Classifying Pakistan's Diverse Languages Through Speech Using Deep Neural Networks

Muhammad Ibad
School of Science and Engineering
Habib University
Karachi, Pakistan
mi08440@st.habib.edu.pk

Eeshal Khalidnadeem Qureshi
School of Science and Engineering
Habib University
Karachi, Pakistan
eq08433@st.habib.edu.pk

Aina Shakeel
School of Science and Engineering
Habib University
Karachi, Pakistan
as08430@st.habib.edu.pk

Abdul Samad
School of Science and Engineering
Habib University
Karachi, Pakistan
abdul.samad@sse.habib.edu.pk

Sandesh Kumar
School of Science and Engineering
Habib University
Karachi, Pakistan
sandesh.kumar@sse.habib.edu.pk

***Abstract*—Languages are a fundamental aspect of human communication. In Pakistan, the linguistic landscape is rich and diverse, with approximately 70–80 languages spoken across the country. Despite the prominence of regional languages, research in language identification has largely been focused on major global languages. This study addresses this gap and explores a deep learning-based system for classifying regional languages through speech. We aim to detect and identify spoken regional languages from audio clips by employing Mel spectrograms as input features and classifying the rich regional languages of Pakistan. Our baseline CNN model achieved an accuracy of 85%, demonstrating its effectiveness in solving this classification problem. By enhancing language detection for Pakistani languages, our project aims to improve access to services and information in native languages, advancing communication and user experience in a multilingual context.**

***Index Terms*—language identification, audio classification, deep learning, multilingual speech, Pakistani languages, signal processing**

I. Introduction

Speech is one of the most natural forms of communication among human beings, serving as the foundation for articulating thoughts, ideas, and stories amongst each other. The act of speaking and using language as a tool allows us to connect, inform, and influence one another. From telling stories and everyday conversations to discourses and disagreements, speech has always been central to human interaction. The diverse number of spoken languages across the entire world reflects the diversity of human culture and its history.

This linguistic diversity is incredibly prominent in Pakistan. The country is home to around 70-80 languages [1] spoken across different areas, each with its own unique phonetic, syntactic, and acoustic characteristics. As per the 2023 census [2] in Pakistan, Table I shows the distribution of languages spoken across Pakistan.

Each language in Pakistan carries with it a rich culture and plays a crucial role in the lives of the millions who speak that language. Urdu, the official national language, is widely used as a lingua franca of Pakistan and is a symbol of unity in a country with such vast linguistic variation. Meanwhile, regional languages like Punjabi, Sindhi, Seraiki, and Pashto are spoken by millions of people in their respective provinces. Together, these four languages account for nearly 90 percent of the population who speak at least one of these languages.

TABLE I
Regional Languages in the 2023 Census of Pakistan

Rank	Language	2023 Census
1	Punjabi	36.98%
2	Pashto	18.15%
3	Sindhi	14.31%
4	Saraiki	12.00%
5	Urdu	9.25%
6	Balochi	3.38%
7	Hindko	2.32%
8	Brahui	1.16%
9	Mewati	0.46%
10	Kohistani	0.43%
11	Kashmiri	0.11%
12	Shina	0.05%
13	Balti	0.02%
14	Kalasha	0.003%
15	Others	1.38%

In a linguistically diverse nation such as Pakistan, developing technologies to accurately recognize and classify regional languages, especially in the context of spoken communication, has huge implications. It can lead to a greater understanding among the locals and improve access to services and information for people in their native languages. Another critical factor in the context of Pakistan is that, despite Urdu being the national language and serving as the lingua franca across the country, only around 9 percent of the population speaks it as their first language. The majority of Pakistanis speak regional languages which also highlights the necessity of

building a regional language classification system for better communication and access to services for speakers of these regional languages.

There has been extensive research in identifying languages through the medium of speech both with and without deep learning methods but mostly for international languages only. Despite the importance of these regional languages in Pakistan, research in the classification of these languages is relatively scarce. This project aims to bridge this gap and contribute to the field of multilingual speech recognition which can have novel applications in Pakistan. Our approach to solving this task involves converting audio snippets into Mel spectrograms, which can store the complexities of particular audio utterances in a visual manner.

II. Research Question

The research question we are addressing is: "Classifying Pakistan's Diverse Languages Through Speech Using Deep Neural Networks" Our problem statement can be formulated as follows:

Given an audio clip containing speech, detect and identify the regional language being spoken

We aim to explore and train deep learning models by using existing models and designing custom neural networks to address this language detection problem. The input to the model will be an audio clip, and the output will be the detected regional language. The regional languages we will be focusing towards are:

- Punjabi
- Sindhi
- Pashto
- Saraiki
- Urdu

The motivation behind this research stems from the under representation of regional languages in automated systems, especially in Pakistan. While widely spoken languages like Urdu and English have seen more progress in detection technologies, smaller regional languages are often ignored. This creates barriers to access to technology, especially for services like voice assistants, customer service, or language-based data analysis.

In addition to bridging this gap, language identification has practical applications across various industries:

- **Routing Customer Queries**: In a multilingual country like Pakistan, automatic language detection can help route customer queries to support representatives who speak the same language, making customer service more efficient.
- **Improved User Experience**: For businesses like banks, telecom companies, or government services, identifying the speaker's language allows systems to respond automatically in that language, providing a more personalized and comfortable experience.
- **Voice Assistants**: Systems like Google Assistant, Siri, or Alexa could benefit from a language identification module for Pakistani languages, enabling the assistant to switch between languages dynamically based on the user's speech.

Our deep learning model aims to address this gap by accurately detecting which regional language is spoken in a given audio sample. Unlike traditional language detection systems that rely on manual feature extraction, we aim to employ modern deep learning models like Convolutional Neural Networks (CNNs) to capture the unique phonetic patterns of each language.

Existing research on language detection largely focuses on global or national languages. By applying deep learning to a broader set of regional languages, our approach seeks to improve detection accuracy and extend language support to underrepresented languages. Hence, the selection of audio features and model architectures suited to regional language identification is central to our research.

III. Literature Review

The rapid advancement of Artificial Intelligence (AI) and Deep Learning (DL) techniques has opened many different possibilities for classifying a particular language through audio speech. By synthesizing findings from a range of studies, this review aims to provide a comprehensive understanding of the current work done in language classification. The papers that we have reviewed can be classified into two categories. A category of using Log-mel Spectrograms as a feature extracted from audio samples and other of extracting MFCC as a feature from the samples. A brief review of key papers from both categories is presented in this section.

A study [3] proposes a novel deep learning ensemble architecture for spoken language identification from speech signals. This architecture combines the classification principles of three distinct models: Deep Dumb Multi-Layer Perceptron (DDMLP), Deep Convolutional Neural Network (DCNN), and Semi-supervised Generative Adversarial Network (SS-GAN). The final output combines the three models' output by applying ensemble learning using the Choquet integral. The model is evaluated using four benchmark datasets which include many Indic languages such as Marathi, Sanskrit, Hindi, Kannada, Malayalam, Tamil, and Telugu. The study uses eight sets of features extracted from the audio signals to obtain useful information for the rest of the working pipeline. They are as follows: MFCCs, Spectral bandwidth, Spectral contrast, Spectral roll-off, Spectral flatness, Spectral Centroid, Polynomial Centroid, Polynomial features, and Tennotz. These features are fed into the aforementioned models across the four datasets. The study achieves impressive results across the four datasets with the IIIT Hyderabad dataset's overall accuracy reaching 95%, the highest achieved among all the datasets. IIT Madras dataset also achieves an accuracy of 81.51%, showing competitive performance of multiple languages derived from the same roots. The researchers acknowledge that while impressive results were obtained, future work can be done to use

lesser computational-intensive architectures to achieve similar accuracy.

A similar research [4] focused on language identification with four distinct machine learning models: Artificial Neural Networks (ANN), Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). They aimed to determine which model could best identify the languages, using a dataset collected from multiple speakers. The authors compiled a dataset of 7,200 voice samples from 180 speakers representing five regional accents from Gilgit-Baltistan, Pakistan: Shina, Balti, Burushiski, Wakhi, and Khuwar. The dataset included recordings in both Urdu and English, with the specific aim of enhancing accent recognition for applications in the banking sector of Pakistan. The dataset was carefully balanced to ensure enough representation of each accent and speaker. To handle the audio data, they used Mel Frequency Cepstral Coefficients (MFCCs) as the main feature extraction method. MFCCs are commonly regarded as one of the most successful approaches for capturing the distinctive properties of speech signals, particularly in tasks associated with language recognition. These features transform the audio into a format that is suitable for machine learning models to analyze and distinguish between languages. Among the four models studied, the ANN model performed the best, with an accuracy of 88.53% on the English dataset and 86.58% on the Urdu dataset. The ANN model also had the lowest Root Mean Squared Error (RMSE) among the models, demonstrating greater performance in this classification evaluation. This study contributes significantly to the field of language recognition in voice-based systems, particularly for applications in the banking industry. By focusing on accents from a distinct region like Gilgit-Baltistan, the work provides valuable insights for developing language classification systems that are capable of handling accent variation.

Another research [5] evaluates the performance of a language identification (LID) system that uses hybrid feature extraction techniques combined with a feed-forward back propagation neural network (FFBPNN) classifier on two different learning algorithms "trainlm" and "trainscg". The primary focus is on the effectiveness of hybrid feature extraction techniques like MFCC, PLP, and RASTA-PLP in enhancing identification rates. The LID system's performance was evaluated using a user-defined database consisting of four languages: Tamil, Malayalam, Hindi, and English. The paper's key findings include a combination of various feature extraction techniques that significantly improve identification accuracy with a peak accuracy of 94.6% achieved using the MFCC + RASTA-PLP feature extraction approach. In summary, The findings suggest that the integration of hybrid features with a sophisticated neural network such as FFBPNN provides a powerful approach to accurately identify languages from audio.

As we expanded our review to the second category which is using Log-mel Spectrograms as the features, we came across a study [6] that developed a novel deep learning model using a convolutional neural network (CNN). This model works by first converting audio files into spectrogram images. Spectrograms capture the frequency content of sound over time, creating a visual representation that can be processed by a CNN through detecting patterns. The dataset provided to the model consisted of audio files, each containing an utterance spanning approximately 10 seconds. They used 4 different datasets, the largest one including only three languages: English, German, and Spanish, containing a total of 73620 audio samples. For the hyper-parameters used for the model, the study states that the trial and error method was used to determine the hyper-parameters. The first convolutional layer (Conv2D) in the first block uses 32 filters, each with a kernel size of 7x7, and a stride of 1. Furthermore, they used ReLU as the activation function, and Categorical cross-entropy was used as the model's loss function. Lastly, the output activation function was set to softmax to normalize the outputs. The results from this study were an astounding 98% during preliminary testing while further rigorous testing resulted in an overall accuracy of 98.9%. The use of Log-Mel Spectrograms and CNNs seems a good fit to test our initial data of Pakistani regional languages as referenced in the paper, the CNN model consistently outperforms other techniques, achieving a high F1-score. Lastly, the study provides some future direction on expanding the model to handle different languages and feeding large amounts of data by augmenting the current dataset, both of which will be explored further in our research.

In addition, we came across a study [7] that also used a Convolutional Neural Network (CNN) to identify spoken language. Their model was designed to accommodate three languages: German, English, and Spanish. The core of their approach involved using filter banks; a representation of the audio signals; as the main input to the CNN model. These filter banks effectively captured the frequency characteristics of the audio, which were necessary for distinguishing between the different languages. The dataset used in this work included 36,810 audio samples for training, with a range of data augmentation techniques such as pitch shifting, speed perturbation, and noise injection used to improve model robustness. The model performed wonderfully, reaching an accuracy of 91.30% on the validation dataset. Furthermore, precise metrics such as precision, recall, and F1 scores were produced for each language class to provide a complete picture of the model's classification ability. This work demonstrates the power of CNNs in handling multi-class classification tasks for language identification. By effectively processing spectrograms and leveraging data augmentation, the study provides a robust baseline for future research in language recognition using deep learning techniques. Its findings directly inform and support ongoing research in language identification, particularly in the use of CNNs for phoneme detection across multiple languages.

Another novel study [8] we came across presents an approach to language identification known as Language Identification for Audio Spectrograms (LIFAS) which used CNNs to classify languages based on spectrograms derived from raw audio signals. VoxForge, an open-source corpus is utilized

TABLE II
SUMMARY OF LITERATURE REVIEW

Reference	Year	Model Architecture	Feature Extraction	Dataset	Data Augmentation	Accuracy
[3]	2021	CNNs	MFCC	7,000 samples	N/A	95% (Overall Accuracy)
[4]	2021	ANNs	MFCC	7,200 samples	Pitch shifting, speed variation, noise injection	0.886 (English), 0.865 (Urdu) (F1 Score)
[4]	2021	SVMs	MFCC	7,200 samples	Pitch shifting, speed variation, noise injection	0.835 (English), 0.817 (Urdu) (F1 Score)
[4]	2021	KNNs	MFCC	7,200 samples	Pitch shifting, speed variation, noise injection	0.862 (English), 0.821 (Urdu) (F1 Score)
[4]	2021	RF	MFCC	7,200 samples	Pitch shifting, speed variation, noise injection	0.861 (English), 0.813 (Urdu) (F1 Score)
[5]	2020	FFBPNN	MFCC combined with Rasta-PLP	200 samples	N/A	94.6% (Trainim algo) (Overall Accuracy)
[5]	2020	FFBPNN	MFCC combined with Rasta-PLP	200 samples	N/A	70.6% (Trainscg algo) (Overall Accuracy)
[6]	2021	CNNs	Log-mel Spectrograms	73,620 samples	Pitch shifting, speed variation, noise injection	98.9% (Overall Accuracy)
[7]	2019	CNNs	Log-mel Spectrograms	36,810 samples	Pitch shifting, speed variation, noise injection	0.94 (German), 0.92 (English), 0.91 (Spanish) (F1 Score)
[8]	2019	CNNs (Pre-trained Resnet50)	Log-mel Spectrograms	7,000 samples	Pitch shifting, speed variation, noise injection	89% (Overall Accuracy)

as the primary data source. The technique employs CNNs to process spectrograms. The paper emphasizes that this technique requires minimal pre-processing as the spectrograms are created during the training process. The model is designed to classify short audio segments, approximately 4 seconds in length, which is essential for applications that require immediate language detection. For multi-class classification, the model was trained in six languages: English, Spanish, German, French, Russian, and Italian. Each language was represented by 5,000 clips of 60,000 samples each, with a validation set containing 2,000 clips per language. The overall accuracy for the multi-class classification was reported at 89% which shows the effectiveness of the LIFAS technique in language identification. The paper acknowledges limitations in their study, including the reliance on a specific dataset and the potential for misclassification between certain languages.

In addition to the important research relevant to our research question, other papers and surveys were reviewed to gain a deeper understanding of the challenges and methodologies involved in the classification of different languages. A summary of these studies can be found in Table II.

Based on our literature review, it is observed that deep learning models such as CNNs, RNNs, and LSTMs were the most commonly used for this identification task. Previous research on language identification has generally focused on popular international languages. In contrast, we aim to work with Pakistani regional languages. Additionally, feature selection played a critical role in the reviewed studies, in which we have chosen Log-mel Spectrograms as our features for their high accuracy results.

IV. MATERIAL AND METHODOLOGY

In this section, we discuss the major steps of our proposed framework for classifying Pakistan's diverse regional languages through speech, focusing on the dataset, deep learning models, fine-tuning, and experimentation. The details of the dataset are presented in Section IV-A, followed by an in-depth discussion of our models and experimentation in Section IV-B.

A. Data Set

In this section, we discuss the format and pre-processing of our dataset used for training and testing our language identification model. This includes details about the audio samples, different languages involved, and the process of data acquisition.

1) Summary: We have gathered a significant amount of audio clips, each five seconds in length, totaling approximately 5000 samples of total data for all languages. These audio clips will be converted into spectrograms and used for training and testing the model. The audio data was collected from a variety of sources, including Mozilla's Common Voice dataset [9] and

open-source platforms like YouTube. The Mozilla Common Voice dataset is an open-source, multilingual dataset designed to help train machine learning models for speech recognition. It contains voice recordings contributed by volunteers from around the world, encompassing a wide variety of languages and accents. The dataset is publicly accessible and can be accessed from this link.

TABLE III
DEMOGRAPHICS OF THE DATASET

Language	**Age**			**Gender**		
	20-39	40-59	No Info	M	F	No Info
Urdu	92%	1%	6%	53%	25%	22%
Punjabi	62%	9%	29%	67%	1%	32%
Saraiki	26%	38%	36%	64%	-%	36%
Pushto	58%	13%	29%	2%	-%	98%
Sindhi	Mostly			Mostly		

2) Data Acquisition: The primary source of data for the majority of our languages were sourced from Mozilla's Common Voice dataset, which was relatively straightforward to obtain. However, collecting enough data for the Sindhi language posed a challenge, as Mozilla's dataset lacked sufficient coverage for this language. To get balanced data across all languages, we manually gathered audio from Sindhi YouTube channels, downloading videos and segmenting the audio into five-second clips. This manual and programmatic effort allowed us to complete the Sindhi dataset. The data at this stage seems sufficient, but if more data is needed after model testing, we will explore data augmentation techniques, like pitch shifting, speed variation and noise injection, similar to those found in the literature. We may also acquire additional samples from YouTube using the method explained above.

3) Metadata: The dataset includes audio recordings from a variety of different speakers and recording environments. We tried to ensure that the recordings used were primarily from native speakers to capture authenticity and linguistic diversity. Table III summarizes the demographic information for each language represented in the dataset.

The selection of these five languages can be justified by data from the 2023 census, which shows that nearly 90% of Pakistan's population speaks at least one of these languages. We believe this provides a solid foundation for our research. Adding more regional languages, such as Balochi, which has fewer available datasets, would have resulted in an unbalanced dataset.

B. Models

Deep learning models are inspired by the structure and functioning of the human brain, making them sophisticated and powerful tools for a wide range of applications across diverse fields. We found from our literature review that deep learning models are quite popular in audio classification as discussed in Section III. Therefore, we experimented with three deep learning models on the extracted features. This section presents the details of each experiment. The following figure 1 presents an overview of our pre-processing and experimentation with the selected deep-learning models.

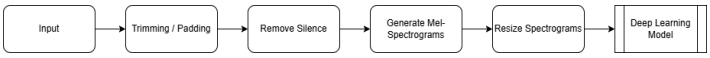

Fig. 1. Pre-processing Overview for Deep Learning Models

Pre-processing: The pre-processing phase is designed to prepare raw audio samples for deep learning models by converting it into features that can be effectively analyzed by neural networks. Robust pre-processing is required to ensure uniformity between languages and eliminate irrelevant variations such as silence or inconsistent audio lengths. The problem is simplified into a categorical classification task by encoding the dataset labels. Each audio sample is converted into a spectrogram to extract distinctive features that could aid in the classification of diverse languages. Below, we detail the major pre-processing steps:

a) Resampling & Normalization: The input audio data is first resampled to a standard sampling rate of 16 kHz for consistency across all samples. Additionally, the audio signals is normalized to ensure uniform amplitude ranges, which can help mitigate biases during training.

b) Trimming & Padding: To standardize audio lengths, each audio file is trimmed to remove silence at the beginning and end using energy thresholds. If the audio is shorter than the desired length, padding is added to ensure consistency in input size across all samples.

c) Silence Removal: Further silence removal is applied to eliminate irrelevant pauses within audio samples. This step helps reduce noise and focus the model on meaningful speech content.

d) Mel-Spectrogram Generation: The processed audio is then converted into Mel-Spectrograms using a standard configuration of 128 Mel bands and a window size of 25ms with a 10ms hop. The Log Mel-Spectrogram is used as it emphasizes lower-frequency information, which is crucial for speech features.

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

Lastly, the generated spectrograms are resized to a fixed dimension of 224 x 224 pixels to match the input size requirements of the deep learning models. This resizing also enabled compatibility with standard pre-trained models when required. Figure 2 illustrates the final spectrogram generated from a Punjabi audio sample. The dataset is divided into 80% training and 20% testing data. The preprocessed spectrograms are then used as inputs to deep-learning models for classification.

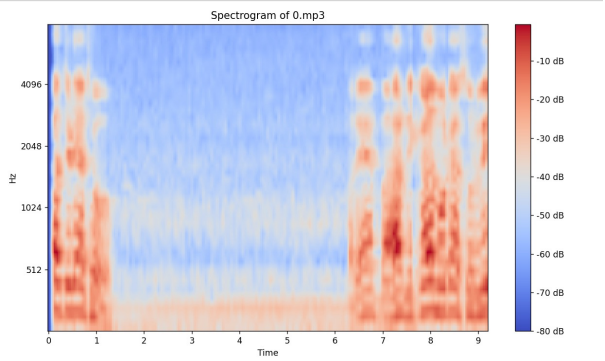

Fig. 2. Spectrogram generated from a Punjabi audio sample

1) Convolutional Neural Network: The baseline CNN model [10] was designed to classify the spectrogram of the audio data into five language categories. The network includes five convolutional layers, each followed by batch normalization and max-pooling layers for feature extraction and dimensionality reduction. After the convolutional blocks, the feature map was flattened into a one-dimensional vector, which was fed into two dense layers. The first dense layer has 256 neurons with *ReLU* activation, and a dropout rate of 0.2 was applied to prevent overfitting. The final dense layer consists of 5 neurons with a *softmax* activation function, outputting class probabilities corresponding to the five languages.

2) ResNet50: We utilized ResNet50, a deep CNN known for its residual learning framework, for the classification of spectrograms. The model leverages residual blocks to address the vanishing gradient problem in deep networks, ensuring stable and efficient training. The base ResNet50 model was initialized without pre-trained weights to allow for training from scratch on the given dataset. Following the base model, a *GlobalAveragePooling2D* layer was added to reduce the feature map dimensions. This was followed by fully connected and dropout layers to classify the input into one of the five language categories.

3) AlexNet: The AlexNet-inspired architecture was implemented to classify the spectrogram representation. The model begins with five convolutional layers with filter dimensions ranging from 32 to 96. Following the convolutional and max-pooling layers, the feature map is flattened and passed through two dense layers with 2048 and 1024 neurons respectively. The final dense layer contains 5 neurons with *softmax* activation to output class probabilities.

All the models were trained using the Adam optimizer with a learning rate of 0.001. The loss function used was sparse categorical cross-entropy, appropriate for multi-class classification. The models were trained for a maximum of 20 epochs, with early stopping applied to monitor the validation loss and prevent overfitting. Additionally, a checkpoint callback was used to save the best-performing model during training.

V. Results

This section analyzes the outcomes of the conducted experiments and presents a comparative evaluation of the different model architectures.

A. CNN

The baseline CNN model achieved an unexpectedly high accuracy of 85% after fine-tuning it by varying the learning rate and its schedule. This performance could be attributed to the model's hyper-parameters, which were likely pre-optimized for classifying spectrograms.

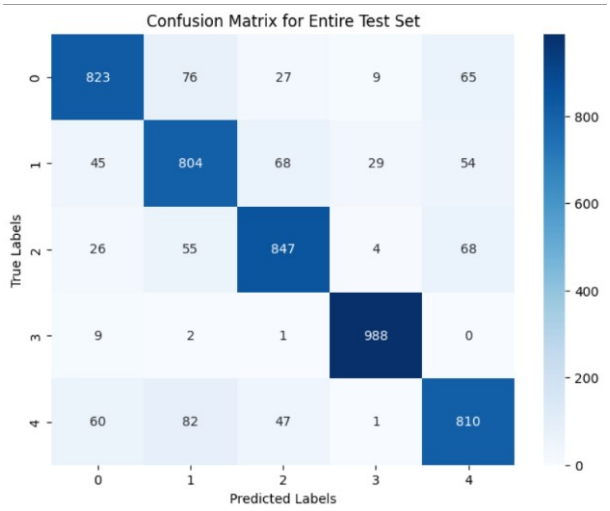

Fig. 3. Confusion Matrix for baseline CNN Model

The confusion matrix illustrates the distribution of true and predicted labels across different classes, highlighting the model's accuracy and potential misclassification. The diagonal elements represent correctly classified instances, while off-diagonal elements indicate misclassification.

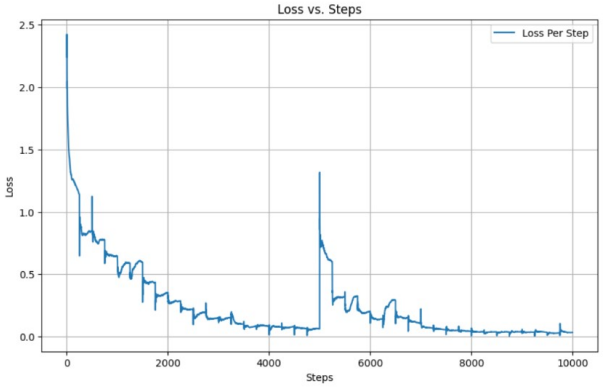

Fig. 4. Loss vs Steps for baseline CNN Model

This plot shows the decrease in loss over training steps, indicating how well the model learns from the data. The plot indicates that the model's training was divided into batches due to resource limitations. This is evident from the spikes

in the loss curve, which likely correspond to the start of new training epochs or batches.

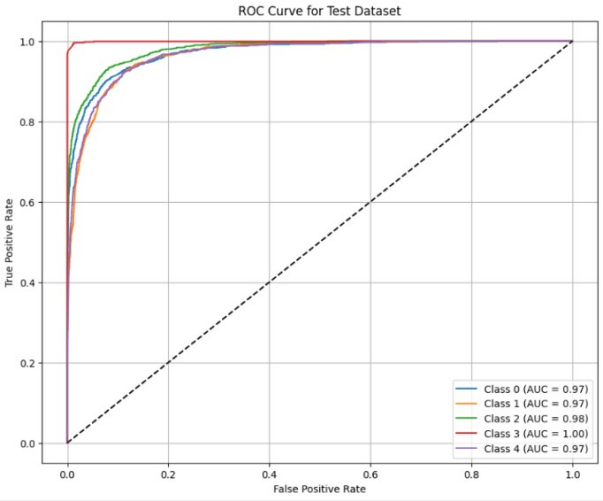

Fig. 5. ROC of baseline CNN Model

Class 3 (Sindhi) has the highest AUC, indicating the best performance. Class 0 (Punjabi) and Class 4 (Urdu) have slightly lower AUC values but still demonstrate strong performance.

TABLE IV
CLASSIFICATION REPORT FOR THE CNN MODEL

Languages	Precision	Recall	F1-Score
Punjabi	0.85	0.82	0.84
Pushto	0.79	0.80	0.80
Saraiki	0.86	0.85	0.85
Sindhi	0.96	0.99	0.97
Urdu	0.81	0.81	0.81

B. *ResNet50*

The ResNet50 model demonstrated a high accuracy of 81%.

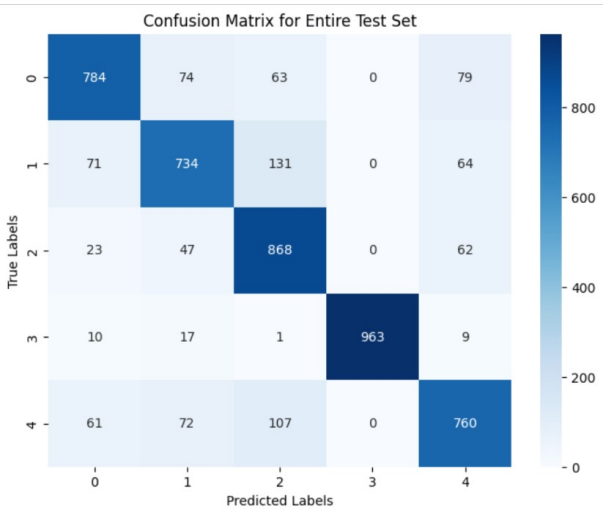

Fig. 6. Confusion Matrix for ResNet50

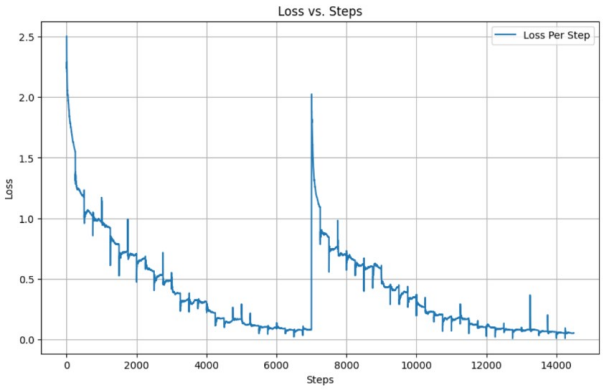

Fig. 7. Loss vs Steps for ResNet50

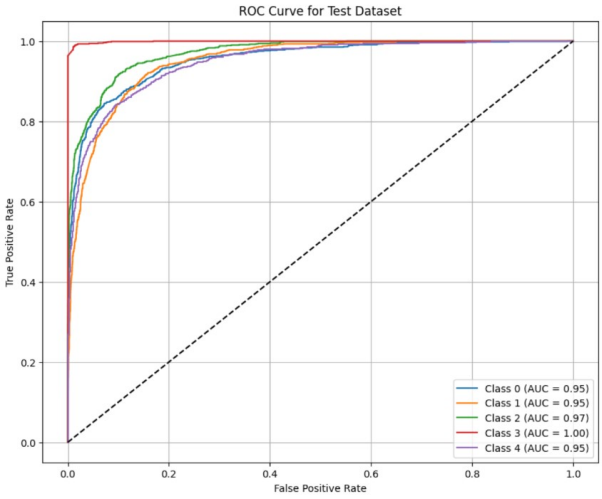

Fig. 8. ROC of ResNet50

The ROC curve indicates that the model performs well for all classes, especially Class 3 (Sindhi) having the highest AUC value similar to the results of other models.

TABLE V
CLASSIFICATION REPORT FOR RESNET50

Languages	Precision	Recall	F1-Score
Punjabi	0.83	0.78	0.80
Pushto	0.78	0.73	0.76
Saraiki	0.74	0.87	0.80
Sindhi	1.00	0.96	0.98
Urdu	0.78	0.76	0.77

C. *AlexNet*

TheAlexNet model demonstrated an accuracy of 81%.

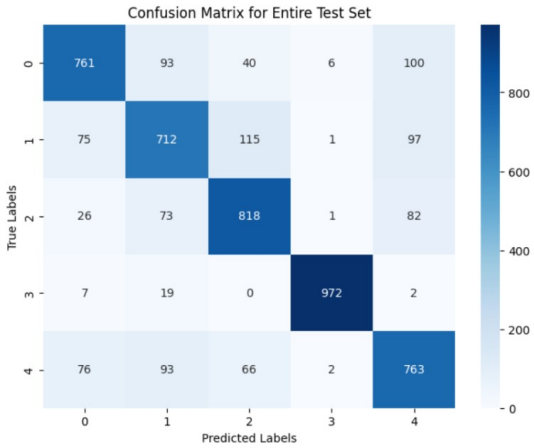

Fig. 9. Confusion Matrix for AlexNet

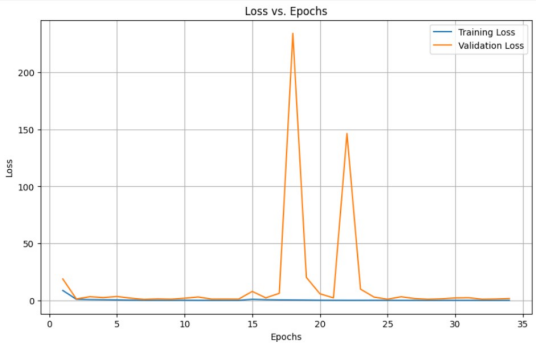

Fig. 10. Loss vs Epochs for AlexNet

The training loss generally decreases over epochs, indicating that the model is learning from the training data. However, there are significant fluctuations, especially towards the end, which might suggest overfitting.

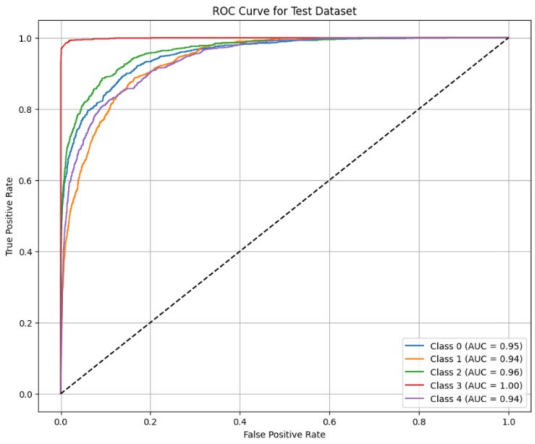

Fig. 11. ROC of AlexNet

TABLE VI
CLASSIFICATION REPORT FOR ALEXNET

Languages	Precision	Recall	F1-Score
Punjabi	0.81	0.76	0.78
Pushto	0.72	0.71	0.72
Saraiki	0.79	0.82	0.80
Sindhi	0.99	0.97	0.98
Urdu	0.73	0.76	0.75

D. Summary of Experiments

TABLE VII
ALL EXPERIMENTS SUMMARY

Model	BS:32	BS:32, LRS	BS:16	BS:16, LRS
CNNs	84%	85%	80%	82%
ResNet50	64.5%	82%	59.6%	74%
AlexNet	64%	81%	72%	71%

BS: Batch Size, LRS: Learning Rate Scheduler.

TABLE VIII
FINAL RESULTS SUMMARY

Model	**Accuracy (%)**
CNN	85%
ResNet50	82%
AlexNet	81%

VI. DISCUSSION

Our study demonstrates that the task of classifying Pakistan's diverse languages through speech can be effectively addressed using relatively simple deep-learning models. The baseline CNN model performed surprisingly well, achieving a high accuracy of 85%. This result suggests that the problem of regional language classification, as formulated in our study, does not necessarily require the complexity of deeper architectures like ResNet50 or AlexNet. However, it also highlights the effectiveness of using Mel spectrograms as input features for such tasks.

A. Model Performance Analysis

The ResNet50 and AlexNet architectures, while still achieving commendable results, exhibited signs of overfitting. This overfitting could be attributed to the comparatively smaller dataset size and limited data augmentation techniques. Deeper models such as ResNet50, designed to handle more complex feature hierarchies, may have been unnecessarily complicated for this specific problem, which primarily involves the phonetic and acoustic distinctions between regional languages. The superior generalization of the CNN model shows the alignment of model complexity with the problem's requirements.

B. Dataset Influence

Interestingly, Sindhi emerged as the best-classified language among the five languages considered. A possible explanation for this could be the difference in data sources. While most

languages utilized the Mozilla Common Voice dataset, the Sindhi dataset was manually curated from YouTube channels. This diverse data source likely introduced more natural variations in speech, enabling the model to learn robust features for Sindhi. On the other hand, the uniformity of Mozilla's dataset may have limited the variability and richness needed for accurate classification of other languages. This observation suggests potential quality issues or biases within the Mozilla dataset for certain languages.

C. Comparisons with Prior Research

Compared to similar studies in the domain of language identification, our results align with findings that simpler models can achieve high accuracy when trained on well-processed spectrogram features. However, most prior research has focused on international languages, where deeper architectures like ResNet50 and ensemble methods have proven more effective due to the complexity of distinguishing between languages with shared linguistic roots. Our study, focusing on linguistically distinct regional languages, suggests that a simpler baseline model can suffice without the computational overhead of more complex architectures.

D. Implications

These results are promising for practical applications, as they indicate the feasibility of deploying lightweight models for language classification in resource-constrained environments. With additional enhancements, such as data augmentation or a broader dataset, this framework can be scaled to include more languages and dialects, providing inclusive and accessible language identification solutions in multilingual contexts.

VII. Future Work

Our study provides a foundational framework for language identification of Pakistan's diverse regional languages using deep learning. While the ResNet50 model has demonstrated promising accuracy, there are several avenues for future research and enhancements:

1) Exploration of Additional Models: While CNNs have shown superior performance in our experiments, we plan to extend our analysis by training and testing other deep-learning architectures. Comparing the results across multiple architectures will provide deeper insights into the optimal model for this task.

2) Dataset Expansion: The dataset used in this study, though representative, can be further expanded to include more samples for each language. We also plan to include other regional languages, such as Balochi and Kashmiri, once sufficient audio data is available. Data augmentation techniques, such as pitch shifting, speed variation, and noise injection, can also be explored to synthetically increase the dataset size.

3) Expanding Language Classes for Better Accuracy: Future work could involve increasing the number of language classes to simultaneously learn multiple languages or dialects. This approach could enhance the generalizability of the framework to unseen data and facilitate the identification of linguistic overlaps between related languages, ultimately improving classification accuracy.

4) Incorporating Advanced Feature Extraction: While we use spectrograms as input features, experimenting with other audio feature representations such as mel-frequency cepstral coefficients (MFCCs), or wavelet transforms could further improve classification accuracy.

5) Real-Time Language Detection: Another direction for future work is to adapt the model for real-time applications, such as automatic transcription systems or multilingual voice assistants. This would require optimization for inference speed and robustness in noisy environments.

6) Broader Application Scenarios: The developed framework could be extended to applications such as speaker identification, dialect classification, or sentiment analysis for regional languages. It could also serve as a stepping stone for creating accessible tools for speech-to-text and translation services in native languages.

7) Evaluation Across Diverse Conditions: Future work could also involve evaluating the models under diverse conditions, such as varying recording quality, background noise levels, and speaker accents, to assess real-world applicability and robustness.

By pursuing these directions, we aim to further refine our framework, contribute to the growing field of multilingual language processing, and support the preservation and computational accessibility of regional languages in Pakistan.

VIII. Conclusion

This study demonstrates the feasibility of using deep learning models to classify Pakistan's regional languages through speech. By leveraging Mel spectrograms as input features, we achieved a baseline accuracy of 85% using a CNN model, highlighting its effectiveness for this task. Deeper architectures like ResNet and AlexNet showed signs of overfitting, indicating that simpler models are more suitable given the dataset size and complexity. Sindhi's superior performance suggests dataset quality plays a critical role. Our findings underscore the potential for lightweight, efficient systems to bridge the gap in linguistic representation and improve access to services in multilingual contexts.

References

[1] "POPULATION BY MOTHER TONGUE — Pakistan Bureau of Statistics," www.pbs.gov.pk. https://www.pbs.gov.pk/node/97

[2] resourceExport, "Language data for Pakistan," CLEAR Global, Feb. 10, 2020. https://clearglobal.org/resources/language-data-for-pakistan/

[3] "FuzzyGCP: A deep learning architecture for automatic spoken language identification from speech signals — Request PDF." Accessed: Oct. 18, 2024. [Online]. Available: https://www.researchgate.net/publication/347518701_FuzzyGCP_A_deep_learning_architecture_for_automatic_spoken_language_identification_from_speech_signals

[4] "Development of a regional voice dataset and speaker classification based on machine learning — Journal of Big Data — Full Text." Accessed: Oct. 18, 2024. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00435-9

[5] "A Language Identification System using Hybrid Features and Back-Propagation Neural Network - ScienceDirect." Accessed: Oct. 18, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0003682X19313672

[6] “Spoken Language Identification Using Deep Learning - Singh - 2021 - Computational Intelligence and Neuroscience - Wiley Online Library.” Accessed: Oct. 18, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/5123671

[7] “Spoken Language Recognition Using CNN — IEEE Conference Publication — IEEE Xplore.” Accessed: Oct. 18, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9031923

[8] S. Revay and M. Teschke, “Multiclass Language Identification using Deep Learning on Spectral Images of Audio Signals,” May 10, 2019, arXiv: arXiv:1905.04348. doi: 10.48550/arXiv.1905.04348.

[9] “Common Voice by Mozilla,” commonvoice.mozilla.org. https://commonvoice.mozilla.org/en/datasets

[10] “GitHub - hubertmaka/Spoken-language-detection: The aim of the project is to design and build a model that recognizes language from a given sound sample. The assumption is a given number of different languages that the model will be able to recognize.,” GitHub, 2024. https://github.com/hubertmaka/Spoken-language-detection/tree/main