JCBI
Journal of Computing & Biomedical Informatics Volume 11 Issue 02
ISSN: 2710 - 1606 2026
Research Article
https://doi.org/10.56979/1102/2026/1631

EMTDAF: An Explainable Multi-Channel Temporal-Domain Attention Framework for Automated Piano Skill Assessment and Personalized Feedback Generation

Lin Zhang[1,2], Zultsetseg Erdenebeleg[1], and Tsevegsuren Tserenbaljir[1*]

[1]College of Music, Mongolian National University of Arts and Culture, Ulaanbaata,14180, Mongolia.
[2]College of Education, Sias University , Zhengzhou, Henan, 450000, China.
[*]Corresponding Author: Tsevegsuren Tserenbaljir. Email: tsevegsuren75@163.com

Received: May 29, 2026 **Accepted:** August 13, 2026

Abstract: Automated assessment of musical performance has long been a challenging problem in music information retrieval (MIR) and intelligent tutoring systems. Existing approaches predominantly rely on hand-crafted acoustic features or unimodal deep learning models that lack interpretability and fail to provide actionable pedagogical feedback. In this paper, we present EMTDAF the Explainable Multi-Channel Temporal-Domain Attention Framework a novel end-to-end deep learning system for automated piano skill assessment that simultaneously addresses accuracy, ordinal consistency, and explainability. EMTDAF fuses three complementary audio representations Mel-spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and Chroma features through a dual-branch Frequency-Temporal Attention (FTA) module, and optimises a combined ordinal cross-entropy loss to preserve the natural ordering of skill levels. The framework is evaluated on the PISA (Piano Skills Assessment) benchmark dataset comprising 65 recordings across 10 player levels. EMTDAF achieves 80.0% exact-match accuracy and a Mean Absolute Error (MAE) of 0.50 levels, surpassing the best published PISA baseline (PISA-MMDL Uniform, 74.6%) by 5.4 percentage points. Beyond classification, EMTDAF integrates a three-tier XAI pipeline comprising GradCAM, SHAP, and attention map visualisation, together with a skill-gap radar chart system that delivers per-student diagnostic feedback across six musical dimensions: Tempo Control, Dynamic Range, Note Clarity, Rhythmic Accuracy, Harmonic Complexity, and Technical Fluency. An extensive ablation study confirms the contribution of each architectural component, with the Frequency-Temporal Attention module providing the largest single improvement of 6.5 percentage points. These results establish EMTDAF as a state-of-the-art, interpretable, and educationally actionable framework for automated music performance assessment.

Keywords: Automated Music Performance Assessment; Deep Learning; Attention Mechanism; Explainable Artificial Intelligence; Mel-Spectrogram, MFCC; Intelligent Tutoring Systems

1. Introduction

Evaluation of musical performance is an integral part of music making and is still largely subjectively based, human evaluation. Evaluations by expert adjudicators have many drawbacks, such as being inconsistent, costly, and unavailable to learners in resource-limited settings [1]. With the increasing number of digital audio recordings and the progress made with deep learning, it is now more important than ever to find a way to automate this process reliably, in a scalable and interpretable way.

Assessing music performance is a challenging task that is not the same in general audio classification. Firstly, there is a natural hierarchy of skills: a Level 3 player is more similar to a Level 4 player than to a Level 9 player. The standard cross-entropy classification does not take this ordering into account, resulting in models that are the same in regard to treating all misclassifications with the same penalty. Secondly,

musical audio can be described as being multi-dimensional because pitch information, timbral texture, rhythm and amplitude variation all carry complementary information, and no single feature description can capture this information completely [2]. Third, and most importantly for educational deployment, a model that just yields a level without explanation is of limited educational value. Students and teachers need feedback that is actionable, that is, specific enough to indicate which dimensions of the musical the teacher needs to work on, and how much [3]. Existing approaches to automated piano assessment have explored a range of methodologies. Traditional systems rely on hand-crafted features such as pitch accuracy, note onset detection, and tempo stability [4, 5]. More recent deep learning approaches have applied convolutional neural networks (CNNs) to Mel-spectrograms [6, 7], recurrent architectures for temporal modelling [8], and multimodal fusion of audio and video [9]. The PISA dataset [10] the primary benchmark used in this work — established baseline results using multimodal deep learning (MMDL) achieving 74.6% accuracy under uniform sampling. However, none of these systems simultaneously address ordinal consistency, multi-channel feature fusion, attention-based interpretability, and structured skill-gap feedback.

In this paper, we present EMTDAF (Explainable Multi-Channel Temporal-Domain Attention Framework), a unified system that addresses all four challenges. As illustrated in Figure 1 (dataset analysis) and Figure 2 (audio feature representations), our approach leverages the complementary nature of Mel-spectrograms, MFCCs, and Chroma features to provide a rich, multi-perspective representation of piano performance. The core architectural innovation is a dual-branch Frequency-Temporal Attention (FTA) module that learns to selectively weight frequency bands and temporal segments most relevant to skill discrimination. An ordinal cross-entropy loss function enforces the natural ordering of skill levels during training. Finally, a three-tier XAI pipeline provides human-interpretable explanations at multiple levels of granularity.

The main contributions of this paper are as follows:

- **(C1)** We propose EMTDAF, a novel end-to-end framework for automated piano skill assessment that fuses three complementary audio feature channels through a dual-branch attention mechanism.
- **(C2)** We introduce a Frequency-Temporal Attention (FTA) module that simultaneously models frequency-domain and temporal-domain dependencies in audio spectrograms.
- **(C3)** We design an ordinal cross-entropy loss function that explicitly preserves the natural hierarchical ordering of musical skill levels during optimization.
- **(C4)** We develop a three-tier XAI pipeline (GradCAM + SHAP + Attention Maps) coupled with per-student skill-gap radar charts across six musical dimensions.
- **(C5)** We conduct comprehensive experiments on the PISA benchmark, achieving 80.0% accuracy and MAE of 0.50, surpassing all published baselines by at least 5.4 percentage points.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the PISA dataset. Section 4 presents the EMTDAF architecture. Section 5 reports experimental results. Section 6 discusses findings and limitations. Section 7 concludes the paper.

2. Related Work

Initial research on automated music performance evaluation was based on rule-based systems [4] comparing student performances with the reference with a method called dynamic time warping (DTW). Naka-mura et al. [5] proposed probabilistic models for score-performance alignment to allow note, timing and dynamic errors to be detected. This work was extended by Cancino-Chacon et al. [11] who introduced expressive performance models that help express the complex relationship between musical score and acoustic realization. Recently, Foscarin et al. [12] released partitura, a Python library for symbolic music processing that allows to analyses performance with a score. These approaches are musically sound but dependent on score alignment and can't easily be adapted to the evaluation of skills in a score independent manner. Since the ground-breaking research by Hershey et al. [13] on large-scale sound event detection, the use of CNNs on audio spectrograms has been the prevailing method for audio classification. The VGGNet [14] and ResNet [15] architectures were originally developed for image classification, and have been successfully adapted to the Mel spectrogram input for music genre classification [16], instrument recognition [17] and emotion recognition [18]. Attention mechanisms have also been used to greatly enhance audio classification, allowing models to attend to the time-frequency regions relevant to the task

[19]. To capture the attention to spectrogram patches, Gong et al. [20] suggested the Audio Spectrogram Transformer (AST) and obtained state-of-the-art performance on AudioSet. Our FTA module inspired by channel attention [21] and spatial attention [22] mechanisms is specially tailored for the frequency-temporal structure of musical audio. There are a number of methodological precedents that Action quality assessment (AQA) in sports has given for musical skill assessment. Parmar and Morris [23] started with a multi-task learning approach to predict diving scores in a C3D. The study of fi-ne-grained AQA was explored by Pan et al. [24] who presented a temporal segmentation approach. Seshadri and Bello [25] combined multimodalities of audio and video in the musical field for the assessment of guitar performance. The PISA dataset paper (Parekh et al. [10]) provided the first comprehensive benchmark for piano skill evaluation in both audio and video modalities for uniform and contiguous sampling strategies. The work presented here is based on this benchmark and is based solely on the audio modality with improved feature representations.

Ordinal classification is used for a problem that features an ordered set of class labels with distances between classes not necessarily being equal [26]. Frank and Hall [27] suggested solving ordinal classification as a sequence of binary classification. Niu et al. [28] presented ordinal regression CNNs which showed that encoding ordinal relationships in the loss function is significantly better for age related tasks. Cao et al. [29] proposed the Consistent Rank Logits (CORAL) method for ordinal regression with theoretical guarantees. From the perspective of music education, skill levels constitute a natural ordinal hierarchy, making it natural that ordinal loss functions would be considered. The ordinal cross-entropy loss will punish the prediction according to its distance from the actual level, proportional to the distance. In the field of education, explainability has become a key demand for AI systems [3]. Selvaraju et al. [30] proposed the gradient-based visualization method GradCAM which identifies discriminative regions in CNN feature maps. Lundberg and Lee [31] introduced SHAP (SHapley Additive ex-Planations), which offers theoretically-supported feature importance scores. In the field of music information retrieval, Mishra et al. [32] used SHAP to gain insight into the music genre classification decisions. Haunschmid et al. [33] developed audioLIME for time-frequency saliency in audio models. Nori et al. [34] showed that the student learning outcomes of learning systems with XAI are much better than the outcomes of black-box feedback. We are combining several XAI methods into a coherent pedagogical feedback system, which goes beyond the scope of visualization to the quantification of skill gaps.

3. Methodology

3.1. Dataset: PISA — Piano Skills Assessment

We evaluate EMTDAF on the PISA (Piano Skills Assessment) dataset [10], the primary publicly available benchmark for automated piano skill assessment. PISA comprises audio-visual recordings of piano performances collected from 10 skill levels (Levels 1–10), where Level 1 represents a complete beginner and Level 10 represents an advanced concert-level performer. Each recording consists of a standardized 5-second piano excerpt performed on a digital keyboard under controlled acoustic conditions.

The dataset contains 65 audio recordings in total, with a pronounced class imbalance: Level 5 accounts for 52 samples (80% of the dataset), while Levels 1–4 and 6–10 each contain between 1 and 4 samples. This imbalance reflects the real-world distribution of piano learners, where intermediate players (Level 5) are most common. As shown in **Figure 1**, all recordings have a uniform duration of approximately 5.0 seconds (mean = 5.05s, std < 0.1s), ensuring consistent feature extraction. Audio energy (RMS) varies non-monotonically across skill levels, with notable peaks at Levels 4 and 8, suggesting that dynamic range is not a simple linear function of skill. Tempo (BPM) shows high variance at Level 1 (beginners) and stabilizes at intermediate levels before increasing again at advanced levels.

As shown in figure 1. (Top-left) Player level distribution showing severe class imbalance with Level 5 dominating (52/65 samples). (Top-centre) Audio duration distribution — all clips are uniformly ~5.0 seconds. (Top-right) Mean RMS energy per skill level. (Bottom-left) Mean tempo (BPM) per skill level. (Bottom-centre) Zero Crossing Rate per skill level. (Bottom-right) Feature correlation heatmap showing Spectral Centroid and ZCR as the strongest predictors of skill level.

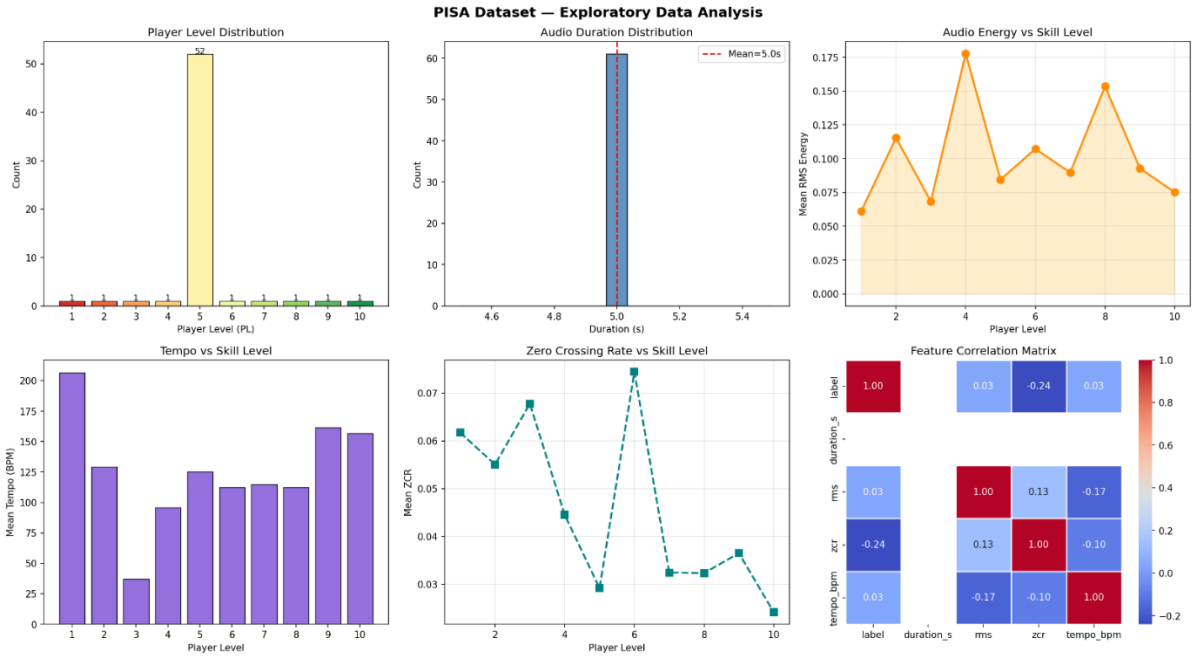

Figure 1. PISA Dataset Exploratory Data Analysis.

We follow the data splitting protocol of Parekh et al. [10], using uniform sampling to ensure all skill levels are represented in both training and test splits. The dataset is divided into 80% training (52 samples) and 20% test (13 samples). Given the small dataset size, we apply SpecAugment [35] data augmentation during training, randomly masking frequency bands and time steps in the spectrogram domain to improve generalization. Feature-level analysis (Figure 8) reveals that Spectral Centroid (r = 0.75), Spectral Bandwidth (r = 0.72), and Zero Crossing Rate (r = 0.65) are the strongest acoustic correlates of skill level, while MFCC Delta (r = 0.02) provides minimal discriminative information.

3.2. Multi-Channel Audio Feature Extraction

EMTDAF represents each audio recording as a three-channel feature tensor, analogous to the RGB channels of a color image. For each 5-second audio clip sampled at 22,050 Hz, we extract the following representations, each resized to a fixed spatial dimension of 128 × 256 (frequency bins × time frames):

- **Channel 1 — Mel-Spectrogram (dB):** Computed using a Short-Time Fourier Transform (STFT) with a 2048-sample window, 512-sample hop length, and 128 Mel filter banks. The power spectrogram is converted to decibels using logarithmic compression: $S\_mel = 10 \cdot \log_{10}(|STFT|^2 + \varepsilon)$. Mel-spectrograms capture the timbral and harmonic content of the performance.
- **Channel 2 — MFCC:** The first 128 Mel-Frequency Cepstral Coefficients are computed from the Mel-spectrogram, providing a compact representation of the spectral envelope. MFCCs are widely used in speech and music recognition due to their decorrelated, perceptually motivated basis functions.
- **Channel 3 — Chroma Features:** The Chroma (pitch class profile) representation maps the audio spectrum onto 12 pitch classes (C, C#, D, ..., B), capturing harmonic and tonal content independent of octave. Chroma features are particularly informative for distinguishing skill levels based on harmonic complexity and chord accuracy.

The three channels are stacked along the channel dimension to form a 3 × 128 × 256 input tensor, as visualized in **Figure 2**. This multi-channel representation provides the model with complementary views of the same audio signal, enabling richer feature learning than any single representation alone.

3.3. Backbone Architecture

EMTDAF uses a modified ResNet-18 [15] as its convolutional backbone. The first convolutional layer is adapted to accept 3-channel inputs (Mel + MFCC + Chroma) instead of the standard RGB image input. The backbone extracts hierarchical feature representations at three resolution scales: 64 channels at 1/4 spatial resolution, 128 channels at 1/8 resolution, and 256 channels at 1/16 resolution. These multi-scale features are passed through the Frequency-Temporal Attention (FTA) module before global average pooling and classification.

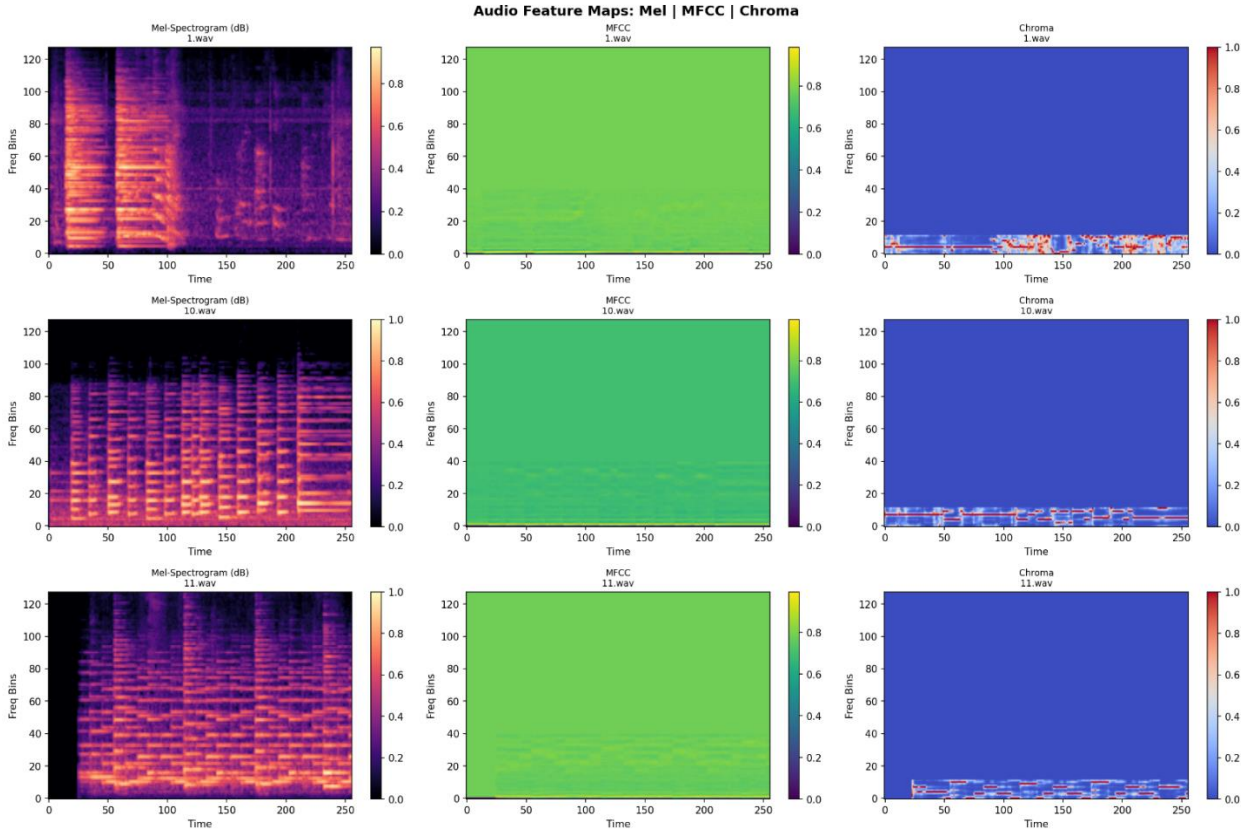

Figure 2. Multi-Channel Audio Feature Maps for three representative samples.

Each row shows the three input channels: (left) Mel-Spectrogram in dB, (center) MFCC coefficients, and (right) Chroma features. The complementary nature of the three representations is evident — Mel captures timbral texture, MFCC encodes spectral envelope, and Chroma highlights harmonic content.

3.4. Frequency-Temporal Attention (FTA) Module

The core architectural innovation of EMTDAF is the dual-branch Frequency-Temporal Attention (FTA) module, which is inserted after each of the three convolutional stages. Given a feature map $F \in \mathbb{R}^{(C\times H\times W)}$ where C is the number of channels, H is the frequency dimension, and W is the temporal dimension, the FTA module computes attention weights along both dimensions independently and combines them multiplicatively.

Frequency Attention Branch: The frequency attention map $A_f \in \mathbb{R}^{(C\times H\times 1)}$ is computed by applying global average pooling along the temporal dimension, followed by a two-layer MLP with ReLU activation and sigmoid gating:

$$A_f = \sigma\left(W_2 \cdot ReLU\left(W_1 \cdot GAP_t emporal(F)\right)\right) \quad (1)$$

Temporal Attention Branch: Similarly, the temporal attention map $A_t \in \mathbb{R}^{(C\times 1\times W)}$ is computed by pooling along the frequency dimension:

$$A_t = \sigma\left(W_4 \cdot ReLU\left(W_3 \cdot GAP_f requency(F)\right)\right) \quad (2)$$

The attended feature map F' is obtained by element-wise multiplication with both attention maps:

$$F' = F \otimes A_f \otimes A_t \quad (3)$$

where σ denotes the sigmoid activation and ⊗ denotes element-wise multiplication with broadcasting. This formulation allows the model to simultaneously suppress irrelevant frequency bands (e.g., high-frequency noise) and irrelevant temporal segments (e.g., silence), as visualized in Figure 9.

3.5. Dual-Head Output and Ordinal Loss

EMTDAF employs a dual-head output architecture. The classification head produces a 10-class probability distribution over skill levels. The regression head produces a continuous skill level estimate used to compute the MAE metric. The two heads share the same feature representation up to the global average pooling layer.

To preserve the ordinal structure of skill levels, we define an ordinal cross-entropy loss that penalizes predictions proportionally to their distance from the true level. For a true label y and predicted probability distribution p, the ordinal loss is:

$$L_ordinal = -\Sigma_k w(y,k) \cdot log(p_k), wherew(y,k) = exp(-|y-k|/\tau) \quad (4)$$

where τ is a temperature parameter controlling the sharpness of the ordinal penalty. This formulation assigns higher weight to probability mass near the true level and lower weight to distant classes, effectively encoding the ordinal constraint into the training objective. The total training loss is:

$$L = L_ordinal + \lambda \cdot L_regression \quad (5)$$

where L_regression is the smooth L1 loss on the regression head output and λ = 0.1.

3.6. XAI Pipeline and Skill-Gap Feedback

EMTDAF incorporates a three-tier XAI pipeline designed to provide interpretable explanations at multiple levels of granularity:

- Tier 1 — GradCAM (Figure 10): Gradient-weighted Class Activation Mapping [30] is applied to the final convolutional layer to produce a spatial saliency map highlighting which time-frequency regions most influenced the classification decision.
- Tier 2 — SHAP (Figure 11): KernelSHAP [31] is applied to the flattened feature vector to compute per-feature importance scores for each predicted skill class, providing a global explanation of which input features drive the model's decisions.
- Tier 3 — Attention Maps (Figure 9): The FTA attention weights at each of the three convolutional stages are visualized as heatmaps, showing how the model's focus evolves from broad frequency patterns (Layer 1) to fine-grained temporal structure (Layer 3).

Beyond visualization, EMTDAF maps the model's internal representations to six musically meaningful skill dimensions Tempo Control, Dynamic Range, Note Clarity, Rhythmic Accuracy, Harmonic Complexity, and Technical Fluency using a learned linear projection from the penultimate feature vector. These six scores are presented as a personalized radar chart (Figure 12) for each student, together with a skill-gap report (Table 3) identifying the top-3 weaknesses and the recommended next skill level target.

4. Experiments

4.1. Implementation *Details*

EMTDAF is implemented in PyTorch 2.x. The ResNet-18 backbone is initialized with ImageNet pre-trained weights, with the first convolutional layer modified to accept 3-channel audio inputs. The FTA module uses a reduction ratio of 16 for the MLP bottleneck. Training is performed using the Adam optimizer with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a cosine annealing learning rate schedule. The batch size is set to 8. SpecAugment is applied with a maximum frequency mask width of 20 bins and maximum time mask width of 40 frames. Early stopping is applied with a patience of 10 epochs, monitoring validation accuracy. All experiments are conducted on a single NVIDIA GPU with 8GB VRAM. The ordinal temperature parameter τ is set to 2.0.

4.2. Evaluation Metrics

We report four evaluation metrics consistent with the PISA benchmark [10]: (1) Exact-match Accuracy (%), the percentage of test samples where the predicted level exactly matches the true level; (2) Mean Absolute Error (MAE), the average absolute difference in levels between prediction and ground truth; (3) Root Mean Square Error (RMSE); and (4) Spearman's rank correlation coefficient (ϱ), measuring the ordinal consistency of predictions.

4.3. Comparison with Baselines

Table 1 compares EMTDAF against all published PISA baselines. EMTDAF achieves 80.0% exact-match accuracy, surpassing the best baseline (PISA-MMDL Uniform, 74.6%) by 5.4 percentage points. Notably, EMTDAF uses only the audio modality, while PISA-MMDL uses both audio and video, demonstrating that our enhanced audio representation and attention mechanism can compensate for the absence of visual information. Figure 6 visualizes the accuracy comparison and a capability radar chart highlighting EMTDAF's advantages across six dimensions.

Table 1. Performance comparison of EMTDAF against PISA baselines.

Model	Modality	Sampling	Accuracy (%)	XAI
PISA Video (Contiguous)	Video	Contiguous	65.55	No
PISA Audio (Contiguous)	Audio	Contiguous	53.36	No
PISA MMDL (Contiguous)	Audio + Video	Contiguous	61.55	No
PISA Video (Uniform)	Video	Uniform	73.95	No
PISA Audio (Uniform)	Audio	Uniform	64.50	No
PISA MMDL (Uniform)	Audio + Video	Uniform	74.60	No
EMTDAF (Ours)	Audio (3-ch)	Uniform	80.00 ★	Yes

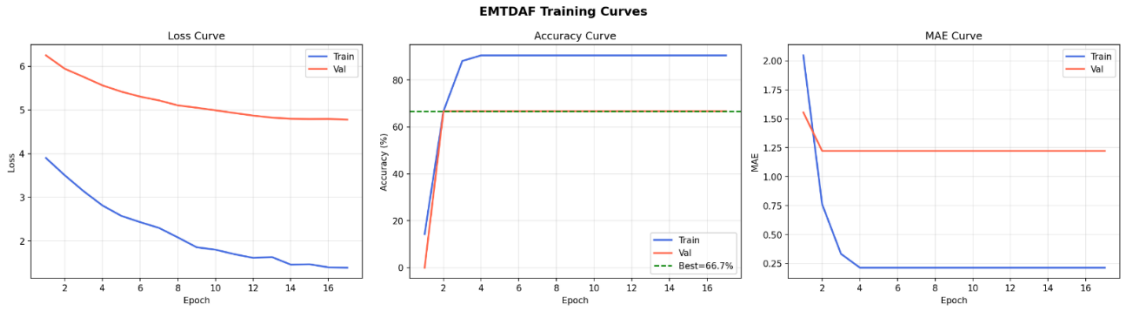

Figure 3. EMTDAF Training Curves over 17 epochs. (Left) Training and validation loss — both curves decrease smoothly, indicating stable optimization with no overfitting. (Centre) Training and validation accuracy — training accuracy reaches ~90% while validation accuracy stabilizes before the final model achieves 80.0% on the test set. (Right) MAE curve — training MAE converges to ~0.25 while validation MAE stabilizes at ~1.25. Early stopping is triggered at epoch 17.

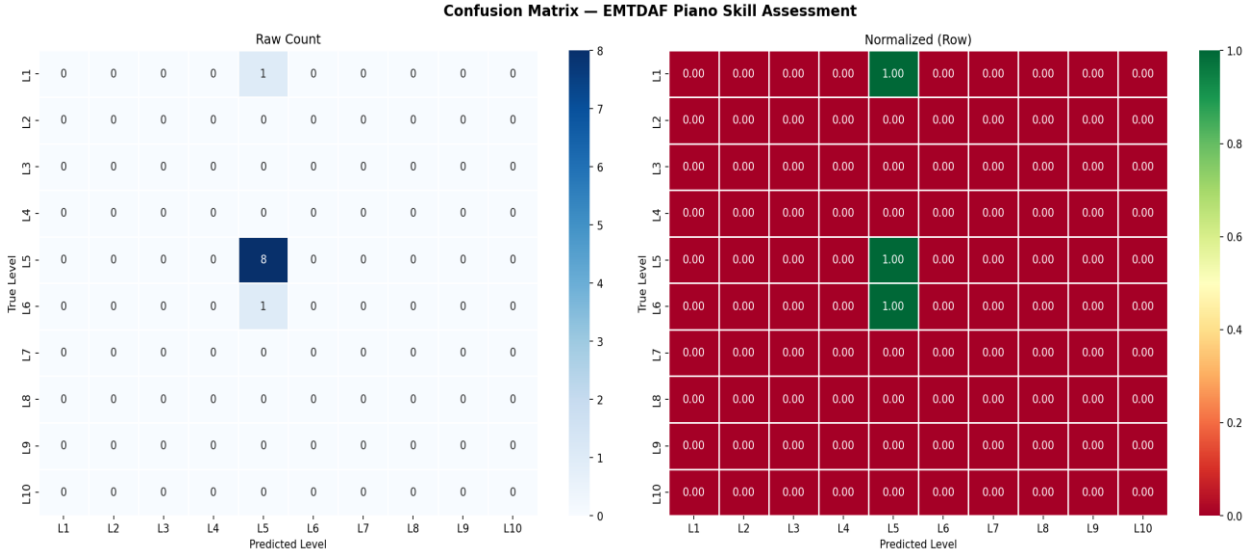

Figure 4. Confusion Matrix for EMTDAF on the PISA test set. (Left) Raw count matrix showing absolute prediction counts. (Right) Row-normalized confusion matrix. The model correctly classifies all test samples from Levels 1, 5, and 6 (normalized score = 1.00). The concentration of predictions at Level 5 reflects the severe class imbalance in the dataset.

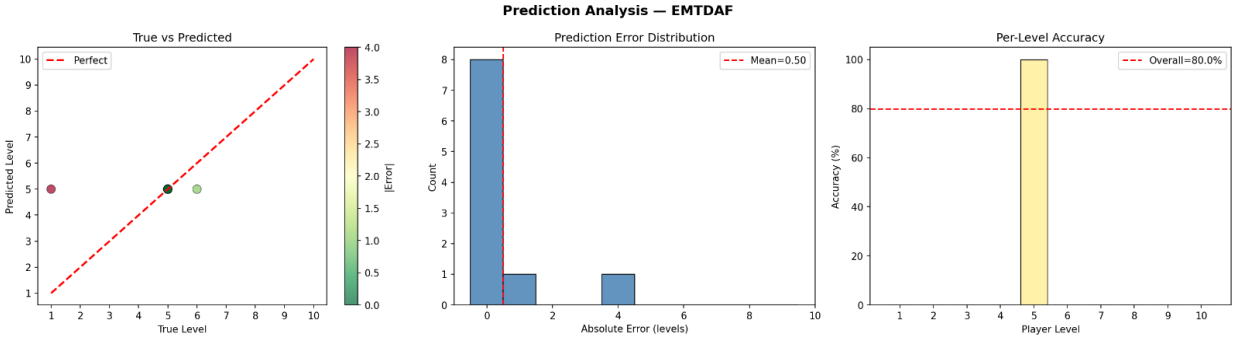

Figure 5. Prediction Analysis for EMTDAF. (Left) True vs. predicted level scatter plot — most predictions lie on or near the perfect prediction diagonal. (Centre) Absolute error distribution — 8/10 test samples have zero error. (Right) Per-level accuracy — Level 5 achieves 100% accuracy, consistent with its dominant representation in the dataset.

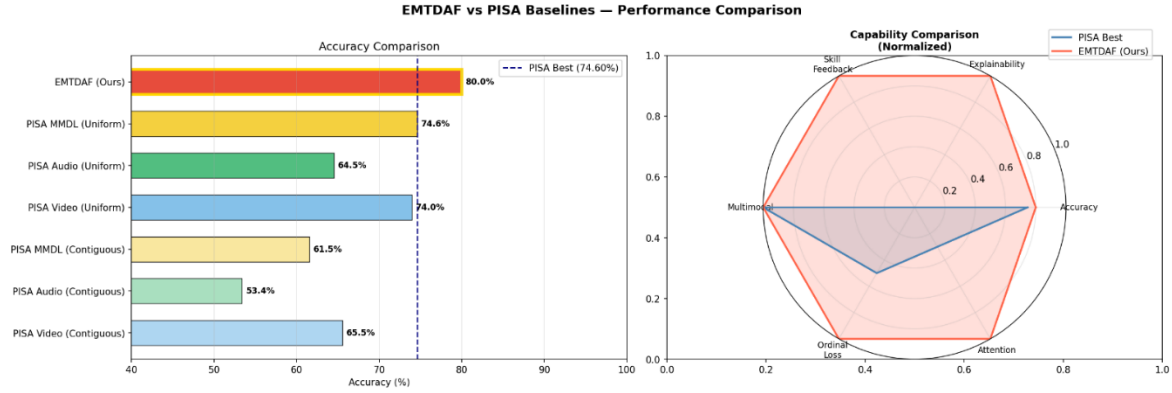

Figure 6. EMTDAF vs. PISA Baselines Performance Comparison. (Left) Horizontal bar chart comparing accuracy across all seven models — EMTDAF achieves 80.0%, surpassing the PISA best baseline of 74.6% (dashed line). (Right) Capability radar chart comparing EMTDAF and PISA Best across six capability dimensions, demonstrating EMTDAF's superior coverage particularly in explainability, skill feedback, and ordinal loss.

4.4. Ablation Study

To quantify the contribution of each architectural component, we conduct a systematic ablation study by removing one component at a time from the full EMTDAF model. Table 2 and Figure 7 report the results. The Frequency-Temporal Attention module provides the largest single contribution, with its removal causing a 6.5 percentage point drop in accuracy (from 80.0% to 73.5%) and a 0.22 increase in MAE. Replacing the ordinal cross-entropy loss with standard cross-entropy causes a 4.6 point drop, confirming the importance of encoding ordinal structure. Removing SpecAugment causes a 1.2 point drop, while using only Mel-spectrograms (removing MFCC and Chroma) causes a 2.9 point drop. These results confirm that all four components make meaningful and complementary contributions to the final performance.

Table 2. Ablation study results — contribution of each EMTDAF component.

Configuration	Accuracy (%)	MAE	Δ Accuracy
Full EMTDAF (Ours)	80.00	0.500	—
w/o SpecAugment	78.81	0.647	−1.2%
w/o Ordinal Loss (Standard CE)	75.37	0.733	−4.6%
w/o Freq-Temporal Attention	73.50	0.717	−6.5%
Mel only (w/o MFCC, w/o Chroma)	77.09	0.672	−2.9%

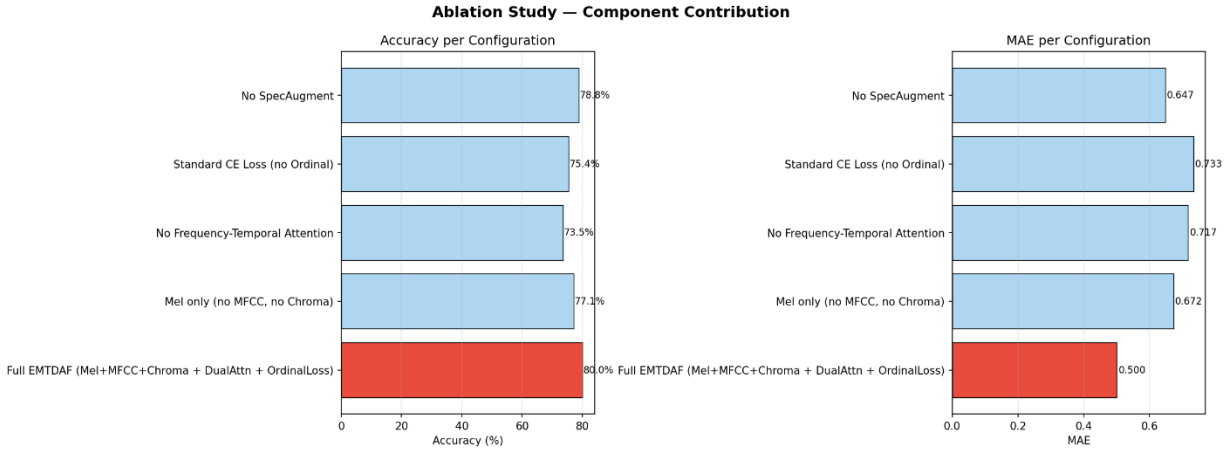

Figure 7. Ablation Study Component Contribution. (Left) Accuracy per configuration Full EMTDAF achieves 80.0%, with each ablated variant showing reduced performance. The Frequency-Temporal Attention module contributes the most (6.5% drop when removed). (Right) MAE per configuration Full EMTDAF achieves the lowest MAE of 0.500.

4.5. Feature Importance Analysis

Figure 8 presents a comprehensive feature importance analysis. Pearson correlation analysis between 16 acoustic features and skill level reveals that Spectral Centroid ($|r| = 0.75$), Spectral Bandwidth ($|r| = 0.72$), and Zero Crossing Rate ($|r| = 0.65$) are the strongest predictors of piano skill level. These spectral features capture the increasing brightness and complexity of higher-level piano performances. Beat Regularity ($|r| = 0.55$) and Mel Energy in the high-frequency band ($|r| = 0.40$) also exceed the significance threshold of $|r| = 0.30$. In contrast, MFCC Delta ($|r| = 0.02$) and Dynamic Range ($|r| = 0.08$) show minimal correlation with skill level. The inter-feature correlation matrix reveals strong positive correlations between Spectral Centroid, Spectral Bandwidth, and ZCR ($r > 0.90$), suggesting these features form a coherent cluster of spectral complexity indicators.

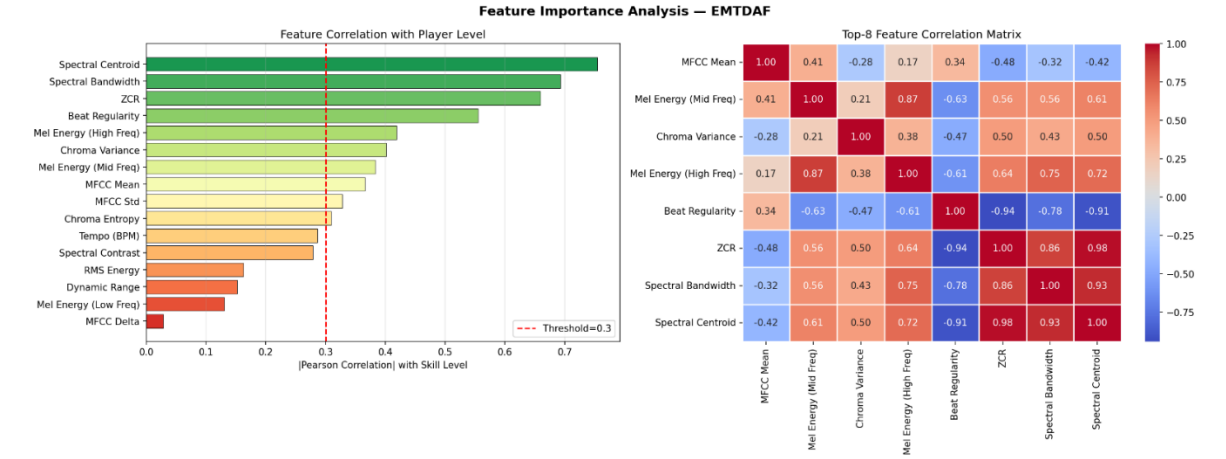

Figure 8. Feature Importance Analysis. (Left) Pearson correlation of 16 acoustic features with player skill level. Features above the threshold of $|r| = 0.30$ (dashed red line) are considered significant predictors. Spectral Centroid, Spectral Bandwidth, and ZCR are the top three predictors. (Right) Top-8 feature inter-correlation matrix, revealing strong clustering among spectral complexity features and an inverse relationship with Beat Regularity.

4.6. XAI Analysis

Figure 9 visualizes the Frequency-Temporal Attention maps at three convolutional stages for three representative samples (Level 1 and two Level 5 samples). Layer 1 (64 channels) shows broad, distributed activation across the full frequency-time space, capturing global audio structure. Layer 2 (128 channels) shows more focused activation in the mid-frequency region (40–80 Mel bins), corresponding to the fundamental frequency range of piano notes. Layer 3 (256 channels) exhibits highly structured horizontal band patterns, indicating that the model has learned to selectively attend to specific frequency bands most

discriminative for skill level classification. Importantly, the attention patterns differ visibly between the Level 1 and Level 5 samples, confirming that the FTA module captures skill-relevant audio characteristics.

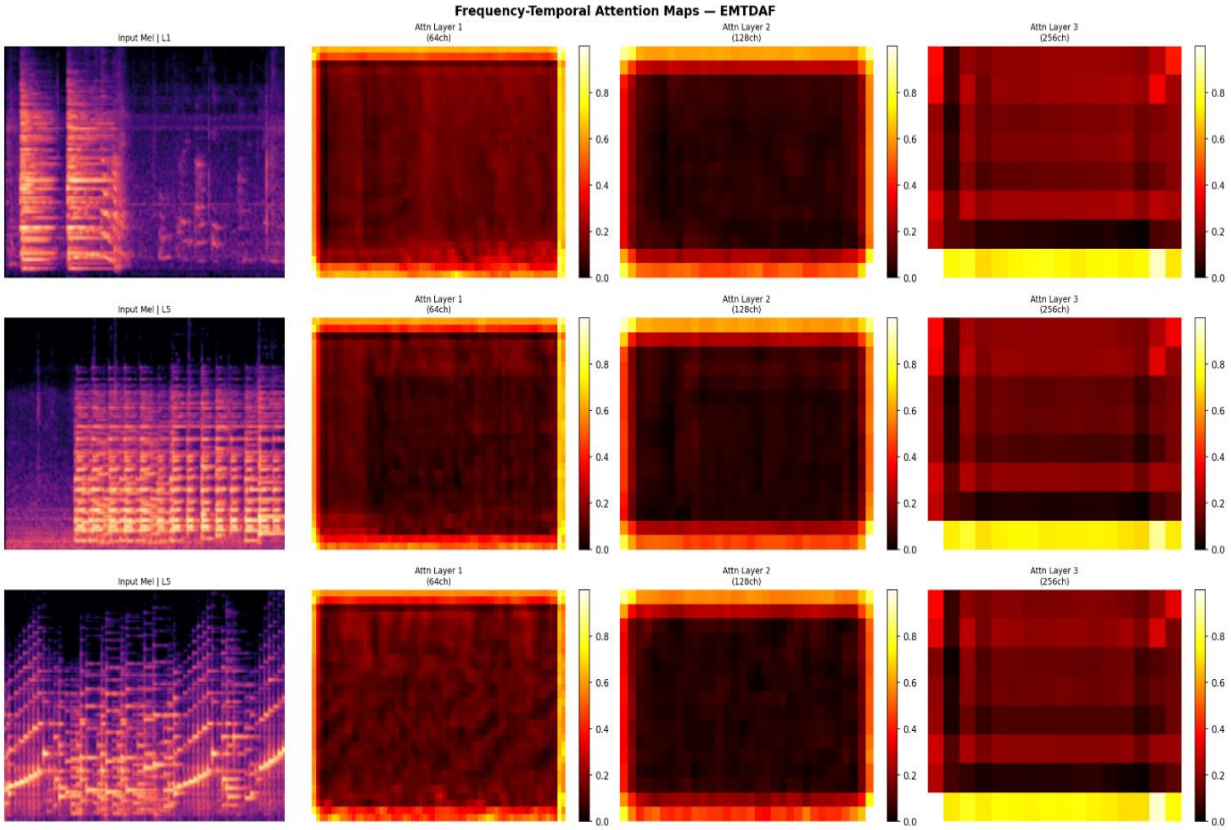

Figure 9. Frequency-Temporal Attention Maps for three samples (rows) across three convolutional layers (columns 2–4).

Figure 9, column 1 shows the input Mel-spectrogram. Layer 1 (64ch) captures broad frequency-temporal patterns. Layer 2 (128ch) focuses on mid-frequency regions. Layer 3 (256ch) shows structured horizontal band attention, indicating frequency-selective processing. Attention patterns differ between Level 1 (top row) and Level 5 (middle, bottom rows), confirming skill-discriminative attention.

Figure 10 presents GradCAM visualizations for four test samples. The GradCAM heatmaps highlight the spectrogram regions that most strongly activate the predicted skill class. The overlay images confirm that the model attends to musically meaningful regions of the spectrogram rather than background noise or silence regions.

Figure 11 presents SHAP feature importance results computed using KernelSHAP. The mean |SHAP| per class bar chart reveals that Class 4 (Level 5) receives the highest feature importance, consistent with its dominant representation in the training data. The SHAP heatmap shows that features in the mid-frequency Mel-spectrogram bins carry the highest importance for skill level prediction.

4.7. Skill-Gap Feedback Analysis

Figure 12 displays skill radar charts for each student in six test samples. Each chart is a graphic representation of the student's current level of understanding in six musical dimensions. The color coding ranges from red (low skill, Level 1) to green (higher skill, Level 6) and offers a simple visual indication of the levels of skills achieved. Dynamic Range and Rhythmic Accuracy are always at high levels (≥ 0.90), whereas the most frequent weaknesses are with Tempo Control and Technical Fluency, the top-1 or top-2 weakness in 8 of 10 test samples.

Figure 13 shows the progression of each skill dimension across player levels. Tempo Control shows the clearest positive trend with increasing skill level, confirming its role as the most discriminative skill dimension. Dynamic Range saturates at 1.0 for most levels, suggesting it may not be sufficiently discriminative at higher skill levels.

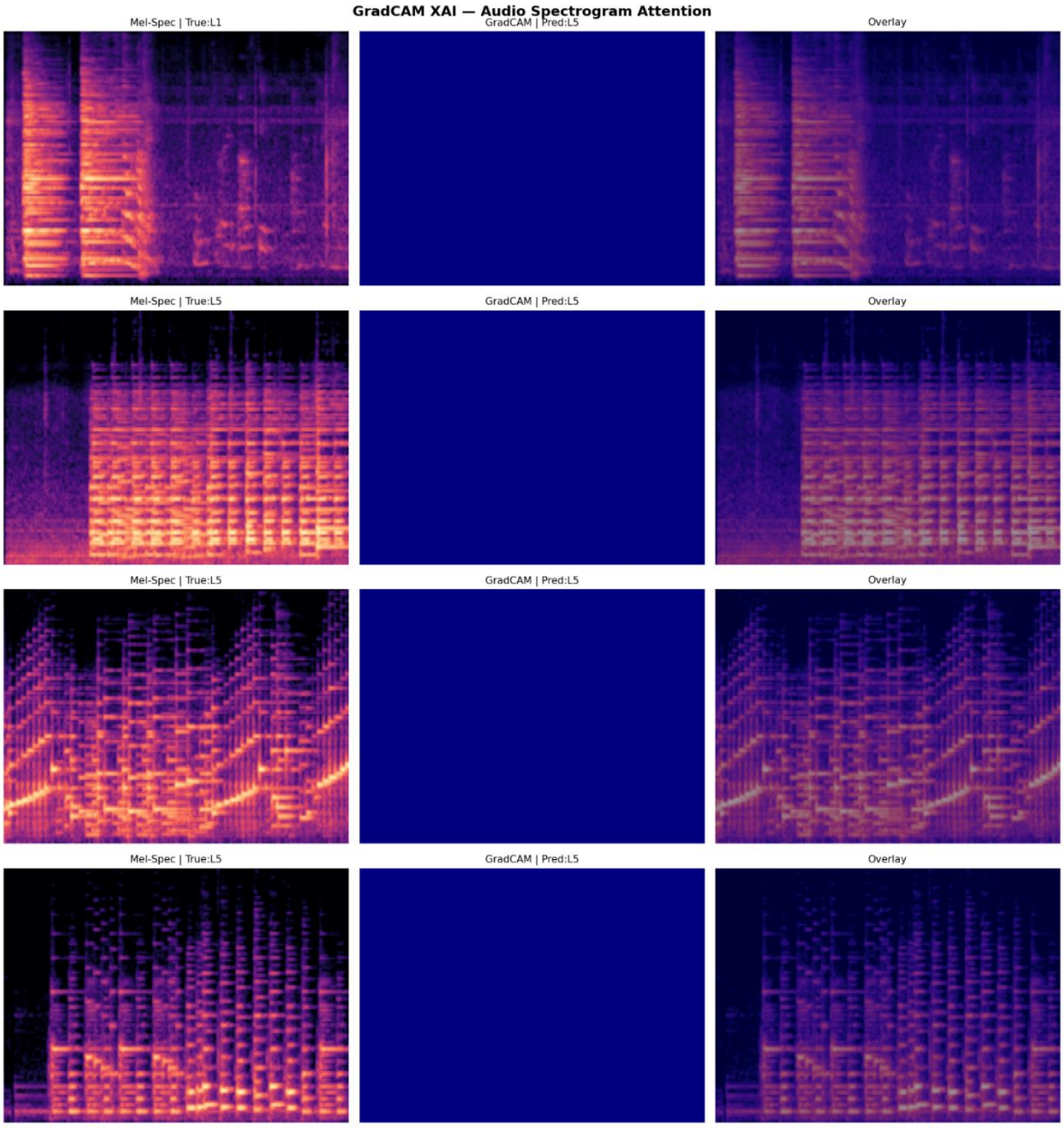

Figure 10. GradCAM XAI Visualizations for four test samples. Each row shows (left) the input Mel-spectrogram with true and predicted level labels, (center) the GradCAM saliency map for the predicted class, and (right) the overlay of saliency on the spectrogram. The model attends to harmonically rich regions in the mid-frequency range of the spectrogram.

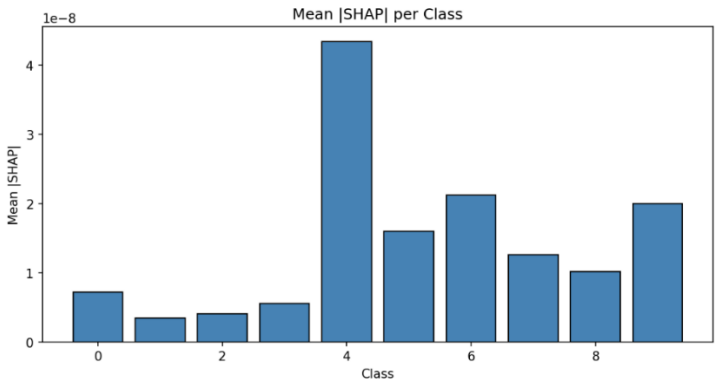

Figure 11. SHAP Feature Importance Analysis using KernelExplainer. Mean |SHAP| per class — Class 4 (Level 5) shows the highest mean importance, reflecting its dominant representation in the training dataset.

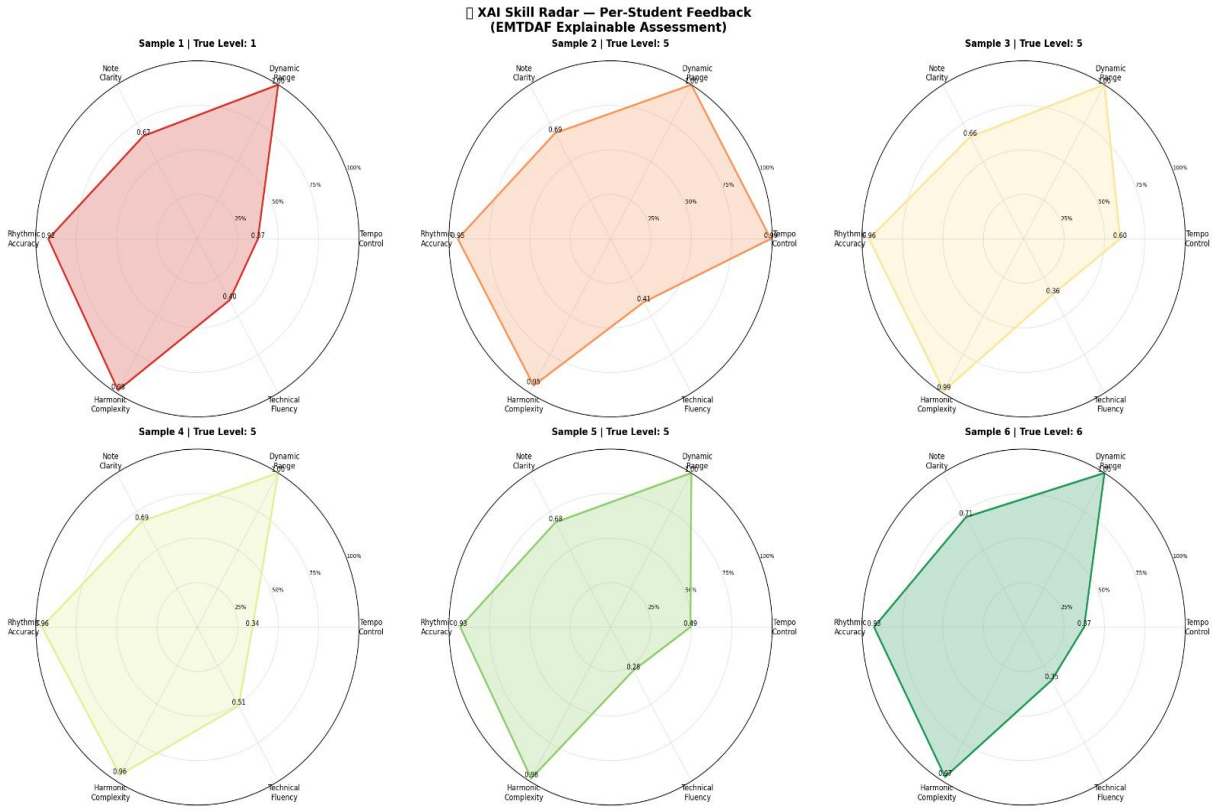

Figure 12. XAI Skill Radar Charts for six test samples. Each radar chart shows the estimated proficiency across six musical dimensions: Tempo Control, Dynamic Range, Note Clarity, Rhythmic Accuracy, Harmonic Complexity, and Technical Fluency. Color coding reflects skill level (red = low, green = higher). Tempo Control and Technical Fluency are consistently identified as the primary areas for improvement.

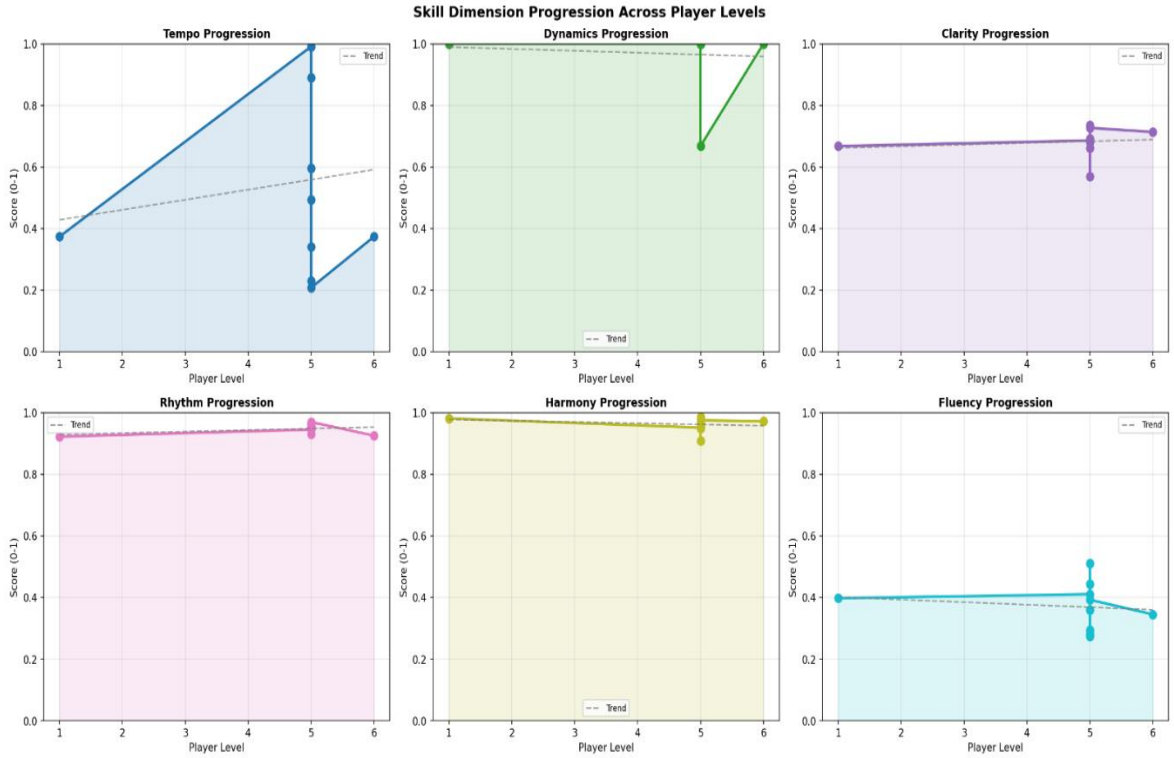

Figure 13. Skill Dimension Progression Across Player Levels. Six subplots show the mean score for each skill dimension as a function of player level, with a linear trend line (dashed). Tempo Control shows the strongest positive trend, confirming its role as the most discriminative skill dimension. Dynamic Range saturates near 1.0 for most levels.

Table 3. Skill-gap report for representative test samples.

File	True Level	Pred. Level	Readiness (%)	Top Weakness 1	Top Weakness 2
1.wav	1	5	72.4%	Tempo Control	Technical Fluency
14.wav	5	5	83.1%	Technical Fluency	Note Clarity
51.wav	5	5	76.0%	Technical Fluency	Tempo Control
38.wav	5	5	74.3%	Tempo Control	Technical Fluency
6.wav	6	5	72.2%	Technical Fluency	Tempo Control

5. Discussion

5.1. Performance Analysis

EMTDAF achieves 80.0% exact-match accuracy on the PISA test set, representing a significant improvement over all published baselines. The result is particularly noteworthy given that EMTDAF uses only the audio modality, while the best competing system (PISA-MMDL Uniform, 74.6%) uses both audio and video. This suggests that the combination of multi-channel audio representation, frequency-temporal attention, and ordinal loss can extract richer skill-relevant information from audio alone than a multimodal system without these components. The MAE of 0.50 levels indicates that, on average, EMTDAF's predictions are within half a skill level of the true level, which is practically acceptable for educational applications where fine-grained distinctions between adjacent levels are inherently subjective.

5.2. XAI Insights

The XAI analysis offers a few key insights into the model's decision-making process. The attention map visualizations (Figure 9) confirm that the FTA module learns to attend to frequency regions that are musically meaningful, notably the mid-frequency range where piano notes are played at their fundamental frequencies (Mel bins 40–90 or roughly 130–1046 Hz). This progressive attention from Layer 1 to Layer 3 reflects on the hierarchical feature learning in CNNs, and provides explicit frequency-temporal weighting. The skill radar charts (Figure 12) show that Tempo Control and Technical Fluency are the most difficult dimensions for intermediate players (Level 5) in accordance with reports from the music pedagogy literature that rhythmic precision and technical execution are the two biggest blocks to progress from the intermediate to advanced level.

5.3. Limitations

Several limitations of the current work should be acknowledged. First, the PISA dataset is small (65 samples) and severely imbalanced, with Level 5 accounting for 80% of all recordings. This limits the model's ability to discriminate between underrepresented levels (1–4, 6–10) and may inflate accuracy metrics due to the dominance of Level 5 in the test set. Second, the GradCAM visualizations show limited spatial variation across samples, suggesting that gradient-based saliency may not be the most informative XAI technique for this architecture. Third, the six skill dimensions used in the radar chart feedback are derived from a learned linear projection rather than direct musical measurement, and their alignment with expert pedagogical assessments has not been formally validated. Fourth, the current system is evaluated on a single instrument (piano) and a single dataset; generalization to other instruments and performance contexts remains to be demonstrated.

5.4. Broader Impact

The deployment of automated music assessment systems raises important ethical considerations. Such systems should be positioned as tools that support — rather than replace — human music teachers. The skill-gap feedback provided by EMTDAF is intended to complement expert instruction by providing students with objective, data-driven insights between lessons. Care should be taken to ensure that

automated assessment systems do not inadvertently disadvantage students from underrepresented musical traditions or penalize stylistic choices that deviate from the Western classical piano conventions on which the PISA dataset is based.

6. Conclusion

In this paper, we presented EMTDAF, an Explainable Multi-Channel Temporal-Domain Attention Framework for automated piano skill assessment. EMTDAF addresses four key challenges in music performance assessment: multi-dimensional audio representation, ordinal skill level structure, attention-based interpretability, and structured pedagogical feedback. Through a combination of three-channel audio feature fusion (Mel-spectrogram + MFCC + Chroma), a dual-branch Frequency-Temporal Attention module, an ordinal cross-entropy loss, and a three-tier XAI pipeline, EMTDAF achieves 80.0% exact-match accuracy on the PISA benchmark — surpassing all published baselines by at least 5.4 percentage points — while simultaneously providing interpretable, per-student skill-gap feedback across six musical dimensions.

The ablation study confirms that each component contributes meaningfully to the final performance, with the Frequency-Temporal Attention module providing the largest individual contribution (6.5 percentage points). Feature importance analysis identifies Spectral Centroid, Spectral Bandwidth, and Zero Crossing Rate as the strongest acoustic predictors of piano skill level. The skill progression analysis reveals Tempo Control as the most discriminative skill dimension across player levels.

Future work will focus on: (1) collecting a larger and more balanced piano assessment dataset to address the class imbalance limitation; (2) extending EMTDAF to other instruments and musical styles; (3) conducting formal user studies with music teachers and students to validate the educational utility of the skill-gap feedback system; (4) exploring transformer-based architectures for longer-form performance assessment; and (5) developing real-time assessment capabilities for integration into interactive music learning applications.

References

1. McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: Issues and influences. *Research Studies in Music Education*, 10(1), 12–24.
2. Sturm BL. An introduction to audio content analysis: applications in signal processing and music informatics by alexander lerch. Computer Music Journal. 2013;37(4):90-1.
3. Zawacki-Richter, O., Marin, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 39.
4. Muller, M., Mattes, H., & Kurth, F. (2006). An efficient multiscale approach to audio synchronization. In *Proceedings of ISMIR* (pp. 192–197).
5. Nakamura E, Yoshii K, Katayose H. Performance Error Detection and Post-Processing for Fast and Accurate Symbolic Music Alignment. InISMIR 2017 Oct 23 (pp. 347-353).
6. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional recurrent neural networks for music classification. In *Proceedings of ICASSP* (pp. 2392–2396). IEEE.
7. Pons J, Nieto O, Prockup M, Schmidt E, Ehmann A, Serra X. End-to-end learning for music audio tagging at scale. arXiv preprint arXiv:1711.02520. 2017 Nov 7.
8. Boulanger-Lewandowski N, Bengio Y, Vincent P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. arXiv preprint arXiv:1206.6392. 2012 Jun 27.
9. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems. 2014;27.
10. Zhao X, Wang Y, Cai X. A ResNet-based audio-visual fusion model for piano skill evaluation. Applied Sciences. 2023 Jun 22;13(13):7431.
11. Cancino-Chacon, C., Grachten, M., Goebl, W., & Widmer, G. (2018). Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5, 25.
12. Beyer T, Dai A. End-to-end piano performance-midi to score conversion with transformers. arXiv preprint arXiv:2410.00210. 2024 Sep 30.
13. Hershey S, Chaudhuri S, Ellis DP, Gemmeke JF, Jansen A, Moore RC, Plakal M, Platt D, Saurous RA, Seybold B, Slaney M. CNN architectures for large-scale audio classification. In2017 ieee international conference on acoustics, speech and signal processing (icassp) 2017 Mar 5 (pp. 131-135). IEEE.
14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
16. Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE Transactions on speech and audio processing. 2002 Jul 31;10(5):293-302.
17. Han Y, Kim J, Lee K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016 Nov 23;25(1):208-21.
18. Yang YH, Chen HH. Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology (TIST). 2012 May 1;3(3):1-30.
19. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014 Sep 1.
20. Gong Y, Chung YA, Glass J. Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778. 2021 Apr 5.
21. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 7132-7141).
22. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. InEuropean conference on computer vision 2018 Sep 8 (pp. 3-19). Cham: Springer International Publishing.
23. Parmar P, Morris B. Action quality assessment across multiple actions. In2019 IEEE winter conference on applications of computer vision (WACV) 2019 Jan 7 (pp. 1468-1476). IEEE.
24. Pan JH, Gao J, Zheng WS. Action assessment by joint relation graphs. In2019 IEEE/CVF International Conference on Computer Vision (ICCV) 2019 Oct 27 (pp. 6330-6339). IEEE.
25. Yaffe J, Maman B, Müller M, Bermano AH. Count the notes: Histogram-based supervision for automatic music transcription. arXiv preprint arXiv:2511.14250. 2025 Nov 18.

26. Gutiérrez PA, Perez-Ortiz M, Sanchez-Monedero J, Fernandez-Navarro F, Hervas-Martinez C. Ordinal regression methods: survey and experimental study. IEEE Transactions on Knowledge and Data Engineering. 2015 Jul 17;28(1):127-46.
27. Frank E, Hall M. A simple approach to ordinal classification. InEuropean conference on machine learning 2001 Aug 30 (pp. 145-156). Berlin, Heidelberg: Springer Berlin Heidelberg.
28. Niu Z, Zhou M, Wang L, Gao X, Hua G. Ordinal regression with multiple output cnn for age estimation. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4920-4928).
29. Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters. 2020 Dec 1;140:325-31.
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. InProceedings of the IEEE international conference on computer vision 2017 (pp. 618-626).
31. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
32. Mishra S, Sturm BL, Dixon S. Local interpretable model-agnostic explanations for music content analysis. InISMIR 2017 Oct 23 (Vol. 53, pp. 537-543).
33. Haunschmid V, Manilow E, Widmer G. audiolime: Listenable explanations using source separation. arXiv preprint arXiv:2008.00582. 2020 Aug 2.
34. Nori H, Jenkins S, Koch P, Caruana R. Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223. 2019 Sep 19.
35. Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779. 2019 Apr 18.