Golem project

Speech Recognition

by Filip Kopyt and Julia Czosnek

Table of CONTENTS

REQUIREMENTS

- Train and test three different neural networks. Two networks should be implemented from scratch: one using 2D convolutions and the other using 1D convolutions.
- The third network should have a different architecture, which does not necessarily have to be custom-built; a pre-trained model can be used.
- Analyze the impact of various parameters on the training process and final results.
- After selecting the final models, examine their limitations and performance in different scenarios.

Audio data PROCESSING

Resampling

Audio files often have different sampling rates, which can lead to inconsistencies in model training. To standardize the data and reduce computational complexity, we resample all audio signals to 8 kHz, preserving essential speech information.

Voice Activity Detection (VAD)

Silence in audio recordings can introduce noise and unnecessary data for the model. We apply VAD to remove small part of silent segments, focusing on speech content and improving model efficiency.

Padding for Equal Length

Since audio samples vary in duration, a consistent input shape is necessary for batch processing. We pad shorter signals with zeros, ensuring all inputs have the same length without altering the original speech content.

Audio data PROCESSING

Mel spectogram Conversion

Raw waveforms are converted into a log spectrogram to provide a more informative time-frequency representation. This involves computing the Short-Time Fourier Transform (STFT) and applying a logarithmic scale to highlight key spectral patterns.

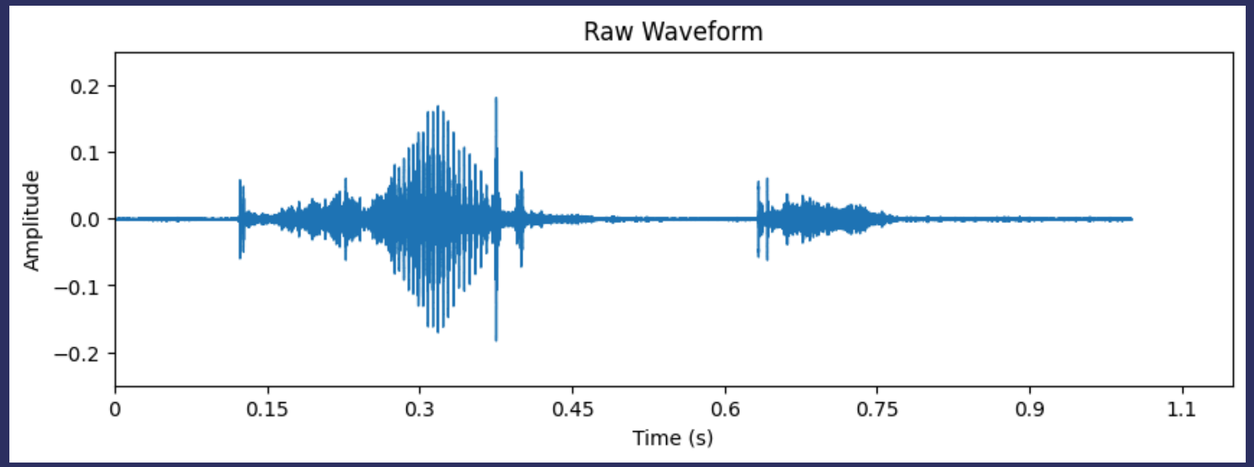
Feature Normalization

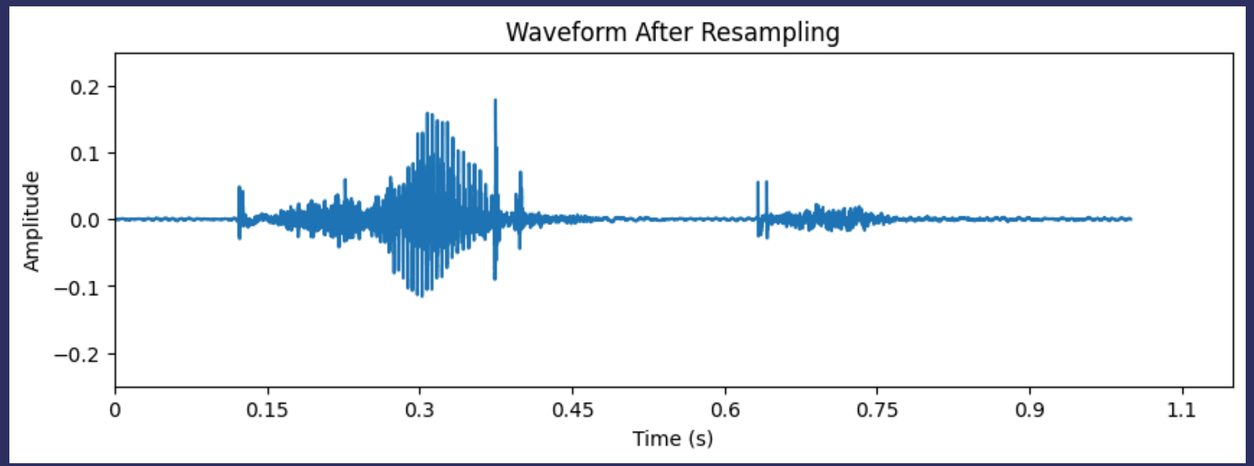
To ensure stable training and prevent large values from dominating, we normalize all spectrogram features. Using mean and standard deviation normalization, we scale the data to a uniform range, improving model performance.

Waveforms

00 Preprocessing

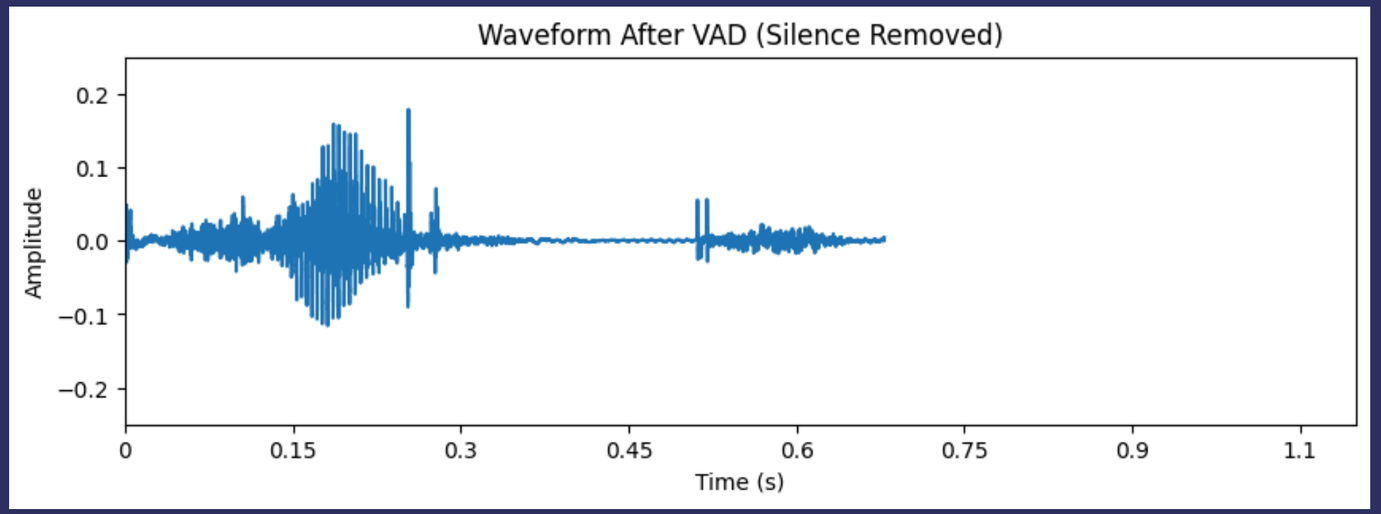
01 Resampling





Waveforms

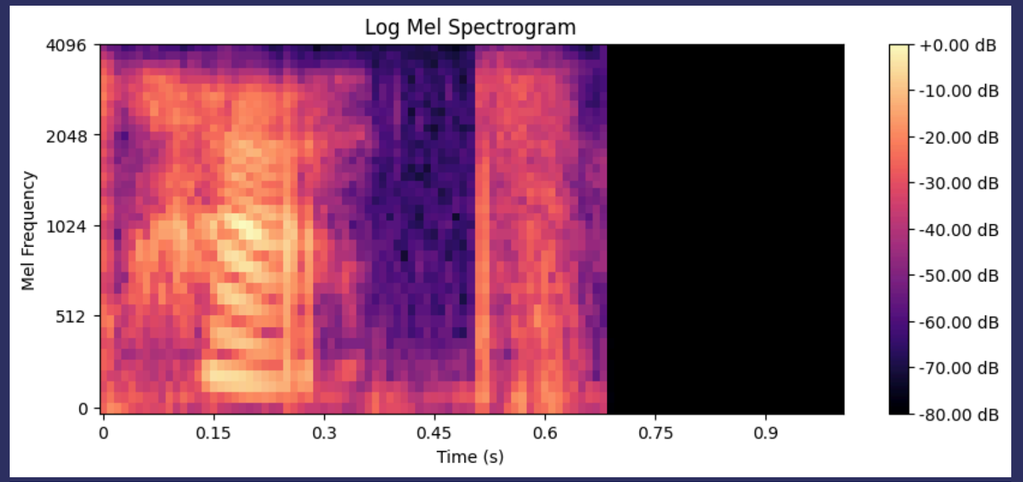
02 Voice Activity Detection

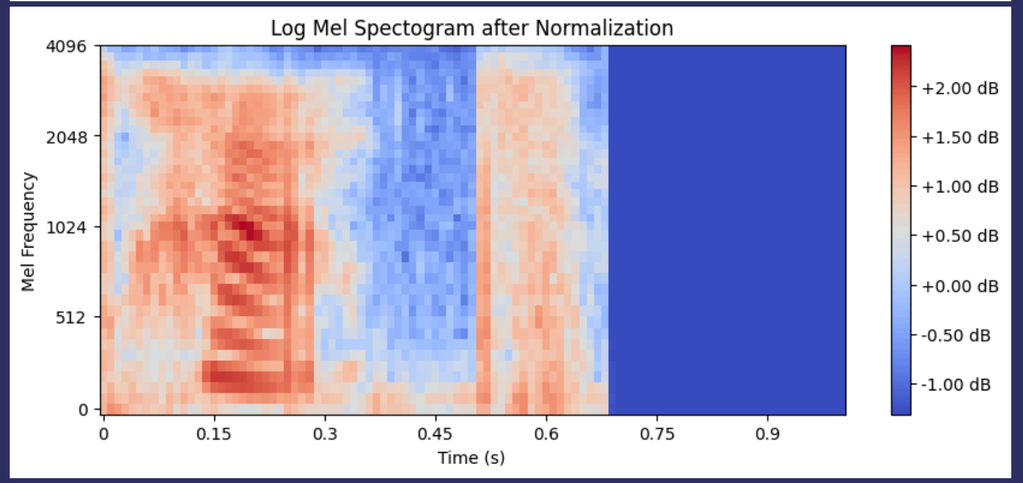


Spectograms

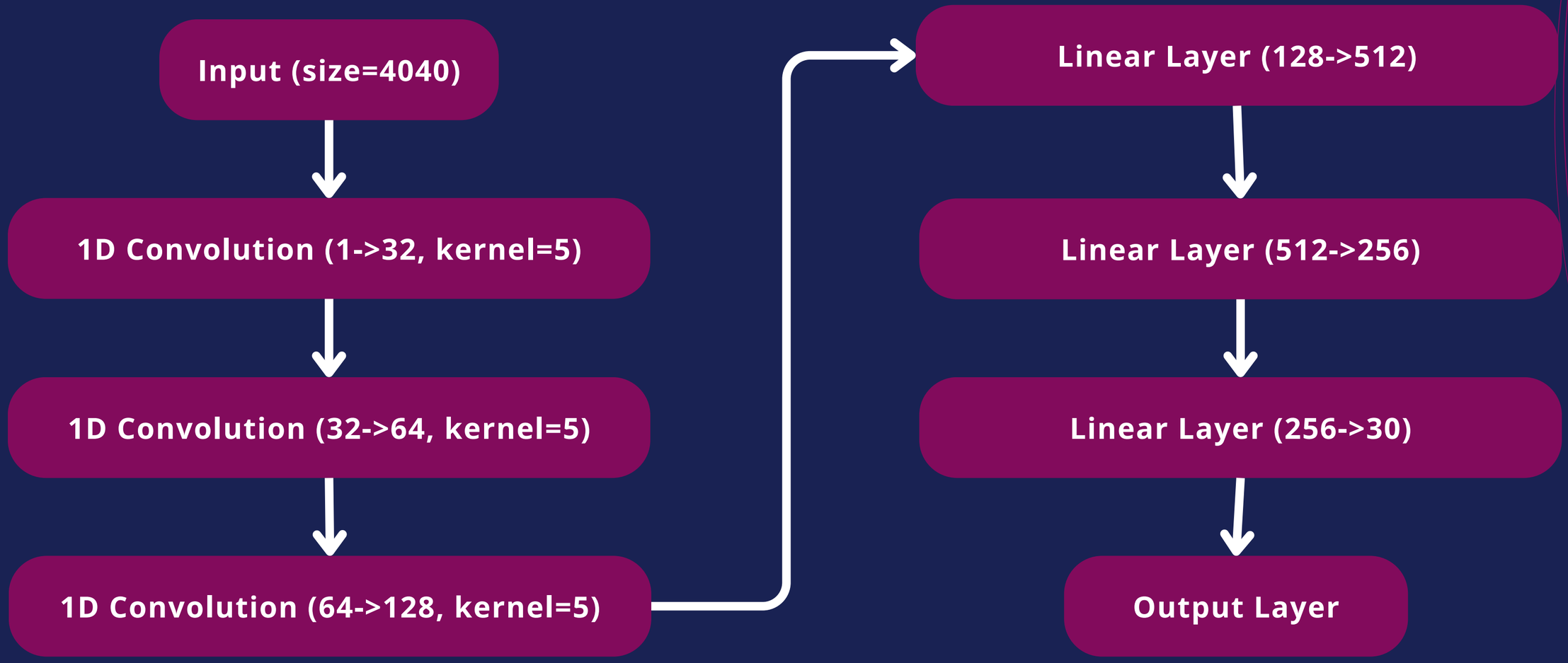
03 Mel spectogram

04 Feature Normalization

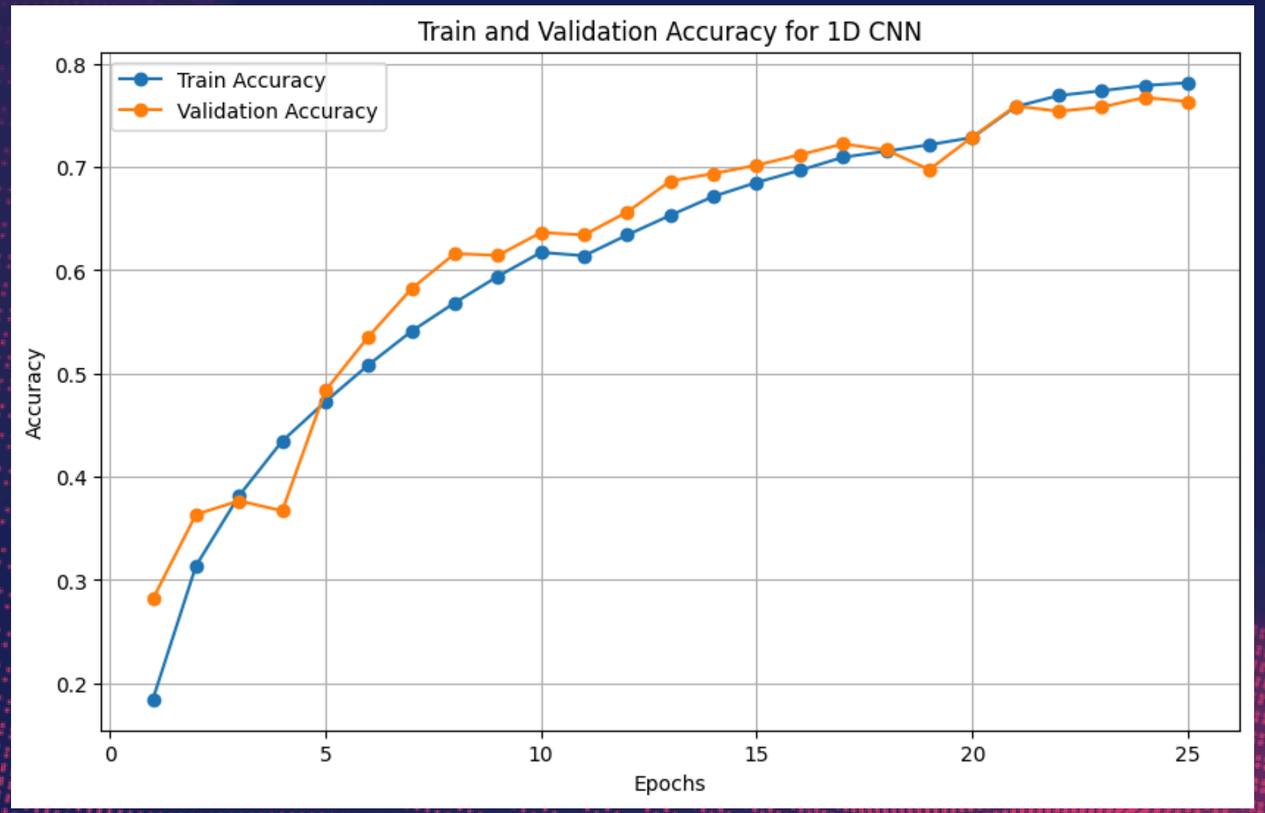




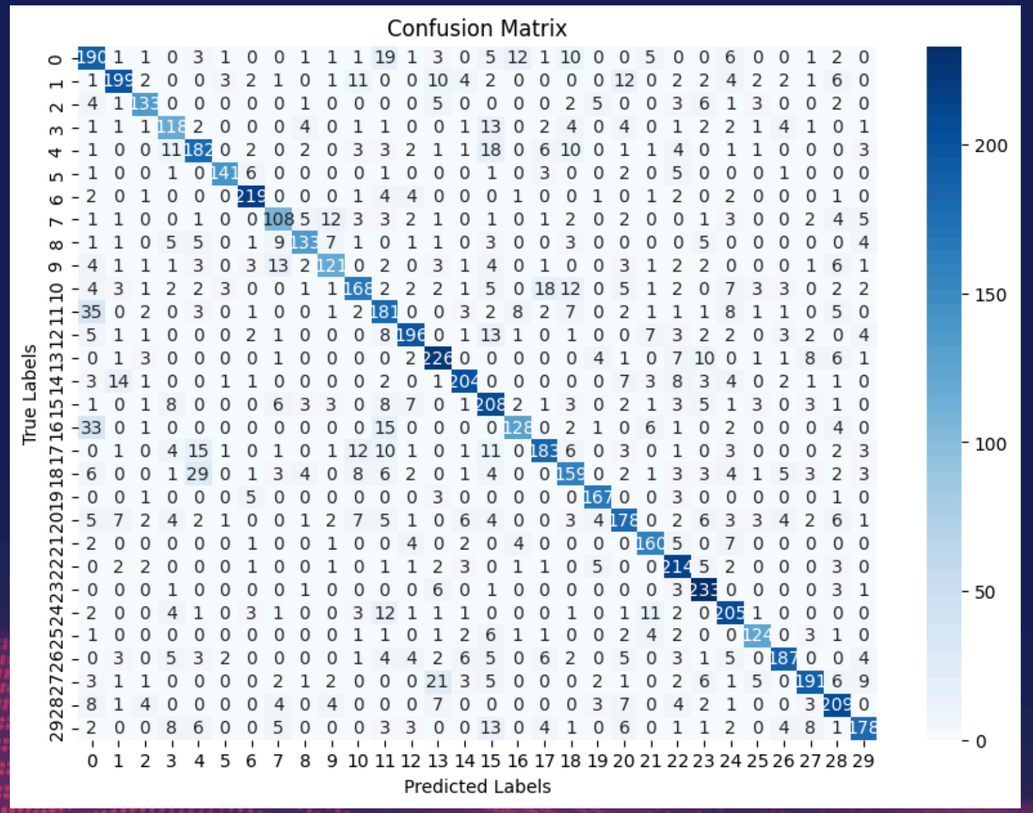
1D CNN Architecture



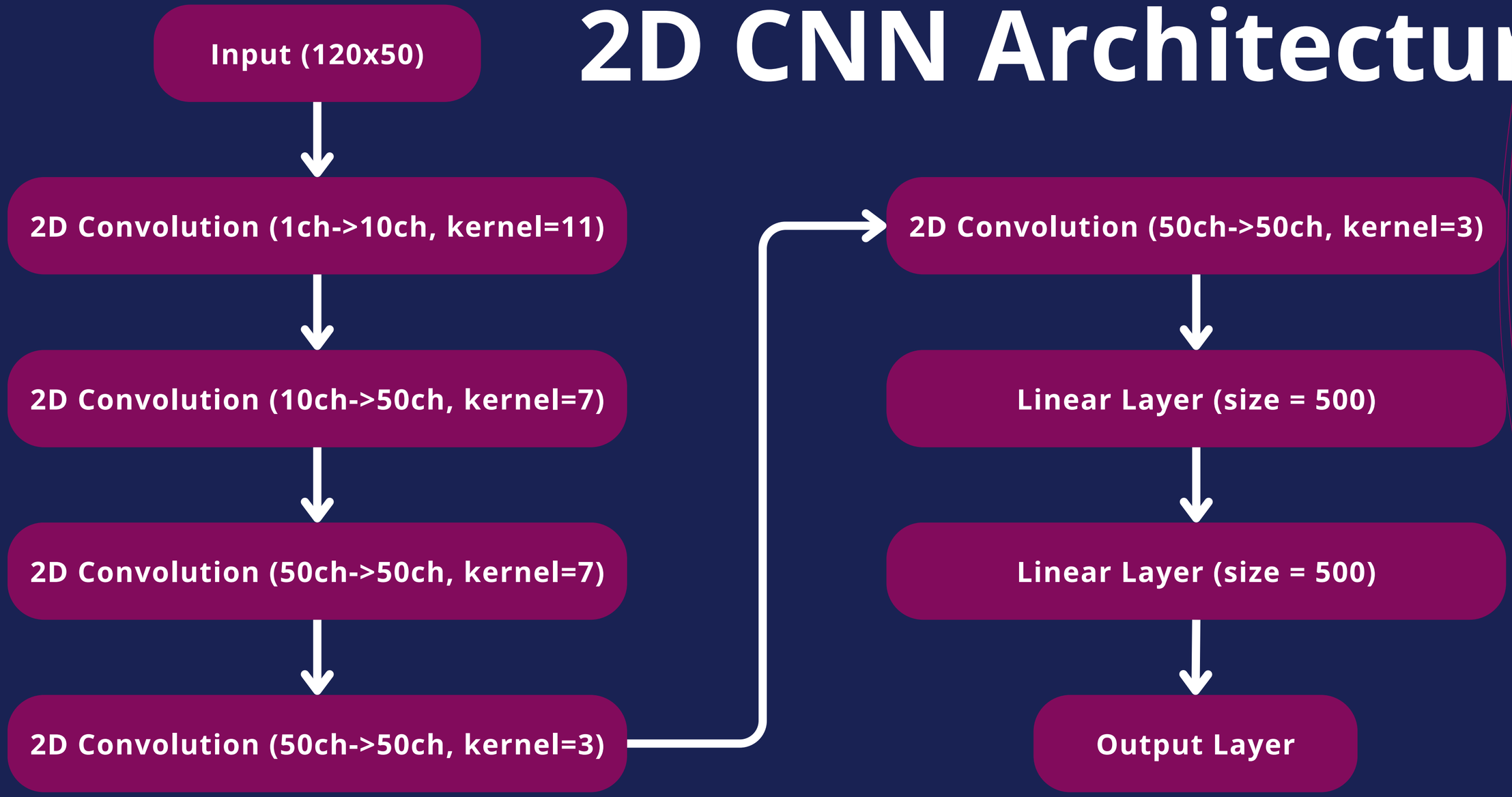
Train vs Val Accuracy



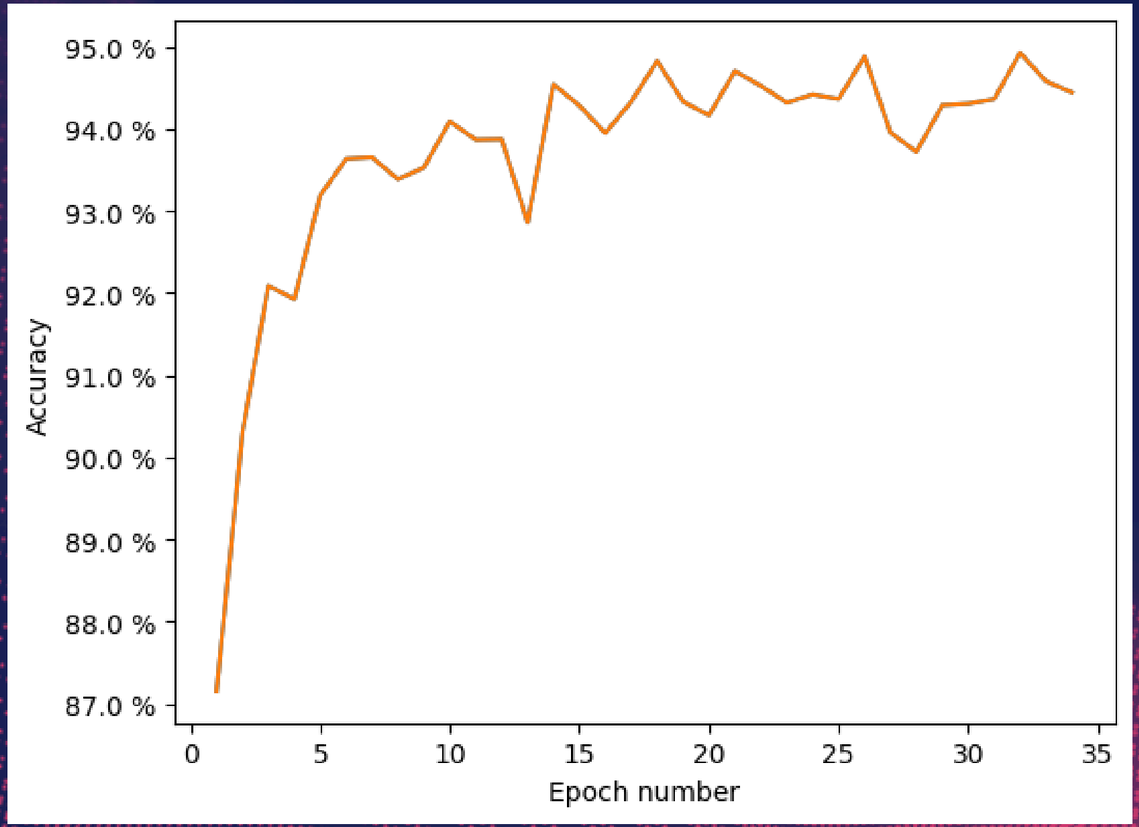
Confusion Matrix



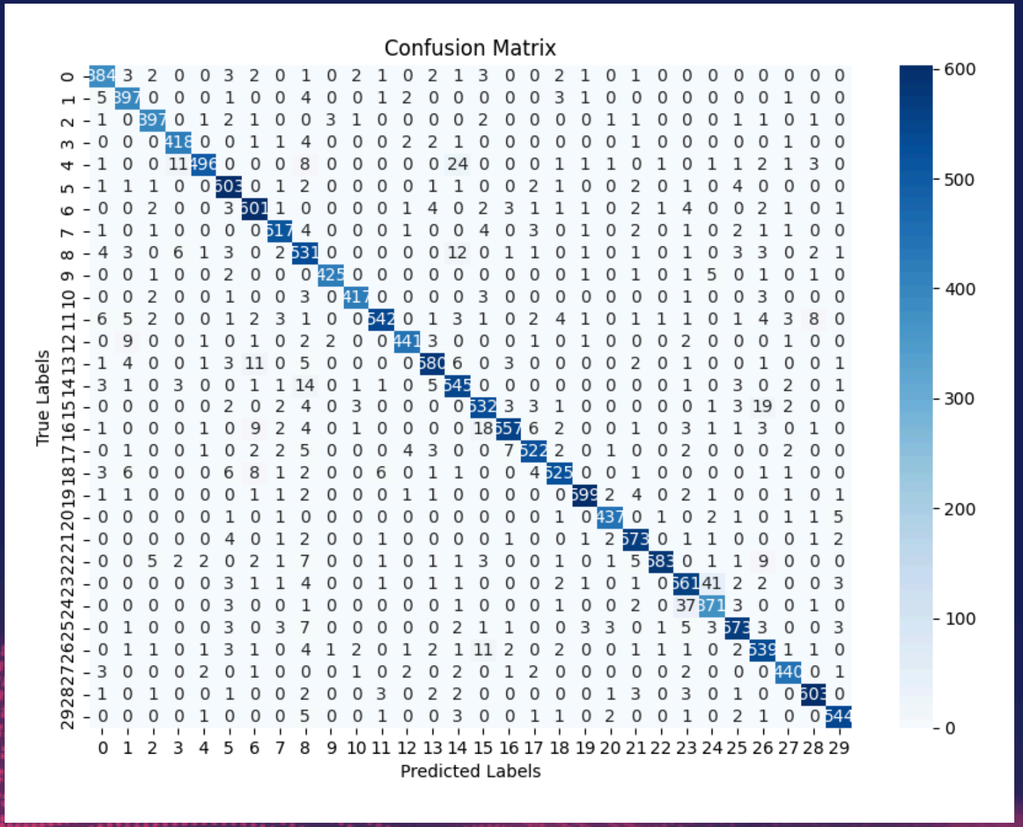
2D CNN Architecture



Validation Accuracy



Confusion Matrix



COMPARISON

ARCHITECTURE	Accuracy	Layers	Epochs
1D CNN	77%	6	25
2D CNN	95.5%	8	20
ResNet18	95.3%	18	18