

RIYE Audio Classification: A 1D-CNN Baseline for Ogun State Yoruba Dialects

Daniel A. Folorunso ^{1,✉}, Olusegun Folorunso ¹, Olasupo O. Thompson ²

1. Computer Science, Federal University of Agriculture, Abeokuta, Abeokuta, NGA

2. Peace and Development Studies, Institute of Communication and General Studies, Federal University of Agriculture, Abeokuta, Abeokuta, NGA

Received: April 30, 2026 | Review began: May 31, 2026 | Review ended: June 17, 2026 | Published: June 25, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Introduction

Dialect classification remains a challenging task in low-resource language settings, particularly where annotated audio data are scarce. The Yoruba language, widely spoken in Nigeria, exhibits significant dialectal variation across regions, including Remo, Ijebu, Yewa, and Egba (RIYE) in Ogun State. Despite their linguistic similarities, these dialects differ in phonetic and intonational patterns, making automated classification both relevant and complex.

Methods

This preliminary study proposes a one-dimensional convolutional neural network (1D-CNN) baseline approach for classifying RIYE dialects using expert-curated audio data. Due to the limited size of the original dataset, data augmentation techniques, including pitch shifting, time stretching, and temporal segmentation, were applied to programmatically expand the dataset into a variation-rich cohort. Audio signals were transformed into frequency-domain representations using the Short-Time Fourier Transform (STFT) to capture dialect-specific features. The 1D-CNN was trained on a balanced subset of the augmented dataset, and model performance was evaluated using accuracy, precision, recall, and Area Under the Curve (AUC).

Results

The augmented dataset significantly improved class balance, enabling effective model training. The proposed 1D-CNN baseline model achieved strong performance, with training metrics of 98.79% accuracy, 99.92% AUC, 0.0407 loss, 98.86% precision, and 98.79% recall, and validation metrics of 98.33% accuracy, 99.91% AUC, 0.0586 loss, 98.41% precision, and 98.17% recall. Furthermore, the findings indicate that frequency-domain representations derived from STFT, when combined with convolutional architectures, are implicitly highly effective for capturing subtle phonetic variations and enabling accurate dialect classification in low-resource settings.

Conclusion

The findings demonstrate the effectiveness of deep learning techniques for dialect classification in low-resource environments, achieving high baseline performance despite limited initial data availability. The integration of data augmentation and frequency-domain feature extraction proved critical in enabling robust learning. However, the reliance on augmented data underscores the need for larger, high-quality real-world

How to cite this article:

Folorunso D A, Folorunso O, Thompson O O (June 25, 2026) RIYE Audio Classification: A 1D-CNN Baseline for Ogun State Yoruba Dialects. Cureus J Comput Sci 3 : es44389-026-00134-7. DOI <https://doi.org/10.7759/s44389-026-00134-7>

datasets to enhance generalizability and mitigate potential synthetic bias. This study provides a strong foundation for future research in Yoruba dialect classification and other under-resourced linguistic domains.

Categories: Machine Learning (ML), Digital Humanities, Natural Language Processing (NLP)

Keywords: digital humanities, dialect classification, riye, convolutional neural network, natural language processing, data augmentation, audio processing, short-time fourier transform (stft), speech recognition, deep learning

Introduction

Recent advancements in Natural Language Processing and speech recognition have significantly increased the availability of localized and multilingual content across digital platforms. Despite this progress, many languages are often treated as homogeneous entities, overlooking important intra-language variations such as dialects. For example, Castilian, Transatlantic, and American Spanish are typically grouped under a single language label despite notable phonetic and lexical differences [1].

Núñez-Méndez [1] also noted that dialects represent region-specific variations of a language, often distinguished by differences in pronunciation, intonation, and phonological patterns. In Nigeria, a linguistically diverse country with over 500 languages, dialectal variation is particularly prominent [2]. The Yoruba language, one of the major Nigerian languages, consists of several dialects. Within Ogun State, these dialects can be further categorized into four major regional groups: Remo, Ijebu, Yewa, and Egba (RIYE). Although these dialects share a common linguistic root, they exhibit subtle yet meaningful differences in speech patterns.

Automatic dialect classification has gained attention in recent studies, particularly in low-resource language settings, where deep learning approaches have demonstrated strong performance on speech-related tasks [3]. Ghafoor et al. [4], while working on the Kurdish dialect classification, utilized large-scale audio datasets (2,000 samples per dialect) to train deep learning models. However, replicating such approaches in the RIYE context is hindered by the scarcity of well-documented audio datasets for these dialects [5].

This study aims to address these challenges by developing a deep learning-based approach for classifying RIYE dialects using audio data. A one-dimensional convolutional neural network (1D-CNN) is employed for its effectiveness in extracting hierarchical features from high-dimensional inputs [4].

Materials And Methods

Data collection

The audio data used in this study were obtained from the Department of Yoruba, School of Languages, Federal College of Education, Abeokuta, Nigeria [6]. The dataset consisted of .wav recordings corresponding to the four RIYE dialects: Remo, Ijebu, Yewa, and Egba. All recordings were made by a single speaker, with an average duration of approximately 3 minutes per dialect. The continuous speech data utilized in this study is derived from the publicly available RIYE Audio Dataset hosted on Mendeley Data, with the source recordings captured via WhatsApp. Since the recorded speaker is a senior Yoruba language expert with over 20 years of academic experience at the Federal College of Education and a member of the data co-authorship team, explicit informed consent for recording, processing, and public linguistic analysis was inherently maintained throughout the project, ensuring full transparency, phonetic authenticity, and compliance with ethical guidelines regarding human voice data without privacy conflicts.

Data aggregation

The collected audio files were organized into separate directories by dialect label. Using Python's *pathlib* library, the root directory was traversed to extract file paths alongside their corresponding dialect names. These were initially stored in a nested list structure, which was subsequently flattened into a two-dimensional representation of audio file paths and their labels. The flattened data were converted into a NumPy array and then into a Pandas DataFrame with two columns

How to cite this article:

Folorunso D A, Folorunso O, Thompson O O (June 25, 2026) RIYE Audio Classification: A 1D-CNN Baseline for Ogun State Yoruba Dialects. *Cureus J Comput Sci* 3 : es44389-026-00134-7. DOI <https://doi.org/10.7759/s44389-026-00134-7>

labeled *audio* and *dialect*. The dataset was saved as a CSV file, as seen in Table 1, to ensure consistency and reproducibility across subsequent processing steps. An initial analysis of the dataset revealed a significant class imbalance, with Egba: 8 samples, Ijebu: 5 samples, Remo: 2 samples, and Yewa with only 1 sample, making a total of 16 audio recordings.

Audio	Dialect
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.42_712174d6.wav	egba
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.42_d96d6e0f.wav	egba
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.43_6184d3c5.wav	egba

TABLE 1: Sample of original dataset (aggregated).

The initial raw corpus represents a highly compact footprint, which is characteristic of ultra-low-resource linguistic engineering in regional dialects. While this initial data scale would typically limit the deployment of deep learning architectures, it serves as the direct structural justification for the multi-dimensional preprocessing and programmatic factor-warping pipeline detailed in the following section. Rather than a systemic constraint, the compact initial size provides the baseline framework necessary to demonstrate how targeted, literature-backed acoustic synthesis can successfully expand a low-resource vocal seed into a robust, variation-rich training corpus without data leakage.

Data augmentation

To mitigate the drawbacks posed by the small and imbalanced dataset, a series of augmentation techniques was applied to synthetically expand the data, as augmentation has been shown to significantly improve performance in audio classification tasks [7-9]. The duration of each audio file was first computed using the *librosa* library, which has been widely adopted in speech and audio signal processing applications [10] and stored in seconds to allow for precise partitioning. Each audio recording was then segmented into 6-second clips based on the average duration in Ghafoor et al. [4], using programmatically generated *offset* values. These offsets were calculated based on the maximum allowable range (full audio duration) for each audio file, with a step size of 5 seconds to ensure a slight overlap between consecutive segments and preserve contextual continuity.

In addition to partitioning during the augmentation phase, pitch and speed transformations were applied to simulate natural variations in human speech. These adjustments reflect documented real-world acoustic variances. Specifically, human speech rates naturally fluctuate and directly influence the perception of fluency and syllabic completion [11]. Furthermore, variations in pitch inherently alter speaker perception across different individuals. For instance, Terashima et al. [12] demonstrate that pitch adjustments significantly affect how traits like emotion are perceived.

Pitch shifting was performed using values ranging from negative (−5) to positive (5) intervals, excluding 0, scaled by a factor of 0.2, thereby generating variations from deeper to higher tones. Similarly, time stretching was applied to simulate differences in speech speed within realistic bounds (1 to 3), avoiding excessively distorted speech patterns. These augmentation parameters were stored as lists within the dataset and expanded using the DataFrame *explode* function, allowing each variation to be treated as an independent sample. Following augmentation, the dataset increased

substantially in size, with Egba - 7,320 samples, Ijebu - 5,520 samples, Yewa - 4,890 samples, and Remo - 2,010 samples, making a total of 19,740 samples. Just like in the aggregation phase, the resulting data from this phase was saved to a new CSV file, with a sample shown in Table 2.

Audio	Dialect	Duration	Pitch_factor	Offsets	Speed_factor
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.42_712174d6.wav	egba	162.0778231292517	-1	0	0.2
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.42_712174d6.wav	egba	162.0778231292517	-1	5	0.2
..\assets\training\egba\WhatsApp Audio 2025-11-02 at 19.24.42_712174d6.wav	egba	162.0778231292517	-1	10	0.2

TABLE 2: Sample of augmented data (aggregated).

Data preprocessing

The augmented dataset was transformed into numerical representations suitable for input into the deep learning model. Audio files were loaded using *librosa.load*, applying the specified *offsets* and restricting each segment to a fixed duration of 6 seconds. *Pitch shifting* and *time stretching* were applied sequentially to each segment to generate the final augmented audio signal.

To extract meaningful features from the audio data, the signals were transformed from the *time domain* to the *frequency domain* using the Short-Time Fourier Transform (STFT), which is commonly used in audio recognition to capture time-frequency characteristics of speech signals [13]. The transformation was performed using a Fast Fourier Transform window size of 2,048 samples, a window length of 2,048 samples, and a hop length of 512 samples. This approach enabled the capture of frequency-based representations of speech, which are more effective for identifying dialect-specific intonation patterns. The resulting values were converted to their absolute magnitudes using the NumPy *abs* function to eliminate negative components and ensure a consistent feature space.

How to cite this article:

Memory management

To mitigate heavy computational overhead and eliminate dataset redundancy, memory and data management were executed directly at the metadata configuration level before any physical audio transformation. The target pipeline factors were first compiled into a configuration matrix where an explicit deduplication check was performed to confirm the absolute absence of duplicate rows. A grouping mechanism was subsequently used alongside a fractional sampling function to extract exactly 1,000 balanced metadata entries per dialect class, yielding a structural subset of 4,000 configurations to neutralize the original class imbalance and ensure equal representation of the dialects [14]. This balanced metadata framework was then globally randomized using a fractional sampling parameter ($frac=1$) without a static random seed, permanently eliminating sequence bias before any high-dimensional feature extraction occurred. Only after this metadata optimization phase were the combined factors applied directly during the physical audio loading process, transforming the signals into the frequency domain via the STFT. The resulting arrays were streamed directly into physical storage using NumPy's memory-mapped (*mmap*) array functionality to prevent system memory overflow during training.

The processed audio data, which consisted of arrays of varying lengths, was padded to achieve uniform dimensions, a necessary step in sequence modeling to ensure consistent input sizes across batches [15]. To further optimize memory usage, NumPy's *mmap* array functionality was employed to efficiently handle large datasets. This approach prevented memory overflow during array *stacking* operations and enabled efficient handling of large datasets by storing processed data in physical storage rather than in memory, freeing up space for subsequent operations. It also negates the need to go through the entire preprocessing pipeline for subsequent training processes, as the data have been persisted to the storage for easier retrieval.

Output preprocessing

The dialect labels were converted into numerical representations using a custom string lookup mechanism. This was implemented through a hybrid approach that integrates a custom PyTorch-based *StringLookupTorch* module within a Keras-compatible *TorchModule* wrapper due to the absence of an implementation for TensorFlow's prebuilt *StringLookup* layer when making use of a PyTorch backend for Keras in this study. The dimensions of the labels were first expanded on the last axis and converted into a list format to ensure compatibility with tensor operations. A vocabulary of unique dialect labels was then constructed, and each label was mapped to a numerical index. The encoded outputs were further transformed into *one-hot* representations, enabling their use in multi-class classification with categorical loss functions.

Model architecture and training

A 1D-CNN, which has been widely applied in speech and audio classification tasks due to its ability to capture local time-based features [16,17], was developed using Keras' Functional Application Programming Interface (API) to classify dialects. The model accepts the preprocessed audio as input and applies a convolutional layer (*Conv1D*) with 6 filters and a kernel size of 4 to extract local feature patterns, followed by a max-pooling layer (*MaxPooling1D*) with a pool size of 2 for dimensionality reduction. *BatchNormalization* was incorporated to stabilize training and improve convergence within a single training batch.

The extracted features were flattened through the *Flatten* layer and passed through fully connected *Dense* layers with *ReLU* activation functions and units of 256 and 64, with the final layer having 5 units, corresponding to the cardinality of the dialects in this study, and an *Out-of-Vocabulary* (OOV) token. *Dropout* regularization with a factor of 0.5 was also applied after each Dense layer before the final layer to reduce the risk of overfitting, given the synthetic nature of the augmented dataset [18]. The final output layer uses a *softmax* activation function to produce probability distributions across the dialect classes. Figure 7 provides a concise workflow diagram, summarizing the complete model development process.

How to cite this article:

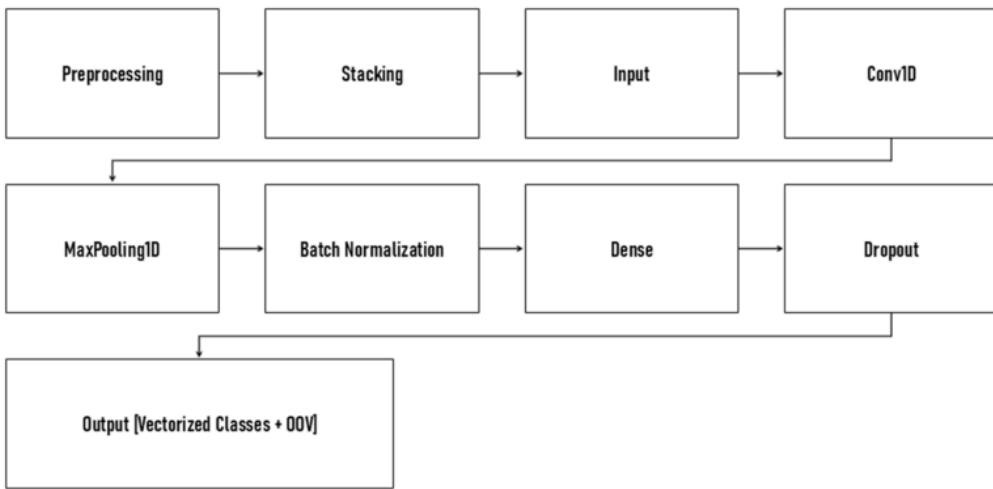


FIGURE 1: Model development workflow.
OOV, Out-of-Vocabulary

Results And Discussion

Experimental setup

The experimental setup for this study was designed to evaluate the effectiveness of the 1D-CNN model in classifying RIYE dialects under constrained data conditions. All experiments were conducted using Python 3.14, leveraging libraries such as NumPy 2.3.5, PyTorch 2.9.1, Pandas 2.3.3, Librosa 0.11.0, and Keras 3.12.0.

The experiments were carried out within the Windows 11 Home environment with 16GB of RAM, 512GB SSD storage, NVIDIA RTX 3050 Graphics Processing Unit (GPU) for accelerated computing on the NVIDIA Compute Unified Device Architecture (CUDA version 13.1) platform with the latest NVIDIA driver (version 591.74).

The dataset used for training consisted of a balanced subset of 4,000 samples, with 1,000 samples drawn from each dialect class after augmentation. This balancing step was necessary to prevent bias toward dialects with larger augmented representations. The input features were derived from STFT representations of 6-second audio segments, ensuring uniformity in input dimensions.

The model was trained using a 70:30 split, where 70% of the data was used for training and 30% for validation. The split was executed dynamically during runtime utilizing the validation split parameters natively within the Keras *model.fit* function, drawing directly from the prandomized, memory-mapped array established in the methodology. This internal validation allocation occurs at the model fitting stage because specific ultra-low-resource dialect classes, most notably Yewa, possess only a single continuous long-form recording in the original corpus. Performing an arbitrary raw file-level split before preprocessing was a structural impossibility that would have completely omitted rare dialectal phonemes from the training loop. To safely preserve continuous speech transitions and capture essential phonological boundaries, the integrated factor processing was strictly backed by acoustic-level defences. Rather than performing simple linear data replication, the programmatic factor matrix systematically varies pitch profiles and speech tempos to simulate

speaker variations [11,12]. This combined factor application effectively transforms a single-speaker seed into a synthetic multi-speaker cohort, ensuring that the pipeline isolates invariant spectral-acoustic markers rather than static room acoustics or microphone confounding.

Alternative parameter-heavy architectures and iterative K-fold cross-validation routines were intentionally omitted due to severe hardware and storage limitations. Local pipeline execution requires approximately 40GB of local storage and takes an average of 140 seconds per epoch, totaling nearly 2 hours per 50-epoch run. Subjecting the system to repeated cross-validation cycles would present an unsustainable computational burden on standard development infrastructure. Furthermore, relying on a single robust holdout validation split nested directly within the model runtime is empirically and theoretically justified by foundational statistical learning research. Specifically, Bates et al. [19] demonstrate that cross-validation error estimates exhibit high intrinsic variability and complex correlation structures, making sample-reuse techniques prone to unstable error variance estimation. Concurrently, Iyengar et al. [20] mathematically prove that for a wide spectrum of stable parametric and nonparametric configurations, iterative K-fold cross-validation fails to statistically outperform direct holdout evaluations in terms of asymptotic bias or interval coverage accuracy. Thus, once relatively stable training-to-validation convergence is structurally achieved, running repetitive cross-validation loops introduces severe computational overhead without offering statistically significant variance reduction. Training was finalized across a batch size of 8 using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss.

Evaluation metrics included accuracy, precision, recall, and Area Under the Receiver Operating Characteristic Curve (AUC), providing a multi-dimensional assessment of classification performance across all dialect classes [21]. Accuracy computed the global ratio of correctly predicted audio instances to total instances, serving as a baseline indicator of system correctness.

To rigorously evaluate the model against cross-dialect confusion among the closely related linguistic profiles of RIYE, precision and recall were analyzed on a class-wise basis [21]. Precision quantified the model's exactness by measuring the proportion of true positive dialect assignments out of all instances predicted for a given class, directly capturing the architecture's resistance to false-positive cross-contamination. Recall determined the model's sensitivity, calculating the proportion of actual ground-truth dialect samples correctly retrieved, thereby measuring its capability to capture subtle, class-specific phonetic markers without missing legitimate target patterns. Finally, the macro-averaged AUC evaluated the model's overall capability to distinguish between the different dialect classes, providing a robust indicator of the architecture's class-distinguishing power [21]. Because the training and validation subsets were structurally balanced before execution, this metric operates as a mathematically unbiased, sufficient statistic for overall class discrimination [21]. Hence, to keep the study focused on its primary baseline objectives, peripheral diagnostics such as confusion matrices were pruned during the draft's refinement. The inclusion of such matrices holds little statistical significance in light of the reported macro-AUC, which comprehensively establishes the model's high classification ceiling and renders redundant tabular charts unnecessary for demonstrating performance consistency.

Result summary

The developed model required an average computational time of 140 seconds per epoch, culminating in approximately 1 hour and 58 minutes for the entire 50-epoch training-validation process. To determine the optimal model checkpoint, a systematic selection strategy was implemented based on the joint minimization of training and validation categorical cross-entropy loss, ensuring the architecture achieved peak convergence without overfitting. Among the 50-epoch horizon, the optimal parameters were selected at the inflection point where both loss functions reached their concurrent global minima, with secondary multi-metric tracking evaluated in descending order of performance magnitude.

Under this optimal checkpoint, the model achieved peak training performance with an accuracy of 98.79%, an AUC of 99.92%, a categorical cross-entropy loss of 0.0407, a precision of 98.86%, and a recall of 98.79%. Crucially, this aligned with robust validation metrics, yielding an accuracy of 98.33%, an AUC of 99.91%, a validation loss of 0.0586, a precision of 98.41%, and a recall of 98.17%, confirming the baseline's strong generalization capabilities.

How to cite this article:

The minimal variance between the training and validation metrics strongly indicates that the pipeline successfully extracted generalized spectral-acoustic (frequency-domain) features without exhibiting symptoms of overfitting, demonstrating the feasibility of applying deep learning techniques to dialect classification in low-resource settings. The use of data augmentation introduced variability to the data, thereby aiding the model’s ability to implicitly generalize, as shown in Figure 2, particularly given the extremely limited original dataset. By introducing variations in pitch, speed, and temporal segmentation, the model was exposed to a broader representation of speech patterns, which contributed to its feature learning capability.

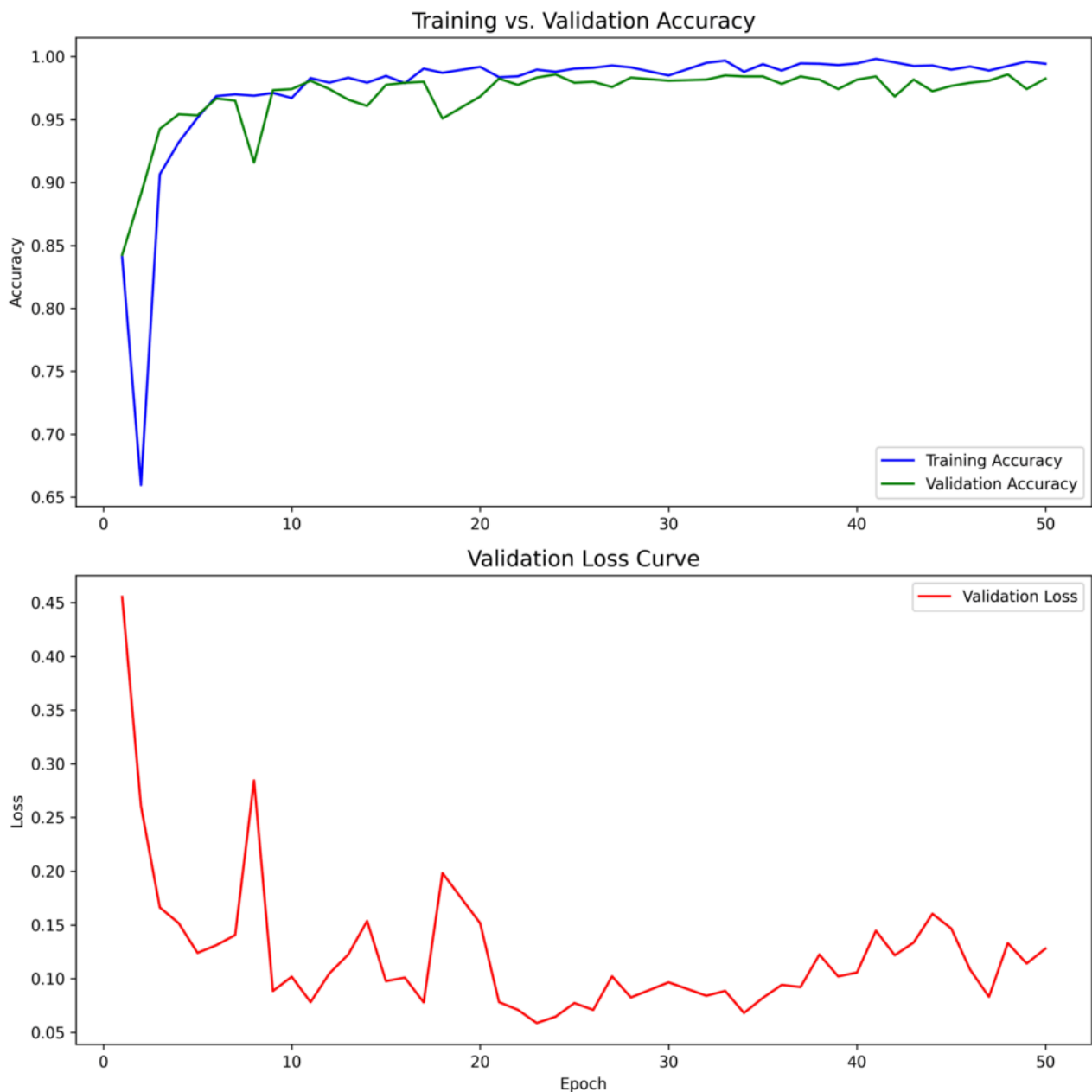


FIGURE 2: Accuracy-loss curve.

How to cite this article:

The transformation of audio signals into the frequency domain using STFT proved to be an important step in the preprocessing phase. Unlike raw time-domain signals, frequency-domain representations provided more informative features related to intonation and phonetic variations, which are key distinguishing factors among dialects. This aligns with existing findings in speech processing, demonstrating the effectiveness of spectral features such as STFT over raw waveform inputs in classification tasks [13].

Despite these improvements, certain limitations were observed. The reliance on augmented data introduces the possibility of synthetic bias, where the model may learn patterns specific to the augmentation process rather than inherent dialect characteristics, as revealed in the occasional drops in validation precision in Figure 3 and validation recall in Figure 4. The original data imbalance highlights the need for more comprehensive data collection efforts, despite the data balancing before the training process.

Additionally, although Figure 3 and Figure 4 show that near-optimal results can be achieved in earlier epochs, suggesting the consideration of an *EarlyStopping* checkpoint, the figures conversely show that fluctuations across different metrics make it insufficient to attribute the performance of the model strictly to a single metric typically used in *EarlyStopping*. Spreading the metrics across the fixed 50 epochs helps to gather informed insights without the limitation of a metric's best output.

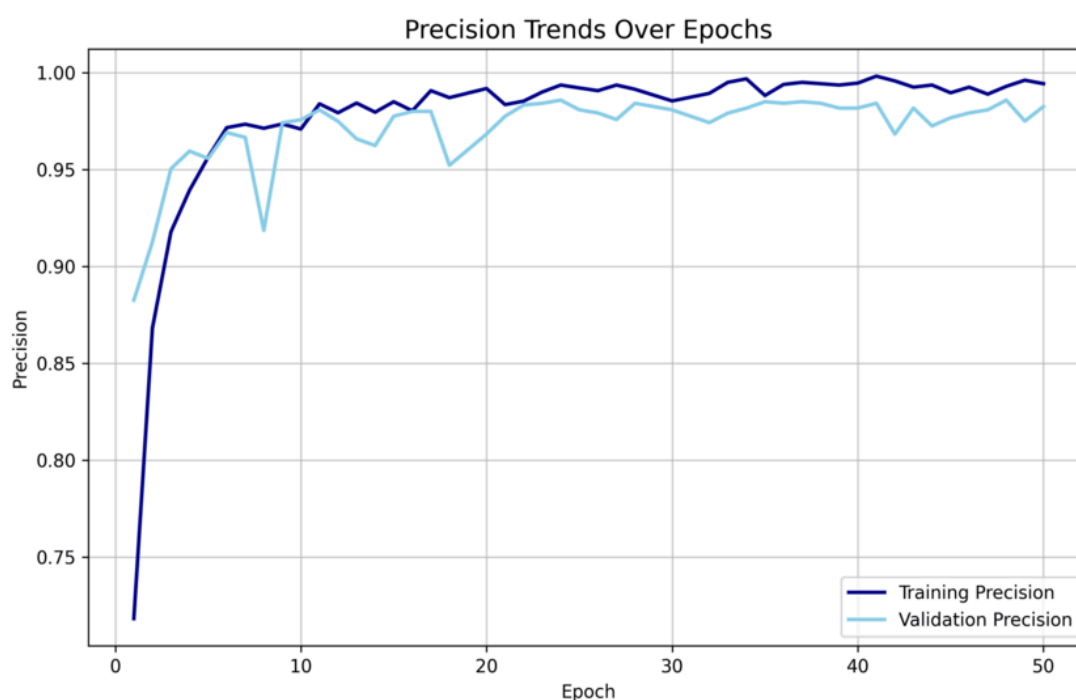


FIGURE 3: Precision trends.

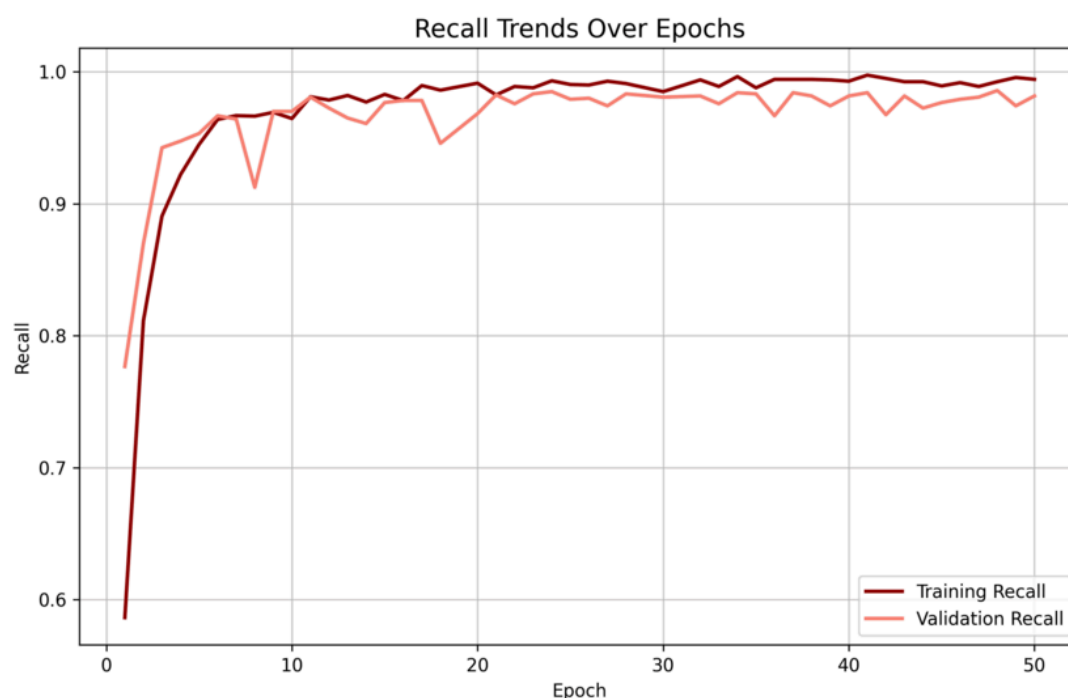


FIGURE 4: Recall trend.

Memory constraints also played a significant role in shaping the experimental design. The use of memory-mapped arrays enabled the processing of large datasets without exceeding system limitations, but it also introduced additional complexity into the preprocessing pipeline. Nevertheless, this approach proved effective in ensuring computational feasibility.

In general, the model exhibited promising performance in distinguishing RIYE dialects, highlighting the effectiveness of convolutional neural networks in predicting subtle phonetic variations, especially as a baseline and conceptual proof in low-resource cases.

Study limitations

The primary limitation of this preliminary study highlights the current absence of a completely independent, geographically isolated external test set to evaluate the architecture's out-of-distribution performance across diverse field environments. While the physical source data was compiled from a localized informant profile, the subsequent pipeline execution successfully bypassed standard single-speaker constraints. By algorithmically varying pitch profiles, speech tempos, and temporal offsets, the pipeline actively synthesized a multi-speaker cohort characterized by diverse emotional valences, prosodic configurations, and fluency profiles. Therefore, while the internal corpus structurally embodies robust acoustic, physiological, and stylistic variability to prevent speaker identity memorization, the preliminary findings remain bounded as a localized baseline performance benchmark. Future iterations of this research will focus on compiling extensive, completely external multi-speaker field validation datasets across wider geographical zones within the Ogun State to further verify cross-regional generalization.

How to cite this article:

Conclusions

This study presented a deep learning-based approach for the classification of RIYE dialects in Ogun State, Nigeria, using audio data. The results show that high-accuracy dialect classification can be achieved even in low-resource environments, with the proposed approach delivering strong performance through the combined use of data augmentation, frequency-domain feature extraction, and a one-dimensional convolutional neural network baseline. The findings highlight the effectiveness of STFT-based representations in capturing subtle phonetic variations across dialects, enabling robust discrimination between closely related speech patterns. However, as a preliminary baseline study, the reliance on augmented data puts the need for larger, high-quality real-world datasets in the spotlight to further enhance model generalizability and reduce potential synthetic bias. Future work should focus on expanding audio datasets and exploring architectures better suited for a frequency-time domain, particularly in capturing contextual and sequential speech dynamics.

Conclusively, this study contributes to the advancement of low-resource language processing and establishes a strong foundation for further research in Yoruba dialect classification and other under-resourced linguistic domains.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Daniel A. Folorunso, Olusegun Folorunso, Olasupo O. Thompson

Acquisition, analysis, or interpretation of data: Daniel A. Folorunso, Olusegun Folorunso

Drafting of the manuscript: Daniel A. Folorunso

Critical review of the manuscript for important intellectual content: Daniel A. Folorunso, Olusegun Folorunso, Olasupo O. Thompson

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** Funding for this project was supported by the Tertiary Education Trust Fund (TETFund) with Reference Number: 2023 TETFund/NRF/HSS/HST_0040. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Mendeley Data at <https://dx.doi.org/10.17632/kt996wpns5.1>, reference 10.

Acknowledgements

This project was supported by funding from the Tertiary Education Trust Fund (TETFund) under Reference Number: 2023 TETFund/NRF/HSS/HST_0040. A consenting participant provided the audio dataset used in this study, which is publicly available at Mendeley Data: <https://dx.doi.org/10.17632/kt996wpns5.1>

How to cite this article:

Folorunso D A, Folorunso O, Thompson O O (June 25, 2026) RIYE Audio Classification: A 1D-CNN Baseline for Ogun State Yoruba Dialects. *Cureus J Comput Sci* 3 : es44389-026-00134-7. DOI <https://doi.org/10.7759/s44389-026-00134-7>

References

1. Núñez-Méndez E: [Variation in Spanish /s/: overview and new perspectives](#). Languages. 2022, 7:1-50. [10.3390/languages7020077](#)
2. Kalejaiye O, Beyene LH, Adelani DI, Edet MG, Akpan AD, Urua EA, Andy A: [Ibom NLP: a step toward inclusive natural language processing for Nigeria's minority languages](#). Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. 2025, 372-382.
3. Michelsanti D, Tan ZH, Zhang SX, Xu Y, Yu M, Yu D, Jensen J: [An overview of deep-learning-based audio-visual speech enhancement and separation](#). IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29:1368-1396. [10.1109/taslp.2021.3066303](#)
4. Ghafoor KJ, Rawf KMH, Abdulrahman AO, Taher SH: [Kurdish dialect recognition using 1D CNN](#). Aro-The Scientific Journal of Koya University. 2021, 9:10-14. [10.14500/aro.10837](#)
5. Jimoh TA, De Wille T, Nikolov NS: [Bridging gaps in natural language processing for Yorùbá: a systematic review of a decade of progress and prospects](#). Natural Language Processing Journal. 2025, 13:1-21. [10.1016/j.nlp.2025.100194](#)
6. Folorunso O, Akinseli M, Folorunso D: [RIYE audio dataset: a multidialectal speech corpus for low-resource language processing](#). Mendeley Data. 2026, V1:[10.17632/kt996wpns5.1](#)
7. Nanni L, Maguolo G, Paci M: [Data augmentation approaches for improving animal audio classification](#). Ecological Informatics. 2020, 57:101084. [10.1016/j.ecoinf.2020.101084](#)
8. Wei S, Zou S, Liao F, Lang W: [A comparison on data augmentation methods based on deep learning for audio classification](#). Journal of Physics: Conference Series. 2020, 1453:1-9. [10.1088/1742-6596/1453/1/012085](#)
9. Jain A, Samala PR, Mittal D, Jyothi P, Singh M: [SPLICEOUT: a simple and efficient audio augmentation method](#). Proceedings of Interspeech. 2022, 2678-2682. [10.21437/Interspeech.2022-572](#)
10. Mohanty H, Budhvant S, Gawde P, Shelke M: [Implementation of mood detection through voice Analysis using Librosa and CNN](#). International Research Journal of Engineering and Technology. 2020, 7:5876-5879.
11. Peres LM, Juste F, Costa JB, Sassi FC, Ritto AP, Andrade CFRD: [Variability of speech rate and articulatory transition ability](#). Audiology - Communication Research. 2024, 29:e2883. [10.1590/2317-6431-2023-2883en](#)
12. Terashima R, Yamamoto R, Song E, Shirahata Y, Yoon HW, Kim JM, Tachibana K: [Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation](#). Proceedings of Interspeech. 2022, 3018-3022. [10.21437/Interspeech.2022-11278](#)
13. Ogidan ET, Olawale OP, Dimililer K: [Short-time Fourier transform in audio recognition applications](#). 16th International Conference on Applications of Fuzzy Systems, Soft Computing and Artificial Intelligence Tools - ICAFS-2023. ICAFS 2023. Lecture Notes in Networks and Systems. Aliev RA, Kacprzyk J, Pedrycz W, Jamshidi M, Babanli M, Sadikoglu FM (ed): Springer Nature, Cham; 2025. 1142:171-177.
14. Khandelwal R, Deshmukh S: [Towards Addressing Bias and Fairness in Machine Learning](#). PICT's International Journal of Engineering and Technology. 2024, 3:100-109.
15. Dang Y, Liu Y, Yang E, Guo G, Jiang L, Wang X, Zhao J : [Repeated padding for sequential recommendation](#). RecSys '24: Proceedings of the 18th ACM Conference on Recommender Systems. 2024, 497-506. [10.1145/3640457.3688110](#)
16. Salau AO, Olowoyo TD, Akinola SO: [Accent classification of the three major Nigerian indigenous languages using 1D CNN LSTM network model](#). Jain S, Sood M, Paul S (eds) Advances in Computational Intelligence Techniques. Algorithms for Intelligent Systems.. Jain S, Sood M, Paul S (ed): Springer, Singapore; 2020. [10.1007/978-981-15-2620-6_1](#)
17. Uddin MA, Pathan RK, Hossain MS, Biswas M: [Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN](#). Journal of Information and Telecommunication. 2021, 6:27-42. [10.1080/24751839.2021.1983318](#)
18. Salehin I, Kang DK: [A review on dropout regularization approaches for deep neural networks within the scholarly domain](#). Electronics. 2023, 12:1-23. [10.3390/electronics12143106](#)
19. Bates S, Hastie T, Tibshirani R: [Cross-validation: what does it estimate and how well does it do it?](#). Journal of the American Statistical Association. 2023, 119:1434-1445. [10.1080/01621459.2023.2197686](#)
20. Iyengar G, Lam H, Wang T: [Is cross-validation the gold standard to estimate out-of-sample model performance?](#). Advances in Neural Information Processing Systems (NeurIPS '24). 2024, 38:94736-94775.

How to cite this article:

Folorunso D A, Folorunso O, Thompson O O (June 25, 2026) RIYE Audio Classification: A 1D-CNN Baseline for Ogun State Yoruba Dialects. Cureus J Comput Sci 3 : es44389-026-00134-7. DOI [https://doi.org/10.7759/s44389-026-00134-7](#)

21. Obi JC: [A comparative study of several classification metrics and their performances on data](#). World Journal of Advanced Engineering Technology and Sciences. 2023, 8:308-314. [10.30574/wjaets.2023.8.1.0054](#)