Research Article BEU Fen Bilimleri Dergisi 15(1): 312-321 (2026) DOI: 10.17798/bitlisfen.1803512

BİTLİS EREN UNIVERSITY
FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129 / e-ISSN: 2147-3188

A CNN–NCP BASED HYBRID DEEP LEARNING MODEL FOR SPEECH-DRIVEN GENDER CLASSIFICATION

Sevda OLGUN [1*], Caner BALIM [2], Nevzat OLGUN [2]

[1] *Afyon Kocatepe University, Institute of Sciences, Afyonkarahisar, Türkiye*
[2] *Afyon Kocatepe University, Software Engineering Department, Afyonkarahisar, Türkiye*
*** Corresponding Author:** sevdaolgun33@gmail.com

Article Info

Received: *October 14, 2025*
Revised: *December 9, 2025*
Accepted: *February 3, 2026*

Keywords
Speech-based gender classification,
Deep learning,
Neural Circuit Policies (NCP).

ABSTRACT

Speech is one of the most natural and effective forms of human communication, carrying both linguistic and non-linguistic information. It plays a crucial role in many applications such as gender classification, biometric authentication, and personalized human-computer interaction. This study aims to investigate the contribution of a hybrid deep learning model based on Neural Circuit Policies (NCP), inspired by biological neural systems, for gender classification on Turkish speech data, by evaluating its performance in terms of accuracy and computational efficiency in comparison with conventional recurrent models. Mel-Frequency Cepstral Coefficients (MFCC) and log-Mel spectrogram features are combined to simultaneously capture the spectral and temporal properties of speech signals. These features are learned as low-level acoustic patterns via Conv1D layers. Long-term temporal dependencies are modeled using Liquid Time Constant (LTC) cells defined within the NCP architecture. To evaluate the generalizability of the model, the experiments were conducted under a speaker-independent setup, and ablation studies were performed by removing different components of the architecture to clearly assess the contribution of the NCP component. Cross-validation was applied on the Mozilla Common Voice 12.0 Turkish dataset during the experiments. The Conv1D+NCP model achieved 99.29% accuracy and 99.28% F1-score, while the LSTM-based model yielded slightly lower results. The NCP-based model offers high performance and computational efficiency with fewer parameters, making it a powerful alternative for real-time applications.

1. INTRODUCTION

Speech is one of the most natural and effective means of communication among humans. Air from the lungs is shaped by the vocal cords and speech organs such as the tongue, lips, and teeth into a sound signal [1]. This sound signal not only conveys verbal content but also contains biometric and emotional characteristics specific to the speaker [2]. Therefore, in addition to linguistic information, speech signals also contain various paralinguistic data such as the speaker's age, gender, identity, emotional state, and ethnicity [3], [4]. Such information has important applications in a wide variety of fields, including security, healthcare, education, advertising, and forensic analysis, as well as in the personalization of human-computer interaction systems [5].

Determining the speaker's gender plays a critical role in voice-based biometric systems. For example, in call center applications, pre-determining gender allows automated response systems to deliver more personalized communication using appropriate tone of voice and language patterns. Similarly, in voice biometrics-based authentication systems, knowledge of gender narrows the search space and improves authentication performance [6]. Moreover, gender is used as an effective discrimination criterion in speaker diarization processes in multi-speaker environments and plays an important role in clustering different speakers [7]. Additionally, gender prediction plays a crucial role in user-specific content recommendation systems in the media and entertainment industry. In such systems, gender information can be used to design a more targeted user experience. In technologies that enable personalized

interaction, such as digital assistants, gender information automatically extracted from the audio signal is used to create a user profile. This information enables processes such as adapting speech style, selecting a form of address, and personalizing content recommendations [8].

Recent studies have shown that machine learning and deep learning-based methods are widely used in audio-based gender recognition. Traditional methods include K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees, and Support Vector Machines (SVM) [9]. These studies typically utilize low-dimensional but effective acoustic features such as Mel Frequency Cepstrum Coefficients (MFCC) [10]. However, these methods often show limited success due to their limited computational capacity and their inability to model the dynamic nature of sequential data.

Advances in deep learning have provided significant progress in audio-based classification problems thanks to structures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN, LSTM) [11]. However, these models require a high number of parameters and long training times. At this point, Neural Circuit Policies (NCP), inspired by biological nervous systems and providing more efficient learning on sequential data, emerged as a promising approach [12], [13]. By modeling variable time constants using Liquid Time-Constant (LTC) cells, the NCP architecture can capture temporal dependencies in speech signals in a more biologically realistic way. This feature offers a significant advantage, especially for time-dependent and variable data such as speech signals.

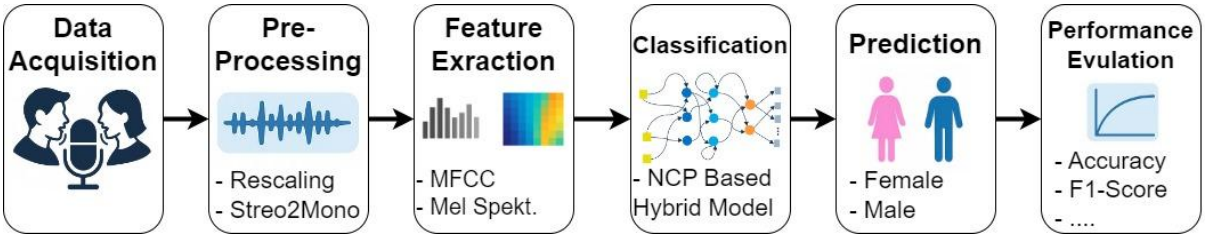

Figure 1. Block diagram of the proposed hybrid deep learning architecture for gender classification using MFCC and log-Mel features.

In this study, gender classification was performed using the Mozilla Common Voice 12.0 Turkish dataset [14]. MFCC and log-Mel spectrogram features extracted from voice recordings were used as input to hybrid deep learning models consisting of Conv1D and NCP layers. The block diagram of the study is shown in Figure 1. In the proposed architectures, the Conv1D layers learn local frequency patterns from the features, while the LTC layer in the NCP-based model models temporal dependencies using a continuous-time approach. Thus, the model processed both low-level spectral features and high-level temporal information to perform gender determination. 5-fold stratified cross-validation was used to validate the proposed model. Experimental results demonstrate that the proposed model performs very well in gender classification using Turkish speech data, achieving higher accuracy than many studies in the literature.

2. RELATED WORK

In recent years, the automatic identification of paralinguistic information such as gender, age, and emotion from speech signals has been the subject of extensive research due to its wide range of applications, from human-computer interaction to biometric security. The methods used for this purpose in the literature vary significantly in terms of feature extraction techniques and classification algorithms.

In traditional approaches, one of the most frequently preferred features has been the Mel Frequency Cepstral Coefficients (MFCC). Younis et al. classified MFCC-based features on 280 audio recordings from the TIMIT dataset using the SVM algorithm and achieved 96.45% accuracy [15]. Similarly, Chaudhari and Kagalkar modeled the MFCC features with the Gaussian Mixture Model (GMM) and then achieved 80.3% accuracy with the SVM classifier [16]. Yücesoy et al. achieved an accuracy of up to 97.76% on the TIMIT dataset using a text-independent approach using MFCC and GMM [17]. In a study by Bakır, the effect of different MFCC dimensions on classification performance was evaluated using German speech data, and the most successful results were obtained with the HMM method and

high MFCC dimensions. Furthermore, increasing the number of words in the speech samples positively affected recognition success [18].

Similar trends have been observed in studies comparing machine learning methods. Zaman et al. compared various algorithms using 20 statistical features extracted from speech signals; they achieved an accuracy rate of 96.4% with the CatBoost algorithm [19]. Munoli et al. compared MFCC-based features with CatBoost, XGBoost, SGD, and decision trees, achieving accuracy values in the range of 88–90% [20]. These findings indicate that while classical methods provide high accuracy under certain conditions, they become limited as the dataset grows larger, or task complexity increases.

On the other hand, the development of deep learning methods has significantly accelerated voice-based gender classification. Mohammed et al. compared Bayesian, K-NN, GMM, SVM, and deep learning-based models using speech data from different languages and demonstrated that deep learning models provide higher accuracy [21]. Kone et al. achieved gender prediction with 91% accuracy using a deep learning architecture with five hidden layers [22].

Focusing on child speech, Safavi et al. used methods such as GMM–UBM, GMM–SVM and i-vector + PLDA for gender and age classification; the highest accuracy was reported as 79.18% in the frequency range of 0.9–2.6 kHz [23]. This result shows that age and speech characteristics have a significant effect on the classification success.

In general, the literature demonstrates that MFCC and its derivatives are widely used in feature extraction, achieving high success rates when supported by powerful machine learning algorithms and deep learning architectures. However, most classical methods fail to adequately capture the temporal dependencies of sequential data, while deep learning models face practical limitations due to their high parameter counts and training costs.

In this study, unlike traditional MFCC feature-based machine learning and deep learning-based approaches in the literature, a new NCP-based hybrid model with low neuron numbers is proposed on Turkish speech data.

3. MATERIALS AND METHODS

This section presents details regarding the dataset used in the study, the feature extraction process, the structure of the proposed model, and the classification process. In this study, conducted using Turkish audio data, an NCP-based hybrid model inspired by biological neural systems was proposed to classify speaker gender. Additionally, an LSTM-based hybrid model was implemented for comparison, and the performance of both architectures was evaluated in detail.

3.1. Dataset

This study used the Turkish dataset from Common Voice version 12.0, an open-source speech data collection created by the Mozilla Foundation [14]. Audio files were recorded in MP3 format, and metadata was recorded for each recording. Each recording in the dataset includes demographic information such as age and gender.

Table 1. Sample texts and genders in the data set used in the study.

Text	Gender
Gülmeyin!	Female
Evet, iyiyim.	Female
Sen rahatına bak	Female
Mustafa: “Baş üstüne beyim…” dedi.	Female
Yüzlerce kiloluk bir ağırlık taşıyormuş gibi aşağıya çekilen elini uzattı.	Female
Param!	Male
Bu çok ani oldu.	Male
Bugün sizin gününüz değil, ne diye kaldınız?	Male

To ensure reliability, only recordings clearly labeled as male or female were considered in this study; samples with missing, incorrect, or inconsistent labeling information were excluded. The dataset,

comprised of 22,651 unique speakers, included 11,142 male and 11,509 female gender-labeled recordings. Each recording corresponds to the utterance of short phrases randomly selected by volunteers. Table 1 presents sample sentences from the dataset.

The audio signal of the text “Evet, iyiyim.” voiced by the female speaker in Table 1 and the audio signal of the text “Bu çok ani oldu.” voiced by the male speaker are given in Figure 2.

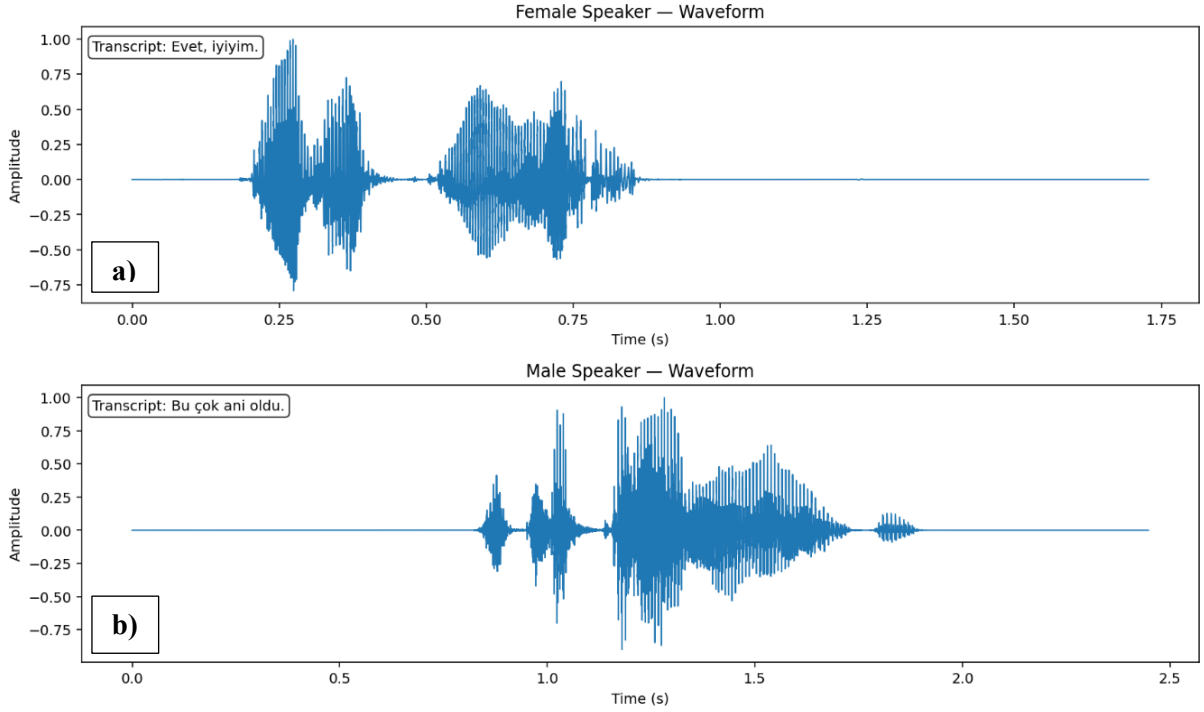

Figure 2. Example waveforms of speech recordings: (a) female speaker and (b) male speaker.

3.2. Preprocessing and Feature Extraction

The raw version of the Mozilla Common Voice 12 Turkish speech dataset is not suitable for direct classification due to variations in recording conditions and devices. Therefore, preprocessing steps were first implemented. All audio files were converted to single-channel (mono) format and the sampling frequency was rescaled to 16 kHz. This standardized the time resolution of data recorded from different devices and ensured consistency throughout the model's learning process.

Applying standardized raw audio data obtained in the preprocessing step directly to classification algorithms often leads to poor generalization performance due to high dimensionality and sensitivity to noise. Therefore, MFCC (Mel-Frequency Cepstral Coefficients) and log-Mel spectrogram features were extracted to achieve dimensionality reduction while preserving semantic and structural information from audio signals and reducing noise components.

MFCC is based on the computation of cepstral coefficients through mel-scale filter banks that mimic the logarithmic frequency perception of the human auditory system. This method concisely represents critical acoustic features such as formant structures, timbre, and speaker-specific spectral envelopes [24]. In contrast, the log-Mel spectrogram presents the energy distribution of the signal in a detailed time-frequency domain and captures biometric and prosodic cues such as gender, mood, and speaking rate [25]. Furthermore, the log-Mel spectrogram effectively reflects non-phonetic but communicatively important components such as stress, pauses, and intonation.

In this study, a (168.1)-dimensional feature vector was created by combining the (40.1)-dimensional MFCC and (128.1)-dimensional log-Mel features extracted for each audio signal to more robustly represent both the spectral and temporal characteristics of the speech signal. This allows the model to learn both low-level acoustic cues and high-level phonetic information simultaneously.

3.3. Proposed Hybrid CNN–NCP Model

In this study, a specifically designed hybrid deep learning architecture was used to process MFCC and log-Mel features extracted from speech signals. One-dimensional CNN layers, placed sequentially at the input of the model, capture short-term local frequency patterns, revealing the characteristic structure of the signal. Following these layers, normalization, pooling, and dropout operations were applied to reduce dimensionality and prevent overfitting. The features obtained from the CNN layers were processed using different approaches to model temporal dependencies. In the scenario based on the NCP architecture, the features were transferred to Liquid Time-Constant (LTC) cells, and sequential data were modeled using a continuous-time approach. This method offers a more biologically realistic representation compared to discrete-time structures such as LSTM and offers significant advantages in terms of both computational efficiency and generalization performance due to its low parameter count [26]. In the alternative scenario, an LSTM layer is used instead of LTC. LSTM cells stand out for their ability to preserve long-term dependencies through gating mechanisms and mitigate gradient vanishing [27]. However, their higher parameter count and longer training time requirements compared to NCP pose a relative disadvantage in terms of computational cost.

In both models, the outputs from the temporal layer were reduced to a one-dimensional vector using Global Average Pooling and then transformed into high-level abstract representations using fully connected (Dense) layers. In the final stage, the sigmoid-activated output layer labels each voice sample with a probability value of belonging to either "female" or "male" classes. Figure 3 shows the general structure of the NCP-based hybrid model. For a fair comparison, the performance of the two different approaches was compared by replacing the NCP layer with an LSTM with the same number of units within the same structure.

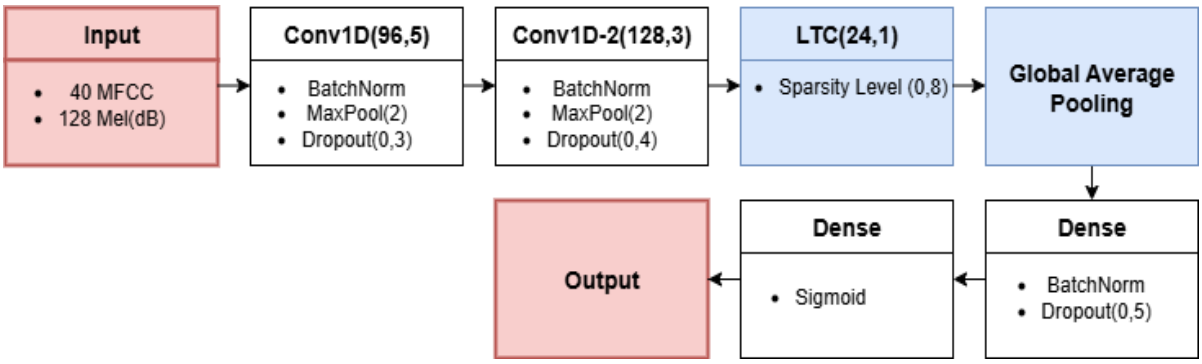

Figure 3. Overall architecture of the proposed NCP-based hybrid model

3.4. Performance Metrics

To comprehensively evaluate the performance of the proposed model, different classification metrics were used. First, the accuracy metric, which indicates the model's correct prediction rate, was considered as the primary success metric. Furthermore, the precision metric indicates how accurately the model predicts positive classes, while the recall metric reveals its success in capturing true positive examples. The F1 score, which provides a balanced summary of precision and recall, provides a reliable performance indicator, especially in situations where false positives and false negatives are both significant. Table 2 presents the performance metrics and their equations. In the formulas presented in Table 2, the terms TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) denote true positives, true negatives, false positives, and false negatives, respectively.

Table 1. Calculation of performance metrics used in classification.

Metric	Formula
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	$Precision = \frac{TP}{TP + FP}$
Recall	$Recall = \frac{TP}{TP + FN}$
F1 Score	$F1 = 2.\frac{Precision.Recall}{Precision + Recall}$

3.5. Stratified and Subject based k-Fold Cross-Validation

In this study, the Stratified k-Fold cross-validation method was used to increase the generalizability of the model and to assess its accuracy results more reliably. K-Fold cross-validation is a widely used statistical validation method to evaluate the generalization performance of a machine learning or deep learning model. In this method, the entire dataset is divided into equal folds. Each fold is used as the test data once, while the remaining k−1 folds are used for training. This process is repeated k times, with each subset treated as the test data. In each round, the model is trained from scratch and its performance is measured on different test sets. This allows us to assess whether the model is overfitting and its consistency across different data splits.

In addition to stratified k-fold cross-validation, a subject-based (speaker-independent) evaluation was conducted using the Leave-One-Subject-Out (LOSO) cross-validation strategy to prevent data leakage and to measure the true generalization capability of the model. In LOSO, all speech samples belonging to a single speaker are held out as the test set, while the samples from all remaining speakers are used for training. This process is repeated for each speaker in the dataset, ensuring that no speaker appears in both the training and test sets at any iteration. Compared to fold-based subject grouping methods, LOSO provides a stricter and more realistic evaluation protocol, particularly for speaker-related tasks such as gender classification.

In this study, k = 5 was selected for stratified k-fold cross-validation to maintain balanced class distributions between training and test sets. For the subject-based evaluation, LOSO cross-validation was applied at the speaker level. After completing all folds and all LOSO iterations, the performance metrics obtained from each run were averaged to report the final results. This combined evaluation strategy allows for a comprehensive assessment of both class-balanced performance and speaker-independent generalization.

4. RESULTS AND DISCUSSION

In this study, extensive experiments were conducted on the Mozilla Common Voice 12.0 Turkish dataset to comparatively evaluate the performance of the proposed Conv1D+NCP-based hybrid model against alternative architectures, including Conv1D+LSTM-based models, under identical experimental settings. The training process was performed using a Binary Cross-Entropy loss function and the Adam optimization algorithm. Dropout and batch normalization layers were also used to reduce the risk of overfitting. Training sessions were limited to 20 epochs, and overfitting prevention strategies such as early stopping and ReduceLROnPlateau were utilized. 5-fold stratified cross-validation was applied to reliably assess the generalizability of the model.

Table 3 shows the performance metrics for each fold in the NCP-based hybrid model. Average accuracy was 99.29%, and average F1 score was 99.28%. Accuracy was above 98% across all folds. The highest accuracy was 99.65% in Fold 1, and the lowest was 98.61% in Fold 3.

Table 3. Performance Achievements with 5-Fold Cross Validation using NCP-based Model.

Fold	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
1	99.65	99.64	99.51	99.78
2	99.18	99.17	99.68	98.65
3	98.61	98.59	98.13	99.06
4	99.45	99.44	99.73	99.15
5	99.58	99.57	99.55	99.60
Average	99.29	99.28	99.32	99.25

Table 4 shows the performance metrics for each fold in the LSTM-based hybrid model. Average accuracy was 99.20%, and average F1 score was 99.19%. Accuracy was above 98% across all folds. The highest accuracy was 99.78% in Fold 4, and the lowest was 98.70% in Fold 1.

Table 4. Performance Achievements with 5-Fold Cross Validation using LSTM-based model.

Fold	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
1	98.70	98.66	99.86	97.49
2	98.92	98.90	99.01	98.79
3	99.01	98.99	98.75	99.24
4	99.78	99.78	99.82	99.73
5	99.60	99.60	99.51	99.69
Average	99.20	99.19	99.39	98.99

The low variance between folds indicates that both models provide consistent results across different data splits. Examining the results presented in Tables 3 and 4, both the NCP-based and LSTM-based hybrid models achieve accuracy above 98% across all folds, with NCP-based outperforming the LSTM-based model in all scores except average precision. This clearly demonstrates that the proposed model not only provides high accuracy but also delivers reliable and generalizable results across different data splits.

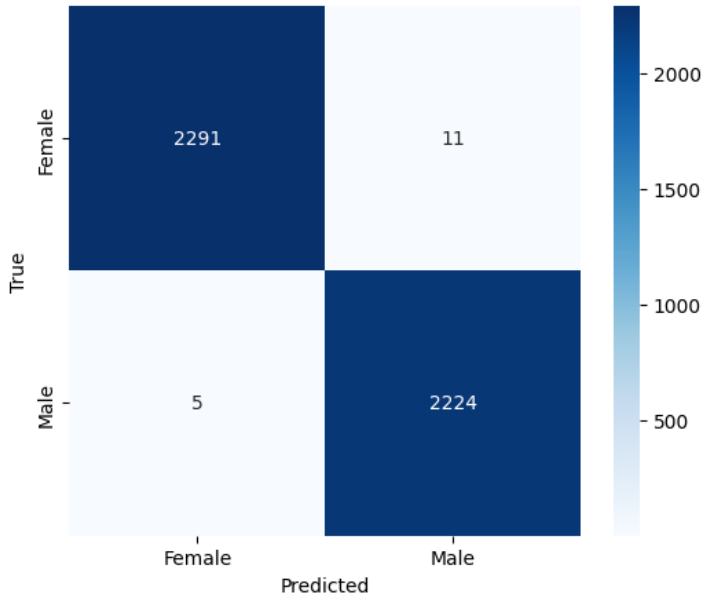

Figure 4. Confusion matrix of the proposed NCP-based model for Fold 1, showing classification results between female and male speakers.

Figure 4 shows the confusion matrices of Fold 1, where the proposed NCP-based model performed best. In Fold 1, the model misclassified only 11 of 2,291 female samples and 5 of 2,224 male samples. A qualitative examination of the misclassified samples revealed that most of the incorrect predictions occurred in recordings with low signal-to-noise ratios or in samples where the frequency distributions

of male and female voices converged. This observation shows that the model performed with high accuracy for most cases, but its performance weakened slightly in noisy or borderline samples.

Additional experiments were conducted to isolate the contribution of both the temporal modeling component and the acoustic feature type. In these experiments, Conv1D-based architectures were evaluated with different back-end structures, including a simple RNN, a Dense layer without explicit temporal modeling, LSTM, and the proposed NCP. To ensure a fair comparison, all models were trained under the same experimental settings and evaluated using identical performance metrics. Moreover, MFCC and log-Mel features were tested separately, as well as in combination, to analyze their individual and joint impact on classification performance. The comparative results of these experiments are summarized in Table 5, highlighting how different architectural choices and feature representations affect overall accuracy and F1-score. The results show that LSTM and NCP architectures achieve higher accuracy and F1-scores compared to simple RNN and Dense structures. Moreover, while MFCC and log-Mel features used individually yield competitive results, the choice of architecture appears to be more decisive for performance than the type of acoustic features.

Table 5. Performance Comparison of Conv1D-Based Architectures Using MFCC and log-Mel Features

Model	Features	Accuracy (%)	F1-score (%)	Model
Conv1D+ RNN	MFCC + log-mel	98.91	98.89	Conv1D+ RNN
Conv1D+ Dense	MFCC + log-mel	98.97	98.95	Conv1D+ Dense
Conv1D+ LSTM	MFCC	99.19	99.18	Conv1D+ LSTM
Conv1D+ LSTM	log-mel	99.20	99.19	Conv1D+ LSTM
Conv1D+ NCP	MFCC	98.79	98.76	Conv1D+ NCP
Conv1D+ NCP	log-mel	98.82	98.81	Conv1D+ NCP

To evaluate the model's performance under a speaker-independent test scenario, experiments were conducted on different Conv1D-based architectures using the LOSO validation approach. An analysis of the dataset revealed that it contains speech recordings from 16 distinct speakers. To ensure a balanced evaluation across speakers, a fixed number of 139 speech samples were randomly selected from each speaker's recordings. For one speaker, only 21 speech samples were available, and this original distribution was preserved, while the remaining 15 speakers were downsampled to 139 samples each. The results are presented in Table 6. An examination of the table shows that all models achieve an accuracy above 0.80; however, the proposed model attains an accuracy of 87.19%. While the Conv1D+LSTM model yields the highest accuracy, the Conv1D+NCP model demonstrates a more balanced classification performance under the subject-based LOSO setting, achieving the highest F1-score of 92.50%. This indicates that the NCP-based architecture exhibits stronger generalization capability under conditions of high inter-speaker variability. Although the Conv1D+Dense and Conv1D+RNN models produce competitive results, they remain limited compared to more complex temporal models.

Table 6. Classification performances of different Conv1D-based architectures under speaker-independent (LOSO) validation.

Model	Accuracy (%)	F1-score (%)
Conv1D+ RNN	82.10	88.06
Conv1D+ Dense	82.28	87.91
Conv1D+ LSTM	88.76	91.80
Conv1D+ NCP	87.19	92.50

5. CONCLUSION

In this study, a hybrid deep learning architecture based on CNN and NCP, combining MFCC and log-Mel spectrogram features, is proposed to perform gender classification on Turkish speech data. Experiments with 5-fold stratified cross-validation yielded an average accuracy of 99.29% and an average F1-score of 99.28% for the NCP-based model. The LSTM-based version of the proposed model

with the same number of units demonstrated similar high performance with an average accuracy of 99.20% and an F1-score of 99.19%. Both models achieved accuracy above 98% across all folds.

Comparisons conducted within the scope of ablation and comparative experiments using different Conv1D-based back-end structures revealed that temporal modeling approaches consistently outperform non-temporal baselines. Among these, the NCP-based model achieved superior results in all evaluation metrics except precision. Furthermore, with 133,995 parameters, the NCP-based model outperformed the LSTM-based model, which comprised 135,233 parameters, thereby demonstrating higher computational efficiency while maintaining comparable classification performance.

In addition, speaker-independent experiments conducted using the LOSO validation strategy showed that the Conv1D+NCP model achieved an accuracy of 87.19% and the highest F1-score of 92.50%, outperforming the Conv1D+LSTM model in terms of balanced classification performance, which achieved an F1-score of 91.80%. Although the LSTM-based model obtained slightly higher accuracy of 88.76%, the NCP-based architecture demonstrated stronger robustness and generalization capability under high inter-speaker variability.

The overall evaluation demonstrates that the proposed NCP-based method performs with high accuracy, strong discrimination, and low error rate for gender classification in Turkish speech data. Furthermore, thanks to its relatively low parameter count of 133,995 parameters and its competitive performance under both stratified and speaker-independent evaluation protocols, it offers a suitable alternative for integration into real-time applications. Future work can be focused on data augmentation strategies, noise-robust feature extraction, and cross-testing across languages to further strengthen the generalizability of the model.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

Artificial Intelligence (AI) Contribution Statement

This manuscript was entirely written, edited, analyzed, and prepared without the assistance of any artificial intelligence (AI) tools. All content, including text, data analysis, and figures, was solely generated by the authors.

Contributions of the Authors

Conceptualization, C. Balım and N. Olgun; Data curation, S. Olgun and C. Balım; Formal analysis, C. Balım and N. Olgun; Investigation, S. Olgun and C. Balım; Methodology, S. Olgun and C. Balım; Resources, S. Olgun and C. Balım; Software, S. Olgun and C. Balım; Supervision, N. Olgun and C. Balım; Validation, C. Balım; Visualization, S. Olgun and N. Olgun; Writing – original draft, S. Olgun; Writing – review & editing, C. BALIM and N. Olgun. All authors have read and approved the final version of the manuscript.

REFERENCES

[1] F. Altunbey Özbay and E. Özbay, “Ses verilerinden cinsiyet tespiti için yeni bir yaklaşım: Optimizasyon yöntemleri ile özellik seçimi,” Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, vol. 38, no. 2, pp. 1179–1192, 2022, doi: 10.17341/gazimmfd.938294.

[2] S. Safavi, M. Russell, and P. Jančovič, “Automatic speaker, age-group and gender identification from children’s speech,” Comput Speech Lang, vol. 50, pp. 141–156, 2018, doi: 10.1016/j.csl.2018.01.001.

[3] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, “Age group classification and gender recognition from speech with temporal convolutional neural networks,” Multimed Tools Appl, vol. 81, no. 3, pp. 3535–3552, 2022, doi: 10.1007/s11042-021-11614-4.

[4] S. Chaudhary and D. K. Sharma, “Gender Identification based on Voice Signal Characteristics,” in 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 869–874. doi: 10.1109/ICACCCN.2018.8748676.

[5] S. J. Chaudhari and R. M. Kagalkar, "Methodology for gender identification, classification and recognition of human age," Int. J. Comput. Appl, vol. 975, p. 8887, 2015.

[6] M. Alsulaiman, Z. Ali, and G. Muhammad, "Gender Classification with Voice Intensity," in 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, 2011, pp. 205–209. doi: 10.1109/EMS.2011.37.

[7] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans Audio Speech Lang Process, vol. 14, no. 5, pp. 1557–1565, 2006, doi: 10.1109/TASL.2006.878256.

[8] E. H. Alkhammash, M. Hadjouni, and A. M. Elshewey, "A Hybrid Ensemble Stacking Model for Gender Voice Recognition Approach," Electronics (Basel), vol. 11, no. 11, 2022, doi: 10.3390/electronics11111750.

[9] S. Hızlısoy, E. Çolakoğlu, and R. S. Arslan, "Speech-to-Gender Recognition Based on Machine Learning Algorithms," International Journal of Applied Mathematics Electronics and Computers, vol. 10, no. 4, pp. 84–92, 2022, doi: 10.18100/ijamec.1221455.

[10] J. Ahmad, M. Fiaz, S. Kwon, M. Sodanil, B. Vo, and S. W. Baik, "Gender identification using mfcc for telephone applications-a comparative study," arXiv preprint arXiv:1601.01577, 2016.

[11] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," Aug. 2019. doi: 10.1109/JSTSP.2019.2908700.

[12] M. Lechner, R. Hasani, A. Amini, T. A. Henzinger, D. Rus, and R. Grosu, "Neural circuit policies enabling auditable autonomy," Nat Mach Intell, vol. 2, no. 10, pp. 642–652, 2020, doi: 10.1038/s42256-020-00237-3.

[13] N. Olgun, "Lazer işaretleri ile yapay zeka temelli hedef analizi (Artificial intelligence based target analysis with laser signals)," Frat University, Turkey, 2022.

[14] "Mozilla Common Voice (2022) Common Voice.," https://commonvoice.mozilla.org/tr/datasets.

[15] H. A. Younis et al., "Multimodal age and gender estimation for adaptive human-robot interaction: A systematic literature review," Processes, vol. 11, no. 5, p. 1488, 2023.

[16] B. K. Munoli, K. A. K. Jain, P. Kumar, A. R. PS, and others, "Human voice analysis to determine age and gender," in 2023 International conference on recent trends in electronics and communication (ICRTEC), 2023, pp. 1–4.

[17] E. Yücesoy and V. V Nabiyev, "Gender identification of a speaker using MFCC and GMM," in 2013 8th International Conference on Electrical and Electronics Engineering (ELECO), 2013, pp. 626–629. doi: 10.1109/ELECO.2013.6713922.

[18] Ç. Bakır, "Alman Dili Üzerinde Konuşmacı Cinsiyetinin Otomatik Olarak Belirlenmesi," Academic Platform - Journal of Engineering and Science, vol. 4, no. 2, 2016, doi: 10.21541/apjes.49291.

[19] 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), 2021, pp. 338–343. doi: 10.1109/COMPSAC51774.2021.00055.

[20] B. K. Munoli, K. A. K. Jain, P. Kumar, A. R. P. S, and Ashwini, "Human Voice Analysis to Determine Age and Gender," in 2023 International Conference on Recent Trends in Electronics and Communication (ICRTEC), 2023, pp. 1–4. doi: 10.1109/ICRTEC56977.2023.10111890.

[21] A. A. Mohammed and Y. F. Al-Irhayim, "An overview for assessing a number of systems for estimating age and gender of speakers," Tikrit Journal of Pure Science, vol. 26, no. 1, pp. 94–100, 2021.

[22] V. S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, and M. S. M, "Voice-based Gender and Age Recognition System," in 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), 2023, pp. 74–80. doi: 10.1109/InCACCT57535.2023.10141801.

[23] S. Safavi, M. Russell, and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," Comput Speech Lang, vol. 50, pp. 141–156, 2018, doi: 10.1016/j.csl.2018.01.001.

[24] Y. Zhao and X. Shu, "Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC)," Sci Rep, vol. 13, no. 1, p. 20398, 2023, doi: 10.1038/s41598-023-47118-4.

[25] K. Donuk and D. Hanbay, "Konuşma Duygu Tanıma için Akustik Özelliklere Dayalı LSTM Tabanlı Bir Yaklaşım," Computer Science, vol. Vol:7, no. Issue:2, pp. 54–67, 2022, doi: 10.53070/bbd.1113379.

[26] M. Lechner, R. Hasani, A. Amini, T. Henzinger, D. Rus, and R. Grosu, "Neural circuit policies enabling auditable autonomy," Nat Mach Intell, vol. 2, pp. 642–652, Aug. 2020, doi: 10.1038/s42256-020-00237-3.

[27] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, "Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review," Polymers (Basel), vol. 16, no. 18, 2024, doi: 10.3390/polym16182607.