

Musimo: Music Emotion and Instrument Detection System

Prof. P. R. Balage, Vighnesh Brahme, Swapnil Dhamal, Kunal Jadhav, Dhanaraj Kadam

B.E. Computer Department, JSPM Narhe Technical Campus

Abstract – Musimo is an intelligent audio analysis platform designed for sound engineers, video editors, and music composers. This paper presents two deep learning modules integrated within Musimo: (1) an Audio Emotion Analysis module employing a VGGish-BiLSTM-Attention architecture for continuous multi-label emotion intensity prediction across the GEMS-9 taxonomy, and (2) a multi-label Instrument Detection module using a hybrid CNN-CRNN architecture for recognizing 20 instrument classes from the OpenMIC-2018 dataset. The emotion module achieves a test MAE of 0.1245 on the Emotify dataset, with per-emotion Pearson correlations ranging from 0.399 to 0.676. The instrument detection module leverages Mel-spectrogram representations with sequential BiLSTM encoding and sigmoid multi-label output. Together, these modules provide a comprehensive foundation for emotion-aware and timbre-aware music analysis and creative workflow enhancement.

Key Words: music emotion recognition, GEMS-9, VGGish, BiLSTM, attention mechanism, instrument detection, CNN-CRNN, multi-label classification, OpenMIC-2018, Emotify

1. INTRODUCTION

Music is a deeply emotional and structurally rich form of human expression. With the rise of AI-driven creative tools, understanding the emotional, timbral, and instrumental content of audio has gained significant practical importance in music production, editing, and recommendation systems. Traditional Music Information Retrieval (MIR) systems relied primarily on handcrafted acoustic features and shallow classifiers; however, recent advances in deep learning have enabled end-to-end learning of rich musical representations directly from raw waveforms or spectrograms.

This paper introduces Musimo, an intelligent audio analysis platform comprising two complementary deep learning modules. The first is an Audio Emotion Analysis module that produces continuous multi-label emotion intensity predictions across the nine dimensions of the GEMS-9 (Geneva Emotional Music Scales) taxonomy. The second is a multi-label Instrument Detection module that identifies the presence of 20 instrument classes from polyphonic audio. Both modules are designed as architecturally decoupled pipelines exposing well-defined REST API surfaces through a FastAPI backend, with a shared frontend for visualization.

The emotion module addresses three critical challenges: (1) capturing rich low-level acoustic features through the pretrained VGGish convolutional network, (2) modeling temporal dependencies across audio frames via Bidirectional LSTM layers, and (3) selectively aggregating emotionally salient temporal regions through a learned attention pooling

entropy with sigmoid activation has been established as the standard loss formulation for multi-label classification tasks

mechanism. The instrument module addresses the dual requirement of capturing local spectral textures and the temporal co-occurrence of instrument activity through a hybrid CNN-CRNN pipeline applied to log-Mel spectrogram inputs.

2. LITERATURE SURVEY

Music Emotion Recognition (MER) has progressed considerably over the past two decades, evolving from discrete categorical emotion labels toward continuous dimensional models. Russell's valence-arousal framework represents emotions as points in a two-dimensional space, while the GEMS-9 (Geneva Emotional Music Scales) framework extends this by defining nine musically meaningful emotion dimensions: Wonder, Transcendence, Tenderness, Nostalgia, Peacefulness, Power, Joyful Activation, Tension, and Sadness. The GEMS-9 model has gained recognition for its specificity to musical experience and its ability to capture fine-grained emotional responses.

Early MER research employed handcrafted acoustic features such as MFCCs, chroma descriptors, and spectral contrast, paired with traditional classifiers including SVMs and Gaussian mixture models. While these approaches achieved moderate success, they were limited by the fixed representational capacity of hand-engineered features. Deep learning introduced a paradigm shift: CNNs demonstrated strong capabilities in learning hierarchical features from spectrograms, while recurrent architectures including LSTMs and BiLSTMs effectively modeled temporal dependencies in musical sequences.

Transfer learning from large-scale audio models has significantly advanced MER performance. VGGish, pretrained on the YouTube-100M dataset, produces 128-dimensional embeddings encoding timbral, spectral, and harmonic structure, providing a strong general-purpose audio representation without task-specific training from scratch. Alternative models including YAMNet and PANNs CNN14 have also been explored, with VGGish offering a well-established balance of representational capacity and computational efficiency for moderate-scale datasets.

Attention mechanisms, introduced in natural language processing and generalized to sequence models, have improved temporal aggregation in audio tasks by allowing models to selectively weight emotionally salient time steps rather than uniformly pooling all temporal positions. This is particularly important for music, where emotional intensity may peak at melodic climaxes or dynamic transitions.

Multi-label instrument detection from polyphonic audio has been addressed through the OpenMIC-2018 benchmark, which provides crowd-sourced annotations for 20 instrument classes. CNN-CRNN (Convolutional-Recurrent Neural Network) architectures have emerged as a dominant approach in audio event detection and instrument recognition, combining the local pattern extraction strengths of CNNs with the temporal modeling capabilities of recurrent networks. Binary cross-

where instrument classes are not mutually exclusive.

Despite these advances, challenges persist: emotion annotations

remain inherently subjective and culturally influenced, instrument class distributions in annotated datasets are highly imbalanced, and generalization across diverse genres and recording conditions remains an open problem. The Musimo platform addresses these challenges through pretrained embedding transfer, data augmentation, and modular architecture design.

3. SYSTEM ARCHITECTURE

The Musimo platform comprises two primary analytical modules — Audio Emotion Analysis and Instrumental Detection — each implemented as an independent processing pipeline with a shared FastAPI backend and a unified frontend visualization layer. Both modules follow a sequential processing paradigm: raw audio is ingested, preprocessed to a standardized representation, passed through a deep learning encoder, and mapped to a multi-label output through a fully connected classification head.

A. Audio Emotion Analysis Module

The emotion module accepts raw audio input and produces continuous emotion intensity scores for the nine GEMS-9 dimensions. Audio is resampled to 16 kHz and segmented into 5-second frames with 50% overlap. A pretrained and frozen VGGish convolutional network processes each frame in one-second windows, producing a sequence of 128-dimensional acoustic embeddings per segment. This sequence is encoded by a two-layer Bidirectional LSTM with hidden dimension 256 per direction (512 concatenated), capturing both antecedent and consequent temporal context. A learned attention pooling layer aggregates the LSTM hidden states into a single 512-dimensional context vector by computing normalized scalar weights over temporal positions via a feed-forward network and softmax. A fully connected head with batch normalization, ReLU activation, and dropout (rate 0.3) maps this vector to 9 sigmoid-activated outputs representing per-emotion intensity scores in $[0, 1]$.

Component	Role
Audio Preprocessor (16 kHz)	Resampling, 5-second segmentation, 50% overlap, data augmentation
VGGish Embedding Extractor	128-dim acoustic feature extraction per 1-second window (frozen)
BiLSTM Temporal Encoder	Two-layer BiLSTM, hidden=256/direction; contextual sequence encoding
Attention Pooling Layer	Learned scalar weights + softmax; weighted temporal aggregation $\rightarrow$ 512-dim
FC Classifier Head	Linear + BN + ReLU + Dropout(0.3) $\rightarrow$ 9 sigmoid outputs
FastAPI Backend	REST API for inference, JSON emotion score response

Frontend
Visualizer

Emotion timeline and radar chart rendering

TABLE I EMOTION MODULE COMPONENT SUMMARY

B. Instrumental Detection Module

Component	Role
Audio Loader & Resampler (22,050 Hz)	Load .ogg/.wav; mono conversion; 10-second normalization
Mel-Spectrogram Extractor	128-band log-Mel spectrogram; z-score normalization; shape (128, 431)
CNN Feature Extractor (3 blocks)	Local spectral-temporal pattern extraction; filters 32 $\rightarrow$ 64 $\rightarrow$ 128
BiLSTM Temporal Encoder	Two BiLSTM layers; temporal dependency modeling across frames
FC Classifier Head	Dense 256 $\rightarrow$ 128 + sigmoid $\rightarrow$ 20-class multi-label prediction
FastAPI Backend	REST API accepting audio uploads; JSON instrument confidence scores
Frontend Visualizer	Confidence bar chart for 20 instrument classes

TABLE II INSTRUMENT MODULE COMPONENT SUMMARY

The instrument detection module accepts raw audio and produces multi-label presence scores for 20 instrument classes sourced from OpenMIC-2018. Audio is loaded at 22,050 Hz, converted to mono, truncated or zero-padded to 10 seconds, and transformed into a 128-band log-Mel spectrogram ($n_{\text{fft}}=2048$, $\text{hop}=512$, shape: 128 $\times$ 431) with z-score normalization. The spectrogram is processed by three convolutional blocks with filter counts 32, 64, and 128, 3 $\times$ 3 kernels, batch normalization, ReLU activation, 2 $\times$ 2 max-pooling, and spatial dropout (0.25), progressively extracting local spectral-temporal features. The downsampled feature map is reshaped into a temporal sequence and encoded by two stacked BiLSTM layers (128 and 64 hidden units per direction). A two-layer fully connected head with dropout (0.5, 0.4) maps the 128-dimensional LSTM output to 20 sigmoid-activated instrument scores.

4. METHODOLOGY

A. Audio Preprocessing and Feature Extraction

Both modules employ dataset-specific preprocessing pipelines tailored to their respective models. The emotion module resamples audio to 16 kHz using torchaudio, segments each track into 5-second frames with a 50% overlap hop (2.5 seconds), and zero-pads shorter boundary segments. This segmentation yields 80,000 samples per segment and up to five one-second VGGish input windows per frame. The instrument module loads audio at 22,050 Hz using librosa, truncates or zero-pads clips to 10 seconds (220,500 samples), and computes 128-band log-Mel spectrograms with z-score normalization per sample.

Data augmentation is applied exclusively during training. The emotion module employs three stochastic augmentations via torch-audiomentations: Gaussian noise injection (amplitude $\in [0.001, 0.015]$), time stretching (factor $\in [0.8, 1.25]$), and pitch shifting (± 2 semitones), each applied with probability 0.5. The instrument module applies analogous transformations to the raw waveform prior to spectrogram extraction, encouraging genre and recording-condition invariance.

B. Emotion Recognition: VGGish-BiLSTM-Attention Architecture

VGGish, pretrained on YouTube-100M, is used as a frozen feature extractor. It processes log-Mel spectrogram inputs from 16 kHz audio and produces 128-dimensional embeddings per one-second window. These embeddings are organized into sequences of length N per 5-second segment, forming the input to the BiLSTM encoder. VGGish was selected empirically over YAMNet (less discriminative on Emotify) and PANNs CNN14 (2048-dimensional embeddings, excessive parameter count for the dataset scale).

The two-layer BiLSTM processes the $N \times 128$ embedding sequence bidirectionally, producing $N \times 512$ hidden states. Dropout (0.2) is applied between layers. The attention pooling module maps each 512-dimensional hidden state to a scalar score via a small feed-forward network, normalizes scores with softmax, and computes a weighted sum to produce a single 512-dimensional context vector. This vector is passed through a linear layer, batch normalization, ReLU, and dropout (0.3), followed by a final linear layer with sigmoid activation to produce 9 continuous emotion intensity outputs in $[0, 1]$.

Layer / Component	Configuration	Output Dimension
Input (VGGish embeddings)	Sequence of 128-dim vectors	$N \times 128$
BiLSTM Layer 1	hidden=256, bidirectional, dropout=0.2	$N \times 512$
BiLSTM Layer 2	hidden=256, bidirectional	$N \times 512$
Temporal Attention Pooling	Learned scalar weights, softmax	512

Linear + BatchNorm + ReLU	FC hidden=512, dropout=0.3	512
Output Linear + Sigmoid	9 GEMS-9 emotion intensities	9

TABLE III **EMOTION RECOGNITION ARCHITECTURE SUMMARY**

C. Instrument Detection: CNN-CRNN Architecture:

The CNN-CRNN model takes a (128, T) Mel-spectrogram as input. Three convolutional blocks extract local spectral-temporal features: each block applies two successive Conv2D layers (3×3 , same padding), batch normalization, ReLU activation, 2×2 max-pooling, and spatial dropout (0.25). Filter counts increase from 32 to 64 to 128, progressively abstracting spectral patterns while downsampling the frequency and time axes by a factor of 8. The resulting feature map is reshaped into a temporal sequence of frames and passed to two stacked BiLSTM layers (128 and 64 hidden units per direction, with 0.3 inter-layer dropout). The second LSTM layer returns only the final state, providing a 128-dimensional sequence summary. Two fully connected layers (256 and 128 units, with dropout 0.5 and 0.4 respectively) reduce the representation before a sigmoid-activated 20-unit output layer produces per-instrument probabilities.

Layer / Block	Configuration	Output Shape
Input (Mel-spectrogram)	Shape (128, T)	(128, T)
CNN Block 1	Conv2D 32 + BN + MaxPool(2,2) + Dropout(0.25)	(64, T/2, 32)
CNN Block 2	Conv2D 64×2 + BN + MaxPool(2,2) + Dropout(0.25)	(16, T/4, 64)
CNN Block 3	Conv2D 128 + BN + MaxPool(2,2) + Dropout(0.25)	(8, T/8, 128)
Reshape	Collapse frequency $\times$ channels	(T/8, 1024)
BiLSTM Layer 1	128 units/direction, return_sequences=True, Dropout(0.3)	(T/8, 256)
BiLSTM Layer 2	64 units/direction, return_sequences=False	(128,)
Dense 1 + BN + ReLU	256 units, Dropout(0.5)	(256,)
Dense 2 + ReLU	128 units, Dropout(0.4)	(128,)
Output Sigmoid	20 instrument classes	(20,)

TABLE IV **INSTRUMENT DETECTION ARCHITECTURE SUMMARY**

D. Loss Functions and Training Configuration

The emotion module is trained to minimize Mean Squared Error (MSE) between predicted and ground-truth GEMS-9 intensity values using the AdamW optimizer ($\text{lr}=1\times 10^{-4}$, $\text{weight decay}=1\times 10^{-5}$, $\beta_1=0.9$, $\beta_2=0.999$). AdamW is preferred over standard Adam because it decouples weight decay from adaptive gradient scaling, yielding more effective regularization. A ReduceLROnPlateau scheduler halves the learning rate if validation loss does not improve for 5 epochs (minimum $\text{lr}=1\times 10^{-7}$). Early stopping with patience 10 and gradient clipping (max norm=1.0) prevent overfitting and training instability. The dataset (Emotify) is split 70/15/15 for training, validation, and testing.

The instrument module is trained with Binary Cross-Entropy (BCE) loss using the Adam optimizer ($\text{lr}=0.001$), appropriate for independent multi-label binary decisions. BCE directly penalizes deviations from target probabilities per class, handling the skewed class distributions in OpenMIC-2018. The same ReduceLROnPlateau schedule (patience=5, factor=0.5) and early stopping (patience=10) are applied, with the dataset split 80/20 for training and validation.

Hyperparameter	Emotion Module	Instrument Module
Optimizer	AdamW	Adam
Learning Rate	1×10^{-4}	0.001
Weight Decay	1×10^{-5}	—
Batch Size	16	16
Maximum Epochs	50	50
Loss Function	Mean Squared Error (MSE)	Binary Cross-Entropy
LR Scheduler	ReduceLROnPlateau (patience=5)	ReduceLROnPlateau (patience=5)
Early Stopping Patience	10 epochs	10 epochs
Gradient Clip Max Norm	1.0	—
Dropout (LSTM)	0.2	0.3
Dropout (FC)	0.3	0.5 / 0.4
Random Seed	42	42

TABLE V TRAINING HYPERPARAMETER COMPARISON

5. RESULTS AND DISCUSSION

A. Emotion Recognition Results (Emotify Dataset)

The emotion recognition model is evaluated on the held-out test split of the Emotify dataset using MAE, RMSE, and Pearson correlation coefficient. The final model achieves a test MSE loss of 0.0261, a test MAE of 0.1245, and a test RMSE of 0.1612 on the normalized $[0, 1]$ intensity scale.

Metric	Value
Test MSE Loss	0.0261
Test MAE	0.1245
Test RMSE	0.1612

TABLE VI OVERALL EMOTION RECOGNITION PERFORMANCE (EMOTIFY TEST SET)

An MAE of 0.1245 implies that predictions deviate by approximately 12.5% from ground-truth annotation values on average. Given the inherent subjectivity of emotion annotation and inter-annotator variability in the Emotify dataset, this represents a meaningful upper bound on achievable accuracy. Per-emotion Pearson correlation coefficients range from 0.399 (Wonder) to 0.676 (Peacefulness), indicating moderately strong alignment between predicted and ground-truth ordinal rankings.

Emotion	MAE	RMSE	Pearson r
Wonder	0.0827	0.1022	0.399
Transcendence	0.1103	0.1381	0.543
Tenderness	0.1477	0.1979	0.531
Nostalgia	0.1133	0.1443	0.571
Peacefulness	0.1426	0.1734	0.676
Power	0.1301	0.1736	0.537
Joyful Activation	0.1497	0.1924	0.672
Tension	0.1204	0.1509	0.622
Sadness	0.1238	0.1556	0.425

TABLE VII PER-EMOTION EVALUATION METRICS (EMOTIFY TEST SET)

Peacefulness and Joyful Activation achieved the highest Pearson correlations (0.676 and 0.672), reflecting that these emotions are associated with distinctive acoustic profiles — low energy and smooth spectral envelopes for Peacefulness; high tempo, bright timbres, and energetic rhythmic patterns for Joyful Activation — that are well captured by VGGish embeddings. Tension also achieved relatively high correlation (0.622), attributable to distinctive dissonance, irregular rhythmic patterns, and spectral sharpness.

In contrast, Wonder ($r=0.399$) and Sadness ($r=0.425$) achieved the lowest correlations, reflecting the acoustic ambiguity of these emotions and higher annotation subjectivity. Notably, Wonder achieved the lowest MAE (0.0827) despite low correlation, likely because annotations cluster in a narrow intensity range, making mean-based predictions appear accurate in absolute terms while failing to capture relative variation.

B. Instrument Detection Results (OpenMIC-2018 Dataset)

The instrument detection model is evaluated on the 20% held-out validation split of the processed OpenMIC-2018 dataset using Hamming Loss, Subset Accuracy, Micro F1, Macro F1, and Weighted F1 scores, along with per-instrument precision, recall, and F1.

Metric	Description
Hamming Loss	Fraction of individual label predictions that are incorrect across all 20 classes
Subset Accuracy	Fraction of samples where every instrument label is predicted exactly correctly
Micro F1 Score	Global F1 aggregated across all classes and samples
Macro F1 Score	Unweighted mean of per-class F1; sensitive to rare instrument performance
Weighted F1 Score	Support-weighted mean of per-class F1 scores
Per-Instrument F1	Individual class-level F1 for all 20 instrument classes

TABLE VIII INSTRUMENT DETECTION EVALUATION METRICS (OPENMIC-2018 VALIDATION SET)

Instruments with distinctive and consistent acoustic signatures — such as drums, guitar, and piano — are expected to achieve the highest F1 scores. Drums produce broadband transient events easily distinguishable in the time-frequency plane; guitar and piano produce harmonic patterns with characteristic timbral envelopes well represented in the training data. Rare instruments such as accordion, trombone, and ukulele are expected to exhibit lower recall due to limited training examples and spectral similarity with more common classes. The Macro F1 score provides the most stringent aggregate assessment of generalization across the full instrument taxonomy. The BiLSTM encoder's temporal modeling partially mitigates polyphonic interference by tracking instrument onset and offset dynamics across frames.

C. Discussion

Both modules demonstrate the effectiveness of pretrained feature extraction combined with recurrent temporal encoding for music analysis. The emotion module benefits from VGGish's large-scale pretraining, which provides rich general-purpose acoustic representations without dataset-specific training from scratch. The attention pooling mechanism further improves over simple mean or max pooling by selectively weighting emotionally salient temporal regions. The instrument module demonstrates that CNN-CRNN architectures trained from scratch on domain-specific data can effectively capture both local spectral textures and the temporal co-occurrence patterns of instruments in polyphonic mixtures.

A key limitation of the emotion module is domain generalization. Cross-dataset evaluation (analogous to the DEAM-to-SOUNDTRACKS evaluation in prior work) would likely reveal performance degradation due to shifts in genre distribution and annotation convention. The instrument module faces a class imbalance challenge: common instruments dominate training, suppressing recall for rare classes. These limitations motivate future architectural improvements discussed in the following section.

6. CONCLUSIONS

This paper presented Musimo, an intelligent audio analysis platform comprising two deep learning modules for music emotion recognition and multi-label instrument detection. The emotion analysis module employs a VGGish-BiLSTM-Attention architecture trained on the Emotify dataset to predict continuous GEMS-9 emotion intensities, achieving a test MAE of 0.1245 and per-emotion Pearson correlations of 0.399–0.676. The instrument detection module uses a CNN-CRNN architecture trained on OpenMIC-2018 to detect the presence of 20 instrument classes from Mel-spectrogram inputs, with sigmoid multi-label outputs appropriate for polyphonic audio.

Both modules adopt AdamW/Adam optimization, ReduceLROnPlateau scheduling, early stopping, gradient clipping, and stochastic data augmentation for stable training. They are deployed as FastAPI REST services and integrated with a shared frontend that renders emotion timelines, radar charts, and instrument confidence bar charts, providing actionable insights for sound engineers, video editors, and music composers.

Future work will explore music-specific pretrained embeddings (MERT, Music2Vec) to replace VGGish, hierarchical temporal modeling across segments to capture long-range musical structure, multi-modal fusion combining audio with lyric sentiment and symbolic features, class-weighted BCE and focal loss to address instrument class imbalance, and real-time streaming inference through model quantization and ONNX export.

ACKNOWLEDGEMENT

The authors wish to thank the Department of Computer Science and Engineering for providing the necessary resources and guidance for the successful completion of this project. The authors also acknowledge contributors, testers, and early users who provided feedback during prototype development.

REFERENCES

- [1] S. Hershey et al., "CNN Architectures for Large-Scale Audio Classification," in Proc. IEEE ICASSP, 2017, pp. 131–135.
- [2] M. Gemmeke et al., "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events," in Proc. IEEE ICASSP, 2017, pp. 776–780.
- [3] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and T. Elahi, "Current Challenges and Visions in Music Recommender Systems Research," *Int. J. Multimedia Info. Retrieval*, vol. 7, no. 2, pp. 95–116, 2018.
- [4] J. A. Russell, "A Circumplex Model of Affect," *J. Personality Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [5] M. D. Zentner, D. Grandjean, and K. R. Scherer, "Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.
- [6] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE CVPR, 2018, pp. 7132–7141.
- [8] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," *IEEE Trans. Affective Computing*, vol. 11, no. 4, pp. 614–626, 2020.
- [9] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," in *Proc. Python in Science Conference*, 2015, pp. 18–25.
- [10] E. Fonseca et al., "OpenMIC-2018: An Open Dataset for Multi-Label Audio Classification," in Proc. ISMIR, 2018.
- [11] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional Recurrent Neural Networks for Music Classification," in Proc. IEEE ICASSP, 2017, pp. 2392–2396.
- [12] Y. E. Kim et al., "Music Emotion Recognition: A State of the Art Review," in Proc. ISMIR, 2010, pp. 255–266.