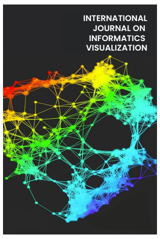
INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv



A Multi-Feature Fusion Approach for Dialect Identification Using 1D CNN

Sarkhel H.Taher Karim [a,*], Karzan J. Ghafoor [a], Ayub O. Abdulrahman [a], Karwan M. Hama Rawf [a]

[a] *Computer Science Department, College of Science, University of Halabja, Halabja 46018, Kurdistan Region, Iraq*
Corresponding author: [*]sarkhel.kareem@uoh.edu.iq

Abstract: The phonological variety of Kurdish, a language with several dialects, poses a distinct problem in automatically identifying dialects. This study examines and evaluates several sound criteria for identifying Kurdish dialects: Badini, Hawrami, and Sorani. We deployed a dataset including 6,000 samples and utilized a mix of 1D convolutional neural networks (CNN) and fully connected layers to conduct the identification job. Our study aimed to assess the efficacy of different sound characteristics in accurately identifying dialects. We employed the Mel-frequency Cepstral Coefficients (MFCC) and other features such as the Mel spectrogram, spectral contrast, and polynomial features to extract the sound characteristics. We conducted training and testing of our models utilizing both individual characteristics and a composite of all features. Our analysis revealed that the identification task achieved excellent accuracy rates, suggesting a promising potential for success. We achieved 95.75% accuracy using MFCC combined with a Mel spectrogram. The accuracy improved by including contrast in the MFCC feature extraction process to 91.42%. Similarly, using poly_features resulted in an accuracy of 90.83%. Remarkably, accuracy reached a maximum of 96.5% when all the attributes were combined.

***Keywords*— Kurdish dialect identification; sound features; 1D convolutional neural networks; spectral contrast; polynomial features.**

Manuscript received 25 Sep. 2023; revised 13 Feb. 2024; accepted 12 Mar. 2024. Date of publication 30 Sep. 2024.

I. Introduction

Speech-processing methods are now used in numerous fields in present-day society. Dialect identification refers to determining the regional accent to which a specific linguistic variant belongs [1]. Compared to other available communication modes, people often view speaking as the simplest, quickest, and most natural connection method [2]. Dialect refers to the linguistic variance within a population that has developed in response to various environmental factors. Because of its usefulness in many contexts, including voice recognition and forensics [3], Dialect recognition (DR) has gained significant attention recently. Normalizing voice samples for a speech recognition system requires various skills, including identifying a speaker's dialect or accent [4].

Kurdish is spoken with regional variations in four central Middle Eastern countries [5]. Sorani, Badini, and Hawrami are the three most widely spoken varieties of Kurdish in the Kurdistan region of Iraq [3]. Most Kurdish speakers live in communities spread throughout Turkey, Iraq, Iran, and Syria. More than 40 million people are estimated to speak Kurdish [6]. Identifying a Kurdish Dialect Identification (DID) system is a significant challenge due to the need for clear separations between various Kurdish dialects.

The ability to identify a speaker's dialect is a subset of language recognition. For differentiation from the standard language from which it evolved [7]. In this context, language recognition techniques are appropriate because dialect identification is similar to language recognition. Deep learning refers to a complete collection of artificial neural networks that use deep architecture and advanced methodologies to process language recognition [8] effectively. Convolutional Neural Networks, or CNNs for short, are a subcategory of deep neural networks that effectively develop voice recognition applications [9].

Most modern scientific research has concentrated on using acoustic characteristics in sound, speaker, and language identification. Several factors influence the effectiveness of automated speech recognition systems: gender, age, and geographical differences in speech patterns [10]. The most current approach in dialect identification uses a hybrid system combining Mel-frequency cepstral coefficients (MFCC) and pitch feature candidates for feature extraction, together with a traditional machine-learning algorithm for dialect classification [11]. Using numerous distinct hybrid feature extraction techniques is essential to enabling the progress of a higher-quality system.

Regarding linguistic variety, Kurdish Sign linguistic (KuSL) recognition and Dialect Recognition Systems (DRS) aim to improve communication access. DRS tackles elements influencing voice recognition and suggests system integration to fit dialects. By adopting a modified CNN architecture, KuSL recognition achieves excellent accuracy in spotting Kurdish signals. Both domains depend on feature extraction, pattern recognition, and deep learning, even if they use distinct data kinds—images for signs and audio waves for languages. This parallel quest presents a complete approach to linguistic variety and fascinating opportunities for unified frameworks [12], [13].

This work uses numerous feature extraction techniques based on convolutional neural networks (CNN) to identify the dialects of the Kurdish language, employing the shortest speech of a sound sample. The aim is to evaluate these aspects' efficiency. The paper's novelty introduces a novel dialect identification technique wherein convolutional neural networks (CNNs) are included in the feature extraction process. This hybrid feature extraction method improves the accuracy of dialect recognition by using a convolutional neural network (CNN) to identify unique, detailed patterns distinct to every dialect. Combining conventional characteristics such as Mel-frequency cepstral coefficients (MFCCs), Mel spectrograms, and spectral contrast attributes with convolutional neural networks (CNNs) that assess spectrogram representations of voice input results in state-of-the-art performance. This method dramatically lessens the demand for thorough human feature building. This creative technique offers a workable means of differentiating dialects throughout a broad spectrum of linguistic and cultural environments.

Dialect identification systems vary depending on the language being examined, the parameters of the dataset utilized, the techniques used for feature extraction, and the classifiers used to identify dialects based on the words being used. Researchers have efficiently classified dialects across several languages by combining machine-learning techniques with many feature extraction methods. Based on unidentified spoken utterances, such research [14] aimed to pinpoint the three dialects of the Telugu language. Researchers applied different spectral and prosodic levels of the speech stream to identify the several dialects. Spectral qualities (MFCC, Delta MFCC, and Delta-Delta MFCC) and prosodic elements of speech signals are considered in the system application based on the Gaussian Mixture Model and Hidden Markov Model. For the Prosodic+MFCC feature, the obtained values for the GMM and HMM models were 88.40%, respectively. Moreover, a strategy for accepting the several dialects spoken in Indonesia was established. First, a spoken expression consists of preprocessing, normalizing, and framing devices. Moreover, the Mel frequency cepstral coefficients (MFCC), a well-known method for obtaining acoustic features, help us retrieve the audio signals' properties. Knowledge is learned, and dialect traits are categorized in the research using a deep recurrent neural network (DRNN). The training set came with an overall accuracy of 87.0%. Applied to dialects not previously encountered, the evaluation of the testing set shows an accuracy rate of 85.4% [15].

Regarding phonetic and grammatical features, Arabic is close to the Kurdish language. From this perspective, the closeness of both languages shows the linguistic influence. Like any other language, scholars have turned to a broad spectrum of feature extraction methods to accurately categorize the several Arabic dialects spoken in several Arab nations. For example, a novel technique was developed to detect repeating sequences (motifs) indicative of each Arabic dialect, extracting several dialects' unique characteristics straight from the voice signal. Within the framework of motif extraction, the scientists applied an effective parameter-free method called Scalable Time series Ordered search Matrix Profile (STOMP). From every theme, the researchers extracted 12 MFCCs to create features. Training a Gaussian Mixture Model-Universal Background Model (GMMUBM) as a classifier using these coefficients came next. Attached in [16] is the remarkable maximum degree of accuracy of 70%.

An emotion identification system's success depends on the traits obtained and the classifier applied for emotional detection. Mel-frequency cepstral coefficients (MFCC), Mel-frequency spectrograms (MEL), and chroma include important audio qualities in the feature vector. Two sets of features taken together create feature vectors. Emotions were categorized using many machine-learning techniques [17]. Furthermore, Mel spectrograms are also crucial for identifying dialects as they translate unprocessed audio inputs into visual representations that successfully depict the spectrum properties vital for human auditory perception. Mel spectrograms—based on the Mel scale corresponding to human auditory perception of speech sounds—greatly help capture prominent phonetic characteristics. As such, this method significantly helps to classify dialects precisely. Examining the physical level of acoustic properties allows one to find the phonetic changes in speech. Log mel-spectrograms of CNN and LSTM kinds help to categorize Turkish dialects depending on their sound and speaking quality. Within the framework of the phonotactic method, LSTM neural networks have extensive applications in language modeling. Results reveal [7] that there is 85.1% accuracy for Turkish dialect recognition.

The identification and distinction of dialects depending on auditory characteristics depends critically on spectral contrast features. By measuring variations in spectral amplitude, the researchers can record various speech patterns and phonetic differences between several languages. These features improve the ability of machine learning systems to classify and differentiate languages using acoustic attributes correctly. [18] presents the results of evaluating several aggregation systems for acoustic features used in environmental sound classification (ESC) tasks to identify the most compelling feature aggregate strategies for tackling the difficult challenge of improving the classification accuracy of environmental sounds. Extensive experimentation has shown that a feature combination consisting of MFCC, Log-mel Spectrogram, Chroma, Spectral Contrast, and Tonnetz can reach state-of-the-art classification accuracy on the ESC dataset (85.6%) and the Urban Sound 8K dataset (93.4%). Another article compared SVM, MLP, and KNN in terms of their accuracy in identifying emotions conveyed via the Saudi dialect of spoken language. Anger, joy, sorrow, and apathy were the four feelings examined. The classification was performed using spectral features. The highest KNN accuracy was 68.57% using MFCC and spectral contrast. Accuracy for both SVM and MLP prediction was improved by adding the mel spectrogram features (77.14% and 71.43%, respectively) [2].

II. Materials and Method

The technique used for dialect recognition classification with a Convolutional Neural Network (CNN) model encompasses a range of feature extraction methods, including Mel-frequency Cepstral Coefficients (MFCCs) and spectral features. Subsequently, these characteristics are input into a customized convolutional neural network (CNN) architecture to extract features and perform classification. The model goes through training, optimization, and evaluation using datasets labeled with dialect information. This enables the model to identify dialects by analyzing the extracted characteristics, providing valuable insights into the linguistic variety present in spoken language. Metrics like F1-Score are part of the assessment process, which helps determine how well the model is doing. The model's effectiveness is assessed using the F1-Score, a statistic often used to classify tasks. The F1-Score objectively evaluates the model's accuracy in classifying cases by considering precision and recall. It is computed as the harmonic means of recall and precision, where recall gauges how well the model can capture all relevant cases, and precision indicates the accuracy of optimistic predictions.

The F1-Score is well-suited for our dialect identification assignment since it considers both false positives and false negatives, which is essential when there are imbalances among the various dialect classes. The F1-Score may be calculated using the following formula:

$$F1 - Score \ = \frac{2*Precision*Recall}{recision+Recall} \quad (1)$$

The block diagram of the suggested dialect recognition model is shown in Figure 1.

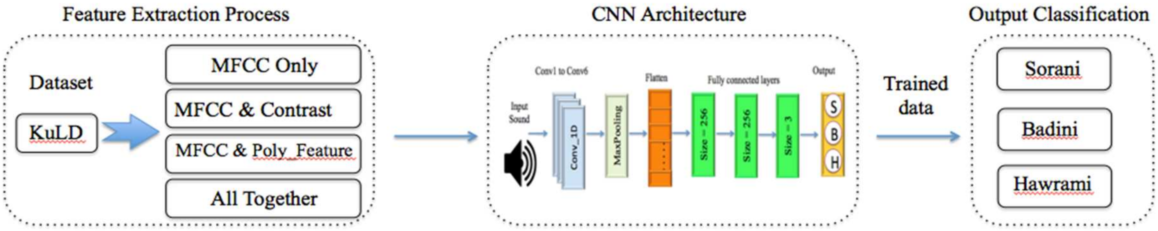


Fig. 1 Shows a block schematic of the suggested model.

A. Research Contribution

Researchers have conducted studies investigating the identification of dialects, including various languages, databases, and methodologies for extracting features. Nevertheless, a continual obstacle persists in differentiating dialects from brief fragments of speech. This study seeks to fill this void by examining the efficacy of convolutional neural networks (CNNs) in extracting distinctive characteristics to identify dialects from voice samples. This technique aims to enhance performance using convolutional neural networks (CNNs) to combine information and capture unique phonetic properties and patterns particular to different dialects from spectral representations. Moreover, Convolutional Neural Networks (CNNs) seek to diminish the dependence on manually designed characteristics, which might result in a resilient and flexible model for identifying languages with a scarcity of voice data. This work may have consequences for applications such as voice recognition and speaker discrimination, mainly when lengthier speech segments are limited datasets, including recordings of dialectal speech.

B. Dialectal Speech Dataset

The dataset used in this study is the Kurdish Language Dialects (KuLD) dataset, which was collected by a team of instructors from the Computer Science Department at the University of Halabja. The data collection process took several months. Throughout each phase of the data collection process, exact conformity to established rules, processes, and standards was maintained. This included considering the ages and genders of the speakers included in the dataset. 2000 samples were recorded for each Sorani, Badini, and Hawrami dialect. The dataset has a total period of 6000 seconds, with each sample having an exactly one-second duration [19].

C. CNN Architecture

The proposed model employed the KuLD dataset of sound data to classify Kurdish dialects using a one-dimensional convolutional neural network (1-D CNN). The primary data included in this study has been categorized as Sorani, Badini, and Hawrami samples, which scholarly experts assigned. CNN has been increasingly recognized and used in computer vision and audio processing. These areas include tasks not affected by the specific location of patterns on spectrogram images. Consequently, CNN has emerged as a viable and effective approach for accurately categorizing spectrogram features [20]. Feature extraction and classification are fundamental to convolutional neural networks (CNNs). It is essential to extract optimal features to accurately categorize signal processing [21]. The suggested 1D CNN model has twelve layers, including an input layer, six 1-D convolution layers, a MaxPooling layer, a flattened layer, and three fully connected layers. The schematic representation of the suggested one-dimensional convolutional neural network (1D CNN) model is shown in Figure 2.

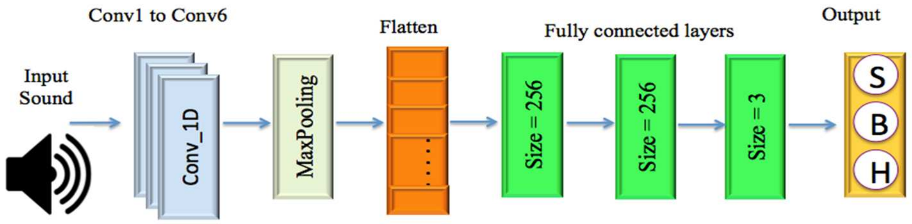


Fig. 2 Illustrates the overall structure of a 1-D CNN model.

As a result, the parameters of the presented technique have been precisely tuned to achieve a notable level of accuracy in categorizing the Kurdish dialect. Table 1 describes the specific parameters of the proposed Convolutional Neural Network (CNN) model for each feature extraction.

TABLE I
SHOWS AN OVERVIEW OF THE LAYERS USED IN THE CNN MODEL

Layer	MFCC only		MFCC and Contrast		MFCC and MEL		MFCC and Poly_Feature		All together	
	Output Shape	Param	Output Shape	Param	Output Shape	Param	Output Shape	Param	Output Shape	Param
conv1d_1 (Conv1D)	(None, 125, 128)	640	(None, 132, 128)	640	(None, 174, 128)	640	(None, 128, 128)	640	(None, 263, 128)	640
conv1d_2 (Conv1D)	(None, 122, 128)	65664	(None, 129, 128)	65664	(None, 171, 128)	65664	(None, 125, 128)	65664	(None, 260, 128)	65664
conv1d_3 (Conv1D)	(None, 119, 32)	16416	(None, 126, 32)	16416	(None, 168, 32)	16416	(None, 122, 32)	16416	(None, 257, 32)	16416
conv1d_4 (Conv1D)	(None, 116, 32)	4128	(None, 123, 32)	4128	(None, 165, 32)	4128	(None, 119, 32)	4128	(None, 254, 32)	4128
conv1d_5 (Conv1D)	(None, 113, 128)	16512	(None, 120, 128)	16512	(None, 162, 128)	16512	(None, 116, 128)	16512	(None, 251, 128)	16512
conv1d_6 (Conv1D)	(None, 110, 128)	65664	(None, 117, 128)	65664	(None, 159, 128)	65664	(None, 113, 128)	65664	(None, 248, 128)	65664
max_pooling1d_1 (MaxPoolining 1D	(None, 18, 128)	0	(None, 19, 128)	0	(None, 26, 128)	0	(None, 18, 128)	0	(None, 41, 128)	0
flatten_1 (Flatten)	(None, 2304)	0	(None, 2432)	0	(None, 3328)	0	(None, 2304)	0	(None, 5248)	0
dense_1 (Dense)	(None, 256)	590080	(None, 256)	622848	(None, 256)	852224	(None, 256)	590080	(None, 256)	1343744
dense_2 (Dense)	(None, 256)	65792	(None, 256)	65792	(None, 256)	65792	(None, 256)	65792	(None, 256)	65792
dense_3 (Dense)	(None, 3)	771	(None, 3)	771	(None, 3)	771	(None, 3)	771	(None, 3)	771

D. Features Extraction

Sound feature extraction is essential for sound identification jobs. By isolating pertinent attributes from audio signals, these features offer significant data that may be utilized for categorization and recognition [22]. Here are some frequently employed methods for extracting sound features:

1) MFCC Features: The MFCC (Mel-Frequency Cepstral Coefficients) approach is based on how people hear sound frequencies between 20 and 20,000 Hz. The phrase MFCC is an acronym that represents four essential elements: Mel, frequency, cepstral, and coefficients, which concisely summarize the fundamental concepts of this approach. The MFCC process derives a set of 40 or more coefficients from the audio signal. This number has been determined sufficient for accurately representing speech samples, as described in references [23] and [24]. Figure 3 depicts the process of Mel-frequency cepstral coefficients (MFCC).

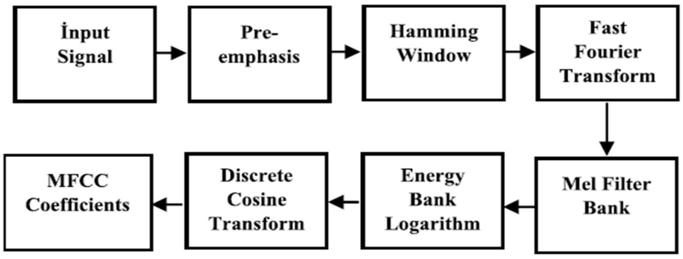


Fig. 3 MFCC feature extraction process

2) Mel Spectrogram: Representing the power spectrum of audio data on the Mel scale, the Mel spectrogram is a major tool for feature extraction. More specifically, it entails the creation of the Mel spectrogram [25], which, depending on human auditory perception, catches the most relevant sound qualities. The method starts with a small frame segmenting of the audio stream and then generates the spectrogram using the discrete Fourier transform (DFT). The Mel Filterbank is then used to calculate the energy within every Mel filter; finally, the logarithm of energy is computed for every filterbank [26].

3) Spectral Contrast Features: Previous studies [27] have shown how well the Spectral Contrast method distinguishes audio genres. This method considers several facets of every sub-band, including spectral peaks, troughs, and their variations [28]. It separates audio frames into sub-bands and measures the contrast in energy using an average energy level at the most significant point (peak energy) and the lowest point (valley energy) within each sub-band. To evaluate its capability in detecting COVID-19 instances, we investigated cough audio recordings using the spectral contrast feature extraction technique. Denoted as SC_k, the mathematical formula for Spectral Contrast is shown in equation (2) [28].

$$SCk = Peakk - V\,alleyk \qquad (2)$$

where k indicates a sub-band, $Peak_k$ is the maximum value and $V\,alley_k$ is the minimum value in each sub-band [22], [29].

4) Poly-feature: A feature extraction method known as poly-feature models the spectral envelope's Poisson fit. It reveals the general form of the audio stream, therefore offering an understanding of the global features of the dialects [22]. In our model, fitting a second order (quadratic) polyn to the audio signal computes these characteristics.

III. Results and Discussion

This section presents the outcomes of our classification experiment, which employed a 1D-CNN model and various feature combinations, including MFCC, Mel Spectrogram, Poly-feature, and Contrast. The information provided in Table 2 pertains to the performance of various feature sets and their impact on the accuracy of language classification.

TABLE II
Accuracy of different feature extraction used in

Feature extracted	1D-CNN Results %
MFCC	91.92%
MFCC + Mel Spectrogram	95.75%
MFCC + Poly-feature	90.83%
MFCC + Contrast	91.42%
MFCC + Mel Spectrogram + Poly-feature + Contrast	96.5%

A. Feature Integration and Model Performance

With just the MFCC data, the suggested model had an accuracy of 91.92%. These first results set the basis for the categorization. By including Mel Spectrogram elements into the MFCC model, its performance was much enhanced, and its accuracy rate was outstanding—95.75%. This mix improved the model's capacity to gather significant spectrum information efficiently, raising classification efficiency.

Integration of the poly-feature with the MFCC produced a 90.83% performance level. Although this mix did not surpass the Mel Spectrogram in terms of accuracy, it underlined the need for feature selection in maximizing classification outcomes. Together with MFCC, the Contrast function produced an accuracy of 91.42%. This mix showed that more than the Contrast feature alone is needed, even if it offered little benefit over MFCC.

The best model came from mixing MFCC, Mel Spectrogram, Poly-feature, and Contrast. Using all accessible features, this integrated model produced an accuracy of 96.5%. The combinatorial effect of incorporating several auditory characteristics allows the 1D-CNN language classification model to reach such outstanding accuracy.

We investigated the confusion matrices, producer accuracy (precision), and F1-scores for every class in great detail to show the potency of the whole feature combination chosen for our model. Table 3 shows these outcomes.

TABLE III
Confusion matrices and accuracy for all combined features.

No.	Badini	Hawrami	Sorani	Classification (support)	Producer Accuracy (Precision)	Recall	F1-score
Badini	402	14	7	423	97%	95%	96%
Hawrami	4	371	8	383	96%	97%	96%
Sorani	7	2	385	394	96%	98%	97%

B. Model Architecture and Feature Integration for all Four Features

The proposed model performed admirably in categorizing Badini, Hawrami, and Sorani languages, aided by acoustic features. Integrating MFCC, Mel Spectrogram, Poly-feature, and Contrast features into the convolutional neural network architecture produced excellent results, demonstrating their synergistic impact.

C. Producer Accuracy and F1-Score

The high producer accuracy numbers attest to the model's precision in classification. In contrast, the high recall rates demonstrate its ability to capture a significant fraction of the relevant language classes in the dataset. The model's effectiveness in finding a balance between precision and recall is confirmed by its harmonious F1-scores.

our model's precision (producer accuracy) is particularly remarkable. The model's ability to accurately categorize instances of the Badini language while minimizing false positives was demonstrated, with producer accuracy reaching an impressive 97%. Furthermore, the model's ability to accurately recognize Hawrami language samples was shown by the Hawrami class's producer accuracy of 96%. Moreover, the overall accuracy for Sorani, the third language in our classification task, was 96%, demonstrating the model's reliability in identifying Sorani language instances.

The F1-scores for all three language classes were higher than 96%, demonstrating how successfully our model handled the trade-off between recall and precision. In multiclass classification problems, striking this balance is crucial to ensuring the model can distinguish between several classes while minimizing false positives.

D. Discussion

The experimental findings reveal that feature selection and integration greatly enhance the given model's performance. With an accuracy of 96.5%, MFCC, Mel Spectrogram, Poly-feature, and Contrast, taken together, proved to be the most successful in precisely distinguishing the Badini, Hawrami, and Sorani languages.

This suggests that every feature set provides unique information to the model, enabling it to capture a broad spectrum of language qualities of interest. The Mel Spectrogram is essential for increasing the model's ability to differentiate across languages, as seen by the considerable accuracy gained from the feature set introduction. These findings show the efficiency of the chosen characteristics and the used models in this work for dialectal speech categorization. Figure 4 shows the chosen feature's stability almost in the 12th epoch.

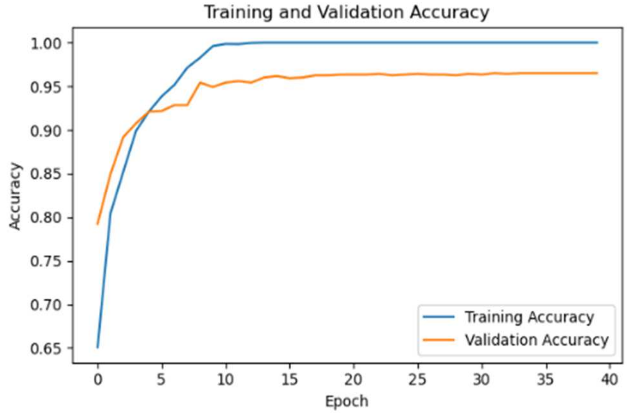


Fig. 4 Stability of (MFCC + Mel Spectrogram + Poly-feature + Contrast) all together.

The great accuracy attained shows the possibility of the suggested method for precisely differentiating and identifying between many Kurdish dialects. The idea behind the better performance of the suggested approach is that the many kinds of features record distinct facets of the voice signal. While Mel spectrograms may record spectrum information not acquired by MFCCs, MFCCs are well-known to be robust against noise and other distortions. Meanwhile, polyn features may capture more complicated links between the original features [30]. [31], contrast features can record the relationship between the present and prior features.

Combining a Mel spectrogram with other sound characteristics in machine learning may often provide more accurate results than utilizing single features such as contrast or Poisson features. This is so because, for sound identification activities, the Mel spectrogram records significant spectrum information about the audio input, which may be rather vital [32].

IV. Conclusion

This article presents a technique for categorizing Kurdish dialects using a one-dimensional Convolutional Neural Network (CNN) and fully linked layers. From the voice recordings, we derived four distinct types of features: MFCCs, Mel spectrograms, contrast features, and polynomial features. Our analysis revealed that the highest level of accuracy was attained by utilizing all four categories of characteristics in conjunction. The suggested technique attained a test set accuracy of 96.5%. The accuracy obtained using various feature extraction approaches, such as only utilizing MFCCs, MFCCs with contrast normalization, and MFCCs with polynomial features, is lower than this. The results indicate that the suggested method is a highly favorable strategy for classifying Kurdish dialects. The proposed approach has several potential applications. It can be utilized to create a system that can recognize the specific dialect of a speaker in real time or transcribe voice recordings in several dialects. The system has the potential to be utilized for the creation of instructional materials for individuals who speak Kurdish or for the development of instruments aimed at safeguarding various Kurdish dialects.

For our future research, we intend to investigate the use of additional characteristics, such as prosodic and phonetic features, to categorize Kurdish dialects. We also intend to create an expanded collection of Kurdish dialect speech recordings to enhance the training and evaluation of more intricate models and investigate the application of transfer learning to categorize Kurdish dialects.

References

[1] N. J. Ibrahim, M. Y. Idna Idris, M. Y. @ Z. Mohd Yusoff, N. N. Abdul Rahman, and M. Izzi Dien, "Robust Feature Extraction Based on Spectral and Prosodic Features for Classical Arabic Accents Recognition," Malaysian Journal of Computer Science, pp. 46–72, Dec. 2019, doi: 10.22452/mjcs.sp2019no3.4.

[2] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition From Saudi Dialect Corpus," IEEE Access, vol. 9, pp. 127081–127085, 2021, doi: 10.1109/access.2021.3110992.

[3] A. Al-Talabani, Z. Abdul, and A. Ameen, "Kurdish Dialects and Neighbor Languages Automatic Recognition," ARO-The Scientific Journal of Koya University, vol. 5, no. 1, pp. 20–23, Apr. 2017, doi:10.14500/aro.10167.

[4] K. J. Ghafoor, K. M. Hama Rawf, A. O. Abdulrahman, and S. H. Taher, "Kurdish Dialect Recognition using 1D CNN," Aro-The Scientific Journal of Koya University, vol. 9, no. 2, pp. 10–14, Oct. 2021, doi:10.14500/aro.10837.

[5] P. A. Abdalla et al., "A vast dataset for Kurdish handwritten digits and isolated characters recognition," Data in Brief, vol. 47, p. 109014, Apr. 2023, doi: 10.1016/j.dib.2023.109014.

[6] S. Badawi, A. M. Saeed, S. A. Ahmed, P. A. Abdalla, and D. A. Hassan, "Kurdish News Dataset Headlines (KNDH) through multiclass classification," Data in Brief, vol. 48, p. 109120, Jun. 2023, doi: 10.1016/j.dib.2023.109120.

[7] G. Işik and H. Artuner, "Derin Öğrenme Mimarilerinde Akustik ve Fonotaktik Öznitelikleri Kullanan Türkçe Ağız Tanıma," Bilişim Teknolojileri Dergisi, vol. 13, no. 3, pp. 207–216, Jul. 2020, doi:10.17671/gazibtd.668023.

[8] L. Deng, & D.Yu, "Deep learning: methods and applications," Foundations and trends® in signal processing, vol. 7, no, (3–4), pp. 197-387, 2014.

[9] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533–1545, Oct. 2014, doi:10.1109/taslp.2014.2339736.

[10] S. Safavi, M. Russell, and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," Computer Speech &amp; Language, vol. 50, pp. 141–156, Jul. 2018, doi:10.1016/j.csl.2018.01.001.

[11] R. Rahmawati and D. P. Lestari, "Java and Sunda dialect recognition from Indonesian speech using GMM and I-Vector," 2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA), Oct. 2017, doi: 10.1109/tssa.2017.8272892.

[12] K. M. Hama Rawf, A. A. Mohammed, A. O. Abdulrahman, P. A. Abdalla, and K. J. Ghafor, "A Comparative Study Using 2D CNN and Transfer Learning to Detect and Classify Arabic-Script-Based Sign

Language," Acta Informatica Malaysia, vol. 7, no. 1, pp. 08–14, 2023, doi: 10.26480/aim.01.2023.08.14.

[13] Hama Rawf, Karwan Mahdi, Ayub Othman Abdulrahman, and Aree Ali Mohammed. 2024. "Improved Recognition of Kurdish Sign Language Using Modified CNN" Computers 13, no. 2: 37. doi:10.3390/computers13020037.

[14] S. Shivaprasad and M. Sadanandam, "RETRACTED ARTICLE: Dialect recognition from Telugu speech utterances using spectral and prosodic features," International Journal of Speech Technology, vol. 27, no. 2, pp. 515–515, Jun. 2021, doi: 10.1007/s10772-021-09854-8.

[15] B. Tawaqal and S. Suyanto, "Recognizing Five Major Dialects in Indonesia Based on MFCC and DRNN," Journal of Physics: Conference Series, vol. 1844, no. 1, p. 012003, Mar. 2021, doi:10.1088/1742-6596/1844/1/012003.

[16] M. Moftah, M. W. Fakhr, and S. El Ramly, "Arabic dialect identification based on motif discovery using GMM-UBM with different motif lengths," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Apr. 2018, doi:10.1109/icnlsp.2018.8374397.

[17] U. Garg, S. Agarwal, S. Gupta, R. Dutt, and D. Singh, "Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Sep. 2020, doi:10.1109/cicn49253.2020.9242635.

[18] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," Applied Acoustics, vol. 158, p. 107050, Jan. 2020, doi: 10.1016/j.apacoust.2019.107050.

[19] Rawf, Karwan M. Hama, Sarkhel H. Taher Karim, Ayub O. Abdulrahman, and Karzan J. Ghafoor. "Dataset for the recognition of Kurdish sound dialects." Data in Brief 53 (2024): 110231.

[20] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," IEEE Access, vol. 7, pp. 7717–7727, 2019, doi: 10.1109/access.2018.2888882.

[21] S. U. Rehman, S. Tu, Y. Huang, and Z. Yang, "Face recognition: A novel un-supervised convolutional neural network method," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), May 2016, doi: 10.1109/icoacs.2016.7563066.

[22] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," Applied Acoustics, vol. 158, p. 107020, Jan. 2020, doi: 10.1016/j.apacoust.2019.107020.

[23] N. Melek Manshouri, "Identifying COVID-19 by using spectral analysis of cough recordings: a distinctive classification study," Cognitive Neurodynamics, vol. 16, no. 1, pp. 239–253, Jul. 2021, doi:10.1007/s11571-021-09695-w.

[24] G. Tzanetakis and P. Cook, "Musimcal genre classification of audio signals," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, Jul. 2002, doi: 10.1109/tsa.2002.800560.

[25] 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, Aug. 1980, doi:10.1109/tassp.1980.1163420.

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," In Proceedings of the 14th python in science conference Vol. 8, pp. 18-25, July 2015.

[27] K. West, and S. Cox, "Finding An Optimal Segmentation for Audio Genre Classification," In ISMIR (pp. 680-685, September 2005.

[28] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," Proceedings. IEEE International Conference on Multimedia and Expo, doi: 10.1109/icme.2002.1035731.

[29] A. Shati, G. M. Hassan, and A. Datta, "COVID-19 Detection System: A Comparative Analysis of System Performance Based on Acoustic Features of Cough Audio Signals," 2023, arXiv preprint arXiv:2309.04505.

[30] F. Alías, J. Socoró, and X. Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," Applied Sciences, vol. 6, no. 5, p. 143, May 2016, doi: 10.3390/app6050143.

[31] M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," IEEE Access, vol. 8, pp. 221640–221653, 2020, doi: 10.1109/access.2020.3043201.

[32] D. T. Pizzo, and S. Esteban, "IATos: AI-powered pre-screening tool for COVID-19 from cough audio samples," 2021, arXiv preprint arXiv:2104.132