DESIGN CREDIT PROJECT

SUPERVISOR : PRATIK MAZUMDER

RHYTHMiQ

MUSIC GENRE CLASSIFICATION USING DEEP LEARNING

ANSHIT AGARWAL [B23CS1087]

KAUSTUBH SALODKAR [B23EE1033]

INTRODUCTION

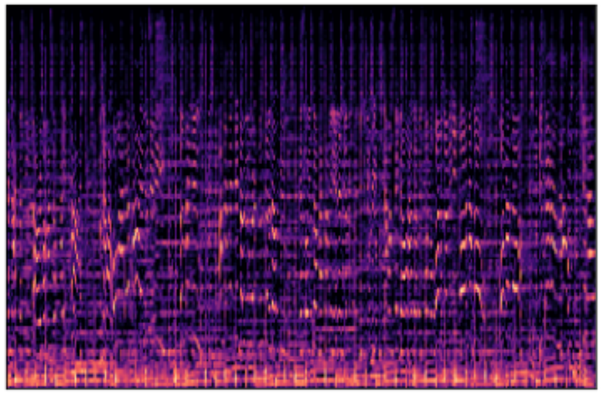
This project investigates the application of machine learning and deep learning techniques for music genre classification using audio-based features. We developed and evaluated five models on the **GTZAN** dataset, focusing on classification accuracy and performance.

Music genre classification plays a crucial role in modern applications such as streaming services, digital music libraries, and recommendation systems. RHYTHMiQ tackles the challenges of overlapping genre characteristics through data-driven methods to enable more reliable and efficient classification.

DATASET

We used the GTZAN Genre Collection, containing 1,000 tracks across 10 genres such as rock, jazz, classical, and hip hop. Each 30-second track was split into 3-second clips, resulting in 10,000 audio samples for training and evaluation.

AUDIO FILE



A MEL SPECTROGRAM

Extracted features like MFCCs and spectral and chroma data derived from the mel spectrogram

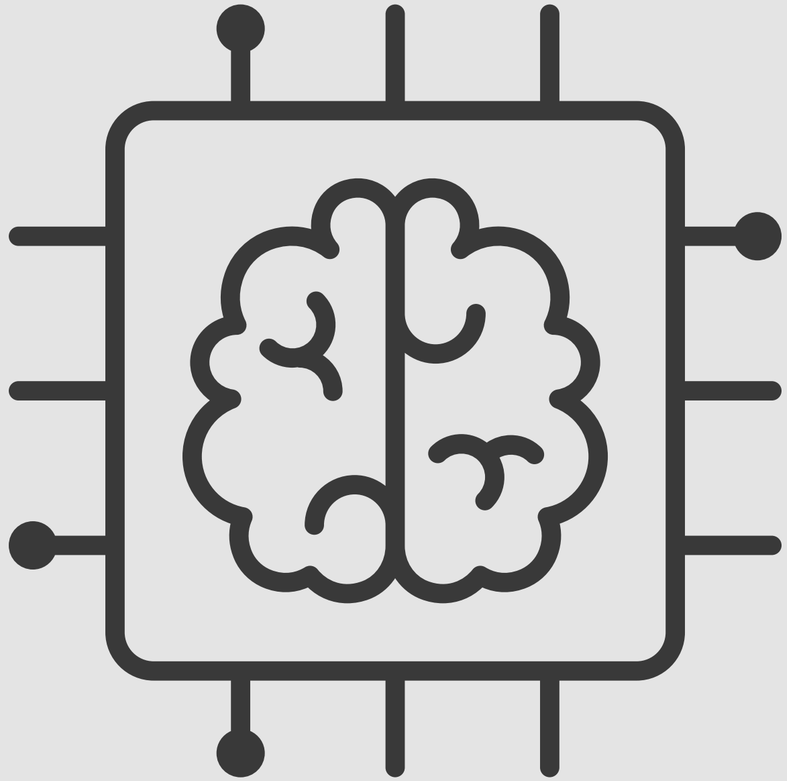
MODELS IMPLEMENTED

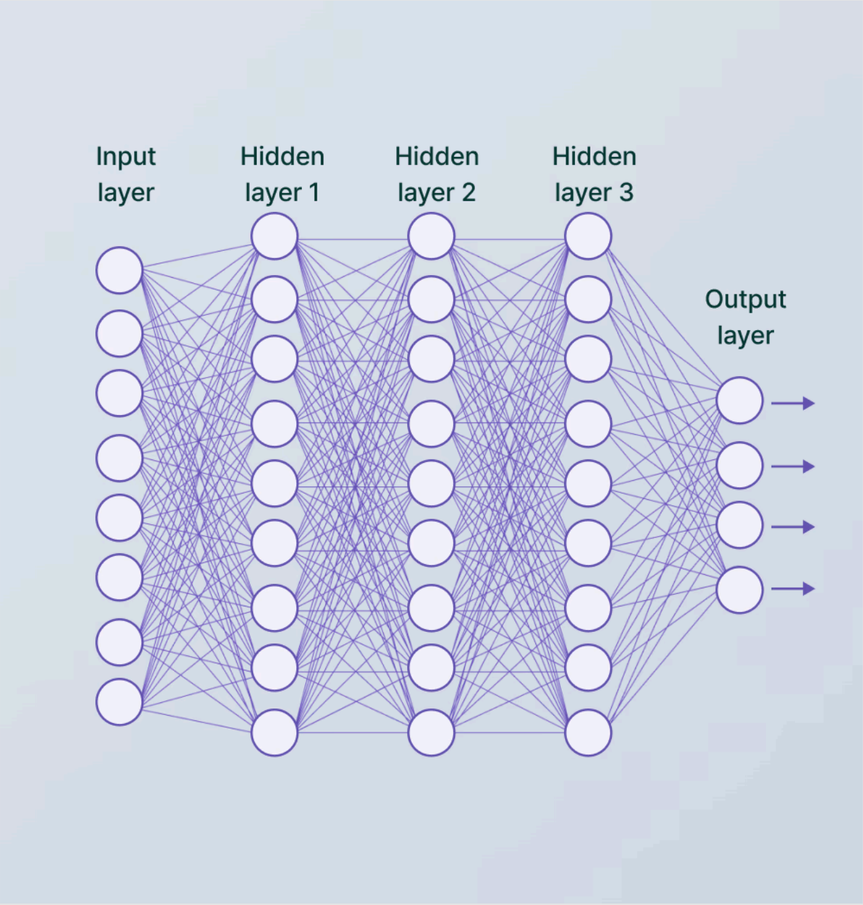
TRADITIONAL ML :

- Support Vector Machine (SVM)

DEEP LEARNING :

- Long Short Term Memory (LSTM)
- Hybrid (LSTM + SVM)
- Convolutional Neural Network (CNN)
- Residual-Gated CNN (Res-Gated CNN)





CONVOLUTIONAL NEURAL NETWORK

TEST ACCURACY: 86.44%

The CNN model emerged as the top performer in our evaluation, demonstrating its strength in learning spatial patterns from audio data.

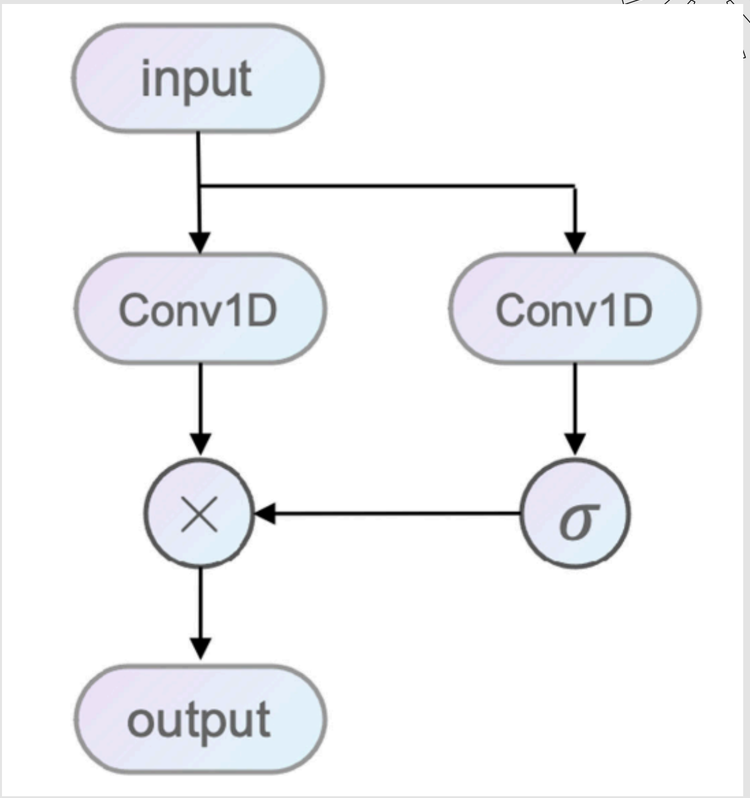
- Architecture:
  - Convolutional layers for feature extraction
  - MaxPooling layers for dimensionality reduction
  - Dense layers for classification
  - Dropout layers (rate 0.3) for regularization
- Key Strengths:
  - Captures genre-specific visual patterns such as rhythm, harmony, and texture.

RESIDUAL-GATED CNN

TEST ACCURACY: 84.18%

The Residual-Gated CNN builds on the CNN architecture by introducing residual connections and gating mechanisms to enhance information flow and learning depth.

- Architecture:
  - Multiple Conv1D layers for deep feature extraction
  - Residual connections to preserve learned information
  - Gating layers (Lambda functions + Multiply operations) for controlled data flow
  - MaxPooling layers for dimensionality reduction
  - Global Average & Max Pooling combined
  - Dense layers with dropout for final classification
- Regularization: Two dropout layers (rate: 0.3)
- Key Strengths:
  - Better gradient flow in deep networks through residual blocks
  - Gating mechanism helps in focusing on relevant features
  - Robust performance with potential for further tuning



1D Gated convolution structure

LSTM

TEST ACCURACY: 75.00%

- Architecture:
  - LSTM (128 units) → LSTM (64 units)
  - Dropout (0.3)
  - Dense (32 units) → Output (10 units)
- Strengths:
  - Learns temporal dependencies in audio
  - Effective for sequence modeling
- Limitations:
  - Short clips offer limited temporal context
  - Slower training and inference

SVM

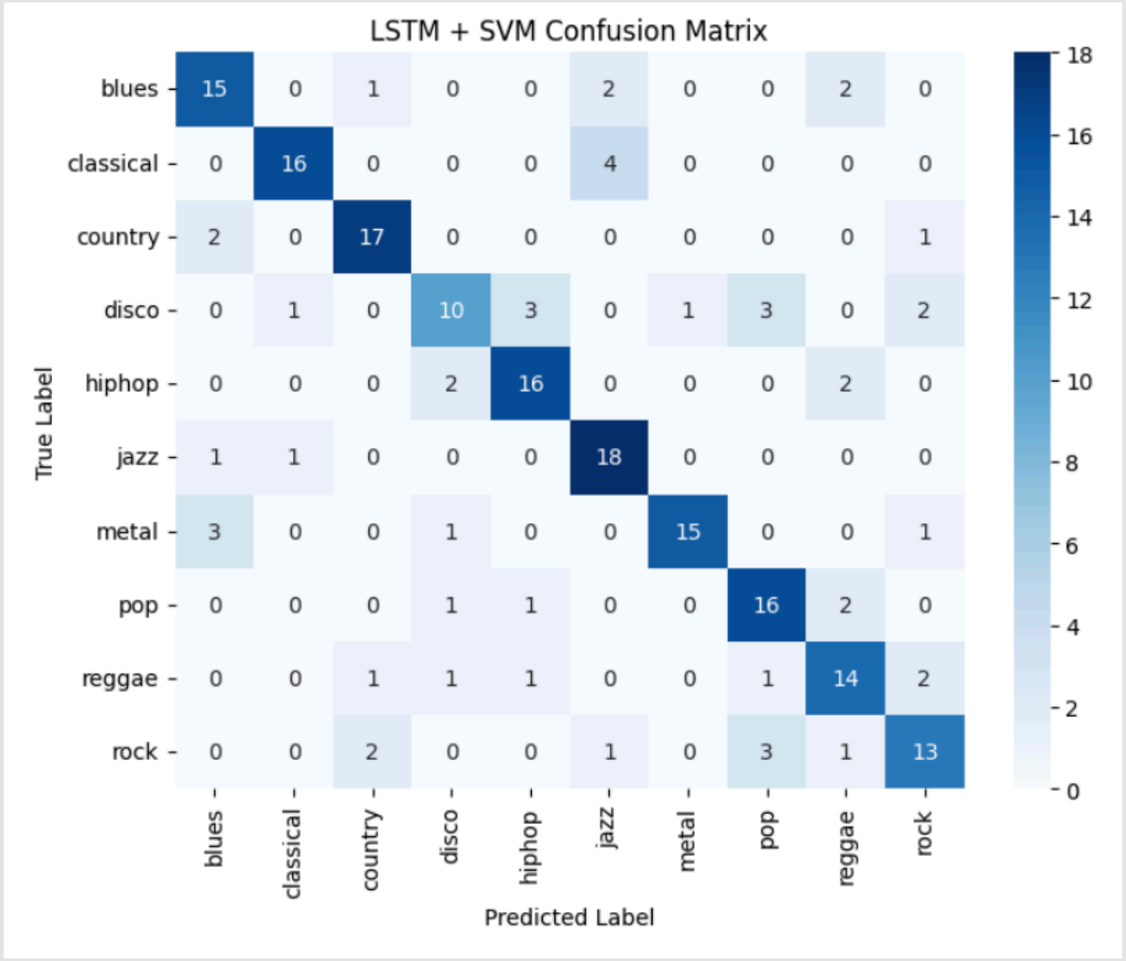
TEST ACCURACY: 77.00%

- Kernel : Radial Basis Function (RBF)
- Strengths:
  - Strong baseline with handcrafted features
  - Simple and interpretable
- Limitations:
  - Cannot model temporal or spatial patterns
  - Less effective for complex, high-dimensional data

LSTM + SVM

TEST ACCURACY: 75.00%

- Architecture:
  - LSTM layers extract temporal features
  - Hidden state output fed into SVM classifier
- Strengths:
  - Combines temporal learning (LSTM) with strong classification (SVM)
  - SVM adds robustness to LSTM's learned features
- Limitations:
  - Complexity increases with hybrid design
  - No significant performance gain over standalone LSTM



PERFORMANCE

Model	Test Accuracy (%)
Convolutional Neural Network (CNN)	86.44
Residual-Gated CNN (Res-Gated CNN)	84.18
Support Vector Machine (SVM)	77.00
Long Short-Term Memory (LSTM)	75.00
LSTM + SVM Hybrid	75.00

CNN-based models outperformed others due to their ability to extract spatial features. LSTM models were less effective for short audio clips where temporal context is limited.

INSIGHTS AND FUTURE WORK

Genres like classical and metal were easiest to classify due to distinct features, while rock and country were commonly confused.

Future work includes using transformers for long-range patterns, optimizing the Res-Gated CNN, trying ensemble methods, and building a real-time app.

CONTRIBUTIONS

Anshit Agarwal: Implemented CNN and Res-Gated CNN, feature extraction, and project architecture.

Kaustubh Salodkar: Implemented SVM, LSTM, hybrid models and handled evaluation and visualization.

THANK YOU