**Project Report: Music Genre Classification using CNN + LSTM**

**Name: Ashutosh Kumar**

**Reg No: 229302357**

**1. Introduction**

This project focuses on building a music genre classification system using deep learning. It utilizes **MFCC (Mel Frequency Cepstral Coefficients)** as audio features and combines **Convolutional Neural Networks (CNN)** and **Long Short-Term Memory (LSTM)** networks to learn spatial and temporal features from audio data. The model is trained using **K-Fold Cross-Validation**, and a **frontend** allows users to predict the genre of an audio file.

**2. Dataset**

* **Dataset Used**: GTZAN Genre Collection
* **Total Classes**: 10 genres (blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock)
* **Audio Length**: 30 seconds per file
* **Audio Format**: .wav

**3. Feature Extraction**

* **Technique**: MFCC (Mel Frequency Cepstral Coefficients)
* **Library**: librosa
* **Parameters**:
  + n\_mfcc=40: Number of MFCCs extracted per frame
  + max\_pad\_len=174: All MFCCs are padded/truncated to this length for uniformity

**4. Model Architecture**

**Combined CNN + LSTM Model:**

* **Input Shape**: (174, 40, 1)

**CNN Layers**:

* Extract spatial patterns from MFCCs
* Includes Conv2D, BatchNormalization, MaxPooling2D, Dropout

**LSTM Layer**:

* Handles temporal dependencies in the MFCC sequence
* LSTM layer after TimeDistributed wrapper applied to CNN outputs

**Final Layers**:

* Flatten → Dense → Dropout → Dense (Softmax for classification)

**5. Training**

* **Batch Size**: 32
* **Epochs**: 50
* **Loss Function**: categorical\_crossentropy
* **Optimizer**: adam
* **Validation Split**: 0.2

**K-Fold Cross-Validation**

* **K = 5**
* Trains 5 models on 5 folds
* Final model can be ensemble or best performing model selected

**6. Prediction**

* **Frontend**: A simple UI built using tkinter or streamlit for browsing a .wav file and displaying the predicted genre
* **Inference**: Loads genre\_model\_fold4.keras and labelencoder.pkl, extracts MFCCs, reshapes them to match model input, and returns top predicted genre

**7. Performance**

* **Training Accuracy**: Up to 99.94%
* **Validation Accuracy**: Varied around 35–40% (indicating overfitting)
* **Suggested Improvements**:
  + Data Augmentation
  + Use Spectrograms or Log-Mel Spectrograms
  + Ensemble all K models
  + Regularization & learning rate decay

**8. Key Libraries**

* tensorflow / keras
* librosa
* sklearn
* numpy, matplotlib
* joblib (for LabelEncoder persistence)

**9. Conclusion**

This project showcases the power of combining CNNs for feature extraction and LSTMs for temporal analysis in audio data. Though the model achieves very high training accuracy, validation accuracy can be improved using regularization, better data diversity, and ensemble techniques.

**10. Future Scope**

* Real-time audio genre detection
* Deployment using Flask + React
* Integration with Spotify API for genre prediction
* Transfer Learning from AudioSet or VGGish embeddings