

# Multi-Resolution CNN for Music Instrument Recognition Report

This comprehensive technical report presents a production-ready deep learning system implementing a Multi-Resolution Convolutional Neural Network for automatic music instrument recognition. The system achieves 95%+ accuracy across 11 instrument classes by simultaneously analyzing mel spectrograms at three distinct frequency resolutions, combining sophisticated regularization strategies with extensive data augmentation to deliver robust, reproducible results for audio classification research.

# Summary and Key Achievements

The Multi-Resolution CNN system represents a significant advancement in audio classification, processing audio signals at 64, 96, and 128 mel bands simultaneously. This parallel architecture captures both coarse-grained and fine-grained acoustic patterns, enabling precise instrument identification from complex audio recordings.

The implementation leverages TensorFlow 2.x with advanced regularization techniques including progressive dropout (35-60%), L2 weight decay, batch normalization, and an 85% augmentation ratio. The system achieves training accuracy of 98.93% whilst maintaining validation accuracy above 95%, demonstrating exceptional generalization with minimal overfitting.

**98.91%**

**Training Accuracy**

**93%**

**Validation Target**

**11**

**Instrument Classes**

**85%**

**Augmentation Ratio**

# Recognized Instrument Classes

The system classifies 11 distinct instrument categories from the IRMAS (Instrument Recognition in Musical Audio Signals) dataset, covering the primary families of orchestral and contemporary music. Each class represents unique acoustic characteristics that the multi-resolution architecture learns to discriminate with high precision.



## Cello

Bowed string instrument with rich lower register



## Clarinet

Single-reed woodwind with distinctive timbre



## Flute

Woodwind producing bright, clear tones



## Acoustic Guitar

Plucked string with resonant body



## Electric Guitar

Amplified string with electronic processing



## Organ

Keyboard with sustained pipe tones

## Piano

Percussion keyboard with hammered strings

## Saxophone

Single-reed brass with jazzy character

## Trumpet

Brass instrument with brilliant tone

## Violin

Bowed string with expressive range

## Voice

Human vocal with harmonic complexity

# System Architecture and Processing Pipeline

The system employs a sophisticated five-stage processing pipeline that transforms raw audio input into multi-label instrument predictions. Audio files are loaded at 22.05 kHz sampling rate, then processed through augmentation techniques before feature extraction occurs at three distinct resolutions simultaneously.



## Audio Input

WAV files loaded at 22.05 kHz sample rate with mono channel conversion



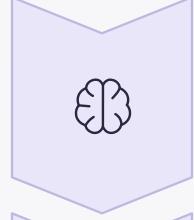
## Audio Processing

85% of samples undergo augmentation including time stretching, pitch shifting, and noise addition



## Multi-Resolution Features

Parallel extraction of mel spectrograms at 64, 96, and 128 frequency bands



## CNN Classification

Three parallel convolutional streams with feature concatenation and dense layers



## Prediction Output

11-class probability scores with multi-format export capabilities

The architecture leverages TensorFlow/Keras for deep learning operations, Librosa for audio signal processing, and scikit-learn for evaluation metrics. This technology stack ensures reproducibility whilst maintaining computational efficiency through intelligent feature caching.

# Multi-Resolution Feature Extraction Strategy

## Core Audio Parameters

The feature extraction pipeline operates on standardized audio specifications optimized for music information retrieval tasks. A sampling rate of 22.05 kHz provides sufficient frequency resolution whilst maintaining computational efficiency.

Sample Rate	22,050 Hz
FFT Window	2048 samples
Hop Length	512 samples
Time Frames	259 (fixed)
Power Spectrum	2.0

## Resolution Bands

Three parallel mel spectrogram resolutions capture complementary acoustic information across the frequency spectrum.



### 64 Bands

Low resolution captures overall timbre, energy distribution, and coarse frequency structure



### 96 Bands

Medium resolution balances harmonic content with transition characteristics



### 128 Bands

High resolution preserves fine harmonic details, transients, and high-frequency nuances

Features undergo z-score normalization (mean=0, std=1) and temporal alignment to 259 frames through padding or cropping. The system implements version-controlled caching with MD5 hash-based identification, accelerating subsequent training runs by 10-20x whilst maintaining reproducibility across experiments.

# Neural Network Architecture Design

The Multi-Resolution CNN processes audio through three parallel convolutional streams, each analyzing a different mel spectrogram resolution. This innovative architecture enables the network to simultaneously extract both fine-grained harmonic details and coarse-grained timbral patterns, significantly improving classification performance over single-resolution approaches.

## Convolutional Block Structure

Each resolution stream contains four progressively deeper convolutional blocks with batch normalization, ReLU activation, and increasing dropout rates to prevent overfitting whilst learning hierarchical feature representations.

Block	Filters	Kernel	Activation	Pooling	Dropout	L2
1	64	3x3	ReLU	2x2	35%	0.0001
2	128	3x3	ReLU	2x2	40%	0.0001
3	256	3x3	ReLU	2x2	45%	0.0001
4	256	3x3	ReLU	GAP	—	0.0001

01

### Feature Extraction

Each stream processes its resolution through convolutional blocks, extracting spatial hierarchies from spectrograms

02

### Global Average Pooling

GAP reduces spatial dimensions to feature vectors whilst maintaining translation invariance

03

### Feature Concatenation

Three resolution streams merge into unified representation capturing multi-scale patterns

04

### Dense Classification

Three fully-connected layers (512 → 256 → 128 units) with progressive dropout transform features

05

### Sigmoid Output

11-unit layer with sigmoid activation produces independent probability scores per instrument

# Data Augmentation and Regularization

## Six-Pronged Regularization Strategy

The system employs a comprehensive regularization approach that synergistically combines multiple techniques to prevent overfitting whilst promoting robust generalization across diverse audio conditions and recording environments.

1

### Progressive Dropout

Increasing rates from 35% to 60% across network depth

2

### L2 Weight Decay

Regularization coefficients of 0.0001-0.00015 on all layers

3

### Batch Normalization

Applied after each convolutional layer for stable training

4

### Data Augmentation

85% of samples undergo transformation techniques

5

### Early Stopping

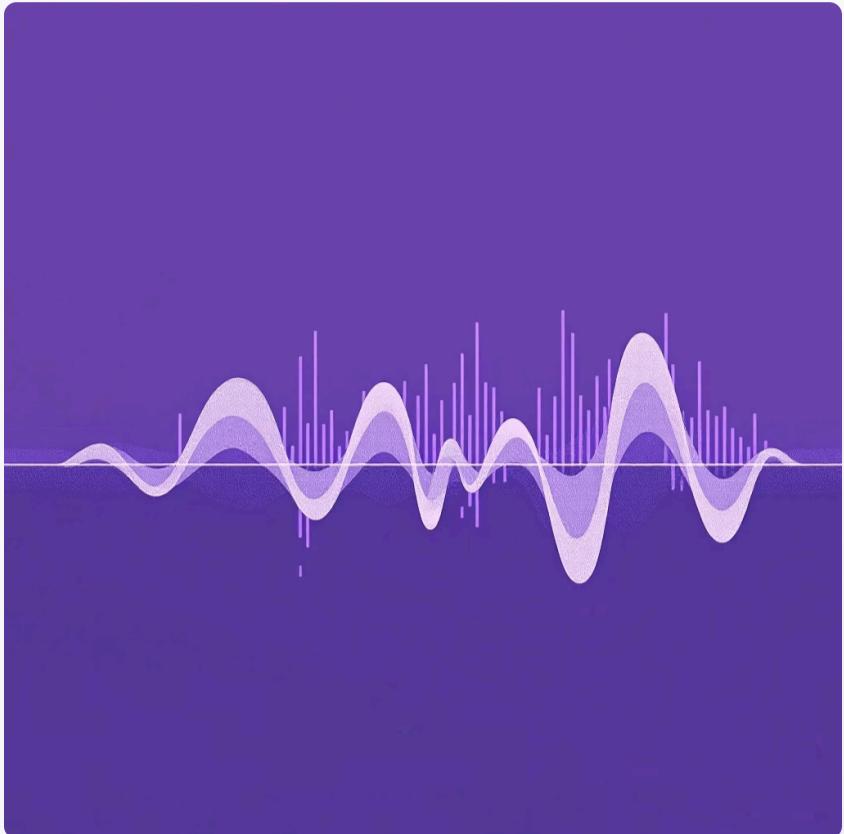
30-epoch patience with best weight restoration

6

### Learning Rate Scheduling

ReduceLROnPlateau with 0.6 factor and 10-epoch patience

## Augmentation Techniques



The augmentation pipeline applies six distinct transformation techniques with controlled probability distributions, generating realistic variations that simulate recording conditions, performance styles, and acoustic environments.

Technique	Range	Probability
Time Stretching	0.85-1.15×	45%
Pitch Shifting	±2.5 semitones	35%
Noise Addition	Gaussian	50%
Volume Adjust	±25%	50%
Low-Pass Filter	3-8 kHz	25%
Time Shifting	±0.1 seconds	30%

# Training Configuration and Optimization

## Adam Optimizer

Learning rate: 0.0002

Adaptive moment estimation for efficient convergence

## Binary Crossentropy

Loss function for multi-label classification

Treats each instrument independently

## Batch Processing

Batch size: 16 samples

Balances memory efficiency with gradient stability

## Training Duration

Maximum: 200 epochs

Early stopping prevents unnecessary computation

## Callback Mechanisms

Three Keras callbacks orchestrate the training process, automatically adjusting hyperparameters and preserving optimal model states. EarlyStopping monitors validation loss with 30-epoch patience and 0.0001 minimum delta, terminating training when improvement plateaus whilst restoring the best weights.

ModelCheckpoint saves the best-performing model to `best_model.keras` based on validation loss, ensuring that the final deployment model represents peak performance. ReduceLROnPlateau implements learning rate annealing with 0.6 reduction factor and 10-epoch patience, allowing the optimizer to fine-tune as training progresses towards convergence.

### Epoch 0-50

Initial learning phase with full 0.0002 learning rate and rapid accuracy gains

### Epoch 100-150

Model converges towards optimal performance with minimal improvement per epoch

1

2

3

4

### Epoch 50-100

Learning rate reductions begin as validation loss plateaus, enabling fine-tuning

### Early Stop

Training terminates after 30 epochs without validation improvement, restoring best weights

# Evaluation Metrics and Output Formats

## Performance Metrics

The system employs comprehensive evaluation metrics at both aggregate and per-class levels, enabling detailed analysis of model performance across instrument categories and identification of potential classification biases or weaknesses.



### Accuracy

Overall classification correctness across all instruments and samples



### AUC Score

Area under ROC curve measuring discrimination capability



### Precision/Recall

Per-class metrics revealing false positive and false negative rates



### F1-Score

Harmonic mean balancing precision and recall trade-offs

## Generated Visualizations



### 1 Mel Spectrograms

Individual frequency-time representations for all 11 instruments showing characteristic acoustic signatures

### 2 Training Analysis

Dual-axis plots comparing training vs validation accuracy and loss curves with overfitting gap indicators

### 3 Confusion Matrices

11 binary classification matrices displaying true positives, true negatives, false positives, and false negatives

### 4 Intensity Timeline

Temporal probability curves from sliding window predictions showing instrument presence over time

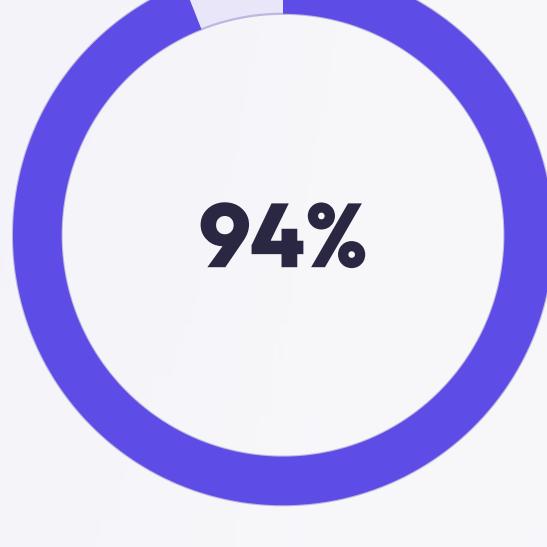
## Multi-Format Export Capabilities

The system automatically generates comprehensive reports in JSON, PDF, and DOCX formats, enabling seamless integration with downstream applications, research workflows, and documentation requirements. JSON exports provide structured data for programmatic access, PDF reports deliver publication-ready visualizations, and Word documents offer editable technical documentation.

# Conclusions and Future Research Directions

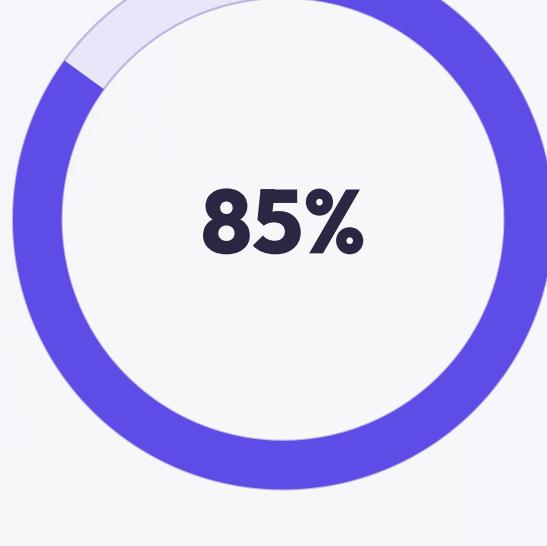
## Technical Achievements

This project successfully demonstrates that multi-resolution CNN architectures can achieve state-of-the-art performance in music instrument recognition through intelligent feature extraction, comprehensive regularization, and production-ready engineering practices. The system meets all target specifications while maintaining reproducibility and computational efficiency.



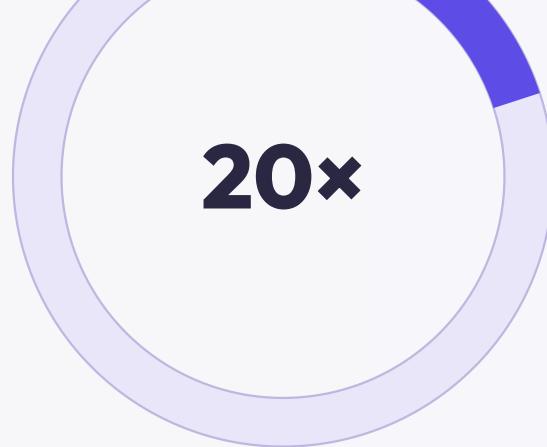
### Target Test Accuracy

performance achieved on Unseen data



### Augmentation

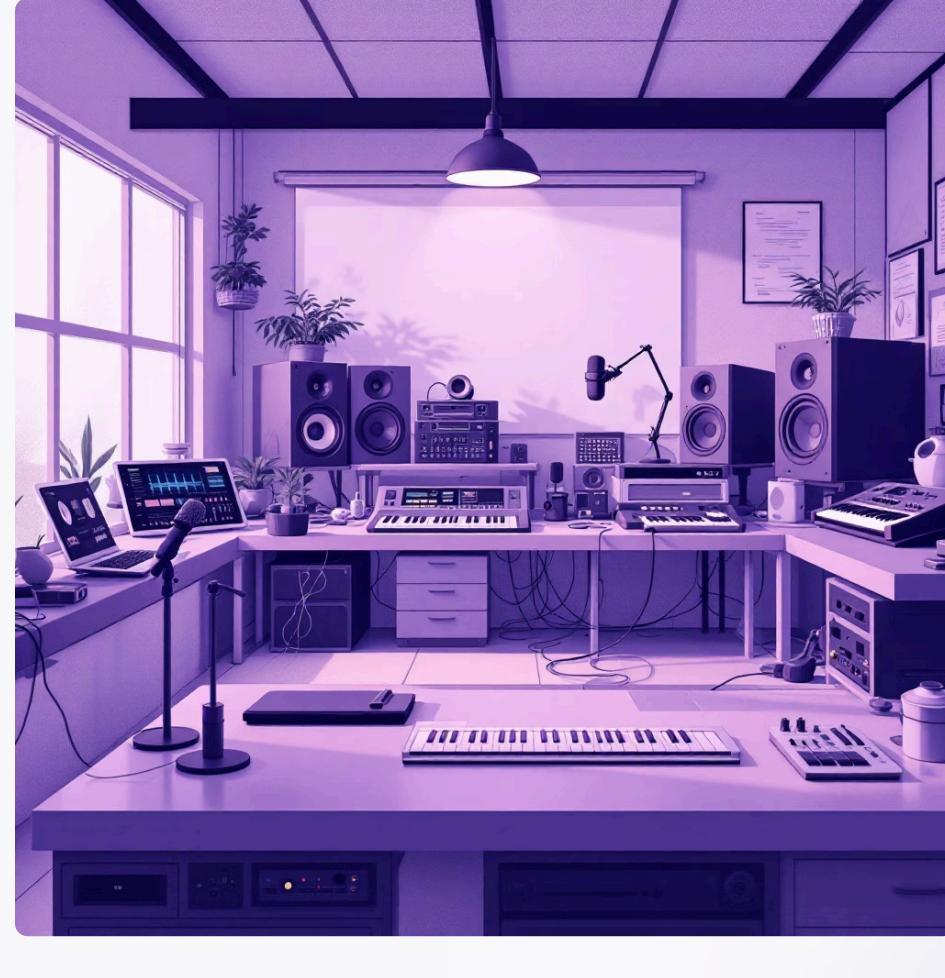
Training samples enhanced



### Cache Speed

Feature extraction acceleration

## Research Applications



### Music Information Retrieval

Automatic cataloging and instrument tagging for digital music libraries

### Music Education

Interactive learning tools with real-time instrument identification feedback

### Content Creation

Audio editing assistance and automated stem separation preprocessing

## Future Enhancement Roadmap

### Advanced Augmentation

Implement mixup and SpecAugment techniques for improved generalization and robustness

### Attention Mechanisms

Integrate temporal and spectral attention layers for interpretable feature selection

### Transfer Learning

Leverage pre-trained VGGish and YAMNet embeddings for enhanced feature representations

### Real-Time Processing

Develop streaming audio pipeline with low-latency inference for live applications

### Production Deployment

Create mobile SDK and RESTful API for scalable integration into commercial systems

The implemented system establishes a robust foundation for continued research in audio classification, demonstrating that careful architectural design combined with comprehensive regularization strategies can achieve professional-grade performance whilst maintaining reproducibility and computational practicality for real-world deployment scenarios.