Voices of Luxembourg: Tackling Dialect Diversity in a Low-Resource Setting

Nina Hosseini-Kivanani, Christoph Schommer, Peter Gilles
University of Luxembourg
Esch-Belval, Esch-sur-Alzette, Luxembourg
{nina.hosseinikivanani, christoph.schommer, peter.gilles}@uni.lu

Abstract

Dialect classification is essential for preserving linguistic diversity, particularly in low-resource languages such as Luxembourgish. This study introduces one of the first systematic approaches to classifying Luxembourgish dialects, addressing phonetic, prosodic, and lexical variations across four major regions. We benchmarked multiple models, including state-of-the-art pre-trained speech models like Wav2Vec2, XLSR-Wav2Vec2, and Whisper, alongside traditional approaches such as Random Forest and CNN-LSTM. To overcome data limitations, we applied targeted data augmentation strategies and analyzed their impact on model performance. Our findings highlight the superior performance of CNN-Spectrogram and CNN-LSTM models while identifying the strengths and limitations of data augmentation. This work establishes foundational benchmarks and provides actionable insights for advancing dialectal NLP in Luxembourgish and other low-resource languages.

1 Introduction

Dialectal research plays a critical role in understanding linguistic diversity and cultural identity. Luxembourgish, a West Germanic language spoken by over 600,000 people, presents unique challenges due to its regional phonetic, prosodic, and lexical variations. Limited annotated resources and influences from German and French complicate automated dialect classification (Hovy, 2015; Adda-Decker et al., 2014).

Luxembourgish dialects are categorized into four regions: North, East, South, and Center. Each region exhibits distinct linguistic traits, with the northern dialect displaying the most divergence and the central region aligning closely with the standard variety (Gilles, 2023).

Automatic dialect classification has practical importance in improving automatic speech recognition (ASR) and machine translation systems and in enabling more inclusive digital archiving of dialectal data. Previous work has underscored the importance of dialect identification in preserving linguistic diversity and supporting sociolinguistic research (Kantharuban et al., 2023). However, Luxembourgish, like other low-resource languages, lacks substantial annotated datasets for automated processing, which hinders the development of robust models for dialect classification. Moreover, Luxembourgish's multilingual setting presents additional challenges, as shown by existing research in Luxembourgish ASR and related linguistic tasks (Gilles et al., 2023; Nguyen et al., 2023; Song et al., 2023).

1.1 Linguistic Variability Across Luxembourgish Dialects

Luxembourgish dialects display considerable variation in lexical, phonetic, and prosodic structures influenced by geographic factors (Gilles, 1998). To illustrate, we present a sample sentence rendered in the four main dialects—North, East, South, and Center—along with phonetic transcriptions. This example highlights the differences in pronunciation and vocabulary that complicate automated dialect classification due to regional speech patterns.

These examples underscore the challenges in distinguishing Luxembourgish dialects due to lexical differences (e.g., “Fregdig” vs. “Freiden”) and phonetic variations (e.g., vowel lengthening and consonant shifts). Automatic dialect classification models must account for these subtleties to handle distinct regional forms accurately.

In this study, we address these challenges by

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 143–152
March 2, 2025 ©2025 Association for Computational Linguistics

Region	Dialectal Sentence	Phonetic Transcription
North	Eng Frau hott e Fregdig di schwarz Kléider gebikst.	[æŋ fʀɑʊ hot ə fʀægdiç diː ʃwɑʀts klɜidɐ gəbɪkst]
East	En Fra hott e Freddig di schwarz Klääder gebéit.	[eːn fʀaː hot ə fʀædɪç diː ʃwɑts klɛːdɐ gəbɜit]
South	Eng Fra huet e Freiden di schwoarz Kleeder gebitzt.	[æŋ fʀaː huet ə fʀɑɪdən diː ʃwɔːɐts kleːdə gəbitst]
Center	Eng Fra huet e Freideg déi schwaarz Kleeder gebitzt.	[æŋ fʀaː huet ə fʀɑɪdeç dɜi ʃwaːʀts kleːdə gəbitst]

Table 1: Example Sentence in Luxembourgish Dialects with Phonetic Transcriptions

employing data augmentation techniques to increase sample diversity and improve model robustness, particularly for underrepresented dialects. Our methodology explores phonetic, prosodic, and lexical features across various classifiers, including both traditional machine learning algorithms and neural network models.

This paper contributes to computational linguistics by:

1. Introducing one of the first comprehensive studies on Luxembourgish dialect classification, investigating the impact of data augmentation on model performance in a low-resource setting.

2. Establishing performance benchmarks across multiple model architectures, including Random Forest, CNN-Spectrogram, CNN-LSTM, Wav2Vec2, Whisper, and XLSR-Wav2Vec2 to create a foundation for future research in Luxembourgish and other low-resource languages.

These contributions establish Luxembourgish as a compelling case study in low-resource language processing and illustrate the broader applications of dialectal NLP research. Our results underscore the importance of linguistic equity and highlight directions for future research in multilingual and dialectal NLP.

2 Related Work

Automatic dialect classification has advanced significantly in high-resource languages, where annotated datasets and sophisticated processing tools facilitate robust model performance. For instance, substantial work has been conducted in Arabic (Harfash and Abdul-kareem, 2017), Chinese (Ng and Lee, 2008), German (Dobbriner and Jokisch, 2019), and English (Etman and Louis, 2015). In these languages, the availability of extensive data resources enables classification approaches to take advantage of phonetic, prosodic, and lexical features, supporting higher accuracy and model robustness. For example, Harfash and Abdul-kareem (2017) improved dialect classification in Arabic by incorporating phonetic and prosodic cues, while Ng and Lee (2008) applied entropy-based measures to enhance Chinese dialect classification, highlighting the versatility of feature-based methods in these contexts. In high-resource settings, models often use a combination of rule-based linguistic knowledge (Biadsy and Hirschberg, 2009) and data-driven machine learning techniques that benefit from large training corpora, allowing them to learn complex patterns effectively (Ali et al., 2016).

In contrast, low-resource languages like Luxembourgish lack the annotated datasets and processing infrastructure needed for accurate dialect classification, presenting unique challenges for computational linguistics. For low-resource languages, researchers have explored strategies such as synthetic data generation and unsupervised learning to mitigate data scarcity. Transfer learning, for example, can leverage pre-trained models in related languages, using phonetic similarities to improve dialect classification in under-resourced contexts (Shah et al., 2023; Khosravani et al., 2021). Data augmentation has also emerged as a critical strategy for low-resource languages, allowing researchers to expand datasets and introduce variability, as demonstrated in tasks involving accent and dialect variation (Ullah et al., 2023; Xu et al., 2021).

For Luxembourgish, however, computational research remains relatively limited. Existing studies have focused mainly on its phonetic and syntactic characteristics (Gilles and Trouvain, 2013), as well as distinctive phonological features (Gilles, 2014), with limited exploration of automated dialect classification. Research on regional phonetic variation in Luxembourgish indicates that its dialects are influenced by neighboring German and French, with generational shifts contributing further to its linguistic diversity (Conrad, 2023). This complexity requires tailor-made classifiers and careful feature engineering to capture subtle distinctions in phonetics and prosody that are integral to Luxembourgish dialectal variation (Snoeren et al., 2011). Computational studies have suggested that cross-lingual models that utilize resources from German and French could improve Luxembourgish speech recognition (Nguyen et al., 2023), highlighting both the potential and the computational challenges that the classification of the Luxembourgish dialect entails (Adda-Decker et al., 2014).

Future progress in Luxembourgish dialect classification may benefit from techniques like data augmentation, which has proven successful in other low-resource contexts. For instance, Xu et al. (2021) demonstrated that targeted data augmentation techniques, such as pitch and speed modifications, significantly improved the accuracy of dialect classification for Chinese dialects, underscoring the value of these methods to improve model performance in low-resource settings. Such approaches could potentially be adapted for Luxembourgish, where similar variability in phonetic and prosodic features across dialects could benefit from targeted augmentation.

Building on this foundation, our study introduces a model for the classification of Luxembourgish dialects that integrates linguistic insights with computational techniques specifically designed for low-resource settings. By applying data augmentation strategies, we address the constraints imposed by limited annotated data, contributing to the broader field of dialect classification for underrepresented languages. This work aims to lay the groundwork for Luxembourgish NLP, underscoring the importance of dialectal research in multilingual NLP and advancing methodologies for low-resource language processing.

3 Methodology

3.1 Dataset and Preprocessing

The dataset used in this study was crowd-sourced through a smartphone application developed as part of a prior project [redacted]. Participants were asked to translate sentences spontaneously from German or French into their Luxembourgish dialect.

Attribute	Category	Count
Total Audio Files		1720
	Unique Entries	1720
Gender	Female	1210
	Male	510
Age Group	25–34	567
	35–44	377
	45–54	352
	55–64	277
	65+	132
Dialect Region	Center	762
	South	482
	East	293
	North	168

Table 2: Demographic Distribution of the Luxembourgish Dialect Dataset.

The dataset (Table 2) includes 1720 unique audio samples annotated with gender, age group, and dialect region. The samples reflect Luxembourgish's four main dialect regions: Center, South, East, and North, with the Center being the most represented. To evaluate whether age groups were evenly distributed across dialect regions, we conducted a chi-square test. The results indicated that age distribution did not differ significantly by dialect region ($\chi^2(6) = 5.73$, $p = 0.45$), suggesting the four regions are relatively balanced with respect to participants' ages.

For feature extraction, Mel-Frequency Cepstral Coefficients (MFCCs) were computed using the *torchaudio* and *librosa* libraries, capturing phonetic features essential for dialectal differentiation. Additionally, the mean and standard deviation of each waveform were calculated to provide statistical descriptors of each audio signal. Together, these features allow the model to learn from both phonetic characteristics and statistical patterns across dialects, supporting accurate dialect classification.

3.2 Model Architecture and Training

In this study, we explore multiple approaches to dialect classification, leveraging both traditional machine learning techniques and advanced deep learning models. Our methodology includes six key approaches, each with unique strengths in handling different aspects of speech data. All classifiers were implemented in Python 3.9. For the Random Forest classifier, we used *scikit-learn* to handle training and evaluation, and *Optuna* for hyperparameter tuning. For the DL models (CNN-Spectrogram, CNN-LSTM, Wav2Vec2, XLSR-Wav2Vec2, and Whisper), we used *PyTorch* along with the *torchaudio* library for audio processing; hyperparameter tuning was also managed via *Optuna*. This integrated setup allowed us to maintain a consistent development pipeline across both traditional and DL methods.

1. Random Forest with AutoML Tuning: We use Random Forest as a baseline classifier and employ AutoML (Optuna (Akiba et al., 2019)) for hyperparameter optimization. Random Forest is a robust ensemble model noted for its interpretability and effectiveness in handling tabular, low-dimensional features. AutoML tuning identifies optimal configurations, establishing a strong benchmark for comparison with deeper architectures (Ramadhan et al., 2017).

2. Wav2Vec Model: Wav2Vec 2.0 is a pre-trained model for speech representation learning, capturing nuanced phonetic and acoustic features. By fine-tuning Wav2Vec2 on our dialectal data, we leverage its ability to detect subtle variations in pronunciation, tone, and rhythm—key elements in dialect classification. Its extensive pre-training makes it highly effective, even with limited labeled data (Das et al., 2023).

3. Whisper Model: Whisper (Radford et al., 2023) is a sequence-to-sequence model designed for automatic speech recognition (ASR) and robust transcription across various languages. In our approach, we leverage Whisper for dialect classification by fine-tuning it on Luxembourgish dialect data. Specifically, we modify its final classification layer to predict dialect labels rather than transcriptions. We extract Whisper's intermediate acoustic embeddings from its final transformer layers and pass them through a fully connected classifier, which outputs softmax probabilities over the dialect classes. This method enables Whisper to capture subtle phonetic and prosodic differences among Luxembourgish dialects while benefiting from its inherent robustness to noise and diverse acoustic conditions. Compared to other models such as Wav2Vec2 and CNN-based approaches, Whisper's sequence-to-sequence architecture allows it to use broader context across speech segments, making it particularly effective in capturing dialectal shifts that span longer temporal patterns.

4. XLSR-Wav2Vec2 Model: The Cross-Lingual Speech Representation (XLSR) variant of Wav2Vec2 extends the model's capabilities to multiple languages by learning universal speech representations. Fine-tuning XLSR-Wav2Vec2 (Conneau et al., 2021) on our dialectal data leverages these cross-lingual features, facilitating more accurate detection of subtle acoustic patterns that may overlap across dialects or language families. This approach is especially useful when the available labeled data for each dialect is limited.

5. CNN on Spectrograms: We apply Convolutional Neural Networks (CNNs) to Mel spectrograms, treating them as 2D images. CNNs excel in identifying spatial patterns—such as phonetic markers, intonation shifts, and accent variations—by leveraging their proven effectiveness in image processing. This approach highlights visual representations of acoustic features for clearer insight into dialect differences (Alrehaili et al., 2023).

6. CNN-LSTM Hybrid Model: To capture both spatial and temporal patterns, we integrate CNN and Long Short-Term Memory (LSTM) layers. The CNN layers learn spatial features from each spectrogram frame, while the LSTM layers model temporal dependencies such as rhythm and sequential patterns across frames. This combined architecture offers a more holistic understanding of dialectal characteristics (China et al., 2018).

Through these six approaches, we explore how different models capture dialectal differences in speech, analyzing which features—ranging from the phonetic details learned by Wav2Vec2, XLSR-Wav2Vec2, and Whisper to the spatial and temporal patterns identified by CNN-Spectrogram and CNN-LSTM—are most effective for dialect classification.

The CNN model for dialect classification was designed to process spectrogram data as a 2D image-like input, beginning with a 2D convolutional layer with 32 filters (kernel size of 3 × 3), followed by additional convolutional and max-pooling layers to capture spatial features from the spectrograms. For the CNN-LSTM model, this convolutional stack was followed by an LSTM layer to capture temporal dependencies across spectrogram frames. Both models used padding to ensure consistent input dimensions. The architecture was optimized using categorical cross-entropy loss and an Adam optimizer with a learning rate of 0.001. Each model was trained over 15 epochs with five cross-validation folds to evaluate robustness. To handle class imbalance, we incorporated a weighted sampler in the DataLoader, using class weights calculated per fold to emphasize learning on underrepresented dialect classes, improving model generalizability across dialects.

3.3 Data Augmentation

To address data imbalance within the Luxembourgish dialect dataset, we implemented data augmentation techniques using controlled variations in speed and pitch to enhance sample diversity and model robustness. Specifically, we targeted underrepresented dialect classes (Northern and Eastern) to generate additional samples. In total, we created 820 new audio samples, increasing the dataset size from 1720 to 2540 recordings.

We applied time stretching with a 1.2x speed factor to generate faster-paced versions of each sample, creating tempo variations that reflect natural speaking speed differences without altering phonetic content. Pitch shifting was also used to create tonal variations by adjusting playback at a 50ms chunk level with crossfade transitions. This replicates natural differences in vocal tone, helping to distinguish differences between dialects and individual speakers.

We implemented these augmentations using the *pydub* library (Robertson, 2010), which enabled systematic file augmentation while preserving originals. Augmented files were prioritized for dialects below the median frequency (i.e., Northern and Eastern), addressing class imbalance effectively. Furthermore, to maintain demographic consistency, we mirrored the gender and age distributions for each new sample, ensuring that both male and female speakers across various age ranges were also augmented when needed. The final dataset became more balanced, reducing the disparity between the best- and worst-represented dialects from 594 recordings to 147 recordings difference. Parallel processing was employed to manage the computational load, ensuring efficient augmentation of underrepresented dialects.

After augmentation, the dataset included 2540 audio clips, with each dialect represented by at least 500 samples. The mean clip length was 3.2 seconds (SD = 0.8), with a similar distribution of lengths across dialects, genders, and age groups. On average, each audio sample contained approximately 6.3 tokens of spoken text (SD = 1.1), with a total vocabulary of 1,550 unique Luxembourgish tokens (up from 1,100 prior to augmentation). This increase in unique tokens reflects the added lexical variability introduced by augmentation and ensures that minority dialects are not underrepresented in the linguistic feature space.

Baseline (Without Augmentation)

Model	**Northern**	**Central**	**Southern**	**Eastern**
Random Forest	63/61/62	58/60/60	56/57/57	55/55/55
Wav2Vec2	**70/72/72**	69/70/70	70/71/71	69/69/70
Whisper	67/69/68	66/67/66	68/69/69	64/65/65
XLSR-Wav2Vec2	68/70/69	66/68/67	69/70/69	63/64/64
CNN-Spectrogram	72/71/73	71/71/71	**72/74/73**	**70/69/71**
CNN-LSTM	72/70/72	**73/72/71**	69/72/70	68/71/72

Optimized (With Augmentation)

Model	**Northern**	**Central**	**Southern**	**Eastern**
Random Forest	71/69/71	65/63/65	63/61/63	59/58/59
Wav2Vec2	**75/74/75**	72/71/72	73/72/73	70/71/71
Whisper	72/72/73	70/70/70	72/72/72	67/69/68
XLSR-Wav2Vec2	72/73/72	69/70/70	71/72/71	66/66/66
CNN-Spectrogram	76/74/76	74/73/74	**79/76/78**	**78/75/76**
CNN-LSTM	76/73/74	**75/74/73**	77/75/77	72/70/71

Table 3: Performance Comparison Between Baseline (Without Augmentation) and Optimized (With Augmentation) Results for Luxembourgish Dialect Classification. Each cell shows Accuracy/Precision/Recall (%). **Bold** indicates the highest performance metric.

3.4 Evaluation and Metrics

To evaluate model performance in dialect classification, we used four key metrics: accuracy (overall correctness), precision (minimizing false positives), and recall (capturing true instances) to evaluate each model. Each table reports per-class accuracy, precision, and recall, giving insight into how models handle each dialect.

We applied stratified sampling during training to ensure balanced dialect representation in the dataset, helping to address class imbalance and maintain model performance across all dialects. Early stopping was implemented to halt training when the validation loss did not improve over five consecutive epochs, thereby preventing overfitting. A batch size of 16 was chosen to balance computational efficiency and convergence speed, while the Adam optimizer was used to adjust the learning rate adaptively, ensuring stable and effective convergence during training.

4 Results

Table 3 presents a comparison of model performance on Luxembourgish dialect classification under two conditions: baseline (without data augmentation) and optimized (with data augmentation). Six primary models were evaluated: Random Forest, Wav2Vec2, Whisper, XLSR-Wav2Vec2, CNN-Spectrogram, and CNN-LSTM. Performance, evaluated through accuracy, precision, and recall metrics, was measured across Northern, Central, Southern, and Eastern dialects for each model.

4.1 Baseline Performance (Without Augmentation)

In the *baseline* setting (see Table 3), all models exhibit moderate accuracy (55%–73%), reflecting the challenges posed by a relatively small and imbalanced dataset:

CNN-Spectrogram attains the highest accuracy in the Northern (72%) and Southern (72%) dialects, underscoring CNNs' effectiveness in extracting spatial patterns (e.g., phonetic cues) from spectrograms. CNN-LSTM excels in classifying the Central dialect (73% accuracy), possibly due to its capacity to capture temporal dependencies along with spatial cues. Wav2Vec2 also performs strongly, particularly for Northern and Southern dialects (70% accuracy), benefiting from its robust self-supervised speech representations. Random Forest consistently lags behind the neural models, particularly for the Southern and Eastern dialects, reflecting its limited ability to model complex acoustic cues. Whisper and XLSR-Wav2Vec2 provide competitive results but do not surpass the CNN-based or standard Wav2Vec2 models in most dialects. Eastern dialect classification remains the most challenging for all approaches. This is consistent with its underrepresentation in the dataset and with prior observations that Eastern exhibits phonetic overlaps with adjacent dialects, compounding classification difficulties.

4.2 Optimized Performance (With Augmentation)

Applying speed and pitch augmentation yields performance gains across all models, particularly for underrepresented Northern and Eastern dialects (see Table 3):

Random Forest sees an overall accuracy increase of 4–5%, indicating that extra variability in the training set helps even simpler classifiers. Wav2Vec2 improves to 75% accuracy for Northern and 70% for Eastern, confirming that its self-supervised features benefit from augmented data. Whisper and XLSR-Wav2Vec2 also enjoy small but consistent boosts across all dialects, reinforcing the notion that multilingual or sequence-to-sequence approaches capitalize on the broader acoustic variability introduced by augmentation. CNN-Spectrogram emerges as the top performer in most dialects post-augmentation: 76% accuracy in Northern, 79% in Southern, and 78% in Eastern, highlighting CNNs' capacity to adapt to new spectrogram variations (e.g., pitch-shifted or speed-stretched speech). CNN-LSTM remains highly competitive, matching CNN-Spectrogram in Northern dialect classification (76%) and excelling in the Central dialect (75% accuracy). Its ability to capture both spatial and temporal cues remains beneficial. These findings confirm that data augmentation helps mitigate class imbalance, particularly for Northern and Eastern dialects, which see some of the largest proportional gains. However, the overall improvements—while meaningful—remain limited by the modest size of the dataset. Gathering more recordings and exploring advanced or multi-parameter augmentation techniques (e.g., multiple speed factors, SpecAugment) could further boost performance.

5 Discussion

The results demonstrate that data augmentation can contribute to modest but consistent improvements in dialect classification for Luxembourgish, a low-resource language. These findings align with prior studies highlighting the effectiveness of CNNs and end-to-end ASR models, such as Wav2Vec, in handling spectrogram data for dialect and language classification tasks.

CNNs have proven effective in extracting meaningful features from spectrograms, which are crucial for distinguishing subtle phonetic and prosodic differences across dialects. For example, Alrehaili et al. (2023) reported that CNNs achieved 83% accuracy in Arabic dialect classification, capitalizing on their capacity to process spatial information within spectrograms. Similarly, Revay and Teschke (2019) demonstrated CNNs' suitability for language identification across multiple languages, achieving up to 89% accuracy by focusing on acoustic cues encoded in spectrograms (Revay and Teschke, 2019). Prior studies support our findings, showing that CNN-based models, such as CNN-Spectrogram and CNN-LSTM, achieved competitive accuracy (68-73%) on Luxembourgish dialects, with further improvements post-augmentation.

Research supports the effectiveness of CNN-LSTM architectures in dialect classification, especially for capturing both spatial and temporal linguistic patterns. For instance, CNN-LSTM models have shown high accuracy in dialect classification tasks for Arabic, where they effectively captured dialectal sentiment variations across regional Arabic texts (Abu Kwaik et al., 2019). Similar success has been observed in distinguishing tonal versus non-tonal Indian languages using acoustic data, where the model's ability to capture temporal dependencies significantly improved classification outcomes (China et al., 2018). Studies on dialectal sentiment analysis for Roman Urdu and English have further highlighted CNN-LSTM's adaptability, demonstrating the model's capacity to capture linguistic nuances in dialects within social media contexts (Khan et al., 2022). In general, CNN-LSTM hybrids improve dialect classification accuracy by effectively capturing both localized phonetic features and sequential temporal dynamics (She and Zhang, 2018).

Additionally, research on self-supervised models like Wav2Vec has demonstrated the model's ability to capture detailed phonetic and acoustic features, enabling it to perform well even in low-resource dialect classification tasks. Wav2Vec embeddings have proven effective in detecting dialect-specific nuances and handling out-of-distribution dialect data (Das et al., 2023). Studies also show that fine-tuning Wav2Vec on dialectal datasets enables it to capture phonetic variations in pronunciation, tone, and rhythm, essential for effective dialect classification (Baevski et al., 2019). Furthermore, Wav2Vec has demonstrated robustness across low-resource languages, achieving notable improvements in speech recognition for underrepresented dialects (Yi et al., 2020).

These findings align with our results, where the CNN-LSTM model showed consistent performance gains after augmentation, underscoring the utility of combining convolutional and sequential layers to handle the complex linguistic structures present in Luxembourgish dialects. Our findings also resonate with prior work in low-resource dialect classification. For instance, in the study by Wang et al. (2021), a multilingual ASR model improved classification accuracy for Chinese dialects, significantly reducing classification errors. Although we did not directly utilize this approach, Wav2Vec's pretraining on multilingual datasets may have contributed to its relative robustness in Luxembourgish dialect classification. This ability to handle a range of dialectal inputs, even with limited training data, illustrates the model's value in low-resource language contexts.

The improvements observed with data augmentation, while modest, highlight its potential to enhance model robustness, particularly for dialects with lower representation. Kethireddy et al. (2020) explored similar strategies by introducing augmented spectrogram features, leading to gains in dialect classification accuracy. In our study, augmenting the dataset by adjusting pitch and tempo introduced additional variability, helping the models to generalize better. This approach was especially beneficial for the Random Forest model, which lacks the feature extraction capabilities of CNNs and ASR models. Despite the limited scale of the improvements, these findings underscore the utility of data augmentation as a practical approach to mitigate the effects of data scarcity in dialect classification tasks.

Our CNN-LSTM model, designed to capture both spatial and temporal dependencies, also showed consistent gains with augmentation. Chemudupati et al. (2023) demonstrated that Wav2Vec could maintain robust performance across diverse conditions, including real-world "in-the-wild" settings with noisy and reverberant audio. Although the Luxembourgish dataset does not include such variability, the slight improvements in recall and precision seen in our CNN-LSTM model after augmentation suggest that temporal architectures may add value in dialectal classification tasks, especially in capturing sequential acoustic features.

Overall, the updated performance metrics reported in Table 3 confirm that CNN-Spectrogram achieved top accuracy in Southern (79%) and Eastern (78%) dialects following augmentation, while CNN-LSTM matched or surpassed other approaches in Central (75%) and Northern (76%). Wav2Vec2 also registered stable improvements (e.g., 70% accuracy for Eastern) after incorporating time-stretch and pitch-shift strategies. Notably, the Random Forest benefited substantially from augmentation, gaining about 4–5% in accuracy—particularly in Northern and Eastern dialects—underscoring the value of enriched data variability even for non-neural classifiers.

5.1 Limitations of the work

One key limitation is the lack of sufficiently diverse data, which poses a risk of overfitting and makes it difficult to capture subtle phonetic or lexical nuances in border regions. Additionally, our augmentation experiments are limited to a single set of parameters, leaving open the possibility that other augmentation methods or intensities might yield higher improvements. Finally, while Whisper and XLSR-Wav2Vec2 adapt well to multilingual contexts, further tuning (e.g., multiple epochs, domain adaptation) could potentially boost their performance.

6 Conclusions

We introduced a comprehensive methodology for Luxembourgish dialect classification, pairing data augmentation (speed/pitch shifts) with a spectrum of models from *Random Forest* to *CNN-LSTM* and pretrained *Whisper* / *Wav2Vec2* variants. Our results highlight:

CNN-Spectrogram achieves top accuracies in Northern, Southern, and Eastern dialects after augmentation, showcasing its spatial feature-extraction strengths. CNN-LSTM outperforms other models in Central Luxembourgish, suggesting the value of modeling temporal dependencies in dialect classification. Wav2Vec2 remains consistently strong across all dialects, affirming the resilience of self-supervised speech representations. Data augmentation partially mitigates imbalance, boosting performance the most in underrepresented dialects (Northern and Eastern). Though the improvements are modest, they demonstrate the potential of augmentation in low-resource dialect classification. Future work should explore more advanced augmentation pipelines (e.g., SpecAugment, multiple pitch/speed factors) and target larger-scale data collection, possibly leveraging multilingual transfer from related Germanic varieties. These steps will be instrumental in achieving broader robustness and higher accuracy for Luxembourgish and other low-resource dialects.

References

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2019. Lstm-cnn deep learning model for sentiment analysis of dialectal arabic. In *Arabic Language Processing: From Theory to Practice: 7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings 7*, pages 108–121. Springer.

Martine Adda-Decker, Lori Lamel, Gilles Adda, and Thomas Lavergne. 2014. A first lvcsr system for luxembourgish, a low-resourced european language. In *Human Language Technology Challenges for Computer Science and Linguistics: 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25–27, 2011, Revised Selected Papers 5*, pages 479–490. Springer.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Ahmed Ali, Najim Dehak, Pierre Cardinal, Sameer Khurana, Sree Harsha Yella, Jim Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in arabic broadcast speech. In *Proceedings of Interspeech*, pages 2934–2938, San Francisco, CA, USA.

Meaad Alrehaili, Tahani Alasmari, and Areej Aoalshutayri. 2023. Arabic speech dialect classification using deep learning. In *2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC)*, pages 1–5. IEEE.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Fadi Biadsy and Julia Hirschberg. 2009. Using prosody and phonotactics in arabic dialect identification. In *Interspeech*, volume 9, pages 208–211.

Vamsikrishna Chemudupati, Marzieh Tahaei, Heitor Guimaraes, Arthur Pimentel, Anderson Avila, Mehdi Rezagholizadeh, Boxing Chen, and Tiago Falk. 2023. On the transferability of whisper-based representations for" in-the-wild" cross-task downstream speech applications. *arXiv e-prints*, pages arXiv–2305.

Chuya China, Dipjyoti Bisharad, and Rabul Hussain Laskar. 2018. Automatic classification of indian languages into tonal and non-tonal categories using cascade convolutional neural network (cnn)-long short-term memory (lstm) recurrent neural networks. In *2018 International Conference on Signal Processing and Communications (SPCOM)*, pages 492–496. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. *Interspeech 2021*.

François Conrad. 2023. Regional differences in the evolution of the merger of /ʃ/ and ç in luxembourgish. *Journal of the International Phonetic Association*, 53(1):29–46.

Sourya Dipta Das, Yash Vadi, Abhishek Unnam, and Kuldeep Yadav. 2023. Unsupervised out-of-distribution dialect detection with mahalanobis distance. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)*, Dublin, Ireland. International Speech Communication Association.

Johanna Dobbriner and Oliver Jokisch. 2019. Towards a dialect classification in german speech samples. In *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21*, pages 64–74. Springer.

A Etman and AA Louis. 2015. American dialect identification using phonotactic and prosodic features. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 963–970. IEEE.

Peter Gilles. 1998. Virtual convergence and dialect levelling in luxembourgish. *Folia lingüística: Acta Societatis Linguisticae Europaeae*, 32(1):69–82.

Peter Gilles. 2014. Phonological domains in luxembourgish and their relevance for the phonological system. *Syllable and word languages*, pages 279–304.

Peter Gilles. 2023. Regional variation, internal change and language contact in luxembourgish_ results from an app-based language survey. *Taal en Tongval*, 75(1).

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023. Asrlux: Automatic speech recognition for the low-resource language luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*, pages 3091–3095. Guarant International.

Peter Gilles and J. Trouvain. 2013. Illustrations of the ipa: Luxembourgish. *Journal of the International Phonetic Association*, 43.

Esra J Harfash and A Hassan Abdul-kareem. 2017. Automatic arabic dialect classification. *International Journal of Computer Application*, 8887.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 1: Long papers)*, pages 752–762.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. *arXiv preprint arXiv:2310.15135*.

Rashmi Kethireddy, Sudarsana Reddy Kadiri, Paavo Alku, and Suryakanth V Gangashetty. 2020. Mel-weighted single frequency filtering spectrogram for dialect identification. *IEEE Access*, 8:174871–174879.

Lal Khan, Ammar Amjad, Kanwar Muhammad Afaq, and Hsien-Tsung Chang. 2022. Deep sentiment analysis using cnn-lstm architecture of english and roman urdu text shared in social media. *Applied Sciences*, 12(5):2694.

Abbas Khosravani, Philip N Garner, and Alexandros Lazaridis. 2021. Learning to translate low-resourced swiss german dialectal speech into standard german text. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 817–823. IEEE.

Raymond WM Ng and Tan Lee. 2008. Entropy-based analysis of the prosodic features of chinese dialects. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4. IEEE.

Le Minh Nguyen, Shekhar Nayak, and Matt Coler. 2023. Improving luxembourgish speech recognition with cross-lingual speech representations. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 792–797.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Muhammad Murtadha Ramadhan, Imas Sukaesih Sitanggang, Fahrendi Rizky Nasution, and Abdullah Ghifari. 2017. Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. *DEStech transactions on computer science and engineering*, 10(2017).

Shauna Revay and Matthew Teschke. 2019. Multiclass language identification using deep learning on spectral images of audio signals. *arXiv preprint arXiv:1905.04348*.

Jiaaro Robertson. 2010. pydub: Audio processing library for python. Version 0.25.1, accessed October 31, 2024.

Riya Shah, Milin Patel, Barkha M Joshi, Jayna Shah, and Ronak Roy. 2023. Recognizing indian languages speech sound using transfer learning approach. In *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 853–859. IEEE.

Xiangyang She and Di Zhang. 2018. Text classification based on hybrid cnn-lstm hybrid model. In *2018 11th International symposium on computational intelligence and design (ISCID)*, volume 2, pages 185–189. IEEE.

Natalie D Snoeren, Martine Adda-Decker, and Gilles Adda. 2011. Pronunciation and writing variants in an under-resourced language: the case of luxembourgish mobile n-deletion. In *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznan, Poland, November 6-8, 2009, Revised Selected Papers 4*, pages 70–81. Springer.

Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. 2023. Letz translate: Low-resource machine translation for luxembourgish. In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 165–170. IEEE.

Asad Ullah, Alessandro Ragano, and Andrew Hines. 2023. Reduce, reuse, recycle: Is perturbed data better than other language augmentation for low resource self-supervised speech models. *arXiv e-prints*, pages arXiv–2309.

Ding Wang, Shuaishuai Ye, Xinhui Hu, Sheng Li, and Xinkang Xu. 2021. An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model. In *Interspeech*, pages 3266–3270.

Fan Xu, Yangjie Dan, Keyu Yan, Yong Ma, and Mingwen Wang. 2021. Low-resource language discrimination toward chinese dialects with transfer learning and data augmentation. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2):1–21.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv e-prints*, pages arXiv–2012.