Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automatic Identification of Maxaatiri and Maay Somali Dialects from Speech Using Mel-Spectrogram-Based Convolutional Neural Networks

Mohamed Mohamud Ali

mohamedmohamud@simad.edu.so

SIMAD University

Article

Keywords: Somali dialect detection, Maay Somali, Maxaatiri Somali, speech classification, low-resource language, mel-spectrogram, convolutional neural network, deep learning

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10071456/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Automatic dialect identification is important for developing inclusive speech technologies in low-resource languages. Somali includes several major dialect varieties, yet dialect-level speech classification remains underexplored. This study presents a preliminary deep learning approach for classifying Maxaatiri/Standard Somali and Maay Somali speech using mel-spectrogram acoustic features and a convolutional neural network (CNN). A custom audio collection pipeline was developed to extract speech segments from publicly accessible Somali-language YouTube broadcast videos, with Maxaatiri/Standard Somali samples collected from Somali National Television and Maay Somali samples collected from Arlaadi TV. The audio was converted to mono-channel 16 kHz WAV format, processed through silence-based segmentation, divided into fixed-length speech chunks, and transformed into mel-spectrogram representations for model training. The final experimental dataset contained 180 audio segments, balanced across the two dialect classes. The CNN achieved an accuracy of 81.48%, precision of 81.60%, recall of 81.48%, and weighted F1-score of 81.43% on a held-out test set. The confusion matrix showed that the model correctly classified most samples from both dialect categories, although cross-dialect misclassification remained. These findings provide an initial computational baseline for Somali dialect detection and demonstrate the feasibility of deep learning-based acoustic classification for Somali speech. However, because the current dataset was derived from limited broadcast sources, contained a small number of samples, and lacked speaker-level metadata, the reported performance should be interpreted as a preliminary baseline rather than evidence of generalization across all Somali speakers, dialect communities, or recording conditions.

1. Introduction

Speech technologies have become increasingly important in communication, education, public service delivery, translation, media monitoring, and human-computer interaction. However, their development remains uneven across the world's languages. High-resource languages benefit from large annotated speech corpora, established benchmarks, and pretrained acoustic models, while low-resource languages often lack the datasets and evaluation standards needed for reliable speech processing. Public speech datasets such as Common Voice, LibriSpeech, Multilingual LibriSpeech, Libri-Light, and AudioSet have helped accelerate speech and audio research, but many African languages remain insufficiently represented in such resources [1, 12–16]. This imbalance is especially visible in many African languages, where speech technology research is constrained by limited annotated data, dialectal diversity, weak benchmarking infrastructure, and uncertain data availability conditions [1]. As a result, many speech-based systems remain less effective for communities whose languages and dialects are underrepresented in computational research.

Somali is a Cushitic language spoken across the Horn of Africa and the Somali diaspora. Although it is often treated as a single national language, Somali contains important dialectal variation. Classical dialectological and grammatical studies have identified major dialect groupings within Somali, including varieties commonly associated with Maxaatiri/Standard Somali and Maay Somali [2, 8–10]. Maxaatiri,

often linked to Northern or Standard Somali, has a stronger presence in formal education, administration, national broadcasting, and written communication. Maay, spoken mainly in southern Somalia by communities including Digil and Mirifle/Rahanweyn groups, differs from Maxaatiri in phonological, lexical, and morphosyntactic features. These differences can affect mutual intelligibility and create practical communication barriers in education, media, humanitarian communication, and digital services [3, 4].

Despite the social and linguistic importance of Somali dialects, dialect-aware Somali speech technology remains limited. Existing Somali natural language processing research has focused more on written text, classification, language modelling, and general low-resource NLP than on dialect-level acoustic classification. In speech processing, Somali is usually discussed under the wider challenge of African low-resource automatic speech recognition, where data scarcity, speaker variation, dialect diversity, transcription cost, and limited public benchmarks remain central obstacles [1, 5]. This creates a gap between the real linguistic diversity of Somali-speaking communities and the current capacity of computational systems to identify and process dialectal speech variation.

Automatic dialect identification is a related task within spoken language processing. It aims to classify speech according to dialect, accent, or regional variety. This task is useful because dialectal variation can reduce the performance of automatic speech recognition, voice search, speech translation, and other downstream language technologies. If a system cannot detect dialectal differences, it may produce biased or inaccurate outputs for speakers whose speech differs from the standard variety used during model development. Dialect identification can therefore support more inclusive speech systems by enabling dialect-specific preprocessing, routing, model adaptation, or user interface personalization.

Recent developments in deep learning have improved audio and speech classification. One common approach is to convert speech signals into time-frequency representations such as spectrograms or mel-spectrograms and then train convolutional neural networks to identify discriminative acoustic patterns. CNN-based audio models have been widely used because spectrograms preserve both temporal and frequency information in a format suitable for image-like feature learning [6, 7, 17, 18]. Studies in spoken language identification have also shown that deep learning models can classify languages using spectrogram-based features, including log-mel representations, without requiring full speech transcription [6, 11]. This is relevant for low-resource dialect detection because dialect labels may be easier to obtain than full transcriptions, especially when the goal is classification rather than speech-to-text conversion.

For Somali dialect detection, mel-spectrogram-based CNN modelling provides a practical starting point. The method does not require manual transcription and can learn acoustic patterns directly from labelled speech segments. This is useful in a low-resource setting where large labelled speech corpora are not yet available. However, the method also has limitations. If data are collected from different media channels for different dialect classes, the model may learn channel-specific acoustic properties rather than dialectal features. For example, background noise, studio quality, compression format, speaker

style, music, and broadcast identity may influence classification. Therefore, Somali dialect detection studies must carefully describe data sources, preprocessing steps, evaluation design, and possible bias in the dataset.

This study addresses the gap by developing a preliminary deep learning framework for automatic classification of Maxaatiri/Standard Somali and Maay Somali speech. Audio data were collected from publicly accessible Somali-language YouTube broadcast videos using a custom audio extraction and preprocessing pipeline. Maxaatiri/Standard Somali samples were obtained from Somali National Television, while Maay Somali samples were obtained from Arlaadi TV. The extracted audio was converted to mono-channel 16 kHz WAV format, cleaned through silence-based processing, segmented into fixed-length audio chunks, and transformed into mel-spectrogram representations. A convolutional neural network was then trained to classify the speech segments into the two dialect categories.

The contribution of this study is threefold. First, it presents an initial computational baseline for Somali dialect detection using acoustic deep learning. Second, it demonstrates a reproducible pipeline linking public broadcast audio collection, audio preprocessing, mel-spectrogram extraction, CNN training, and metric-based evaluation. Third, it identifies methodological challenges that are especially important for Somali and other low-resource languages, including limited dataset size, source-channel bias, and the need for larger multi-speaker and multi-source dialect corpora. By focusing on Maxaatiri/Standard Somali and Maay Somali, the study contributes an initial experimental baseline for dialect-aware Somali speech technology while also identifying the methodological requirements needed for future generalizable systems, including larger datasets, multiple media sources, speaker-level metadata, and external validation.

2. Results

2.1 Dataset composition and feature representation

The final experimental dataset contained 180 speech audio segments representing two Somali dialect categories: Maxaatiri/Standard Somali and Maay Somali. The dataset was balanced across the two classes, with 90 audio segments assigned to each dialect category. All audio files were converted into mono-channel WAV format at a sampling rate of 16 kHz and divided into fixed-length 10-second segments. The processed audio files were then transformed into mel-spectrogram representations and used as input features for the convolutional neural network model.

Table 1
Dataset summary

Dataset feature	Maxaatiri / Standard Somali	Maay Somali	Total
Audio segments	90	90	180
Audio format	WAV	WAV	WAV
Sampling rate	16 kHz	16 kHz	16 kHz
Audio channel	Mono	Mono	Mono
Segment length	10 seconds	10 seconds	10 seconds
Feature representation	Mel-spectrogram	Mel-spectrogram	Mel-spectrogram

2.2 Model training performance

The convolutional neural network was trained using mel-spectrogram representations extracted from the processed audio segments. During training, the loss decreased across epochs, indicating that the model was learning discriminative acoustic patterns from the input features. However, training loss alone does not provide sufficient evidence of classification performance. Therefore, the model was evaluated using a held-out test set and standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis.

2.3 Test-set classification performance

The model achieved a test accuracy of 81.48%, with a weighted precision of 81.60%, weighted recall of 81.48%, and weighted F1-score of 81.43%. The test loss was 0.4511. These results indicate that the model was able to distinguish between Maxaatiri/Standard Somali and Maay Somali speech with moderate classification performance on the held-out test split.

Table 2
Final test performance of the CNN model

Metric	Score
Test loss	0.4511
Accuracy	81.48%
Precision	81.60%
Recall	81.48%
Weighted F1-score	81.43%

2.4 Class-level performance

The class-level results show that the model performed slightly better on Maxaatiri/Standard Somali than on Maay Somali in terms of F1-score. For Maxaatiri/Standard Somali, the model achieved a precision of 0.80, recall of 0.86, and F1-score of 0.83. For Maay Somali, the model achieved a precision of 0.83, recall of 0.77, and F1-score of 0.80. This suggests that the model identified Maxaatiri/Standard Somali samples with higher recall, while Maay Somali predictions showed slightly higher precision but lower recall.

Table 3
Class-level classification report

Dialect class	Precision	Recall	F1-score	Test samples
Maxaatiri / Standard Somali	0.80	0.86	0.83	14
Maay Somali	0.83	0.77	0.80	13
Weighted average	0.8160	0.8148	0.8143	27

2.5 Confusion matrix analysis

The confusion matrix shows that the model correctly classified 12 out of 14 Maxaatiri/Standard Somali test samples and 10 out of 13 Maay Somali test samples. Two Maxaatiri/Standard Somali samples were incorrectly classified as Maay Somali, while three Maay Somali samples were incorrectly classified as Maxaatiri/Standard Somali.

Table 4
Confusion matrix

Actual / Predicted	Maxaatiri / Standard Somali	Maay Somali
Maxaatiri / Standard Somali	12	2
Maay Somali	3	10

Overall, the findings provide preliminary evidence that mel-spectrogram-based convolutional neural networks can classify Maxaatiri/Standard Somali and Maay Somali speech with reasonable performance under the current experimental conditions. However, the results should be interpreted cautiously because the dataset was small, the test set contained a limited number of samples, and the audio was derived from restricted broadcast sources. The current evaluation should therefore be viewed as an initial baseline rather than definitive evidence of model generalization across Somali speakers, dialect communities, recording environments, or unseen media channels.

3. Discussion

This study developed a preliminary deep learning framework for automatic classification of Maxaatiri/Standard Somali and Maay Somali speech using mel-spectrogram acoustic features and a convolutional neural network. The model achieved a test accuracy of 81.48% and a weighted F1-score of

81.43%, suggesting that acoustic deep learning methods can distinguish between the two Somali dialect categories with moderate performance under the current experimental conditions. These findings provide an initial computational baseline for Somali dialect detection and demonstrate the feasibility of applying deep learning to dialect-aware speech processing in a low-resource language context.

The class-level results indicate that the model performed relatively well for both dialect classes, although performance was not identical. Maxaatiri/Standard Somali achieved a precision of 0.80, recall of 0.86, and F1-score of 0.83, while Maay Somali achieved a precision of 0.83, recall of 0.77, and F1-score of 0.80. This means that the model detected Maxaatiri/Standard Somali samples with slightly higher recall, while Maay Somali predictions showed slightly higher precision but lower recall. The confusion matrix further showed that 12 of 14 Maxaatiri/Standard Somali test samples and 10 of 13 Maay Somali test samples were correctly classified. Misclassification occurred in both directions, with two Maxaatiri/Standard Somali samples classified as Maay Somali and three Maay Somali samples classified as Maxaatiri/Standard Somali.

The results are useful because Somali dialect detection remains underdeveloped in computational speech research. Most existing Somali language technology work is limited by the lack of publicly available speech datasets, limited dialect annotation, and the broader challenges associated with low-resource language processing. By using publicly accessible broadcast audio and a reproducible preprocessing pipeline, this study shows that initial dialect classification experiments can be conducted even when large curated corpora are unavailable. The approach may support future applications in dialect-aware speech recognition, voice-based public service tools, media analysis, educational technologies, and inclusive Somali language processing systems.

However, the findings should be interpreted cautiously. The dataset was relatively small, containing 180 audio segments, with 90 samples for each dialect class. Although the classes were balanced, the data were derived from limited broadcast sources. Maxaatiri/Standard Somali samples were collected from Somali National Television, while Maay Somali samples were collected from Arlaadi TV. This creates a possible source-channel bias. The model may have learned not only dialectal speech patterns but also acoustic characteristics associated with specific media sources, such as studio quality, compression effects, background noise, presenter style, or broadcast formatting. This is a key limitation because a model that performs well on a held-out split may not necessarily generalize to unseen speakers, unseen videos, or non-broadcast speech.

Another limitation concerns dataset partitioning and independence. The filename structure was used to extract approximate video identifiers, and video-level grouping was applied where feasible to reduce leakage between training and testing samples. However, the available dataset contained a limited number of video groups and did not include reliable speaker identifiers. Therefore, the current evaluation should be treated as a preliminary performance estimate rather than definitive evidence of robust dialect generalization.

The study also has methodological limitations related to speaker diversity and linguistic annotation. Speaker-level metadata were unavailable for many recordings, meaning that the study could not explicitly control for speaker identity during dataset partitioning. The current dataset also does not fully document speaker gender, age, region, speaking style, or transcription content. Without these details, it is difficult to determine whether classification errors are caused by dialect overlap, speaker-specific features, background noise, or source-specific recording conditions. Future work should include broader speaker representation, multiple channels per dialect, manual dialect verification, speaker-level splitting, and external validation using audio from unseen sources.

Future research should expand the dataset to include more videos, more speakers, more recording environments, and additional Somali dialect varieties. A stronger design would collect multiple sources for each dialect and evaluate the model using video-level, speaker-level, or channel-level splitting. Future studies should also compare the CNN model with traditional machine-learning baselines such as MFCC-based support vector machines or random forests, as well as more advanced pretrained speech models such as wav2vec 2.0, XLSR, and XLS-R architectures [19–21]. These comparisons would clarify whether the observed performance is specific to the CNN architecture or reflects broader dialectal separability in Somali speech.

This study should be interpreted as a preliminary baseline rather than a fully generalizable Somali dialect identification system. Although the model achieved 81.48% accuracy and an 81.43% weighted F1-score, the evaluation was based on a relatively small dataset of 180 audio segments and a limited held-out test set. In addition, the two dialect classes were collected from different broadcast sources, which may introduce source-channel bias. Therefore, the model may have learned acoustic properties related to media source, recording quality, compression, speaker style, or broadcast format in addition to dialect-specific speech patterns. Future work should address this limitation by collecting larger multi-source and multi-speaker datasets, applying speaker-level and video-level partitioning, and validating the model on external audio sources not used during training.

Overall, the study provides an early but useful step toward Somali dialect-aware speech technology. The results show that mel-spectrogram-based CNN classification can identify Maxaatiri/Standard Somali and Maay Somali speech with promising preliminary accuracy. At the same time, the limitations of dataset size, restricted broadcast sources, speaker-level metadata, and evaluation design indicate that larger and more diverse Somali speech corpora are needed before such models can be considered reliable for real-world deployment.

4. Methods

4.1 Study design

This study used an experimental deep learning design to classify Somali speech audio into two dialect categories: Maxaatiri/Standard Somali and Maay Somali. The workflow consisted of five main stages:

audio collection, audio preprocessing, feature extraction, model training, and performance evaluation. The study used mel-spectrogram representations as acoustic input features and a convolutional neural network as the classification model.

The overall workflow of the proposed Somali dialect detection system is shown in Fig. 1.

4.2 Data sources and audio collection

Audio data were collected from publicly accessible Somali-language YouTube broadcast videos. Maxaatiri/Standard Somali samples were collected from Somali National Television, while Maay Somali samples were collected from Arlaadi TV. A custom Python-based audio collection pipeline was developed to extract audio from selected videos. The pipeline used yt-dlp for audio extraction and FFmpeg-supported processing tools for audio conversion.

The audio collection procedure produced dialect-specific folders corresponding to the two target classes: standard_somali and maay_somali. Each processed audio segment was saved under its respective dialect folder and later loaded by the model training pipeline.

4.3 Audio preprocessing

The extracted audio files were converted into WAV format, resampled to 16 kHz, and converted to mono-channel audio. Silence-based segmentation was applied to reduce non-speech portions and separate speech-containing regions. Audio recordings were initially segmented into 10-second chunks during preprocessing. During model training, each sample was standardized to a fixed 5-second input window. Segments longer than 5 seconds were truncated, whereas shorter segments were zero-padded to ensure uniform model input dimensions.

4.4 Feature extraction

Each standardized audio sample was converted into a mel-spectrogram using librosa [22]. The sampling rate was set to 16 kHz, and 64 mel bands were extracted for each audio sample. The mel-spectrogram was then converted to decibel scale using power-to-decibel transformation. A channel dimension was added to each mel-spectrogram so that it could be used as input to the convolutional neural network.

4.5 Dataset composition

The final dataset contained 180 audio segments, balanced across the two dialect classes. The dataset included 90 Maxaatiri/Standard Somali segments and 90 Maay Somali segments. The class labels were encoded as follows: standard_somali was assigned label 0, and maay_somali was assigned label 1.

4.6 Data splitting

To reduce the risk of data leakage, video identifiers were extracted from filenames and used to group related audio segments during dataset partitioning where feasible. The preferred evaluation design was a video-level group split, because audio chunks originating from the same source video may share acoustic and speaker characteristics. Class balance was preserved during dataset splitting using Scikit-

learn utilities [24], and a random seed of 42 was used to support reproducibility. Because speaker-level metadata were not available, the split could not explicitly control for speaker identity.

4.7 Model architecture

The classification model was a convolutional neural network implemented in PyTorch [23]. The model consisted of two convolutional blocks followed by a fully connected classification layer. The first convolutional layer used 16 filters with a 3 x 3 kernel and padding of 1, followed by a rectified linear unit (ReLU) activation function and a 2 x 2 max-pooling operation. The second convolutional layer used 32 filters with a 3 x 3 kernel and padding of 1, followed by ReLU activation and 2 x 2 max pooling. The resulting feature maps were flattened and passed to a fully connected output layer with two neurons corresponding to the Maxaatiri/Standard Somali and Maay Somali classes.

4.8 Training procedure

The model was trained using cross-entropy loss and the Adam optimizer [25] with a learning rate of 0.001. Training was conducted for 20 epochs with a batch size of 16. After each epoch, the model was evaluated on the validation set. The model checkpoint with the highest validation F1-score was saved and used for final test-set evaluation. All experiments were conducted in Google Colab using the CPU runtime. The implementation used Python with PyTorch for model development, Librosa for audio feature extraction, and Scikit-learn for data splitting and evaluation. No GPU acceleration was used during model training. Training was computationally feasible on CPU because the dataset was small and the CNN architecture was lightweight.

4.9 Evaluation metrics

Model performance was evaluated using test loss, accuracy, precision, recall, F1-score, and confusion matrix analysis. Weighted precision, recall, and F1-score were reported to account for class-level performance across the two dialect categories. The confusion matrix was used to examine the direction and frequency of misclassifications between Maxaatiri/Standard Somali and Maay Somali.

Because the present study reports a preliminary baseline from a small dataset, future experiments should include repeated runs, k-fold cross-validation, and comparisons with traditional machine-learning baselines such as MFCC-based support vector machines, random forests, and logistic regression. These analyses are recommended to assess performance stability and determine whether the CNN provides measurable gains over simpler classifiers.

Conclusion

This study presented a preliminary CNN-based framework for classifying Maxaatiri/Standard Somali and Maay Somali speech using mel-spectrogram acoustic features. The model achieved 81.48% accuracy and an 81.43% weighted F1-score on a held-out test set, showing that deep learning can provide a feasible starting point for Somali dialect detection in low-resource speech technology contexts. However, the findings should be interpreted cautiously because the dataset was small, the test set was

limited, speaker-level metadata were unavailable, and the two dialect classes were collected from restricted broadcast sources. These limitations mean that the model should be viewed as an initial experimental baseline rather than a fully generalizable Somali dialect identification system. Future research should develop larger, speaker-diverse, and multi-source Somali speech corpora; apply video-level and speaker-level evaluation designs; compare CNN performance with traditional and pretrained speech models; and validate results using external audio sources.

Declarations

Competing interests

The authors declare no competing interests.

Funding

The author received no funding for this work.

Author Contribution

M.M.A. conceptualized the study, designed the research methodology, collected and processed the audio data, developed the preprocessing pipeline, implemented the mel-spectrogram feature extraction, built and trained the convolutional neural network model, performed the statistical and classification analysis, prepared the figures and tables, wrote the main manuscript text, revised the manuscript, and approved the final version for submission.

Acknowledgement

The author gratefully acknowledges SIMAD University for providing institutional support and resources that facilitated this research.

Data availability

The raw audio data used in this study were derived from publicly accessible YouTube broadcast videos published by Somali National Television and Arlaadi TV. Due to third-party copyright and platform-use restrictions, the raw downloaded audio files are not redistributed with this article. To support reproducibility, the authors can provide the source video identifiers or URLs, segment-level metadata, preprocessing procedure, feature extraction code, model training code, and trained model files, subject to applicable legal, ethical, and institutional requirements.

Code availability

The code used for audio extraction, preprocessing, mel-spectrogram generation, CNN training, and model evaluation is publicly available at: https://github.com/mohamed-tsx/Somali-Dialect-Detection. The repository includes the preprocessing scripts, model architecture, training workflow, evaluation scripts, and supporting files required to reproduce the experimental procedure.

Ethics declaration

This study used speech data derived from publicly accessible broadcast media and did not involve direct interaction with human participants. No private recordings were collected, and no personal identifying information was extracted, analysed, or reported. The analysis was limited to acoustic dialect classification. Because the study used publicly available media content, raw audio files are not redistributed. The authors obtained institutional confirmation that formal ethical approval was not required for this secondary analysis of publicly accessible broadcast data.

References

1. Imam, S. H. et al. Automatic speech recognition for African low-resource languages: Challenges and future directions. Proceedings of the Sixth Workshop on African Natural Language Processing. (2025).
2. Lamberti, M. *Map of Somali Dialects in the Somali Democratic Republic* (Helmut Buske, 1986).
3. Hassan, D. Somali dialects in the United States: How intelligible is Af-Maay to speakers of Af-Maxaa? Master's thesis, Minnesota State University, Mankato. (2011).
4. CLEAR Global. (n.d.). Language data for Somalia. CLEAR Global.
5. Mohamud, J. H., Thompson, L. A., Ndoye, A. & Besacier, L. Fast development of ASR in African languages using self-supervised speech representation learning. *arXiv preprint arXiv* :210308993. (2021).
6. Singh, G., Sharma, R. & Kaur, M. Spoken language identification using deep learning. Computational Intelligence and Neuroscience, 2021, Article 5123671. (2021).
7. Hershey, S. et al. CNN architectures for large-scale audio classification. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 131–135. (2017).
8. Saeed, J. I. *Somali* (John Benjamins, 1999).
9. Andrzejewski, B. W. *The Declensions of Somali Nouns* (School of Oriental and African Studies, University of London, 1964).
10. Andrzejewski, B. W. *The Case System in Somali* (School of Oriental and African Studies, University of London, 1979).
11. Valk, J. & Alumäe, T. VoxLingua107: A dataset for spoken language recognition. Proceedings of the IEEE Spoken Language Technology Workshop (SLT), 652–658. (2021).

12. Ardila, R. et al. Common Voice: A massively-multilingual speech corpus. Proceedings of the 12th Language Resources and Evaluation Conference, 4218–4222. (2020).
13. Gemmeke, J. F. et al. Audio Set: An ontology and human-labeled dataset for audio events. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 776–780. (2017).
14. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. LibriSpeech: An ASR corpus based on public domain audio books. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206–5210. (2015).
15. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G. & Collobert, R. MLS: A large-scale multilingual dataset for speech research. Proceedings of Interspeech 2020, 2757–2761. (2020).
16. Kahn, J. et al. Libri-Light: A benchmark for ASR with limited or no supervision. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7669–7673. (2020).
17. Piczak, K. J. Environmental sound classification with convolutional neural networks. Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 1–6. (2015).
18. Salamon, J. & Bello, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE. Signal. Process. Lett.* **24** (3), 279–283 (2017).
19. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural. Inf. Process. Syst.* **33**, 12449–12460 (2020).
20. Conneau, A., Baevski, A., Collobert, R., Mohamed, A. & Auli, M. Unsupervised cross-lingual representation learning for speech recognition. Proceedings of Interspeech 2021, 2426–2430. (2021).
21. Babu, A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296. (2021).
22. McFee, B. et al. librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference, 18–25. (2015).
23. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural. Inf. Process. Syst.*, **32**. (2019).
24. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
25. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). (2015).

Figures



Figure 1

Proposed workflow for Somali dialect detection. The pipeline begins with publicly accessible YouTube broadcast videos, followed by audio extraction, WAV conversion to 16 kHz mono format, silence reduction and segmentation, standardization into 5-second model inputs, mel-spectrogram extraction, CNN-based classification, and final prediction into Maxaatiri/Standard Somali or Maay Somali dialect categories.

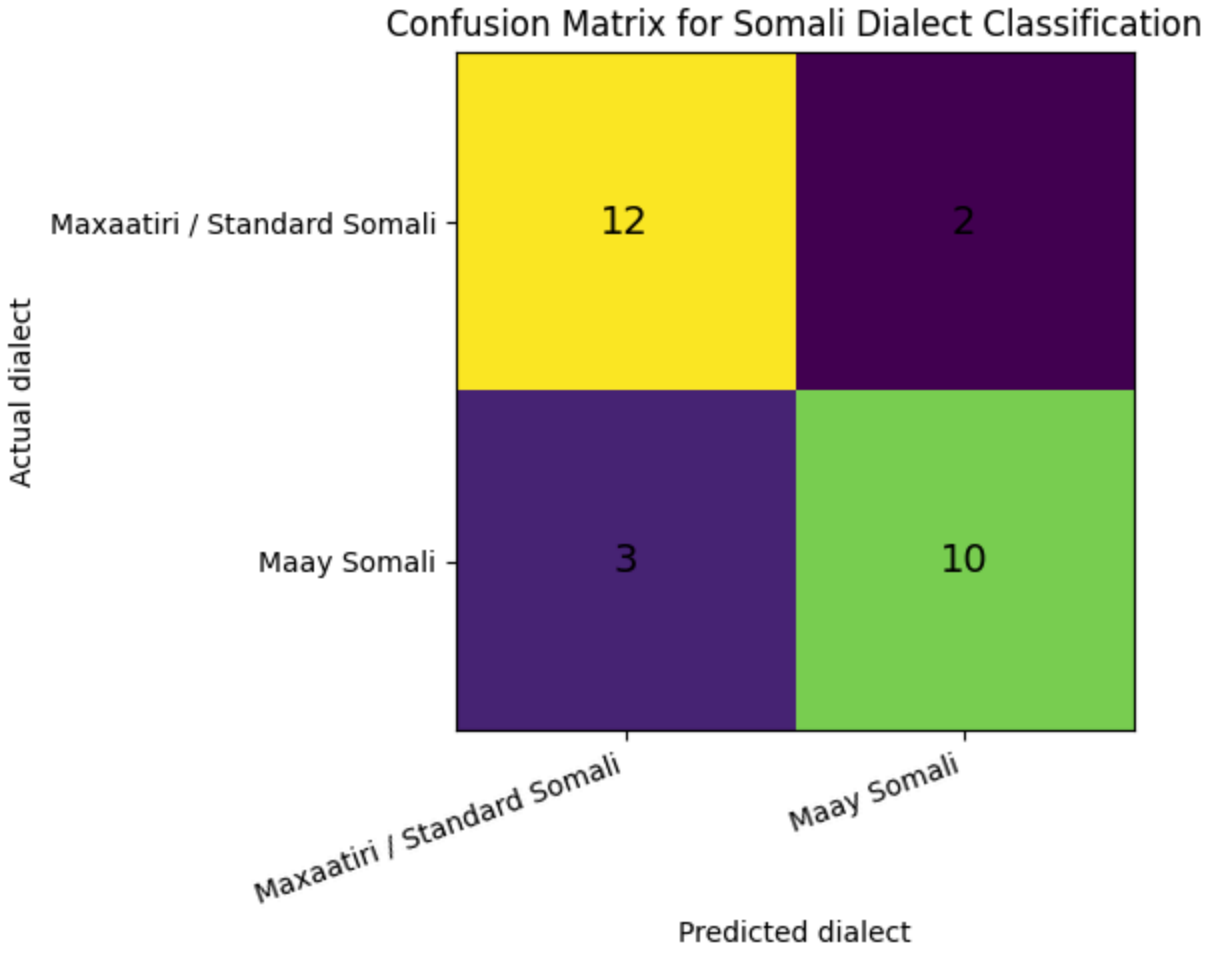


Figure 2

Confusion matrix for Somali dialect classification. The matrix shows the classification performance of the CNN model on the held-out test set. The model correctly classified 12 Maxaatiri/Standard Somali samples and 10 Maay Somali samples, while 2 Maxaatiri/Standard Somali samples were misclassified as Maay Somali and 3 Maay Somali samples were misclassified as Maxaatiri/Standard Somali.