Salsa Subgenre Analysis via Generative Latent Spaces

Attila Barna & Balázs Menkó

Advanced Machine Learning Laboratory

May 16, 2025

Eötvös Loránd University

Outline

1. Introduction to Salsa Music Classification
2. Audio Feature Extraction & Processing
3. Neural Network Architectures
    - Basic LSTM (Model 1)
    - Bidirectional LSTM with Self-Attention (Model 2)
    - Enhanced LSTM with BatchNorm & Global Pooling (Model 3)
4. Feature Importance Analysis
5. Experimental Results & Model Comparison
6. Cross-Modal Feature Visualization
7. Musicological Insights & Conclusion

Introduction

Introduction: Salsa Music Classification

- Salsa music contains diverse subgenres with subtle differences
- Automatic classification remains challenging
- **Research Question:** Can deep learning identify distinguishing acoustic patterns between salsa subgenres?
- Applications: Music recommendation, archival organization, musicological analysis

Key Challenge

Developing models that can capture both temporal dynamics and distinguish subtle rhythmic, harmonic, and timbral patterns specific to each subgenre

Gathering the Data

Gathering the Data

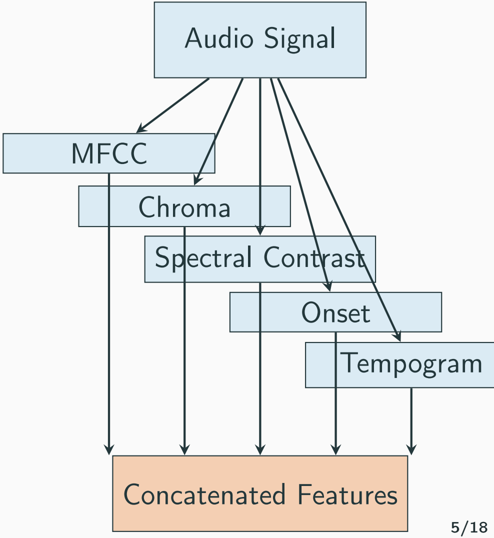
- Four Spotify playlist
  - Salsa Con **Rumba** – 66+19 track (with overlap)
  - **Son** – 44+43 track (with overlap)
  - **Linear** Salsa – 77 track
- Download Spotify playlist → Not allowed to train ML models
- Download only titles and artist to a `.txt` file with `spotipy`
- Use `yt_dlp` and `pytube` to search and download music in `.webm` format

- Problem with `yt_dlp`: not always find the correct music
- Data overview was needed
- **Final dataset**:
  - Rumba – 46 audio files
  - Son – 59 audio files
  - Linear – 67 audio files

Audio Feature Extraction

Audio Feature Extraction

- `librosa` package
- Multiple features extracted from audio:
  - **MFCCs:** (Mel-frequency cepstral coefficients) Timbre and spectral characteristics ($D = 13$)
  - **Chroma:** Harmonic content and tonality ($D = 12$)
  - **Spectral Contrast:** Frequency distribution ($D = 7$)
  - **Onset:** Rhythmic pattern detection ($D = 1$)
  - **Tempogram:** Tempo and rhythmic periodicity ($D = 384$)
- Features concatenated along feature dimension $\rightarrow$ saved to `.npy` file

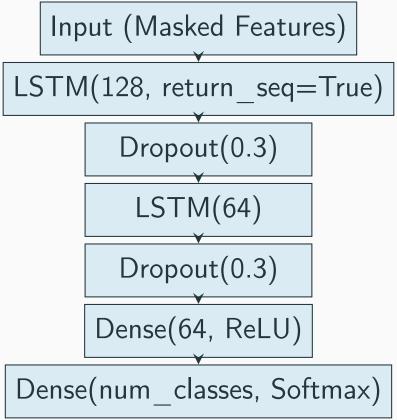


Neural Network Architectures

Model 1: Basic LSTM Architecture

Architecture Details

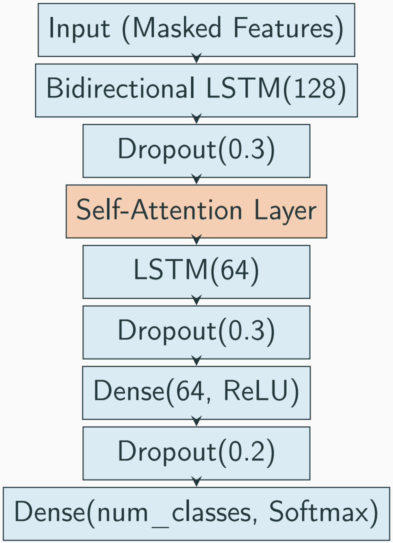
- **Input:** Padded features with masking
- **First LSTM Layer:** 128 units with sequence return
- **Second LSTM Layer:** 64 units
- **Regularization:** 30% dropout after each LSTM
- **Dense Layers:** 64 ReLU units followed by softmax
- **Loss:** Categorical cross-entropy with class weights



Model 2: Bidirectional LSTM with Self-Attention

Architecture Details

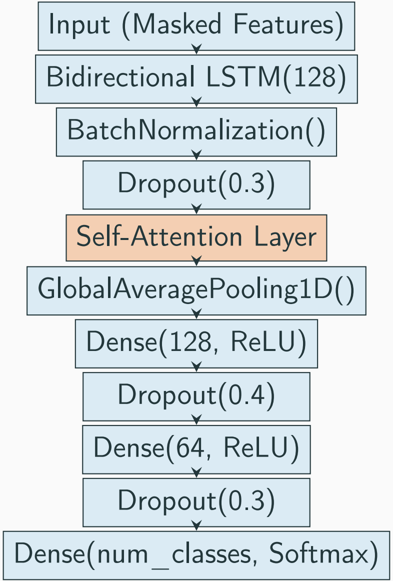
- **Input:** Same as Model 1
- **Bidirectional LSTM:** Processes sequences forward and backward
- **Self-Attention Layer:** Learns to focus on important timesteps
- **Training:** Smaller batch size (16 vs 32)
- **Advantages:**
  - Captures dependencies in both directions
  - Attention mechanism highlights relevant parts



Model 3: Enhanced LSTM with BatchNorm & Global Pooling

Architecture Details

- **Bidirectional LSTM:** Same as Model 2
- **Batch Normalization:** Stabilizes training
- **Global Average Pooling:** Alternative to flattening sequences
- **Deeper Dense Network:** 128 $\rightarrow$ 64 units
- **Heavier Regularization:** 40% and 30% dropout
- **Benefits:**
  - Better gradient flow and faster convergence
  - More robust to overfitting



Key Differences Between Architectures

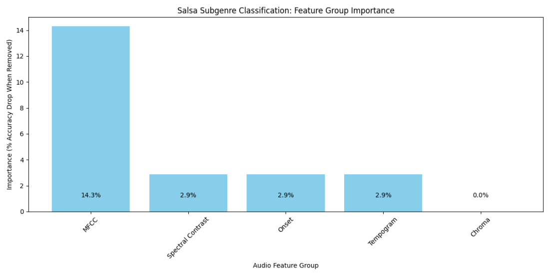
Feature	Model 1	Model 2	Model 3
Bidirectional LSTM	✗	✓	✓
Self-Attention	✗	✓	✓
Batch Normalization	✗	✗	✓
Global Pooling	✗	✗	✓
Dense Layers	1	1	2
Batch Size	32	16	16
Return Sequences	First layer only	First layer only	First layer only
Max Dropout Rate	30%	30%	40%

Architectural Progression

- From simple sequential processing to bidirectional analysis
- Addition of attention for selective feature focus

Feature Importance Analysis

Feature Ablation Analysis



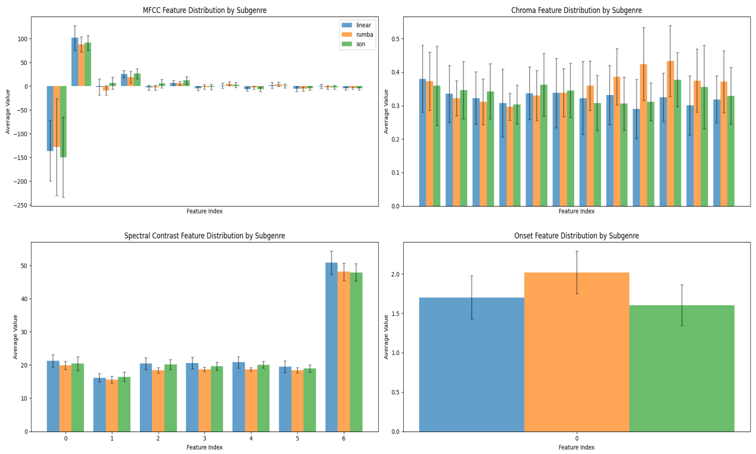
Methodology

- Baseline model trained with all features
- Separately trained, drop features
- Performance drop indicates feature importance

Key Findings

- MFCC features most critical (14.29% drop)
- Spectral Contrast, Onset, and Tempogram equally important, Chroma showed no impact

Feature Distribution by Subgenre

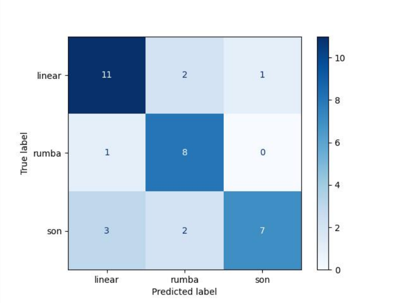


Experimental Results

Model Performance Comparison

Metric	Model 1	Model 2	Model 3
Test Accuracy	74.29%	57.14%	71.43%
Test Loss	1.0525	1.2996	0.6219

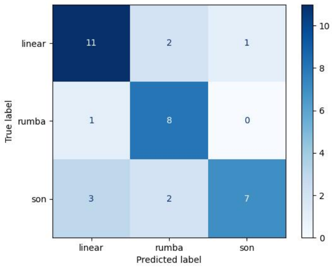
- Model 1 achieves highest accuracy despite simpler architecture
- Model 3 shows lowest loss, suggesting better generalization
- Self-attention improves loss in Model 3 but Model 2 underperforms
- Class weighting helps with imbalanced data



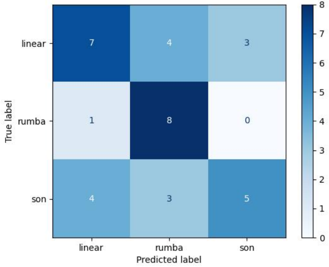
Confusion matrix of Model 1

Confusion Matrices

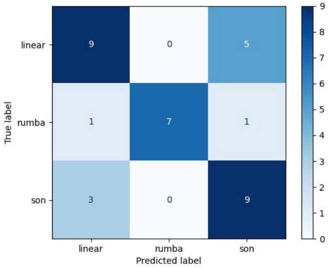
Model 1



Model 2

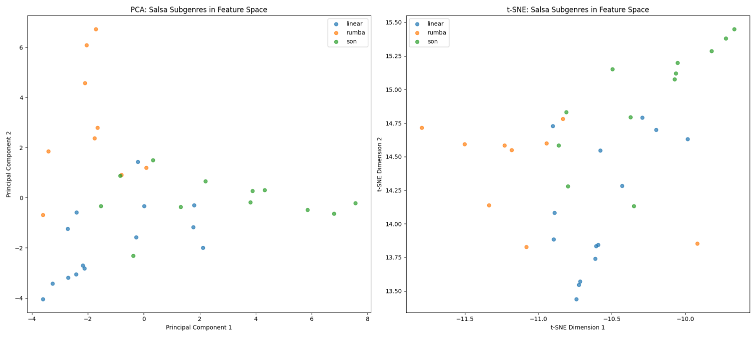


Model 3



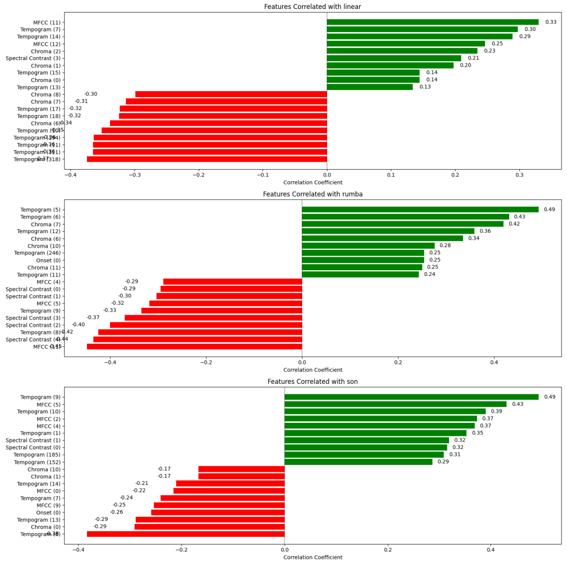
- Model 1 shows best overall classification accuracy
- Model 3 has better precision for certain classes
- All models struggle with some cross-genre misclassifications
- Consistent pattern of confusion between similar subgenres

Feature Space Visualization



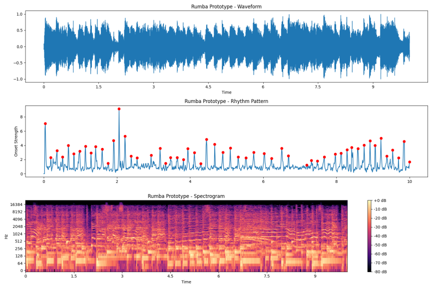
- Dimensionality reduction reveals clustering by subgenre
- PCA shows linear separability between some classes
- t-SNE preserves local structure, showing finer subclass relationships
- Feature space organization matches musicological understanding

Feature Correlation with Subgenres



- Specific features show strong correlation with particular subgenres
- MFCC coefficients significantly more important than Chroma features
- Onset features and Tempogram moderately correlated with rhythmically distinctive subgenres
- Spectral Contrast shows moderate importance for distinguishing subgenres

Representative Audio Examples



Key Audio Characteristics

- Son: Traditional Cuban clave with more space
- Rumba: Complex percussion patterns
- Linear: Emphasis on dance beats (1 and 5)

Example Selection Methodology

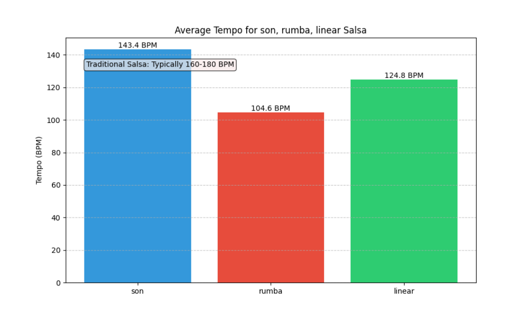
- Correctly classified examples with high confidence
- Attention model used to identify representative segments
- 5-second clips extracted from most salient regions
- Examples validated against musicological understanding

Conclusion

Musicological Insights

Key Findings

- Son distinguished by traditional Cuban rhythm and tres guitar
- Linear salsa shows consistent emphasis on dance beats (1 and 5)
- Timbral features (MFCC) more important than previously thought
- Rhythm patterns (onset, tempogram) provide moderate discriminative power
- Chroma features unexpectedly showed minimal impact on classification



Conclusions

Summary of Contributions

- Three progressive deep learning architectures for salsa subgenre classification
- Attention-based models that highlight musically significant segments
- Comprehensive feature importance analysis revealing distinguishing characteristics
- Musicological insights about salsa subgenre differentiation

Questions?