SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

DEEP LEARNING-BASED AUTOMATIC MUSIC TRANSCRIPTION USING CR-GCN

By

126003238 Selvakarthik S
126003241 Shanthosh Kumar R
126003218 Rithvik L

Project Guide – Dr. Emily Jenifer A

Motivation

- Music is a major stress-buster for billions around the world. While we listen to millions of songs each year, certain tunes stay with us in our minds.
- Musicians often transcribe these melodies into music sheets to aid in composition and performance.
- Our goal is to convert a piece of music into its corresponding notes using a CR-GCN model.
- To make this process accessible to everyone, we plan to develop a web application where users can upload a music file, and our model will generate the corresponding music notes, which will be displayed as a music sheet within the application.

Objective

- To develop an approach to convert an audio file (music) into its corresponding musical notes.
- To identify and leverage efficient feature selection algorithms.
- To select and utilise suitable classification algorithms.
- To construct a model with the highest possible accuracy through rigorous training and testing on large datasets.
- To deploy the model as a back-end in an user-friendly web application.

Base paper

- [1] Xiao, Z., Chen, X., & Zhou, L. (2023). Polyphonic piano transcription based on graph convolutional network. Signal Processing, 212, 109134.
- Doi : https://doi.org/10.1016/j.sigpro.2023.109134

Problem Statement

- To design an Automatic Music Transcription (AMT) system that accurately converts complex polyphonic audio signals into symbolic music representations by capturing note interdependencies and temporal dynamics.
- To deploy the model as a back-end in an user-friendly web application.

ABSTRACT

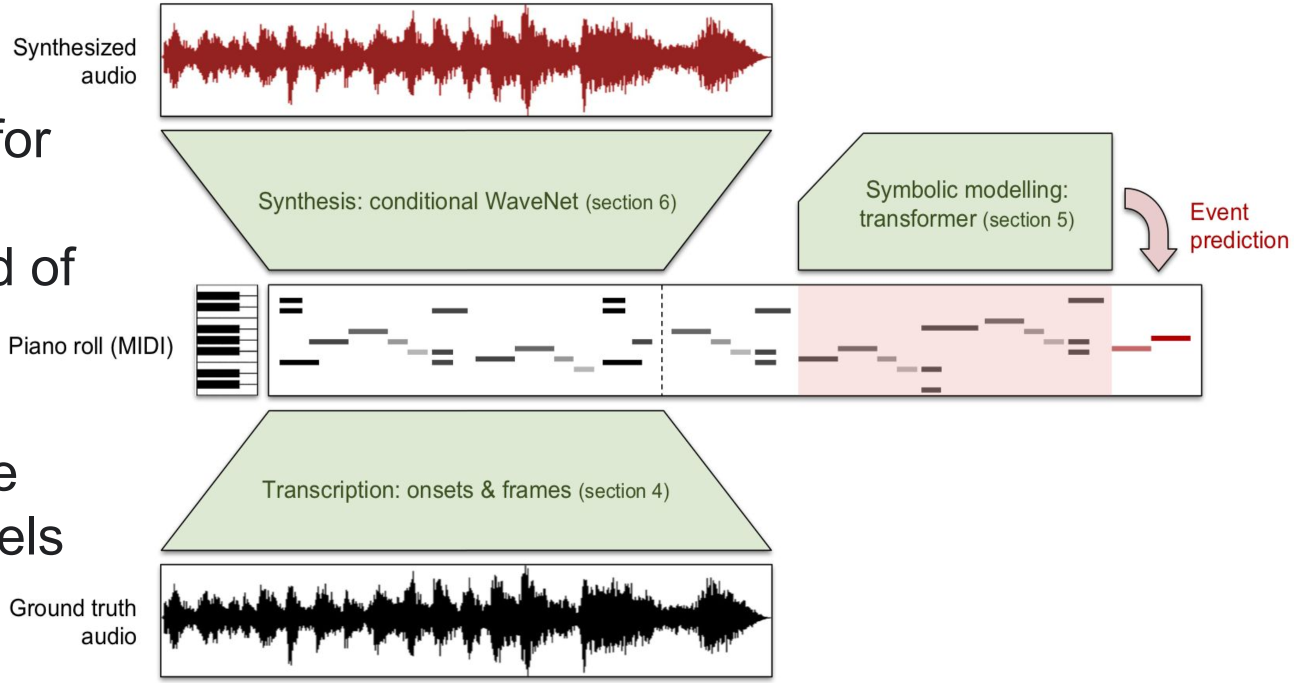
- The task of automatic music transcription (AMT) mainly focuses on converting audio signals to symbolic music representations, facilitating applications such as computational musicology and music analysis.
- One of the biggest problems is when multiple notes are played at the same time, dimension explosion can happen which makes it difficult for accurate music note transcription.
- To overcome this challenge, we have proposed a hybrid deep learning architecture combining Convolutional Neural Network for spatial feature extraction, bidirectional LSTMs or self-attention mechanisms for precise temporal note-level predictions and Graph Convolutional Network for accurate label learning to capture note interdependencies in polyphonic music.
- Experiments on public datasets like MAESTRO, MAPS, GiantMIDI show that the proposed methodology with F1-score of 96.88% is much more superior than existing methodologies like Onset and Frames, Wavenet, Non-Negative Matrix Factorization (NMF).
- The generated music sheets validate the model's accuracy and practical applicability, providing a valuable tool for musicians and researchers.
- By addressing the limitations of prior methods, the proposed approach CR-GCN (Channel Relationship-Based Graph Convolutional Network) represents a step forward in automated transcription technology, making it feasible for large-scale and real-time applications.

LITERATURE

TITLE	MERITS	DEMERITS
A Data-Driven Analysis of Robust Automatic Piano Transcription	The study improved note-onset accuracy to 88.4 F1-score on the MAPS dataset through data augmentation.	Performance on out-of-distribution annotated piano data indicates challenges in generalizing to unseen data.
Automatic Piano Sheet Music Transcription with Machine Learning	BiLSTM architecture is identified as one of the effective for automatic piano music transcription, achieving a top F1-score of 74.80%.	CNNs underperform significantly in music transcription tasks, achieving only an F1-score of 22.85% despite extensive hyper-parameter tuning.
Research on the Recognition of Piano-Playing Notes by a Music Transcription Algorithm	The CRNN algorithm combines CNN and BiLSTM, achieving impressive F1-scores of 84.90%, 92.24%, and 79.27% for frames, notes, and offsets.	The study's results have limited generalizability due to small datatset, with further exploration of different piano note features needed to enhance recognition accuracy.
Multimodal Image and Audio Music Transcription	The multimodal framework combining OMR and AMT improves transcription accuracy by reducing errors up to 40% for more accurate music transcription.	The proposed framework can degrade overall transcription accuracy if either OMR or AMT already performs near-perfectly. Overall Transcription may degrade
Automatic Piano Music Transcription Using Audio-Visual Features	This research uses audio-visual features to improve piano music transcription accuracy by 12.69%, surpassing audio-only systems.	The system's dependency on specialized equipment, like an overhead camera, limits its practicality for general use

DATASET PREPARATION

- **Datasets Used:**
  - **The MAESTRO Dataset**
    - MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) is a dataset composed of about 200 hours of virtuosic piano performances.
    - The audio files are captured with fine alignment (~3 ms) between note labels and audio waveforms.



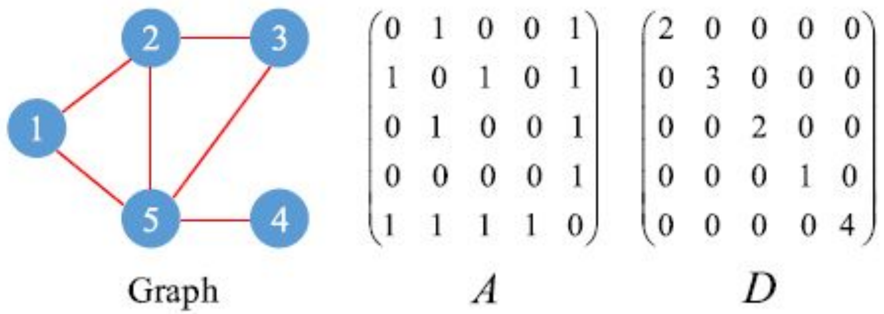
  - **The MAPS Dataset**
    - MAPS (MIDI Aligned Piano Sounds) , is a piano sound dataset dedicated to research on multi-F0 estimation and automatic transcription.
    - The audio is composed of about 31 GB of CD-quality recordings in .wav format.
    - The audio is obtained by means of Virtual Piano softwares and a Yamaha Disklavier, nine settings of different pianos and recording conditions were used.
    - MAPS is freely released with a Creative Commons license.

DATASET PREPARATION

- **Preprocessing:**
  - Parsed MIDI data by extracting note onset times, pitches and velocities.
  - Normalized spectrogram values and applied log-scaling for better learning.
  - Converted MIDI note timestamps to a frame-based representation matching the spectrogram to ensure precise time alignment between MIDI and audio frames.
  - Split audio into fixed-length chunks for efficient training and testing.
- **Data Augmentation:**
  - Transposed MIDI notes within a limited range (e.g., ±2 semitones) and varied note velocities and timings for diversity.
  - Applied time-stretching, pitch-shifting, and added noise to WAV files to improve model robustness.
- **Dataset Characteristics:**
  - The datasets tested specific challenges, such as variations in handling tempo fluctuations, segmenting long recordings, and ensuring musical integrity.

EXISTING SYSTEM

- **Key Features:**
  - Combination of CNN, RNN, GCN
  - Two Stage Learning - Feature Learning and Label Learning
  - Graph Representation of Notes
  - Mapped Classifier with Dot Product



- **CR-GCN (Channel Relationship-Based Graph Convolutional Network)**
  - CNN for spatial features, LSTM for temporal dependencies, and GCN for modeling relationships between notes in a graph-based format.
- **Performance:**
  - **Datasets & Accuracies:**
    - MAESTRO (v2) : Precision - 97.38%, Recall - 96.21%, F1-Score - 96.88%
    - MAPS Dataset : Precision - 84.30%, Recall - 84.65%, F1-Score - 84.48%
- **Advantages:**
  - Joint Feature and Label Learning
  - Scalable to Multi-Label Tasks
  - Improved Performance for Complex Music and Generalizable to Other Domains

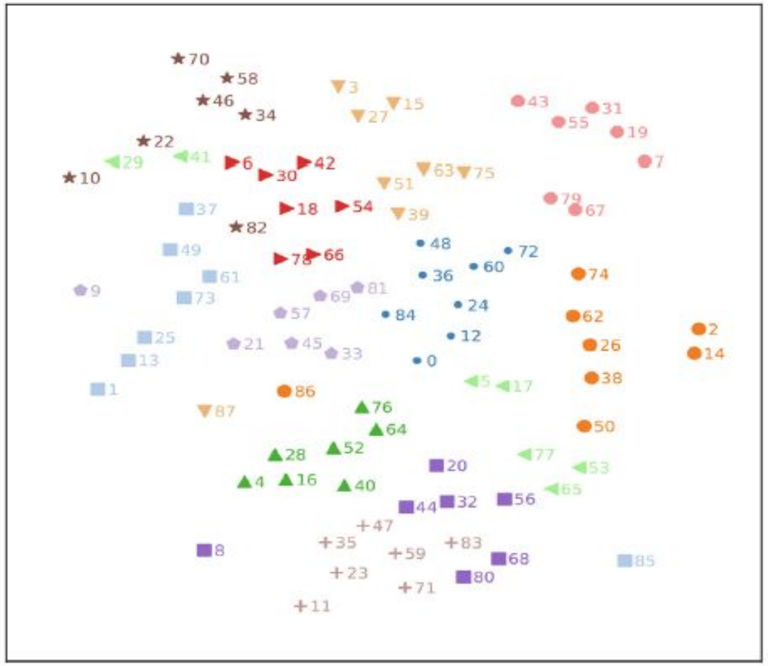
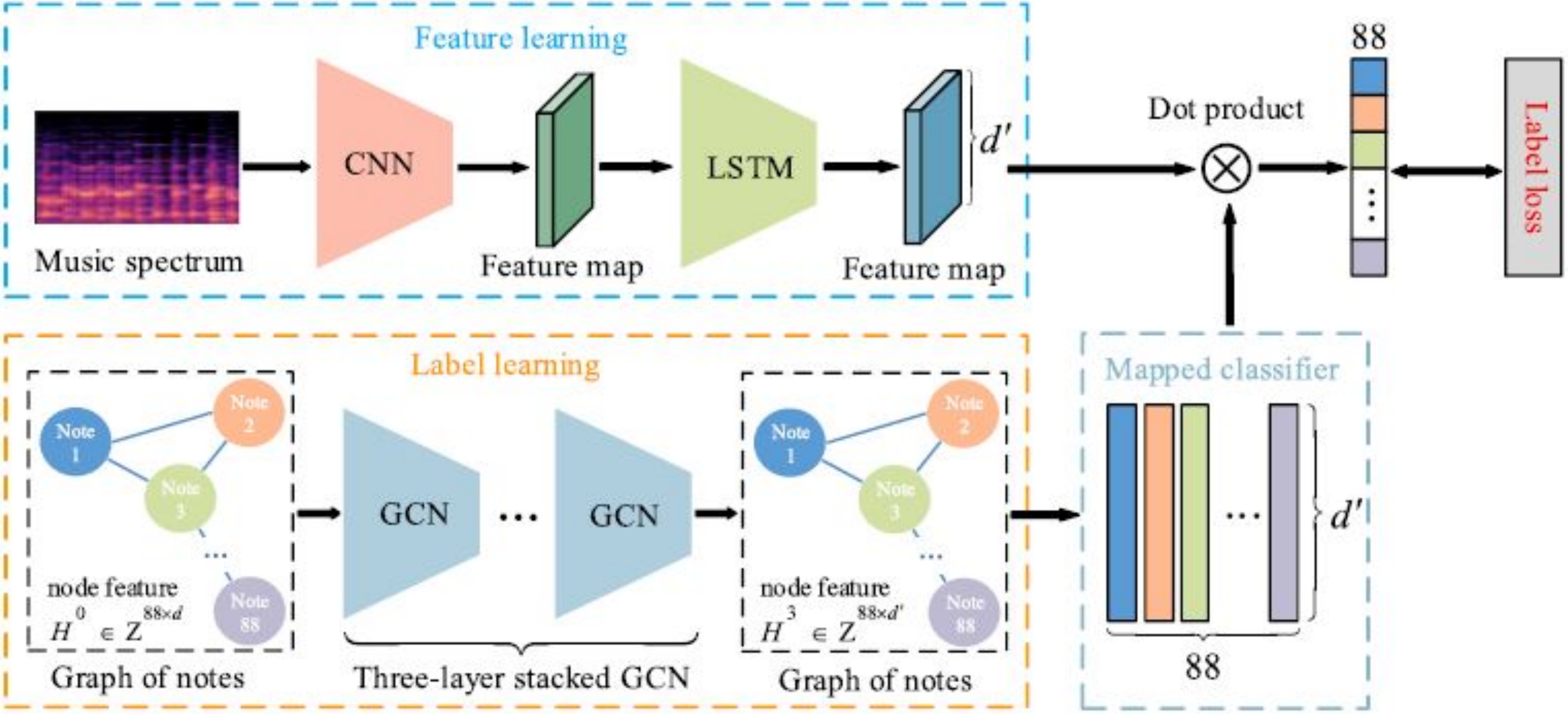
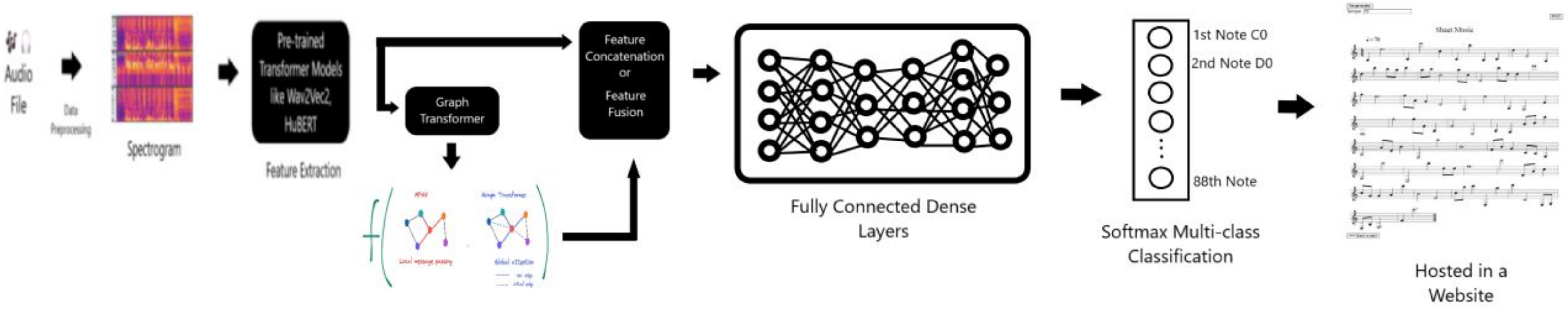


Fig. 8. Visualization of the CR-GCN model based on t-SNE.

EXISTING SYSTEM



PROPOSED SYSTEM (OPTIONAL)



- We present to you a **novel hybrid model** that integrates **pretrained audio Transformers (Wav2Vec2, HuBERT, AST) for feature extraction** with **Graph Transformers (Graphormer, GAT) to model note dependencies**, followed by **fusion techniques** and a **Softmax classifier** for automatic music transcription.
- This approach effectively captures both **temporal and structural musical relationships**, enhancing note prediction accuracy

CONCLUSION

The objective of the project is to integrate the techniques of CNN, RNN, GCN for effective Automatic Music Transcription (AMT). The accuracy outcome of Bi-LSTM, CRNN, CNN and other existing models were analyzed and compared. From the obtained results, we can infer that CR-GCN Model has outperformed other models in terms of metrics and has shown excellent generalization capabilities to unseen data. This highlights the potential of graph-based learning in enhancing music transcription tasks. Future work can focus on refining the model architecture and incorporating additional musical features for further improvements.

REFERENCES

1 Xiao, Z., Chen, X., & Zhou, L. (2023). Polyphonic piano transcription based on graph convolutional network. *Signal Processing*, *212*, 109134.

2 Edwards, D., Dixon, S., Benetos, E., Maezawa, A., & Kusaka, Y. (2024). A Data-Driven Analysis of Robust Automatic Piano Transcription. *IEEE Signal Processing Letters*.

3 Saputra, F., Namyu, U. G., Vincent, D. S., & Gema, A. P. (2021). Automatic piano sheet music transcription with machine learning. *Journal of Computer Science*, *17*(3), 178-187.

4 Guo, R., & Zhu, Y. (2025). Research on the Recognition of Piano-Playing Notes by a Music Transcription Algorithm. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *29*(1), 152-157.

5 de la Fuente, C., Valero-Mas, J. J., Castellanos, F. J., & Calvo-Zaragoza, J. (2022). Multimodal image and audio music transcription. *International Journal of Multimedia Information Retrieval*, *11*(1), 77-84.

6 Wan, Y., Wang, X., Zhou, R., & Yan, Y. (2015). Automatic piano music transcription using audio‐visual features. *Chinese Journal of Electronics*, *24*(3), 596-603.


Thank
you