Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 3 PAGE: 701-708 YEAR: 2024
DOI:10.17798/bitlisfen.1477191

Application of Deep Learning for Voice Command Classification in Turkish Language

Yusuf ÇELİK[1*]

[1]*Department of Computer Engineering, Faculty of Engineering, Munzur University, Tunceli, Türkiye (ORCID: 0000-0002-7859-7543)*

Keywords: Deep Learning, Voice Command Recognition, Neural Network, Feature Extraction

Abstract

In this study, a deep learning model was developed for the recognition and classification of voice commands using the Turkish Speech Command Dataset. The division of training, validation, and test sets was carried out on an individual basis. This approach aims to prevent the model from memorizing and to enhance its generalization capability. The model was trained using Mel-Frequency Cepstral Coefficients features extracted from voice files, and its classification performance was evaluated in detail. The findings indicate that the model successfully classifies voice commands with a high accuracy rate, achieving an overall accuracy of 92.3% on the test set, highlighting the potential of deep learning approaches in voice recognition technologies

1. Introduction

Today, voice and speech recognition systems are gaining increasing importance in many areas such as smart assistants, automatic customer services, and biometric security systems. The foundation of these systems lies in the ability to extract meaningful features from raw voice signals and to make accurate classifications based on these features. This process requires the development of an effective machine learning model [1,2].

Deep learning is revolutionizing many fields today and is providing groundbreaking innovations across numerous sectors [3]. Deep learning techniques are being applied in areas including, but not limited to, image recognition and processing [4,5,21,23,24], natural language processing [6], and robotics [7], thereby expanding the boundaries of machine learning and artificial intelligence technologies.

Deep learning also plays a significant role in the field of voice recognition and processing. Applications such as speech recognition, voice assistants, emotion analysis, and speaker verification have shown significant development thanks to deep learning models [8,9,10]. The success in this field is fundamentally due to the ability to extract effective features from voice signals. In particular, feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC) have achieved great success in capturing the characteristic features of voice signals. MFCC, which is based on the short-term spectral analysis of voice signals' frequency components, is widely used in speech recognition, voice classification, and similar tasks. This technique is crucial for training deep learning models, enabling them to interpret voice data and classify with high accuracy [11,12].

In the literature, MFCC have been frequently used. Deng et al. have proposed a new method for heart sound classification based on MFCC features and convolutional recurrent neural networks, processing heart sound signals entirely with MFCC in their study. This method allows for a detailed analysis of temporal changes in heart sounds, extracting features that, when applied to deep learning techniques, achieve high classification success rates in the early diagnosis of heart diseases [14].

Rejaibi et al. have presented a framework based on deep Recurrent Neural Networks (RNN) for detecting depression from speech and predicting the severity

*Corresponding author: celikyusuf@munzur.edu.tr
Received: 02.05.2024, Accepted: 25.07.2024

level of depression. MFCC was used for the extraction of voice features [15].

Anjana et al. conducted a study on speech recognition across various local languages in India, utilizing MFCC and formant frequencies to distinguish between languages. Comparing two classification algorithms, LDA and SVM, their findings demonstrated that LDA outperformed SVM with a classification accuracy of 93.88% [20].

In the study conducted by Putra et al., an automatic door control system was developed using voice commands "buka" (open) and "tutup" (close). The research utilized MFCC and Convolutional Neural Networks (CNN) for feature extraction and classification of sound signals. The CNN model achieved a success rate of 89% in classifying these commands [22].

The number of existing studies on the classification of Turkish voice commands is limited. The aim of this study is to fill the existing gaps in the recognition and classification of Turkish voice commands and to provide a new contribution in this field. In this study, a voice recognition system was developed using the Turkish Speech Command Dataset with the MFCC feature extraction method followed by a deep learning-based classification model. The dataset comprises voice recordings from various speakers, categorized into different classes. The model's training, validation, and testing were conducted with the dataset divided into three parts: training, validation, and test. The deep learning model, created using various layers such as CNN and Long Short-Term Memory networks (LSTM), aims to successfully classify voice recordings from the dataset into relevant classes using the extracted MFCC features.

2. Material and Method

The study was conducted on the Turkish Speech Command Dataset. MFCC features were extracted from the voice files of this dataset, and these features were used to develop a deep learning model. The extracted MFCC features formed the basis for training the model, after which the model's classification performance was tested in detail. This study was carried out using the Python programming language in the Jupyter Notebook environment. The Keras library was utilized for coding the deep learning model, and the Librosa library was used for audio processing. The methodology of the study is shown in Figure 1.

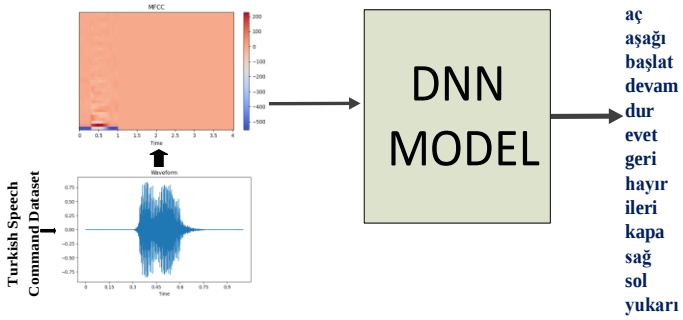

Figure 1. Overview of the Study Methodology

2.1. Dataset Description

The dataset considered in this study is specifically compiled for the recognition and processing of speech commands in Turkish. It has been published on the Kaggle platform by MURAT KURTKAYA under the name "Turkish Speech Command Dataset"[16].

The "Turkish Speech Command Dataset" consists of 14 different speech commands commonly used in daily life. These commands are as follows. The dataset contains a total of 26,484 voice recordings, each lasting 1 second and recorded at a sampling frequency of 16 kHz. The voice recordings are stored in wav format. Each voice recording is uniquely named in the format of "command-personID-specificID", which allows for the effective classification and processing of the recordings. The numerical distribution of the commands included in the dataset is shown in Figure 2.

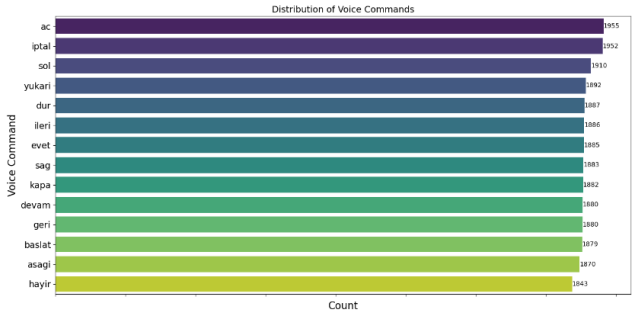

Figure 2. Number of Voice Commands in the Dataset

The "database.xlsx" file contains metadata for each sound recording, playing a crucial role in organizing the recordings by individuals and commands. This file comprises five columns: "id," "person," "type," "gender," "path." The "id" serves as a unique identifier for each sound file. The "person" column codes the identity of the individual who made the recording. "type" indicates the recorded sound command, while "gender" shows the gender of the recorder. "path" specifies the location of the sound

recording. The content of the "database.xlsx" file is detailed in Table 1.

Table 1. "database.xlsx" File Content

id	person	type	gender	path
ac_UXOD_BTMEAPU	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_YJTEPUH	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_CKTQBAP	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_YCJKIMB	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_YJKRCPO	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_HJKNTRU	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_XYGONEW	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_HRWYZVK	UXOD	ac	ERKEK	dataBase\ac\
ac_UXOD_AQNLZBD	UXOD	ac	ERKEK	dataBase\ac\
asagi_UXOD_OQFUBAC	UXOD	asagi	ERKEK	dataBase\asagi\
asagi_UXOD_GSYOHNZ	UXOD	asagi	ERKEK	dataBase\asagi\
asagi_UXOD_QIUOVND	UXOD	asagi	ERKEK	dataBase\asagi\

The dataset uniquely consists of 257 individuals, among which 140 are identified as "ERKEK" and 117 as "KIZ".

2.2. Mel Frequency Cepstral Coefficients (Mfcc)

The Mel-Frequency Cepstral Coefficients (MFCC) method [13], introduced by Davis and Mermelstein, is a technique used to extract rich and meaningful features from audio signals, mimicking the logarithmic response of the human ear to sound frequencies. This method plays a critical role in speech and sound recognition systems by capturing the fundamental components of sound and their variations over time. MFCC are commonly used in speech processing and recognition applications because they model the sound spectrum in a way similar to human auditory perception [17]. The steps of MFCC are illustrated in Figure 3.

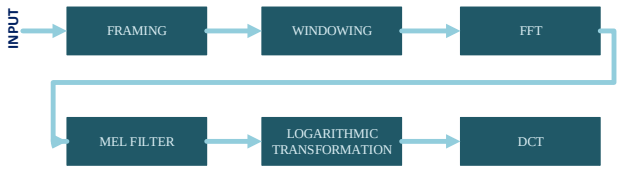

Figure 3. Steps of Mel Frequency Cepstral Coefficients (MFCC)

MFCC primarily consists of the following steps:

- Framing: The audio signal is divided into short time frames. This allows for the analysis of the audio signal in segments. Since the audio signal varies over time, analyzing a long audio recording directly is not practical. Instead, dividing the signal into short-term segments that can be considered quasi-stationary allows for a more detailed examination of each segment's characteristics.
- Windowing: Applying a window function to each frame mitigates spectral leakage that can occur at the edges of the frame (at the start and end points of the signal).
- FFT (Fast Fourier Transform): A fundamental step for analyzing the frequency content of the signal, FFT is used to identify the characteristic features of the audio signal, such as its pitch and harmonics.
- Mel Filter: Reflecting the human ear's greater sensitivity to lower frequencies and decreasing sensitivity towards higher frequencies, the Mel filter allows for a perceptually more relevant representation of the audio signal.
- Logarithmic Transformation: Since the perception of sound intensity by humans is logarithmic, this step ensures the audio signal is represented in a way that is closer to human perception.
- Cepstral Coefficients (DCT - Discrete Cosine Transform): Used to effectively capture the characteristic features of the audio signal, such as the voice tone of the speaker or the unique structure of spoken words [17,18].

2.2 Model Architecture

The model used in this study consists of a deep learning architecture designed for audio classification tasks. Initially, a convolutional layer (Conv1D) is employed to extract fundamental features from the audio signals. This layer is followed by batch normalization, which regulates the learning process and enhances stability. The max pooling (MaxPooling1D) operation reduces the size of the feature map, thus lowering the model's complexity and computational load. Additionally, dropout layers are incorporated to mitigate the risk of overfitting. An LSTM layer is added for processing time-dependent features, aiming to learn the temporal dynamics of the model. The model is trained using categorical cross-entropy loss and the Adam optimization algorithm. This approach allows the model to accurately distinguish between different classes. Further details on the model's architecture and layers are presented in Figure 4.

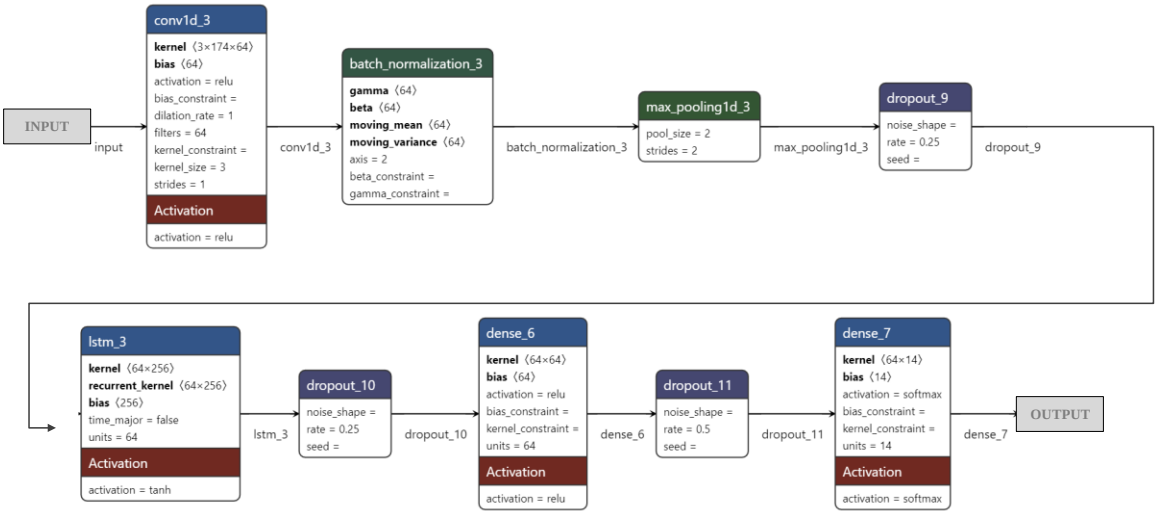

Figure 4. Convolutional Neural Network Model Used in the Study

3. Results and Discussion

To prevent overfitting on the data set, a strategy of separating the training, validation, and test sets on an individual basis has been adopted. Figure 5 illustrates how this distribution was implemented.

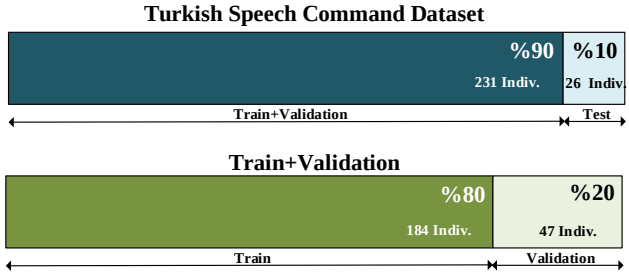

Figure 5. Distribution of the Dataset

Initially, the dataset, consisting of a total of 257 individuals, was divided into two main groups: a test set and a training set. During this division, 10% of the dataset was allocated for the test set, and the remaining 90% was used for the training set. Consequently, 26 individuals were designated for the test set and 231 for the training set.

The 231 individuals in the training set were further divided into training and validation sets. 80% of these individuals were allocated to the training set, and the remaining 20% to the validation set. Thus, the training set included 184 individuals, and the validation set included 47 individuals.

This person-based distribution resulted in 19,490 audio files for the training set, 4,410 for the validation set, and 2,548 for the test set. This methodology aimed to enhance the model's generalization ability and prevent it from memorizing person-specific characteristics. Alternatively, methods such as K-Fold Cross Validation can be used to evaluate the model's performance more reliably. This method measures the model's overall performance more accurately by training and testing it multiple times on different subsets of the dataset. However, in this study, a person-based split method was preferred because this approach aims to prevent the model from memorizing person-specific features and to enhance its generalization ability.

There are various versions of MFCC filter banks, each with different numbers of filters and their amplitudes. A commonly used filter bank, originally developed for speech analysis, consists of 40 filters in the Mel band-pass filter. This filter, like the human ear's perception of speech, aims to extract a non-linear representation of the speech signal. The conventional Mel filter bank is made up of 40 triangular filters. In this study, a Mel filter bank comprising 40 filters was used to extract meaningful features from the audio signals. These features were then used as inputs to the model.

3.1. Evaluation of the Model's Performance

The model was trained on a total of 19,490 training samples and validated with 4,410 validation samples. The validation and loss graphs related to the model are shown in Figure 6.

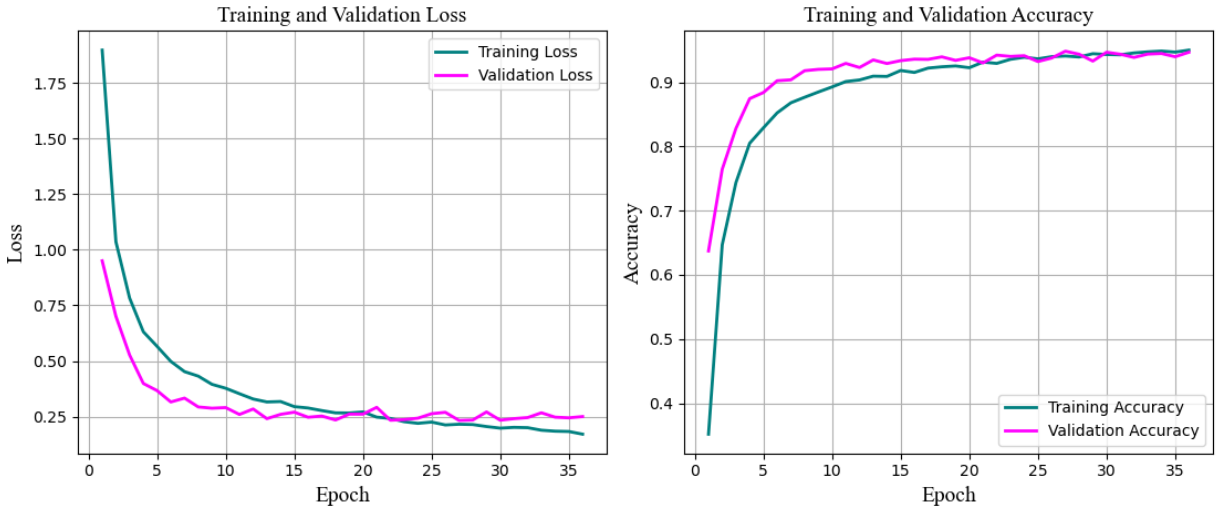

Figure 6. Validation and Loss Graphs of the Model

Training was planned for a total of 150 epochs, but due to the "early stopping" criteria, it was terminated at the 36th epoch. Early stopping halts the training process when the model's performance on the validation set does not improve over a certain period. Throughout the training process, both training and validation losses decreased while accuracy rates increased, indicating the effectiveness of the learning process and an improvement in the model's generalization capability.

The confusion matrix is a table used to evaluate the performance of classification algorithms. For each class, the model includes true positives (correct classification), false positives (misclassification of a sample as this class), false negatives (instances of this class being classified as another class), and true negatives (correctly identifying a non-member of this class) [19]. The classification confusion matrix of the model on the test data is shown in Figure 7.

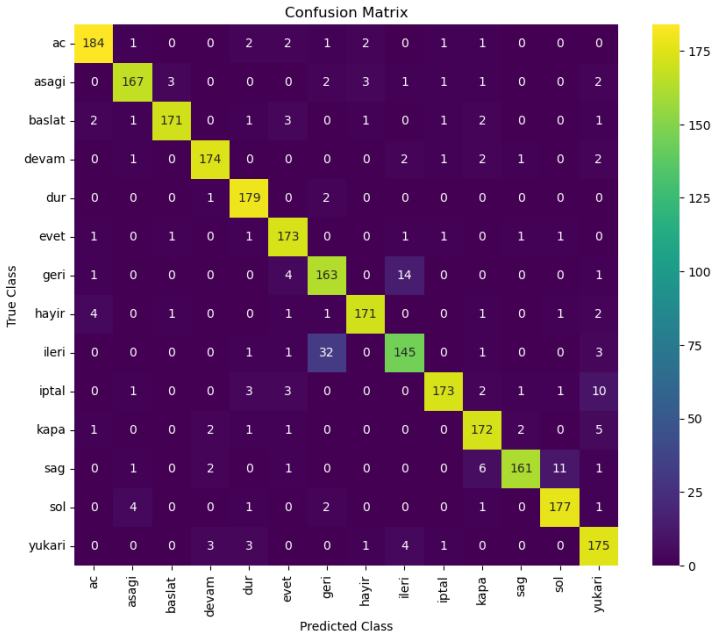

Figure 7. Classification Confusion Matrix on Test Data

The obtained confusion matrix meticulously highlights the performance of our deep learning-based voice command recognition model on the test dataset. The significant numbers along the main diagonal indicate the model's ability to recognize various voice commands with high accuracy, showcasing its robust training and generalization capabilities. Notably, the high classification accuracy rates for commands such as 'ac', baslat, 'devam', and 'dur' demonstrate the model's clear distinction and proper learning of these commands.

However, the low performance between the 'geri' and 'ileri' commands is due to their acoustic similarity. This makes it difficult for the model to distinguish between these two commands. To mitigate the impact

of this low performance on the model's overall success and to improve their performance, several strategies can be applied. Firstly, collecting more data for the 'geri' and 'ileri' commands can help the model learn these commands better. Additionally, using data augmentation techniques can increase the diversity of the existing dataset, thereby improving the model's generalization capability.

The model has demonstrated 92.30% accuracy across 2584 samples in the test set, indicating its capability to classify unseen data with a high accuracy rate. Table 2 includes metrics such as precision, recall, and F1-score for each class. These metrics illustrate how accurately the model predicts the classes and the balance of performance among them. The 'Precision' metric for each class denotes the rate at which the model accurately predicts that class; 'Recall' indicates the proportion of correctly predicted samples among all samples belonging to that class. The 'F1-Score', calculated as the harmonic mean of precision and recall, reflects the model's balanced performance across all classes. The 'Support' column shows the number of samples for each class, representing the sample size used in the calculation of these metrics.

Overall, the model maintains high precision and recall values for most classes. Although there might be performance variations among some classes, the model generally exhibits a balanced performance. The performance on both validation and test sets indicates a high generalization capability and successful avoidance of overfitting.

Table 2. Classification Performance Metrics by Voice Commands

Command	*Precision(%)*	*Recall(%)*	*F1 Score(%)*	*Support*
ac	95	95	95	194
asagi	95	93	94	180
baslat	97	93	95	183
devam	96	95	95	183
dur	93	98	96	182
evet	92	96	94	180
geri	80	89	84	183
hayir	96	94	95	182
ileri	87	79	83	183
iptal	97	89	93	194
kapa	91	93	92	184
sag	97	88	92	183
sol	93	95	94	186
yukari	86	94	90	187

When Table 2 is examined, it is observed that the model demonstrates impressive performance in classes such as 'ac', 'asagi', and 'baslat', with precision and recall values reaching up to 95%. However, in certain classes like ‘geri’ and ‘ileri’, these metrics are comparatively lower, indicating the challenges the model faces in distinguishing these commands. The recall value for the ‘geri’ class is 89%, while for ‘ileri’, this value is 79%, suggesting the model differentiates ‘geri’ more accurately than ‘ileri’.

4. Conclusion and Suggestions

The deep learning model developed in this study possesses an effective ability to recognize and classify Turkish voice commands. The model has achieved a high overall accuracy rate of 92.30%. A methodological innovation of this study is the person-based separation process of the training, validation, and test sets, which has significantly contributed to enhancing the model's generalization capability. This approach has prevented the model from memorizing person-specific features, enabling it to exhibit a robust and generalizable classification performance. Although some classes like ‘geri’ and ‘ileri’ show lower performance, the model generally displays a balanced performance. This study proves the applicability and effectiveness of deep learning models in the development of voice recognition and classification systems. Future work is expected to further improve the model and test it on various voice datasets.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

References

[1] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021.

[2] F. Afandi and R. Sarno, "Android application for advanced security system based on voice recognition, biometric authentication, and internet of things," *in 2020 International Conference on Smart Technology and Applications (ICoSTA)*, Feb. 2020, pp. 1-6.

[3] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ Digital Medicine*, vol. 4, no. 1, p. 5, 2021.

[4] C. Li, X. Li, M. Chen, and X. Sun, "Deep learning and image recognition," in *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT*), July 2023, pp. 557-562.

[5] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya, "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images," *Pattern Recognition Letters* vol. 133, pp. 232-239, 2020.

[6] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," arXiv preprint arXiv:1503.00075, 2015.

[7] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, A review," Cognitive Robotics, 2023.

[8] Z. Bai and X. L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65-99, 2021.

[9] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," *arXiv preprint arXiv:1708.01227*, 2017.

[10] P. Dhakal, P. Damacharla, A. Y. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 504-520, 2019.

[11] M. D. Shakil, M. A. Rahman, M. M. Soliman, and M. A. Islam, "Automatic Isolated Speech Recognition System Using MFCC Analysis and Artificial Neural Network Classifier: Feasible For Diversity of Speech Applications," in *2020 IEEE Student Conference on Research and Development (SCOReD)*, Sept. 2020, pp. 300-305.

[12] H. Dolka, A. X. VM, and S. Juliet, "Speech emotion recognition using ANN on MFCC features," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, May 2021, pp. 431-435.

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

[14] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Networks*, vol. 130, pp. 22-32, 2020.

[15] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.

[16] M. Kurtkaya, "Turkish Speech Command Dataset [Data set]," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/murat-kurtkaya/turkish-speech-command-dataset

[17] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its applications: A Review," *IEEE Access*, 2022.

[18] T. Maka, "Change point determination in audio data using auditory features," *International Journal of Electronics and Telecommunications*, vol. 61, no. 2, pp. 185-190, 2015.

[19] M. Tripathi, "Analysis of convolutional neural network based image classification techniques," *Journal of Innovative Image Processing (JIIP)*, vol. 3, no. 02, pp. 100-117, 2021.

[20] Anjana, J. S., and Poorna, S. S., "Language identification from speech features using SVM and LDA," in *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Mar. 2018, pp. 1-4.

[21] C. Ozdemir and Y. Dogan, "Advancing brain tumor classification through MTAP model: an innovative approach in medical diagnostics," *Medical & Biological Engineering & Computing*, pp. 1-12, 2024.

[22] B. S. P. Laksono, T. Syaifuddin, and F. Utaminingrum, "Voice Recognition to Classify 'Buka' and 'Tutup' Sound to Open and Closes Door Using Mel Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN)," *Journal of Information Technology and Computer Science*, vol. 9, no. 1, pp. 58-66, 2024.

[23] C. Ozdemir, "Adapting transfer learning models to dataset through pruning and Avg-TopK pooling," *Neural Comput & Applic*., vol. 36, pp. 6257–6270, 2024. https://doi.org/10.1007/s00521-024-09484-6

[24] C. Ozdemir, "Classification of brain tumors from MR images using a new CNN architecture," *Traitement du Signal*, vol. 40, no. 2, pp. 611-618, 2023. https://doi.org/10.18280/ts.400219