Deep Learning Assignment

30/05/2025

Bilal Bilican; u951464 (SNR: 2082071)

1. Data Preprocessing

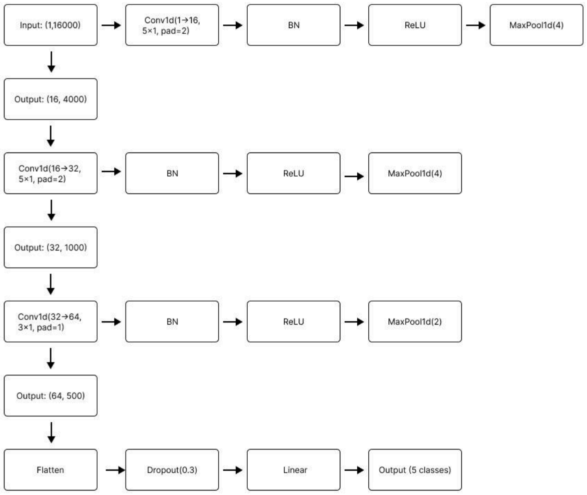
All audio files were resampled to 16,000 Hz for consistency. Labels were extracted from filenames and mapped to class indices. Two input pipelines were used:

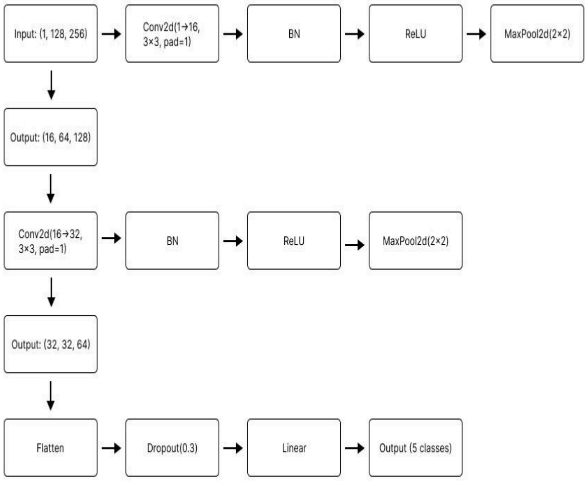
- RawCNN received raw waveforms directly, allowing the model to learn from the time-domain signal.
- MelCNN used log-mel spectrograms with normalization, providing a structured time-frequency representation. Normalization improved stability and training speed.

2. Experiments

We compared two models: RawCNN (1D convolutions on waveforms) and MelCNN (2D convolutions on spectrograms). Both used cross-entropy loss and the Adam optimizer. MelCNN showed better learning due to richer inputs and normalization. No explicit regularization (e.g., dropout) was used, but spectrogram normalization acted as a regularizing factor. Accuracy and F1 scores were tracked across epochs.

3. Architecture Visualization

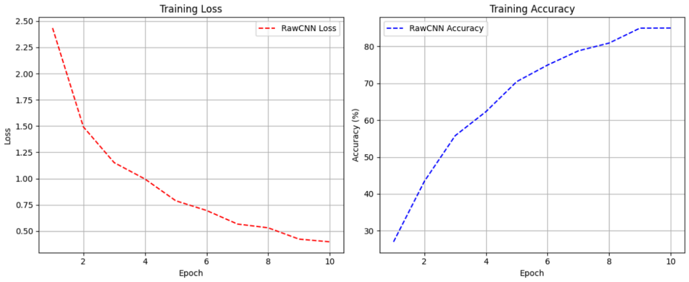


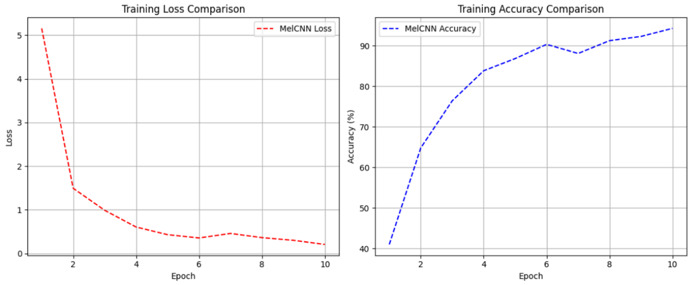


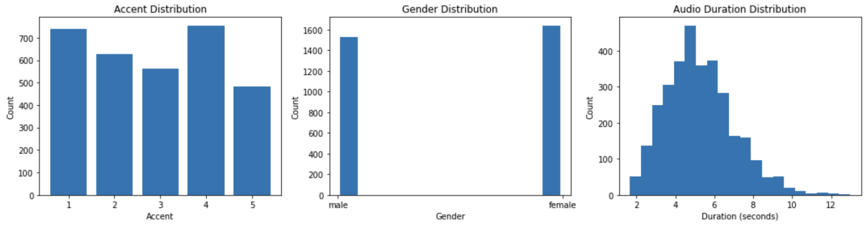
RawWaveformCNN Architecture: Takes a raw waveform of shape (1×16000) and passes it through three 1D convolutional layers with ReLU and batch normalization, interleaved with max pooling layers. The resulting tensor is flattened, regularized with dropout, and classified using a fully connected layer.

MelCNN Architecture: Operates on log-mel spectrogram inputs of shape (1×128×256). It applies two 2D convolutional blocks (Conv → BN → ReLU → MaxPool), reducing the size to (32×32×64). After flattening and dropout, a linear layer produces the final 5-class output.

4. Results







Model	**Accuracy (%)**	**Final loss**	**F1 score (across all accents)**
RawCNN	90.08%	0.397	0.902
MelCNN	95.64%	0.207	0.955

In terms of overall performance, MelCNN outperformed RawCNN both in accuracy and consistency across classes. The final training accuracy for MelCNN reached 94.28% with a loss of 0.207, while RawCNN achieved 85.00% accuracy and a loss of 0.397. A more detailed performance breakdown showed that both MelCNN and RawCNN scored low F1 scores on accent 4, but MelCNN had a very low recall rate with perfect precision while RawCNN had a lower precision rate and mediocre recall. Moreover, while RawCNN performed the worst on accent 1 this was the best F1-score for MelCNN with an F1-score of 0.9984 for accent 1. Data augmentation techniques weren't successful in boosting accuracy or recall for both RawCNN and MelCNN. Gender differences showed that female accents were harder to predict than men.

Error Analysis

RawCNN showed difficulty between accent classes with subtle acoustic differences, particularly misclassifying accent 1 in other accents, which might show some case of overfitting on that accent. In contrast, MelCNN demonstrated great separation between classes, but was misclassifying accent 4 as both accents 3 and 5. In the RawCNN accent 1 was the most common fallback prediction when the model was unsure, which is likely due to model bias.

5. Conclusions

- The MelCNN model clearly outperformed RawCNN in both accuracy and F1 score, highlighting the benefit of preprocessed spectral representations.
- Spectrogram normalization was a key factor in improving convergence and model stability.
- The RawCNN approach, while promising due to its end-to-end nature, proved more difficult to train effectively without extensive tuning or regularization.
- Error analysis showed that certain accent classes remain difficult to distinguish, suggesting that further work should address data imbalance. The data augmentation techniques weren't successful in boosting recall and accuracy, which calls for further investigation of this technique
- Future improvements should include more different validation and test splits, more diverse applications of regularization methods, and possibly hybrid models that combine raw and spectral inputs for robust feature learning.