker profile. In total, the dataset contains 203,847 valid raw speech samples available for analysis. A detailed description of the dataset attributes is presented in Table 1.

Table 1. Description of Common Voice Dataset Attributes

No	Common Voice Dataset		
	Feature	Description	Data Type
1	filename	Relative file path or audio filename (.mp3 format)	Object (String)
2	text	Text transcription of the utterance contained in the audio recording	Object (String)
3	up_votes	Number of validators that declared the audio matches the text	Integer
4	down_votes	Number of validators that declared the audio does not match the text	Integer
5	age	Speaker age range category (example: teens, twenties, thirties)	Object (Categorical)

6	gender	Speaker's gender (male, female, other)	Object (Categorical)
7	accent	Speaker accent (examples: US, Australia, India)	Object (Categorical)

2.2 Preprocessing Data

First, the three subsets of the Common Voice dataset (train, development, and test) were merged into a single unified dataset to provide a comprehensive representation of the overall data population. From the initial 203,847 raw samples, a data cleaning process was performed by removing records with missing values in the essential target attributes, namely age and gender. This preprocessing step resulted in 76,522 valid samples for subsequent analysis. The age distribution was then examined to identify potential class imbalance. Statistical analysis revealed a substantial disparity among age groups, with the twenties category being the most represented, comprising 23,874 samples. In contrast, older age groups were significantly underrepresented, with only 1,705 samples for the seventies category and 248 samples for the eighties category. Following the age grouping process, the dataset was further analyzed to evaluate the balance among class groups defined by the combination of age and gender labels. The resulting class distribution is presented in Table 2.

Table 2. Distribution of Age and Gender Class Groups

Class Group	Number of Samples	Description
Male 20s	19,259	Largest sample group
Male 30s	14,213	–
Male 40s	9,048	–
Male 50s	5,178	–
Female 30s	4,740	–
Male Seniors (60+)	4,733	Combined age groups (60s–90s)
Female 50s	4,643	–
Male Teens	4,415	–
Female 20s	4,122	–
Female 40s	2,294	–
Female Seniors (60+)	1,959	Combined age groups (60s–90s)
Female Teens	1,122	Smallest sample group
Total	**75,726**	–

Although the dataset presents a naturally imbalanced class distribution, the proposed hybrid model is designed to mitigate this issue through its robust architecture. The Transformer attention mechanism and the Squeeze-and-Excitation (SE) blocks integrated into the CNN adaptively highlight salient vocal characteristics of each class, allowing the model to capture discriminative patterns from minority classes without the need for data reduction or aggressive resampling strategies, thereby preserving valuable information and improving generalization.

2.3 Audio Standardization

The standardization parameters used in this study include sampling rate resampling, audio channel conversion, and fixed-duration setting. The detailed configuration of these parameters is presented in Table 3 below.

Table 3. Audio Standardization

Parameter	Configuration Value
Sampling Rate	16,000 Hz (16 kHz)
Channel	1 (Mono)
Duration	4 seconds
Total Samples	64,000 samples

Based on Table 3, the first configuration parameter defined in this study is the sampling rate, set to 16,000 Hz. This value is selected because it is considered optimal for processing human speech. Speech signals are typically well captured at this sampling rate without requiring the higher fidelity used in music processing (e.g.,

44.1 kHz). Using an excessively high sampling rate would significantly increase data size and computational cost without providing substantial improvements in predictive performance. In addition, the audio is converted to a single-channel (mono) format. This decision is made because the task focuses on extracting vocal characteristics such as tone and timbre for age and gender classification, rather than spatial audio information (e.g., left or right channel direction). Converting to mono reduces storage requirements and improves computational efficiency. Next configuration involves standardizing the audio duration to 4 seconds, resulting in a fixed-length input of 64,000 samples per audio file. This duration is chosen as a balanced trade-off: it is long enough to capture meaningful speech characteristics such as speaking style and vocal tone, while still maintaining efficient model training and computation time. To achieve uniform input length, two preprocessing strategies are applied. If an audio recording exceeds 4 seconds, it is truncated by retaining only the initial segment, which is assumed to sufficiently represent the speaker's vocal characteristics. Conversely, if the recording is shorter than 4 seconds, zero-padding (silence padding) is applied at the end of the signal until the required duration is reached. This ensures that all audio inputs have consistent length, enabling stable and efficient training of the deep learning model.

2.4 Feature Extraction

The final stage of the data preprocessing pipeline involves transforming the raw audio signals into a numerical representation that can be effectively processed by deep learning models, as well as preparing the corresponding target labels. Raw audio waveforms have high dimensionality and are not directly suitable for efficient model training. Therefore, the audio data is converted into Mel-spectrogram representations, which provide a more compact and informative feature space for learning. The mathematical formulation used to convert the original frequency unit in Hertz (f) into the Mel scale (m) is given by the following equation (1):

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

m represents the frequency value in the Mel scale, while f denotes the original frequency value measured in Hertz (Hz). The selection of parameter values is crucial to preserve important speech features during the extraction process. The parameter configuration employed in this study is presented in Table 4.

Table 4. Mel-Spectrogram Feature Extraction Parameters

Parameter	Configuration Value	Function
Sample Rate	16,000 Hz	Standard sampling frequency for human speech processing.
N_MELS	128	Number of Mel-frequency bins representing spectral resolution.
N_FFT	2048	Window length used for the Fourier Transform computation.
Hop Length	512	Frame shift interval between consecutive time windows.
Scale (dB)	Logarithmic (dB)	Decibel scale used to enhance weak speech signal components.

The feature extraction parameters were optimized to balance acoustic detail and computational efficiency. A sampling rate of 16 kHz was employed to preserve essential speech information while minimizing storage and processing costs. The combination of N_FFT and N_MELS was selected to produce high-resolution Mel-spectrograms capable of capturing salient vocal characteristics, including pitch and timbre, without significantly increasing training complexity. Subsequently, the Mel-spectrogram amplitudes were transformed into the logarithmic decibel (dB) scale to enhance informative speech patterns and reduce the influence of noise. Finally, the feature matrix was transposed into a [Time, Frequency] representation to align with the sequential processing requirements of the Bi-LSTM component within the proposed hybrid CNN–BiLSTM–Transformer architecture. Alongside feature extraction, a labeling process was conducted to support the multi-output classification framework. Each audio sample was assigned two target labels: age and gender. Age categories were encoded into six numerical classes (0–5), whereas gender was represented as a binary variable (0 for male and 1 for female). The alignment between feature data and corresponding labels was strictly preserved to ensure dataset consistency and reliable model training.

2.5 Proposed Method

2.5.1 Overall Architecture

The proposed architecture aims to improve the accuracy of voice-based age and gender classification by jointly capturing local spectral characteristics, long-term temporal dependencies, and global contextual relationships from speech signals. The model integrates the complementary strengths of Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM) networks, and Transformer encoders into a unified hybrid framework. Figure 1 show model proposed.

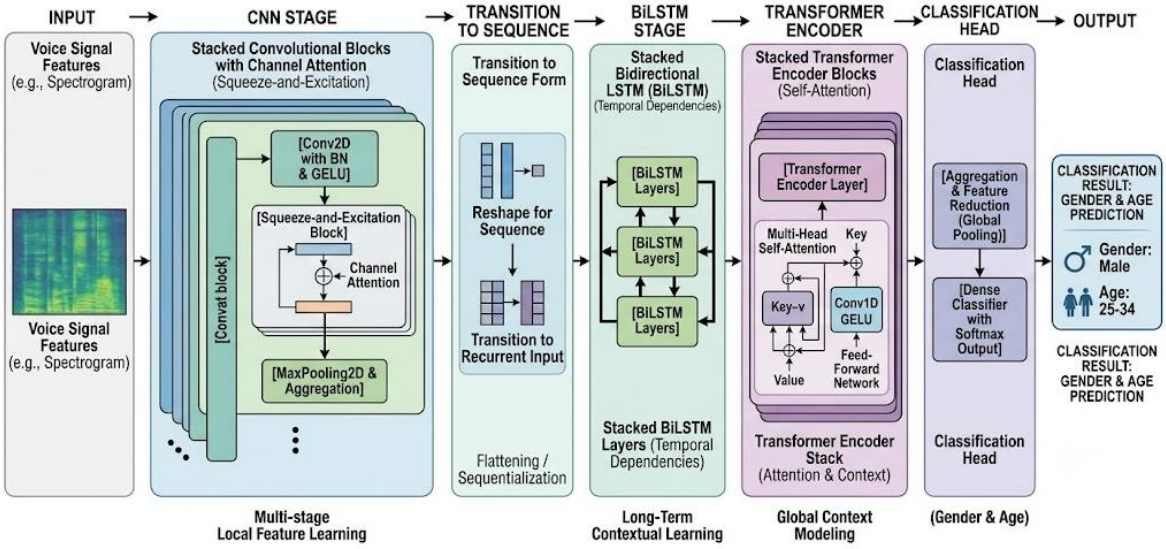

Figure 1. Proposed model CNN-BiLSTM-Transformer for Voice Classification

The proposed framework for voice-based age and gender classification in Figure 1 consists of three sequential feature-learning stages: a modified Convolutional Neural Network (CNN), a Bidirectional Long Short-Term Memory (Bi-LSTM) network, and a Transformer Encoder. Voice-based age and gender classification requires the extraction of complementary information from speech signals. Conventional convolutional neural networks (CNNs) effectively capture local spectral patterns but are limited in modeling long-range temporal dependencies. Conversely, recurrent architectures such as Long Short-Term Memory (LSTM) networks capture sequential information but often struggle to preserve discriminative spatial representations. More recently, Transformer-based models have demonstrated remarkable capability in modeling global contextual relationships; however, they generally require informative feature representations as input and are computationally expensive when applied directly to raw speech features. To overcome these limitations, this study proposes a hybrid CNN-BiLSTM-Transformer architecture that combines the strengths of convolutional feature extraction, bidirectional temporal modeling, and global attention mechanisms within a unified framework. The proposed architecture first extracts discriminative local spectral representations using a modified CNN equipped with channel attention, subsequently models bidirectional temporal dependencies using Bi-LSTM, and finally enhances long-range contextual learning through a Transformer Encoder before performing joint age and gender classification.

2.5.2 CNN Stage: Local Spectral Feature Learning

Speech signals represented as Mel-spectrograms contain abundant local spectral structures, including harmonic components, formant distributions, and energy transitions that are closely associated with speaker age and gender. Conventional CNNs have demonstrated excellent capability in extracting these local discriminative patterns. However, standard convolutional operations assign equal importance to all feature channels, making them susceptible to redundant information and environmental noise. To address this limitation, the proposed CNN incorporates Squeeze-and-Excitation (SE) blocks, enabling adaptive channel-wise attention following equation (2-4).

Given an input Mel-spectogram $X \in \mathbb{R}^{HxWxC}$ (2)

a convolution operation produces $F = \sigma(W * X + b)$ (3)

Where W denotes convolution kernels, b is the bias, $\sigma(.)$ represents the GELU activation. Instead of ReLU, the proposed model employs the Gaussian Error Linear Unit (GELU) activation function.

$$GELU\ (x) = x\Phi(x) = \frac{x}{2}\left(1 + \tanh(\sqrt{\frac{2}{n}}(x + 0.004715x^3)\right) \quad (4)$$

Where $\Phi(x)$ denotes the cumulative distribution function of the standard Gaussian distribution. Compared with ReLU, GELU provides smoother nonlinear transformations and preserves small informative activations

that are frequently observed in speech signals. The SE block first performs global average pooling in equation (5-6)

$$Z_c = \frac{1}{HW}\left(\sum_{i=1}^{H}\sum_{j=1}^{W} F_c(i,j)\right) \quad (5)$$

followed by channel recalibration

$$s = \sigma\,(W_2\delta(W_1 z)) \quad (6)$$

Where W_1 and W_2 are fully connected layers, δ denotes ReLU and σ is the sigmoid activation. The recalibrated feature map becomes $\hat{F}_c = scF_{c.}$ Finally, MaxPooling reduces spatial resolution while Spatial Dropout improves generalization. Compared with conventional CNNs, the proposed CNN stage emphasizes informative frequency channels while suppressing irrelevant responses, thereby improving robustness against background noise and speaker variability. Furthermore, the combination of GELU activation and channel attention facilitates richer nonlinear feature representations without significantly increasing computational complexity. Although CNN effectively extracts spatial representations, recurrent networks require sequential inputs. Therefore, the feature maps produced by CNN are transformed into temporal sequences before entering the Bi-LSTM module. This reshaping operation preserves spectral representations while enabling recurrent modeling of temporal evolution across the speech sequence.

2.5.3 Bi-LSTM Stage: Bidirectional Temporal Modeling

Age and gender characteristics are not determined solely by instantaneous spectral features but also by temporal dynamics, including speaking rate, intonation, and articulation patterns. Standard LSTM processes information only in the forward direction, potentially neglecting contextual information from future frames. The proposed model employs stacked Bi-LSTM layers. Forward hidden states, backward hidden states until final representation describe in equations (7-9).

$$\vec{h}_t = LSTM_f(x_t, \vec{h}_{t-1}) \quad (7)$$

$$\overleftarrow{h}_t = LSTM_b(x_t, \overleftarrow{h}_{t+1}) \quad (8)$$

$$h_t = [\vec{h}_t\,;\,\overleftarrow{h}_t] \quad (9)$$

This concatenation enables simultaneous utilization of past and future contextual information. Compared with unidirectional LSTM, Bi-LSTM captures richer temporal dependencies and improves robustness when speaker characteristics are distributed across multiple speech segments.

2.5.4 Transformer Encoder: Global Context Modeling

Although Bi-LSTM effectively models sequential dependencies, its recurrent nature makes learning very long-range relationships difficult. To alleviate this limitation, the proposed architecture incorporates a Transformer Encoder. The Transformer employs Multi-Head Self-Attention describe in equations (9-10).

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

$$MultiHeadAttentions = Concat(\,head_1, \ldots head_h)W^O \quad (10)$$

Each encoder block further applies residual connection, Layer Normalization, position-wise feed-forward network. Unlike recurrent models, the Transformer directly models interactions between distant temporal positions, allowing the network to identify globally discriminative speech regions associated with speaker age and gender.

2.5.5 Classification Head

The extracted contextual representations must be aggregated into a compact embedding suitable for classification. Global Average Pooling computes in equations (11-13)

$$g = \frac{1}{T}\sum_{t=1}^{T} h_t \quad (11)$$

$$y = softmax(W_g + b) \quad (12)$$

Global Average Pooling significantly reduces the number of trainable parameters while preserving the global semantic information extracted by the Transformer. The final Softmax classifier performs joint multi-class prediction, enabling simultaneous estimation of age and gender within a unified end-to-end framework.

2.6 Experimental Setup

The experiments were conducted on a workstation equipped with an Intel Core i7-8750H processor (2.20 GHz, 12 logical cores), 16 GB RAM, and an NVIDIA GeForce GTX 1060 GPU with 6 GB of VRAM. The

proposed hybrid CNN–BiLSTM–Transformer model was implemented using the TensorFlow deep learning framework with GPU acceleration. A relatively small batch size of 16 was selected to accommodate the 6 GB GPU memory limitation, while the maximum number of training epochs was set to 80. Model regularization was achieved through L2 weight decay with a coefficient of 0.0001, multiple Dropout layers, and label smoothing. The network was optimized using the Adam optimizer with an initial learning rate of 0.0005, which was dynamically adjusted using a Cosine Decay learning rate scheduler throughout training. To address class imbalance and improve classification robustness, the model employed Categorical Focal Cross-Entropy loss with parameters $\alpha = 0.25$ and $\gamma = 2.0$, together with a label smoothing factor of 0.1.

2.7 Evaluation Matrix

The proposed model was evaluated on the test set using accuracy, precision, recall and F1-Score metrics, as defined in Eqs. (2)-(5), to assess its classification performance and generalization capability.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1-Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+\ Recall} \quad (5)$$

3. RESULTS AND DISCUSSIONS

The proposed hybrid CNN–BiLSTM–Transformer architecture was designed to enhance voice-based age and gender classification by simultaneously modeling local acoustic features, sequential temporal patterns, and long-range contextual dependencies in speech signals. By leveraging the complementary advantages of CNNs, BiLSTM networks, and Transformer encoders, the model provides a comprehensive representation of speaker characteristics for more reliable classification. The performance of the proposed model during the learning process was evaluated using the training and validation accuracy curves, which are illustrated in Figure 2

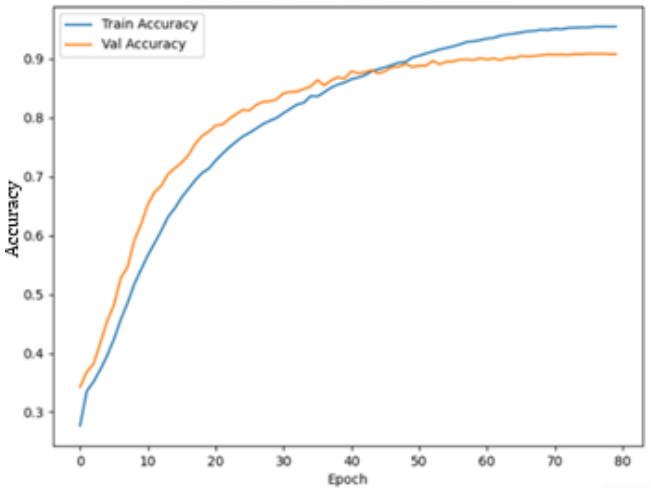

Figure 2. Train vs Val Accuracy during training

As illustrated in Figure 2, the learning curves exhibit stable and convergent training behavior, with no evident indication of overfitting. The training accuracy (blue curve) gradually converges to approximately 95% rather than reaching perfect accuracy, reflecting the effectiveness of employing the complete original dataset together with regularization strategies, including Dropout and Label Smoothing. Meanwhile, the validation accuracy (orange curve) follows a similar upward trend without experiencing significant degradation during the later training epochs and achieves a peak value of approximately 91%. The narrow performance gap of

about 4% between the training and validation accuracies suggests that the proposed model generalizes well to unseen data and possesses strong robustness for age and gender classification from speech signals. After the model training phase was completed, the next step was to evaluate its ability to generalize to previously unseen data. The evaluation was conducted using 15,146 testing samples, representing 20% of the original dataset. Two evaluation metrics were employed: the Confusion Matrix, which provides a visual representation of the prediction distribution across classes, and the Classification Report, which offers a detailed quantitative assessment of the model's performance. These evaluation methods were used to ensure that the proposed model not only achieved high overall predictive accuracy but also maintained consistent and balanced performance across all age and gender categories, thereby reducing potential bias toward majority classes. A visual analysis was performed using the Confusion Matrix, which illustrates the relationship between the actual labels and the predicted labels. Figure 3 presents the confusion matrix of the proposed hybrid model, providing an overview of its classification performance across the different age and gender classes.

Figure 3. Confusion Matrix CNN-BiLSTM-Transformer Model

As shown in Figure 3, the confusion matrix exhibits a highly concentrated dark-green main diagonal, indicating that the number of correctly classified samples is substantially higher than that of misclassified samples. This pattern demonstrates that the proposed model successfully predicts the majority of the test instances with high accuracy according to their true class labels. Moreover, the model shows excellent discriminative capability in distinguishing between male and female speakers, as evidenced by the very limited cross-gender misclassification. The prediction errors are relatively minor and primarily occur between adjacent age groups. For example, some samples belonging to the Male 20s category are occasionally misclassified as Male 30s. Such errors are expected because speakers within neighboring age ranges often share similar acoustic characteristics, including fundamental frequency (pitch) and vocal timbre. These overlapping speech attributes make age discrimination particularly challenging, not only for automatic classification systems but also for human listeners. Overall, the confusion matrix indicates that the proposed hybrid CNN–BiLSTM–Transformer architecture effectively captures both gender-specific and age-related speech patterns while maintaining a low level of classification error. Beyond the confusion matrix analysis, the model performance was further examined using a classification report. Unlike the confusion matrix, which only illustrates the distribution of predicted and actual classes, the classification report provides a detailed evaluation of each class through performance metrics, including accuracy, precision, recall, and F1-score. The classification results of the proposed model are summarized in Table 5.

Table 5. Classification report of the hybrid model

Class	Precision	Recall	F1-Score	Support
Male Teens	0.86	0.87	0.87	883
Male 20s	0.92	0.90	0.91	3,852
Male 30s	0.91	0.87	0.89	2,843

Class	Precision	Recall	F1-Score	Support
Male 40s	0.89	0.91	0.90	1,810
Male 50s	0.91	0.94	0.93	1,036
Male 60+	0.91	0.95	0.93	947
Female Teens	0.92	0.88	0.89	224
Female 20s	0.89	0.92	0.90	824
Female 30s	0.91	0.94	0.92	948
Female 40s	0.92	0.91	0.92	459
Female 50s	0.94	0.95	0.95	928
Female 60+	0.91	0.96	0.94	392
Accuracy	**–**	**–**	**0.91**	**15,146**
Macro Average	**0.91**	**0.92**	**0.91**	**15,146**
Weighted Average	**0.91**	**0.91**	**0.91**	**15,146**

The classification report provides a comprehensive evaluation of the proposed Hybrid CNN–BiLSTM–Transformer model across the twelve age–gender categories. As shown in Table 5, the proposed model achieved an overall classification accuracy of 91%, demonstrating its effectiveness in discriminating age and gender characteristics from speech signals. The macro-average precision, recall, and F1-score were 0.91, 0.92, and 0.91, respectively, indicating balanced performance across all classes regardless of class size. Similarly, the weighted-average precision, recall, and F1-score reached 0.91, suggesting that the model maintained consistent performance while accounting for class imbalance in the dataset. Among the individual categories, the highest performance was observed for the Female 50s class, achieving a precision of 0.94, recall of 0.95, and F1-score of 0.95, while the Female 60+ class obtained the highest recall of 0.96. For male speakers, the Male 50s and Male 60+ classes achieved F1-scores of 0.93, reflecting the model's strong ability to capture age-related vocal characteristics in older speakers. In contrast, relatively lower F1-scores were observed for the Male Teens (0.87) and Female Teens (0.89) classes, which may be attributed to greater variability in adolescent vocal features and differences in the number of training samples. To verify the effectiveness of the proposed hybrid architecture, its classification performance was compared with two widely used baseline models, namely a conventional Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (Bi-LSTM) network. All models were evaluated under the same experimental settings to ensure a fair comparison. The comparative results are presented in Table 6.

Table 6. Comparison with Baseline model

Architecture	Accuracy (%)
Standard CNN	84
Bi-LSTM	74
CNN-BiLSTM-Transformer	91

As shown in Table 6, substantial performance differences are observed among the evaluated models. The conventional CNN achieved an accuracy of 84%, while the standalone Bi-LSTM obtained an accuracy of 74%. In contrast, the proposed hybrid CNN–BiLSTM–Transformer model achieved the highest accuracy of 91%, outperforming both baseline architectures. These results demonstrate that the integration of CNN-based local spectral feature extraction, Bi-LSTM-based bidirectional temporal modeling, and Transformer-based global contextual learning provides a more effective representation of speech signals for accurate age and gender classification. Overall, the classification report demonstrates that the proposed Hybrid CNN–BiLSTM–Transformer architecture provides robust and balanced performance for simultaneous age and gender classification from speech signals, achieving high predictive accuracy across diverse demographic groups despite variations in class distribution. To evaluate the effectiveness of the proposed Hybrid CNN–BiLSTM–Transformer architecture, its performance was compared with several recent state-of-the-art studies on voice-based age and gender classification. The comparison considers multiple aspects, including the dataset used, dataset size, classification method, number of target classes, class labels, and overall classification accuracy. A summary of the comparative analysis is presented in Table 7.

Table 7. Comparison with Previous Studies

Reference	Dataset	Number of Samples	Method	Number of Classes	Class Labels	Accuracy
F. Pasqualle et al[18]	Common Voice, VoX Celeb	7300, 153000	Multi Task Deep Learning	N/A	N/A	**94.36%**
V. Karenina et al[5]	Kaggle Audio Dataset	6,000	CNN-based Deep Learning	6	M/F Child, M/F Teen, M/F Adult	92.00%
E. Yücesoy[11]	Common Voice (Turkish Subset)	8,040	Ensemble Learning (Stacking SVM and Random Forest)	12	M20s, F20s, M30s, F30s, M40s, F40s, M50s, F50s, M60s, F60s, F70s, F80s	**97.41%**
Z. Zhang et al[10]	Common Voice	N/A	Parallel CNN–Transformer	12	M-Teens, F-Teens, M20s, F20s, M30s, F30s, M40s, F40s, M50s, F50s, M60, F60	84.90%
This Study	**Common Voice**	**40,392**	**Hybrid CNN–BiLSTM–Transformer**	**12**	**M-Teens, F-Teens, M20s, F20s, M30s, F30s, M40s, F40s, M50s, F50s, M60+, F60+**	**93.00%**

Table 7 presents a comparative analysis between the proposed Hybrid CNN–BiLSTM–Transformer model and several recent studies on voice-based age and gender classification. The comparison highlights differences in dataset characteristics, classification complexity, model architecture, and predictive performance. At first glance, the study by S. Kavitha et al achieved the highest accuracy of 99.90%. However, this result was obtained from a binary gender classification task involving only two classes (male and female), which is considerably less challenging than multi-class age–gender classification. Similarly, the work of V. Karenina et al. (2023) addressed a six-class problem and reported an accuracy of 92.00%. Among studies involving twelve classification categories, E. Yücesoy achieved an accuracy of 97.41% using an ensemble learning approach. Nevertheless, the model was trained on a Turkish subset of the Common Voice dataset containing only 8,040 samples and focused on age groups ranging from the twenties to the eighties, without including teenage speakers. The study by Z. Zhang et al employed a Parallel CNN–Transformer architecture for a twelve-class classification task based on the Common Voice dataset and achieved an accuracy of 84.90%. Compared with this approach, the proposed Hybrid CNN–BiLSTM–Transformer model attained a substantially higher accuracy of 93.00%, representing an improvement of approximately 8.1 percentage points while addressing the same classification complexity. A notable contribution of the present study is the use of a significantly larger dataset comprising 40,392 speech samples, which is considerably larger than those used in most previous studies. In addition, the proposed model simultaneously captures local acoustic patterns through CNN, bidirectional temporal dependencies through Bi-LSTM, and long-range contextual relationships through the Transformer Encoder. This hybrid feature-learning strategy enables the model to effectively discriminate among twelve age–gender categories while maintaining robust classification performance.

4. CONCLUSION

Based on the experimental results and comprehensive evaluations, it can be concluded that the proposed hybrid Deep Learning architecture, integrating a Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and Transformer Encoder, was successfully designed and implemented to extract both spatial and temporal representations from speech signals. The hybrid framework demonstrated robust performance in addressing the challenges posed by globally diverse multi-accent speech data for age and gender classification. The experimental results achieved an overall classification accuracy of 91%, supported by stable and consistent F1-score values across the evaluated classes. These findings confirm the effectiveness of the proposed approach and fulfill the primary objective of this study by demonstrating that the hybrid architecture significantly outperforms individual Deep Learning models under the same experimental conditions. Specifically, the proposed model surpassed the standalone CNN model, which achieved an accuracy of 84%, and the standalone Bi-LSTM model, which achieved 74% accuracy. The superior performance of the proposed framework can be attributed to its ability to combine complementary feature-learning mechanisms, including spatial feature extraction through CNN, bidirectional temporal dependency modeling through Bi-LSTM, and global contextual representation through the Transformer's self-attention mechanism. This integration enables the model to effectively capture the complex characteristics of human speech associated with age and gender, making the proposed architecture a reliable and effective solution for voice-based demographic classification tasks.

ACKNOWLEDGEMENTS

The authors would like to express their sincere appreciation to Universitas Amikom Yogyakarta and Author 2 for their guidance, support, and valuable contributions to this research and manuscript preparation.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Author 1: Conceptualization, Methodology, Software, Formal Analysis, Writing Original Draft, Visualization, and Project Administration, Author 2: Supervision, Writing – Review & Editing, and Validation, Author 3: Funding Acquisition and Resources, Authors 4-6: Validation and Investigation.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest.

DATA AVAILABILITY

The dataset used in this study was obtained from the Mozilla Common Voice dataset, which is publicly available through Kaggle at https://www.kaggle.com/datasets/mozillaorg/common-voice

REFERENCES

[1] Malavika.M, Afrin Dinusha.J, Mr. Shenbagharaman A, and Dr. B. Shunmugapriya, "Real-Time Gender and Age Detection Using Visual and Vocal Cues," *International Research Journal on Advanced Science Hub*, vol. 7, no. 02, pp. 94–102, Feb. 2025, doi: 10.47392/IRJASH.2025.012.

[2] Muhammed Shameem P, Muhammed Faheem, Muhammed Sahad V V, Nafla Ashraf, and Shad Shaharyar, "Age And Gender Prediction From Human Voice For Customized Ads In E-Commerce," *International Journal of Engineering Research & Technology*, vol. 11, no. 01, Jun. 2023.

[3] M. Maayah, A. Abunada, K. Al-Janahi, M. E. Ahmed, and J. Qadir, "LimitAccess: on-device TinyML based robust speech recognition and age classification," *Discover Artificial Intelligence*, vol. 3, no. 1, p. 8, Feb. 2023, doi: 10.1007/s44163-023-00051-x.

[4] D. T. Adherda, M. Hikmatyar, and Ruuhwan, "GENDER CLASSIFICATION BASED ON VOICE USING RECURRENT NEURAL NETWORK (RNN)," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 17, no. 1, pp. 111–122, Oct. 2023, doi: 10.35457/antivirus.v17i1.3049.

[5] V. Karenina, M. F. Erinsyah, and D. S. Wibowo, "Klasifikasi Rentang Usia Dan Gender Dengan Deteksi Suara Menggunakan Metode Deep Learning Algoritma Cnn (Convolutional Neural Network)," *Komputika : Jurnal Sistem Komputer*, vol. 12, no. 2, pp. 75–82, Sep. 2023, doi: 10.34010/komputika.v12i2.10516.

[6] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Speech-based Age and Gender Prediction with Transformers," *Speech Communication; 15th ITG Conference*, Jun. 2023, doi: 10.30420/456164008.

[7] M. S. Remya, P. Ishwar, and P. Nedungadi, "A Hybrid Cross-Attentive CNN-BiLSTM-Transformer Network for Dysarthria Severity Classification," *Sci. Rep.*, vol. 15, no. 1, p. 42080, Nov. 2025, doi: 10.1038/s41598-025-26049-2.

[8] H. Kumari, H. Kumari, and U. Nawarathne, "Speech Emotion Recognition with Hybrid CNN- LSTM and Transformers Models: Evaluating the Hybrid Model Using Grad-CAM," *International Journal of Research in Computing*, vol. 3, no. 1, p. 56, Jul. 2024, doi: 10.64701/ijrc/345/8907.

[9] J. Jorrin-Coz, M. Nakano, H. Perez-Meana, and L. Hernandez-Gonzalez, "Multi-Corpus Benchmarking of CNN and LSTM Models for Speaker Gender and Age Profiling," *Computation*, vol. 13, no. 8, p. 177, Jul. 2025, doi: 10.3390/computation13080177.

[10] Z. Zhang, R. Li, and K. Chen, "A Parallel CNN-Transformer Framework for Speech Age Recognition," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 29, no. 5, pp. 1137–1144, Sep. 2025, doi: 10.20965/jaciii.2025.p1137.

[11] E. Yücesoy, "Automatic Age and Gender Recognition Using Ensemble Learning," *Applied Sciences*, vol. 14, no. 16, p. 6868, Aug. 2024, doi: 10.3390/app14166868.

[12] S.-H. Noh, "Analysis of Gradient Vanishing of RNNs and Performance Comparison," *Information*, vol. 12, no. 11, p. 442, Oct. 2021, doi: 10.3390/info12110442.

[13] Z. Liao, "Comparative analysis between application of transformer and recurrent neural network in speech recognition," *Applied and Computational Engineering*, vol. 6, no. 1, pp. 524–529, Jun. 2023, doi: 10.54254/2755-2721/6/20230879.

[14] A. Tursunov, Mustaqeem, 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, Sep. 2021, doi: 10.3390/s21175892.

[15] M. S. Remya, P. Ishwar, and P. Nedungadi, "A Hybrid Cross-Attentive CNN-BiLSTM-Transformer Network for Dysarthria Severity Classification," *Sci. Rep.*, vol. 15, no. 1, p. 42080, Nov. 2025, doi: 10.1038/s41598-025-26049-2.

[16] S. Kim and S.-P. Lee, "A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech," *Electronics (Basel).*, vol. 12, no. 19, p. 4034, Sep. 2023, doi: 10.3390/electronics12194034.

[17] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus", Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), European Language Resources Association, pp. 4218–4222, Mar. 2020.

[18] F. Pasquale, "Identity, Gender, Age, and Emotion Recognition from Speaker Voice with Multi-task Deep Networks for Cognitive Robotics," *Cognitive Computation* 2713–2723 (2024). https://doi.org/10.1007/s12559-023-10241-5

.