

Machine Learning-based Classification and Identification of Arabic Dialects through Acoustic Feature of Guttural Sounds.

Dr Hamed Al-Tairi
School of Information Technology
Whitecliffe College
New Zealand
20232006@mywhitecliffe.com

Dr. Seyed Ebrahim Hosseini
School of Information Technology
Whitecliffe College
New Zealand
seyedh@whitecliffe.ac.nz

Dr Shahbaz Pervez Chattha
School of Information Technology
Whitecliffe College
New Zealand
shahbazp@whitecliffe.ac.nz

Abstract— This study addresses the challenges of accurately identifying Arabic dialects due to their phonetic complexity and diversity. By developing a robust machine learning model for accurate Arabic dialect identification, the research focuses on guttural sounds and analyzes their acoustic features within Yemeni, Palestinian, Saudi, and Egyptian Arabic dialects. Utilizing a dataset of guttural sounds, we applied various machine learning models, including CatBoost, LightGBM, and Random Forest, to classify these dialects. Among the models tested, CatBoost demonstrated strong performance and generalization capabilities, effectively distinguishing between the dialects. This study contributes to the improvement of speech recognition technologies and offers valuable insights for linguistic research and communication technology.

Keywords— *Arabic dialect identification; Machine learning; Guttural sounds; Acoustic features; Speech recognition*

I. INTRODUCTION

Speech recognition in natural language processing (NLP) has revolutionized human-computer interaction, offering practical applications such as voice-activated assistants like Siri and Alexa, communication aids for the hearing impaired, and enhanced smart home device interactions. Despite these significant advancements, accurately identifying and processing Arabic dialects remains a formidable challenge due to their phonetic complexity and diversity [1]. This paper aims to address these challenges by developing a robust machine learning model for accurate Arabic dialect identification. Arabic is a macrolanguage with many dialects that can differ significantly in phonetics, vocabulary, and grammar [2]. This diversity poses a unique challenge for speech recognition systems, which need to accurately identify and process these variations to function effectively. Accurate dialect identification is essential for improving the performance of voice-activated assistants, aiding forensic investigations, enhancing socio-linguistic research, optimizing media content localization, and supporting language learning applications.

There have been several recent studies, including those by [3-6], in the domain of Arabic dialect identification that have leveraged machine learning models to tackle the challenge of distinguishing between Arabic dialects. The studies primarily used audio recordings of speech, with specific features extracted such as Mel-Frequency Cepstral Coefficients (MFCC), Temporal Frequency Cepstral Coefficients (TFCC), segmental, and prosodic features. Researchers such as [3-6], have demonstrated the application of various machine learning techniques, from

Hidden Markov Models (HMM) and Deep Neural Networks (DNN) to Support Vector Machines (SVM) and hybrid models. [3] utilized a combination of HMM and DNN, achieving an accuracy of 79% to 86% in accent recognition, although they faced challenges with recording quality variability and distinguishing closely related accents. [4] employed SVM to analyze segmental and prosodic features, effectively capturing dialectal variations without detailing specific limitations. [2] used features such as MFCC and TFCC with various classifiers, finding K-Nearest Neighbors (KNN) using TFCC features to be the most accurate at 76%. [6] explored a hybrid approach combining KNN and Decision Trees (DT), improving speech recognition accuracy by over 10%, though the study used a relatively small dataset.

Other notable studies include [7], who introduced the ADI17 dataset for fine-grained Arabic dialect identification, highlighting its value despite dialect diversity limitations. [8] combined phonetic and lexical features with a multi-class SVM, achieving 59.2% accuracy in dialect identification but faced challenges with code-switching and recording setup variability. [9] proposed a fusion approach with Convolutional Neural Networks (CNN), i-vectors, and transformer models, achieving 82.44% classification accuracy, though data augmentation was not employed. [10] utilized a combination of Gaussian Naïve Bayes, SVM, Recurrent Neural Networks (RNN), and CNN-RNN classifiers, with the CNN-RNN classifier achieving the highest precision, recall, and F1- score. [11] focused on Arabic dialect identification using CNN, achieving 83% accuracy, but noted limitations due to the small dataset size. [12] explored intonation patterns with hybrid LSTM networks, achieving state-of-the-art results but faced potential issues with generalizing to other datasets. [13] applied motif discovery in time series analysis, achieving 62.75% accuracy, while [14] introduced a transformer-based system, achieving 86.29% accuracy on the ADI17 dataset.

Ashifur and his team researcher [15] emphasized the critical need for extensive, high-quality, and publicly available Arabic speech datasets, identifying nine available corpora and discussing their applications in various challenging and realistic environments. They highlighted the importance of incorporating diverse features such as variations in input devices, dialects, age groups, environments, noise levels, and speech disabilities to create more robust Arabic ASR systems. Despite these advancements, existing research primarily focuses on broad dialectal differences without delving into finer phonetic nuances, particularly at the level of phonemes or sounds. This gap in the literature highlights the need for more

detailed exploration of specific phonetic features that can serve as critical markers for dialect identification. By concentrating on guttural sounds and analyzing their acoustic features, namely the first three formants of the vowels that are adjacent to the guttural sounds within Yemeni, Palestinian, Saudi, and Egyptian Arabic dialects, we can investigate how these acoustic features contribute to the classification and identification of Arabic dialects.

Guttural sounds in Arabic, including emphatic sounds such as /tQ/ and /sQ/ (also known as Tafkheem in Arabic), uvulars such as /X/ and /K/, and pharyngeals such as /è/ and /Q/, as well as laryngeals like /P/, are central to its phonetic system. These sounds exhibit significant variance in articulatory and acoustic properties across different dialects. Based on various investigations into Arabic gutturals in different dialects, previous studies have debated whether these sounds should be classified as one group or multiple groups due to their varied articulatory strategies. For example, researchers [16, 17] consider emphatics, pharyngeals, and uvulars as one group produced with the retraction of the tongue root, while [18, 19] argue that only pharyngeals involve tongue root retraction.

These divergent views suggest that guttural sounds could provide distinctive acoustic features that are crucial for dialect identification. This study focuses on the acoustic features of guttural sounds within Yemeni, Palestinian, Saudi, and Egyptian Arabic dialects. Our primary objective is to develop a machine learning model that classifies these dialects based on guttural sound patterns, representing a novel approach in the literature. By analyzing the first three formants (F1, F2, F3) of these sounds, we aim to improve the accuracy of speech recognition technologies. This research contributes to information technology with potential applications in communication technology, linguistic research, and beyond.

Following [20], we propose a comprehensive framework designed to tackle the complexities of Arabic dialect identification. Our approach includes the extraction of phonetic features and the application of various machine learning models such as CatBoost, LightGBM, and Random Forest. Each model is rigorously evaluated using cross-validation techniques to ensure generalizability and reliability. Feature importance analysis and hyperparameter tuning are conducted to optimize model performance. This paper is organized as follows: Section 1 provides a brief background of the study with the aim. Section 2 details the methodology, including data collection, preprocessing, and model training. Section 3 presents the results and Section 4 discusses the results and their implications.

II. MATERIALS AND METHODS

A. Dataset Description

The data utilized in this study were derived from a previous PhD research by [21], which featured eight adult male participants from four different dialects. The study included two Egyptians (E1 from Upper Egypt and E2 from Cairo), one Palestinian from Jerusalem, two Saudis (S1 from Jeddah and S2 from Hail), and three Yemenis (Y1 from

Dhamar, Y2, and Y3 from Hadhramout). At the time of the study, all participants were students at the University of British Columbia in Vancouver, Canada, aged between 24 and 35 years. Their time residing in Vancouver was less than five years, and none reported any hearing or speech impairments.

The dataset used in this research, derived from [21], is particularly relevant because it includes a diverse range of guttural sounds from different dialects, providing a rich source of phonetic variation that is critical for our analysis. These sounds are central to the phonetic systems of the dialects under study, making the dataset ideal for examining the unique acoustic features that distinguish these dialects. The study encompassed twenty-one test conditions, as detailed in Tables 1 and 2, with a focus on the non-round vowels /a:/, /i:/, and /u:/ across various conditions. Monosyllabic words featuring these long vowels were selected for their pronounced and extended formant trajectories relative to their shorter counterparts. The selected words, introduced to participants, contained either a plain coronal, emphatic (coronal), uvular, or pharyngeal at the onset. Subjects were instructed to articulate each target word in their native Arabic dialect within the carrier phrase [ga:lli. . .] “he said to me. . .”.

A total of 1,691 tokens were collected, with 640 tokens each for /a:/ and /i:/ across eight conditions, and 400 tokens for /u:/ across five conditions. Data collection involved simultaneous acoustic and articulatory (ultrasound) recordings at UBC using a Sennheiser MKH 416 P48 microphone and a desktop computer setup. Each participant was shown test words on a MacBook laptop through PowerPoint slides, with each word repeated ten times across two sessions.

Table 1: Test Conditions

Environment	CV:C
Vowels	(/a:/, /i:/) after (/t ^Q /, /t/, /s ^Q /, /s/, /X/, /K/, /è//, /Q/)
	(/u:/) after (/t ^Q /, /t/, /X/, /è//, /Q/)
Total	21 test conditions

Table 2: Word List

Plain	Emphatics	Uvulars	Pharyngeals	Velar
ta:b	tQa:b	Xa:b	èa:b	ka:s
‘repente d’	‘recovere d’	‘failed’	‘like’	‘glas s’
sa:b	sQa:b	Ka:b	Qa:b	ki:s
‘left’	‘targeted’	‘was absent’	‘didn’t like’	‘bag’
si:b	sQi:b	Xi:b	èi:d	ku:b
‘leave’	‘touch’	pseudowor d	‘deviate’	‘cup’
ti:b	tQi:b	Ki:b	Qi:d	
pseudo word	‘perfume’	‘be absent’	‘holiday’	

III. RESULTS

The training results reveal that the CatBoost model achieved the highest validation accuracy of 79.32%, followed closely by LightGBM at 75.64% and XGBoost at 74.22%. The Random Forest and Decision Tree models also performed well on the training data, with accuracies of 78.19% and 71.67%, respectively. Support Vector Machine (SVM) and Logistic Regression lagged behind, with accuracies of 56.66% and 37.68%. In terms of precision, recall, and F1-scores, CatBoost consistently outperformed other models across most classes, indicating its robustness. For instance, CatBoost achieved precision scores of 0.77, 0.68, 0.84, and 0.81 across different classes, which were higher than those of other models. Similarly, its recall and F1-scores were consistently strong, demonstrating its ability to balance precision and recall effectively.

On the test data, there was a noticeable drop in the performance of all models, indicating potential overfitting. CatBoost still led the pack with a validation accuracy of 66.45%, showing better generalization compared to other models. XGBoost and LightGBM followed with accuracies of 63.16% and 61.84%, respectively. The Random Forest and Decision Tree models also experienced performance drops, with test accuracies of 65.79% and 61.18%. The precision, recall, and F1-scores for CatBoost, while lower than on the training data, remained relatively high, reflecting its strong performance. For instance, CatBoost achieved a recall of 0.72, 0.33, 0.68, and 0.72, and F1-scores of 0.69, 0.39, 0.64, and 0.74.

This consistency indicates that CatBoost managed to mitigate overfitting better than other models, maintaining balanced precision, recall, and F1-scores across different class. In summary, CatBoost emerged as the best-performing model, achieving the highest validation accuracy on both training (79.32%) and test (66.45%) datasets. It consistently maintained high precision, recall, and F1-scores across different classes, demonstrating strong generalization capabilities. XGBoost and LightGBM also performed well but were slightly less effective in generalization compared to CatBoost. The significant drop in performance for models like Random Forest, Decision Tree, SVM, and Logistic Regression highlights the challenges of overfitting and the importance of robust model selection and tuning.

Table 3: Training Results for Various Models

Model	Val Acc	Prec (avg)	Reca ll (avg)	F1-Score (avg)
Random Forest	0.7819	0.77	0.75	0.76
Logistic Regression	0.3768	0.39	0.39	0.37
Support Vector Machine	0.5666	0.55	0.56	0.55
Decision Tree	0.7167	0.70	0.69	0.69
XGBoost	0.7422	0.73	0.71	0.72
LightGBM	0.7564	0.74	0.74	0.74
CatBoost	0.7932	0.78	0.77	0.77

Table 4: Test Results for Various Models

Model	Val Acc	Prec (avg)	Reca ll (avg)	F1-Score (avg)
Random Forest	0.6579	0.63	0.57	0.58
Logistic Regression	0.3684	0.35	0.33	0.33
Support Vector Machine	0.4408	0.46	0.49	0.44
Decision Tree	0.6118	0.56	0.56	0.56
XGBoost	0.6316	0.61	0.58	0.59
LightGBM	0.6184	0.60	0.59	0.59
CatBoost	0.6645	0.62	0.61	0.61

A. Hyperparameter Tuning for Models

The tables provided summarize the performance of CatBoost and Random Forest models on both training and test datasets. For the training data, the CatBoost model was optimized with parameters including a learning rate of 0.1, l2 leaf regularization of 3, 500 iterations, and a depth of 8, achieving a best score of 0.7844 and a validation accuracy of 0.8017. The precision, recall, and F1-score per class were [0.79, 0.67, 0.85, 0.82], [0.84, 0.69, 0.77, 0.84], and [0.82, 0.68, 0.81, 0.83], respectively. The Random Forest model, with parameters such as 100 estimators, a minimum samples split of 5, a minimum samples leaf of 1, 'auto' max features, a max depth of 20, and bootstrap set to True, achieved a best score of 0.7674 and a validation accuracy of 0.7819. Its precision, recall, and F1-score per class were [0.79, 0.67, 0.80, 0.80], [0.77, 0.64, 0.77, 0.85], and [0.78, 0.66, 0.79, 0.82], respectively.

For the test data, the CatBoost model retained its optimized parameters and achieved a best score of 0.6817 and a validation accuracy of 0.6579. The precision, recall, and F1-score per class were [0.66, 0.47, 0.57, 0.77], [0.72, 0.39, 0.62, 0.72], and [0.69, 0.42, 0.59, 0.74], respectively. Similarly, the Random Forest model, with its best parameters, achieved a best score of 0.6789 and a validation accuracy of 0.6579. Its precision, recall, and F1-score per class were [0.66, 0.57, 0.62, 0.68], [0.72, 0.22, 0.53, 0.82], and [0.69, 0.32, 0.57, 0.74], respectively. An additional test was performed on the CatBoost model with selected features, resulting in a validation accuracy of 0.6436. The precision, recall, and F1-score per class were [0.64, 0.47, 0.54, 0.79], [0.64, 0.47, 0.59, 0.75], and [0.64, 0.47, 0.57, 0.77], respectively.

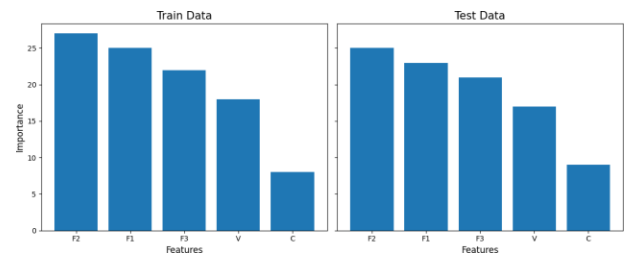


Figure 4: Feature Importance of CatBoost Model

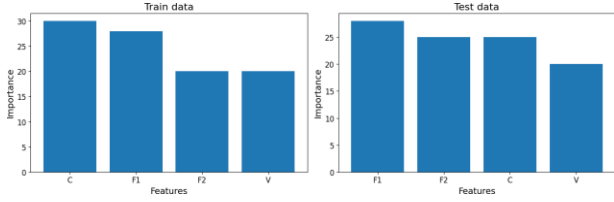


Figure 5: Selected Features in CatBoost Model

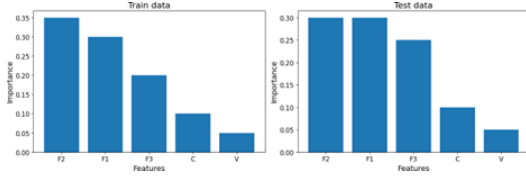


Figure 6: Feature Importance in Random Forest Model

Table 5: Summary of Model Parameters and Score

Data Type	Model	Best Parameters	Best Score	Validation Accuracy
Train Data	CatBoost	{learning_rate: 0.1, l2_leaf_reg: 3, iterations: 500, depth: 8}	0.7844	0.8017
Train Data	Random Forest	{n_estimators: 100, min_samples_split: 5, min_samples_leaf: 1, max_features: auto, max_depth: 20, bootstrap: True}	0.7674	0.7819
Test Data	CatBoost	{learning_rate: 0.1, l2_leaf_reg: 3, iterations: 500, depth: 8}	0.6817	0.6579
Test Data	Random Forest	{n_estimators: 100, min_samples_split: 2, min_samples_leaf: 1, max_features: auto, max_depth: None, bootstrap: True}	0.6789	0.6579

Table 6: Summary of Validation and Test Metrics

Data Type	Model	Precision	Recall	F1-score	Accuracy
2*Train Data	CatBoost	0.79,	0.84,	0.82,	0.7532
		0.67,	0.69,	0.68,	
		0.85,	0.77,	0.81,	
		0.82	0.84	0.83	
	Random Forest	0.79,	0.77,	0.78,	-
		0.67,	0.64,	0.66,	
		0.80,	0.77,	0.79,	
		0.80	0.85	0.82	
2*Test Data	CatBoost	0.66,	0.72,	0.69,	0.6436
		0.47,	0.39,	0.42,	
		0.57,	0.62,	0.59,	
		0.77	0.72	0.74	
	Random Forest	0.66,	0.72,	0.69,	-
		0.57,	0.22,	0.32,	
		0.62,	0.53,	0.57,	
		0.68	0.82	0.74	

IV. DISCUSSION

The results of this study highlight the efficacy of various machine learning models in classifying Arabic dialects based on guttural sound features, with a particular focus on CatBoost, LightGBM, and Random Forest classifiers. The CatBoost model demonstrated superior performance, achieving the highest validation accuracy of 79.32% on the training data and 66.45% on the test data. This indicates CatBoost's robustness in generalizing well to unseen data, making it a suitable choice for complex tasks such as Arabic dialect identification where subtle phonetic variations are crucial.

The success of the CatBoost model can be attributed to its efficient handling of categorical features and its ability to manage overfitting through hyperparameter tuning and cross-validation. The model's precision, recall, and F1-scores were consistently higher than those of other models across different classes, emphasizing its balanced performance. The feature importance analysis revealed that formant frequencies (F1, F2, F3) were the most significant features for dialect classification, corroborating the linguistic hypothesis that guttural sounds play a pivotal role in distinguishing Arabic dialects.

The Random Forest model also performed well, achieving a validation accuracy of 78.19% on the training data and 65.79% on the test data. However, its performance was slightly lower than that of CatBoost, indicating that while Random Forest is effective, it may not capture the intricate patterns in the data as efficiently as CatBoost. Despite these promising results, the study has several limitations. The dataset used was relatively small and homogeneous, with only eight male participants from specific regions. This limited sample size might not fully represent the extensive variability within Arabic dialects, potentially affecting the generalizability of the findings. Additionally, the focus on male speakers could introduce gender bias, limiting the applicability of the models to female speakers.

The implications of these results are significant for both technological and linguistic fields. For speech recognition technology, the findings suggest that incorporating machine

learning models like CatBoost, which effectively handle phonetic features, can substantially improve the accuracy of dialect identification. This can lead to better performance in voice-activated assistants, enhancing user experience and accessibility. In forensic investigations and socio-linguistic research, accurate dialect identification can provide more precise insights into speaker origins and regional language variations. Furthermore, in media content localization and language learning applications, these advancements can facilitate more accurate and culturally relevant translations and educational materials.

Future research should address these limitations by incorporating a larger and more diverse participant pool, including female speakers and individuals from a wider range of Arabic-speaking regions. Furthermore, exploring more advanced deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), might enhance the ability to capture complex acoustic features and improve classification accuracy. In conclusion, this study presents a novel approach to Arabic dialect identification using machine learning models focused on guttural sound features. The findings underscore the importance of feature selection and model tuning in achieving high classification accuracy. While CatBoost emerged as the best-performing model, future work should aim to enhance the dataset's diversity and explore more sophisticated modelling techniques to further improve performance and generalizability. The implications of these findings extend to improving speech recognition technologies, forensic linguistics, socio-linguistic research, media localization, and language learning tools.

V. CONCLUSION

This study has demonstrated the potential of machine learning models, particularly CatBoost, for the classification of Arabic dialects based on the acoustic features of guttural sounds. By focusing on formant frequencies of emphatics, uvulars, and pharyngeals, we have highlighted the significance of these phonetic features in distinguishing between Yemeni, Palestinian, Saudi, and Egyptian Arabic dialects. The CatBoost model achieved the highest accuracy, showcasing its robustness and ability to generalize well to unseen data. This underscores the model's capability to handle the complexity and variability inherent in Arabic dialects, providing a reliable tool for enhancing speech recognition technologies.

VI. LIMITATIONS AND FUTURE WORK

One of the dataset's primary limitations is its relatively small sample size, which might not fully capture the extensive variability within each represented dialect. Additionally, the participant pool is exclusively male, potentially limiting the generalizability of findings across genders. Another challenge arises from the inherent complexity of dialectal Arabic phonetics, where subtle acoustic differences may be difficult to capture and classify accurately.

To mitigate these limitations, future studies are suggested to:

- Include a larger, more diverse participant pool, incorporating female speakers and a wider range of dialects to enhance the dataset's representativeness.

- Employ sophisticated acoustic analysis techniques and machine learning models capable of capturing subtle phonetic variations. This includes the potential use of deep learning approaches that can learn complex patterns from limited data.

REFERENCES

- [1] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, "Arabic speech recognition using end-to-end deep learning," *IET Signal Processing*, vol. 15, no. 8, pp. 521-534, 2021.
- [2] S. Jamal et al., "In the identification of arabic dialects: A loss function ensemble learning based-approach," in *International Conference on Model and Data Engineering*, 2022: Springer, pp. 89-101.
- [3] E. Alsharhan and A. Ramsay, "Robust automatic accent identification based on the acoustic evidence," *International Journal of Speech Technology*, vol. 26, no. 3, pp. 665-680, 2023.
- [4] G. Brown and S. Hellmuth, "Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects," *Speech Communication*, vol. 141, pp. 80-92, 2022.
- [5] K. M. Nahar, O. M. Al-Hazaimeh, A. Abu-Ein, and M. A. Al-Betar, "Arabic dialect identification using different machine learning methods," 2022.
- [6] B. Mouaz, C. Walid, B.-H. Abderrahim, and E. Abdelmajid, "A new framework based on KNN and DT for speech identification through emphatic letters in Moroccan dialect," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1417-1423, 2021.
- [7] S. Shon, A. Ali, Y. Samih, H. Mubarak, and J. Glass, "ADII7: A fine-grained Arabic dialect identification dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: IEEE, pp. 8244-8248.
- [8] A. R. Ali, "Multi-dialect Arabic speech recognition," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020: IEEE, pp. 1-7.
- [9] M. A. Humayun, H. Yassin, and P. E. Abas, "Dialect classification using acoustic and linguistic features in Arabic speech," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, p. 739, 2023.
- [10] M. Abdelazim, W. Hussein, and N. Badr, "Automatic Dialect identification of Spoken Arabic Speech using Deep Neural Networks," *International Journal of Intelligent Computing and Information Sciences*, vol. 22, no. 4, pp. 25-34, 2022.
- [11] M. Alrehaili, T. Alasmari, and A. Aoalshutayri, "Arabic speech dialect classification using deep learning," in *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, 2023: IEEE, pp. 1-5.
- [12] A. A. Alvarez and E. S. A. Issa, "Learning intonation pattern embeddings for arabic dialect identification," *arXiv preprint arXiv:2008.00667*, 2020.
- [13] M. Mofatih, M. Fakhre, and S. El-Ramly, "Spoken Arabic Dialect Identification Using Motif Discovery," *The Egyptian Journal of Language Engineering*, vol. 5, no. 1, pp. 25-36, 2018.
- [14] W. Lin, M. Madhavi, R. K. Das, and H. Li, "Transformer-based Arabic dialect identification," in *2020 International Conference on Asian Language Processing (IALP)*, 2020: IEEE, pp. 192-196.
- [15] A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic Speech Recognition: Advancement and Challenges," *IEEE Access*, 2024.
- [16] S. Davis, "Emphasis spread in Arabic and grounded phonology," *Linguistic inquiry*, pp. 465-498, 1995.
- [17] S. Rose, "Variable laryngeals and vowel lowering," *Phonology*, vol. 13, no. 1, pp. 73-117, 1996.
- [18] M. S. Bin-Muqbil, *Phonetic and phonological aspects of Arabic emphatics and gutturals*. The University of Wisconsin-Madison, 2006.
- [19] S. A. M. Shar, "Arabic emphatics and gutturals: A phonetic and phonological study," 2012.
- [20] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", 2022.
- [21] H. Al-Tairi, "Tongue Retraction in Arabic: An Ultrasound and Acoustic Study," *ResearchSpace@ Auckland*, 2018.