UNIVERSITAS
GADJAH MADA
**An Empirical Analysis on Feature Complementarity and Fusion Depth in Multimodal Deep Learning for
Music Genre Classification**
Daniel Winston Mandela Tulung, Rifki Afina Putri, S.T., M.S., Ph.D.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

# ABSTRACT

## AN EMPIRICAL ANALYSIS ON FEATURE COMPLEMENTARITY AND FUSION DEPTH IN MULTIMODAL DEEP LEARNING FOR MUSIC GENRE CLASSIFICATION

by

Daniel Winston Mandela Tulung 23/516260/PA/22080

Music Genre Classification (MGC) models often struggle to effectively integrate diverse acoustic features. Combining 2D spatial spectrograms with 1D statistical descriptors is commonly assumed to improve accuracy, but the relative value of different descriptors and fusion depths remains unclear, and prior studies are frequently compromised by temporal data leakage. This research conducts a systematic ablation study using a dual-branch CNN–MLP architecture to evaluate the complementarity of three statistical descriptor groups (MFCC, Chroma, Spectral Contrast) and three fusion depths (early, intermediate, late) for multimodal MGC. To assess the robustness of the comparison, 23 configurations were each trained and evaluated on three repeated random track-level splits of the GTZAN dataset (69 model trainings in total), with final predictions aggregated via track-level majority voting. Results show that a Mel-spectrogram-only CNN baseline is already a strong classifier (F1 = 0.854), and that the benefit of adding statistical descriptors is small and uncertain: among 21 fusion configurations, only intermediate (feature-level) fusion of MFCC and Chroma was not below this baseline on any of the three splits, by an average of 1.1 F1 points, which corresponds to roughly one test track. Mean F1 differs little across fusion depths (intermediate 0.848, early 0.844, late 0.835), and all three families average slightly below the Pure CNN baseline, so no fusion depth reliably outperformed a spectrogram-only model. Combining all three descriptor groups provides no measurable advantage over a well-chosen pair. These findings suggest that, once segment-level leakage is mitigated, Mel-spectrograms already capture the large majority of genre-discriminative information, and that statistical descriptors offer at most a small, fusion-depth-dependent complementary signal rather than the substantial gains reported in prior literature. This study contributes a controlled, single-architecture comparison protocol and exploratory guidelines for feature and fusion-depth selection in multimodal audio classification.

**Keywords:** Music Genre Classification, Multimodal Deep Learning, Feature Fusion, Empirical Analysis, Convolutional Neural Network, Multi-Layer Perceptron.