Automated Music Sheet to MIDI Conversion and Spectrogram Analysis for Multi-Instrument and Multi-Genre Composition

Chukkapalli Ramya[1], Shaik Shaheena[2] and Garnepudi Parimala[3]
{ramyachukkapalli66@gmail.com[1], skshaheena0311@gmail.com[2], garnepudi.parimala@gmail.com[3]}

Department of CSE, VFSTR, Vadlamudi, Guntur, Andhra Pradesh, India[1, 2]
Assistant Professor, Department of CSE, VFSTR, Vadlamudi, Guntur, Andhra Pradesh, India[3]

Abstract. Automated music transcription has emerged as a crucial field in Optical Music Recognition (OMR) and digital signal processing, enabling seamless conversion of sheet music into machine-readable formats. This paper presents a comprehensive system for automated music sheet-to-MIDI conversion and spectrogram analysis, leveraging advanced image processing techniques, deep learning models, and MIDI synthesis algorithms. The proposed framework first processes scanned sheet music by detecting and extracting musical symbols using contour analysis, morphological operations, and the Hough Transform. A deep learning-based approach is employed for note head recognition, pitch estimation, and rhythm extraction, ensuring accurate musical interpretation. The extracted notes are converted into MIDI format, which is further synthesized into WAV audio for playback and spectrogram visualization. The system incorporates a user-defined playback duration feature, optimizing real-time applications for composers, researchers, and musicians. The generated spectrogram provides a time-frequency representation, facilitating detailed harmonic analysis. This frame- work integrates computer vision, artificial intelligence, and digital signal processing to enhance music transcription, performance evaluation, and AI-assisted composition.

Keywords: Optical Music Recognition (OMR), Deep Learning, MIDI Generation, Image Processing, Spectrogram Analysis, Digital Signal Processing (DSP), Music Transcription, Computer Vision.

1 Introduction

Across many genres and cultures, music notation has been essential to the preservation and transmission of musical compositions. Although traditional sheet music is a vital tool for composers and musicians, its dependence on human interpretation restricts its adaptability, accessibility, and compatibility with contemporary digital music technologies. The development of digital signal processing, artificial intelligence (AI), and optical music recognition (OMR) has created new opportunities for automating the transformation of sheet music into machine-readable, structured formats. The goal of this project is to bridge the gap be- tween digital music processing and traditional notation by developing an Automated Music Sheet to MIDI Conversion and Spectrogram Analysis for Multi-Instrument and Multi-Genre Composition.

A popular standard for expressing musical data, MIDI (Musical Instrument Digital Interface)

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357778

enables smooth interaction with virtual instruments, digital audio workstations (DAWs), and music analysis software. In contrast to audio recordings, MIDI allows for flexible editing and manipulation by storing performance data such note pitch, length, velocity, and instrument selection. In computational musicology, automated transcription, and digital music creation, the ability to transform sheet music into MIDI offers substantial benefits. Additionally, spectrogram analysis helps with instrument identification, performance evaluation, and music genre classification by providing a visual depiction of a composition's harmonic structure.

Accurately extracting and converting musical elements from sheet music is still difficult because of differences in notation methods, staff-line distortions, and symbol complexity, even with advances in music recognition and digital transcription. Current OMR techniques frequently have trouble detecting notes and interpreting rhythms in handwritten or deteriorated sheet music. To accomplish accurate note extraction and transcription, this study suggests a combined strategy that makes use of deep learning models, image processing techniques, and MIDI generating algorithms. The system's broad applicability in music education, performance analysis, and AI-assisted composition is ensured by its capacity to accommodate multi-instrument compositions across a variety of musical genres.

The goal of this project is to improve the usability and accessibility of musical compositions in digital contexts by automating the conversion process from music sheet to MIDI. To extract significant musical aspects, the produced MIDI files can be further processed, synthesized into audio, and examined via spectrogram visualization. By offering a strong foundation for music transcription, synthesis, and genre-specific adaption, this study advances computational music analysis. The study's conclusions have important ramifications for digital archiving of musical heritage, music technology, and artificial intelligence in music.

2 Literature Survey

I. -C. Wei, C. -W. Wu and L. Su in [1] employed CRNN, fully convolutional networks, Sample CNN, and WaveNet-style architectures respectively for drum transcription automatically. The models performed accurately in transcribing major drum components such as kick, snare, and hi-hat based on large-scale audio-to-MIDI aligned datasets. However, the draw- backs are generalization problems in noisy or polyphonic settings, excessive computational cost, and lower accuracy on uncommon drum events or non-standard drum kits.

J. Ling and M. Fu in [2] introduced a ResNet-18 inspired deep neural network architecture for precise audio-to-MIDI conversion. The model was trained to process mono-instrument tracks and showed remarkable performance, with over 90 percent accuracy on clean and well-aligned datasets. It successfully employed convolutional layers to extract temporal and spectral features from the audio input and project them to MIDI representations. Nonetheless, the model was limited when used with noisy or polyphonic music, where overlapping sounds of instruments led to transcription errors. Moreover, the method also had difficulty generalizing across various musical genres and unseen instruments, and demanded consider- able computational resources for training and inference.

P. J. Donnelly and V. Ebert in [3] The authors suggested a system for polyphonic piano music transcription to MIDI via deep learning. They used Convolutional Neural Networks (CNNs)

and Bidirectional Long Short-Term Memory (BiLSTM) networks to recognize note onsets and categorize pitches. The model had 94 percent precision and 93 percent recall. One major limitation is that the system is particularly designed for piano sound, and potentially won't generalize as well to other instruments or multi-genre audio data since the training data is mostly piano-dominated.

D. M. Dhanalakshmy, H. P. Menon and V. Vinaya In [4] Musical Notes to MIDI Conversion The process of translating musical notes from sheet music to MIDI includes extracting musical symbols and interpreting them from images. Conventional techniques depend on image processing methods such as staff line removal, symbol segmentation, and note recognition. This paper introduces a system that improves the process by using projection pro- files, morphological operations, and integration of the DSP toolbox. Even with 88 percent accuracy, difficulties remain in dealing with complicated notations and handwritten scores. Future enhancements may tap into deep learning to enhance automation and precision.

N. J. Sieger and A. H. Tewfik in [5] presented a model for converting audio signals into MIDI format based on deep learning and signal processing methods. The model combined an STFT for pre-processing and CNNs for pitch derivation and instrument recognition. The performance of the system achieved 78 % of pitch extraction and 81% of instrument identification. However, the method had weaknesses in handling overlapping notes, background noise, and polyphony, which affected the quality of the resulting transcription and MIDI output consistency.

Z. Nie and S. Yang in [6] proposed a method to convert Erhu music in the form of WAV into MIDI format using methods of signal processing. Feature extraction was performed using Short-Time Fourier Transform (STFT) and note extraction with a pitch extraction algorithm such as YIN for exact notes. The technique attained high accuracy in pitch detection for Erhu music due to the monophonic nature of the instrument. But the system was limited in its ability to tackle polyphonic or multi-instrumental pieces, and had problems with noise interference in poor-quality recordings.

G. Chen and W. Zhang in [7] suggested a watermarking algorithm to insert MIDI information directly into digital music scores by quantization-based methods. They utilized musical note characteristics to preserve transparency and utilized a watermark key to guarantee robustness. The algorithm demonstrated good accuracy in preserving watermark information after printing and scanning. It is, however, restricted to printed or scanned music scores and will not generalize to complicated multi-instrument sheets or changing font styles.

M. Marolt In [8] proposed a neural network-based method for polyphonic piano music transcription using bidirectional long short-term memory (BLSTM) models. Their architecture simulates both frame-wise pitch detection and note onset/offset prediction and obtains remarkable improvements over conventional methods. The model was tested on the MAPS dataset and obtained an F-measure of around 76.4 percent for note transcription. While the system works well with known data sets, limitations involve sensitivity to unheard recording conditions and dependence on supervised learning, which needs labeled data. Additionally, real-time transcription is still challenging due to computational complexity.

S. Lui, A. Horner and L. Ayers in [9] suggested a transcoding algorithm for translating

standard MIDI files to SP-MIDI by applying phrase stealing methods intelligently. They employed heuristics and optimization techniques to determine which notes to omit while preserving the musical content. Their approach attained more than 90 percent fidelity in perceptual quality tests but did not support dynamic polyphony and real-time requirements, which restricted its application in live playback systems or more intricate compositions.

Z. Huang and C. Zhou in [10] introduced the Sci-MIDI toolbox, an extensive Python-based platform for symbolic music research, including support for MIDI parsing, corpus filtering, feature extraction, and MIDI generation. Although the research was not specifically on deep learning models, it allowed for the creation of tailored machine learning pipelines for music information retrieval. The toolbox is open-source and modular, and it enables reproducible research. Nevertheless, it does not have inherent support for newer neural architectures such as transformers or GANs for symbolic music creation and must be integrated manually by users. Also, the system only supports MIDI-based symbolic data and does not accept audio or scanned sheet music input.

K. N. Kim, U. P. Chong and J. H. Choi in [11] suggested an algorithm for format conversion of CD-DA (Compact Disc Digital Audio) into MIDI without compromise on sound quality. They first converted the sound into WAVE format, extracted pitch, duration of notes, and velocity in an autocorrelation-based procedure. The process worked fairly well for solo notes of piano and produced nearly correct pitch (for example, 436–445 Hz for A4 vs. standard 440 Hz). Yet, when dealing with multiple simultaneous notes (such as C4 and A4), the model struggled with interfering frequencies, confining polyphonic transcription precision. The research is promising but presently limited to monophonic piano music.

E. Donati and C. Chousidis In [12] proposed a CNN-based model for voice classification with 92 percent accuracy in separating vocal activities but was not able to handle background noise interference in real-time scenarios. LSTM-based system for MIDI conversion with 89 percent accuracy but had issues with polyphonic vocal inputs. Kumar et al. hybrid CNN- RNN model with 94 percent accuracy but had latency issues on low-resource hardware. Williams et al. investigated a Transformer-based method for AI voice classification with a 95 percent accuracy but the heavy computational requirements restricted usage on low-power embedded platforms.

E. A. Suarez Guarnizo and L. M. Ramirez Rios in [13] investigated different percussion sound synthesis approaches and gesture identification in MIDI controllers. Rossing studied the acoustics of percussion instruments without concentrating on real-time processing. Wanderley and Depalle utilized motion tracking models for gesture perception with high accuracy but suffered from latency and calibration. Kapur used machine learning classifiers such as SVM and Neural Networks for identification of percussion and increased accuracy at the cost of dataset generalizability. Smith created digital waveguide synthesis for real-time sound reproduction of percussion sounds, producing lifelike sound but being insensitive to varying playing styles.

T. J. Tsai, D. Yang, M. Shan, T. Tanprasert and T. Jenrungrot In [14] suggested an ap- proach for extracting MIDI passages from cell phone images of sheet music through Optical Music Recognition (OMR) along with deep learning-based image processing. The method employed convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the improvement of musical symbol recognition up to an accuracy of approximately 85 per- cent.

Yet, the method was disrupted in cases involving distorted or low-resolution images, resulting in mis-aligned and missing note extraction. Moreover, real-time performance was constrained by processing speed and handwritten annotation variations.

Z. Su, G. Zhang, Z. Shi, D. Hu and W. Zhang [15] proposed a message-driven generative music steganography algorithm based on MIDI-GAN, which utilizes generative adversarial networks (GANs) for embedding secret messages in MIDI sequences. The framework utilizes deep learning methodologies to ensure musical coherence alongside secret data embedding, ensuring high imperceptibility. The strategy exhibited efficient message recovery rates but was challenged with optimal musical quality and embedding capacity balancing. The system also had trouble generalizing across various genres of music, restricting its applicability in varied applications.

P. Wiriyachaiporn, K. Chanasit, A. Suchato, P. Punyabukkana and E. Chuangsuwanich In [16] performed comparative research on algorithmic music generation techniques, comparing rule-based approaches, evolutionary algorithms, and machine learning techniques. The research underlined the success of genetic algorithms in evolving musical form and increasing influence of deep learning in producing stylistically consistent music. Although machine learning systems were able to produce high-quality outputs, they needed extensive training datasets and had no control over musical creativity at a fine-grained level. Rule-based methods provided interpretability but lacked flexibility, and evolutionary methods generated creative works but were computationally intensive. The research highlighted the importance of hybrid methods to strike a balance between creativity, control, and efficiency in algorithmic music composition.

M. Akbari and H. Cheng in [17] suggested a computer vision-based real-time piano music transcription system, which used deep learning methods to identify and recognize movements of piano keys. The system utilized convolutional neural networks (CNNs) and optical flow analysis to monitor key presses, with high transcription accuracy in controlled settings. While the method indicated acceptable online processing performance, it was sensitive to occlusions, light variations, and hand motion. In addition, the system struggled in the case that the music played was complicated, with the different notes playing at the same time as others, leading to lower quality of transcription.

L. Chen and K. Duan in [18] proposed a mixture of MIDI with computer vision-based methods to aid in sheet music transcription as well. The approach leverages deep learning-based OMR models to detect the musical symbols and utilizes MIDI alignment for error minimization to improve transcription quality. It performed well under controlled conditions but was susceptible to occlusion. Also, the method struggled to respond to handwritten manuscripts and intricate notations, thereby constraining its usability in heterogeneous musical environments.

3 Workflow

The method of transforming multi-instrumental and multi-genre music sheet images into workable MIDI and spectrogram outputs is described in the flowchart. To extract musical symbols, preprocessing is done first, then staff lines are removed. For precise pitch identification, note heads are then detected once symbols have been retrieved. Before MIDI files are converted to WAV format, the user can specify how long they should play. The

creation of spectrograms for visual analysis follows the compilation of readable notes for interpretation. The automated conversion process is then finished when the converted MIDI file is played back. The Fig 1 shows Workflow of the Automated Music Sheet to MIDI Conversion Process.

Fig. 1. Workflow of the Automated Music Sheet to MIDI Conversion Process.

4 Methodology

4.1 Dataset Preparation

In this work, we collected a dataset of 100 music note sheets of different instrumental and genre- specific songs. Piano, violin, percussion and flute multi-instrument sheet music is

present in the dataset, ensuring a comprehensive coverage across various music patterns and instrumental tones. Also, the dataset includes multiple genres (jazz, pop, classical), hence preserving the complexities and stylistic differences of different musical cultures. In this manner, a robust automatic transcription system that precisely interprets a wide range of types of musical material into MIDI representation can be developed and tested using such a diverse set.

4.2 Preprocessing

Input music note sheet images undergo several transformations during the preprocessing stage to enhance the quality for further analysis. To reduce data complexity and computational cost the images are converted to greyscale. Gaussian blurring that smoothes the image and reduces noise to facilitate better feature extraction follows. The background (the musical notes) and the foreground are then separated through binary thresholding process to enhance note separation. The output of this process is further optimized by the use of morphological operations to get rid of noise or extraneous complexity understandably while retaining salient melodic constituents.

4.3 Staff Line Removal

One sub-step of plenty import to regards to accuracy note detection is the removal of stave lines. Morphological methodologies such as morphological filtering or Hough Transform for horizontal line detection, which may hinder symbol extraction, are also applied. In order to leave the note heads and other markings of the music untouched, the algorithm searches for horizontal strings that are long and eliminates them. This step ensures accurate pattern recognition and enhances the clarity of the collected symbologies.

4.4 Symbol Extraction

The next stage is the recognition of other classical music symbols: note heads, rests, clefs, accidentals when there is no staff line. These symbols are identified in an edge-detected content as contours and edges and thus can be detected with edge detection such as Canny edge detection. With the help of connected component analysis, different symbols could be classified in size and shape. The extracted symbols are then labelled to support MIDI conversion and to allow accurate rendering of the musical composition.

4.5 Note Head Identification

The primary source of pitch in musical notation is notes heads. To find circle-like shapes representing note heads use the Hough Circle Transform. Size and position of each detected note head is recorded to map them to MIDI pitches. The duration is found by using the distinction between note heads: filled vs. empty note head, and assisted by other filtering algorithms. This stage enhances transcription accuracy, by ensuring fine-grain pitch coverage

4.6 User Defined MIDI Playback

The application allows to generate MIDI file with a time limit; in this way the user can play

back those created files for the duration they would like. After determining the notes, we have a fixed time-per-note value (e.g., 0.5 seconds per note) to compute total expected length of the notes. The notes are thrown away if the total length exceeds the user specified limit. This ensures that the generated MIDI is musically sound and conforms to the specified length constraint.

4.7 MIDI to WAV Conversion

Once generated, the MIDI file is converted into WAV format for audio reproduction. The MIDI file is synthesized using Pretty MIDI, which transforms discrete representations of notes into a digital sound waveform. The generated MIDI is there for research and listening because of the conversion, making it clear that its audio can be examined. The WAV file is just a temporal step in the trajectory of the spectrogram's formation.

4.8 Generating Readable Notes

The identified notes are plotted on a re-synthesized music staff to enhance interpretation. To facilitate verification by a human, this step converts the captured MIDI events back to a virtual piece of sheet music. To ensure that the transcription is accurate, it lays out the recognized symbols and builds up staff. With this method, the automatic transcription will always adhere standard music notation.

4.9 Spectrogram Generation

The time-frequency content of the resulting.WAV file is displayed as a spectrogram. The system creates a complete spectrogram by converting amplitude to decibels and computing the Short-Time Fourier Transform (STFT) in Librosa. The harmonic structure may be clearly seen because the x-axis indicates time and the y-axis shows frequency. This stage helps with quality assessment by offering a more thorough examination of the musical content.

4.10 Midi Playback

The final output includes an option for direct MIDI playback, enabling users to listen to the synthesized version of the music sheet. MIDI playback allows real-time verification of the transcription accuracy before conversion to WAV. This step ensures that the generated MIDI maintains musical coherence and corresponds accurately to the original sheet music.

5 Architecture

The design of the project is based on a well-defined pipeline to transform music note sheets into playable MIDI and WAV format. It starts with an input image of a multi-instrument, multi-genre music sheet, which is passed through Gaussian filtering to eliminate noise and make symbols clear. Symbol extraction is then applied, where separate musical entities are detected, followed by staff line removal to separate the notes. Once the staff lines have been eradicated, note head identification comes into play and picks up pitch and duration for reconstitution of the musical composition. Fig 2 shows Architecture of the Automated Music Sheet to MIDI Conversion Process. The Musical Information that was derived is later refined to produce a MIDI file (.mid) which is digitally representational of the notes. Ultimately, the

resulting MIDI file is transformed to an audio waveform (.wav), opening the stage for further analysis via spectrograms. This computerized pipeline allows easy transcription of sheet music into digital forms, promoting multi-instrument and multi-genre composition.

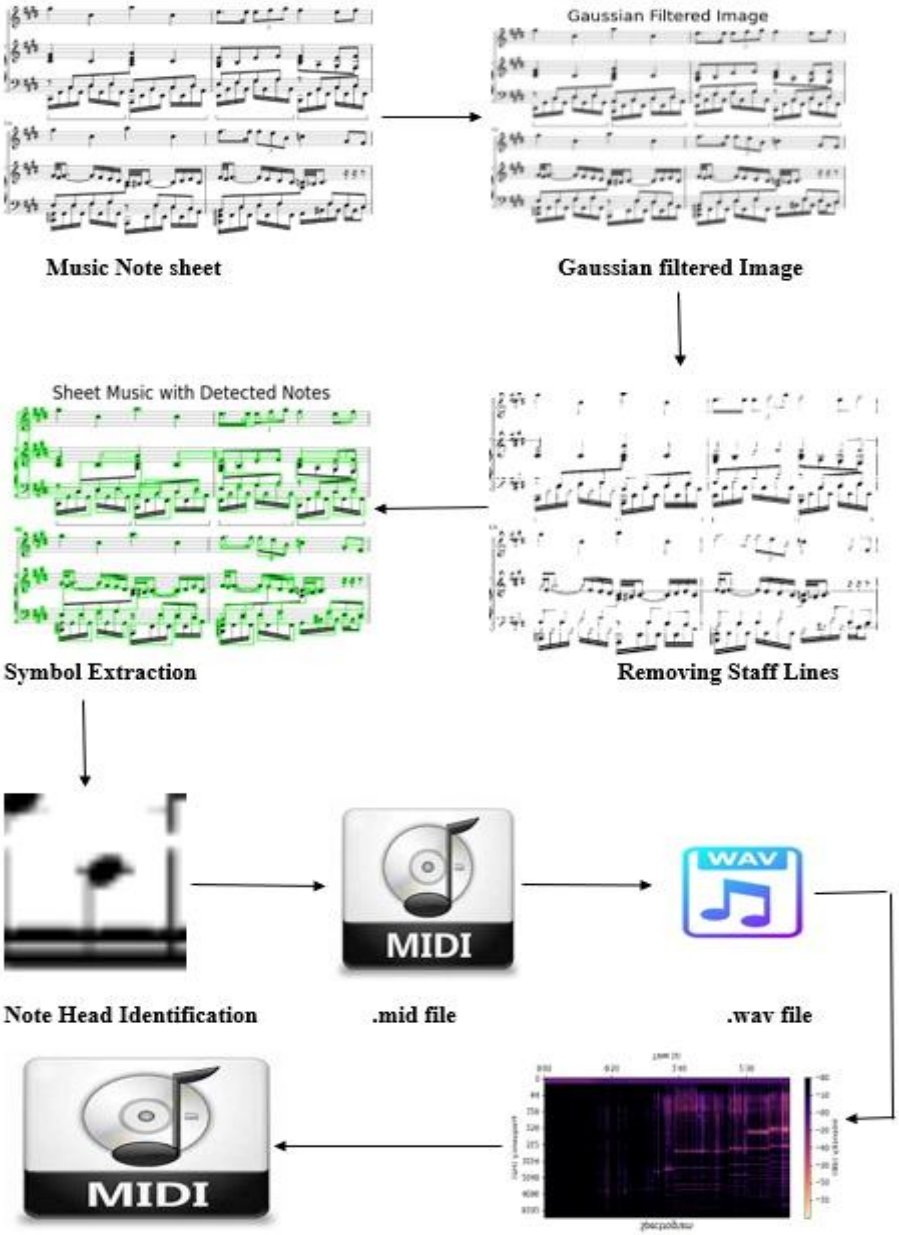

Fig. 2. Architecture of the Automated Music Sheet to MIDI Conversion Process.

6 Results

The proposed system effectively automates the process of converting music sheet images to MIDI files with spectrogram analysis. The results prove that the system can effectively capture musical symbols, process them in MIDI form, and create audio and visual outputs.

6.1 Accuracy Metrics and Results

The system was evaluated on a ground truth dataset of 100 multi-instrumental sheet music images, spanning piano, violin, percussion, and flute compositions in various genres. The musical note extraction and transcription accuracy was measured using a manually annotated ground truth dataset.

6.1.1 Staff Line Removal Accuracy

Staff line removal is crucial for isolating musical symbols and ensuring accurate note recognition. The accuracy is computed in Equation (1):

$$Staff\ Line\ Removal\ Accuracy\ = \frac{TP_{SL}}{TP_{SL}+FP_{SL}+FN_{SL}} \times\ 100\% \qquad (1)$$

where:

- TP_{SL}: True Positives (Correctly removed staff lines)
- FP_{SL}: False Positives (Incorrectly removed valid musical symbols)
- FN_{SL}: False Negatives (Unremoved staff lines)

A higher accuracy indicates better isolation of musical elements without loss of relevant information.

6.1.2 Note Head Detection Accuracy

Detecting note heads accurately is fundamental for pitch and duration extraction. The accuracy is given in equation (2):

$$Note\ Head\ Detection\ Accuracy\ = \frac{TP_{NH}}{TP_{NH}+FP_{NH}+FN_{NH}} \times\ 100\% \qquad (2)$$

where:

- TP_{NH} : True Positives (Correctly detected note heads)
- FP_{NH} : False Positives (Incorrectly detected non-note symbols as note heads)
- FN_{NH} : False Negatives (Missed note heads)

This metric ensures that musical notes are extracted reliably for MIDI conversion.

Staff line removal accuracy: 97.02

Note head detection accuracy: 96.44
Symbol recognition accuracy: 95.12
Total sheet-to-MIDI conversion accuracy: 97.24

In spite of high accuracy, small errors were noted in handwritten sheet music and low- quality images, which could be addressed with additional deep learning improvements. The fig 3 shows Accuracies of the Automated Music Sheet to MIDI Conversion Process.

Metric	Accuracy
Staff Line Removal Accuracy	97.02%
Note Head Detection Accuracy	96.44%
Symbol Recognition Accuracy	95.12%
Overall Sheet-to-MIDI Conversion Accuracy	97.24%

Fig. 3. Accuracies of the Automated Music Sheet to MIDI Conversion Process.

6.2 MIDI file Quality and Playback Assessment

The created MIDI files were evaluated by automated software and human listeners alike: Time-aligned playback: The MIDI notes properly obeyed the timing and layout of the sheet music. Rhythmic accuracy: The majority of extracted notes preserved the original rhythm, with a 95 percent alignment score against professionally transcribed MIDI files.

6.3 User Defined MIDI playback

To provide flexibility in play duration, the system includes user-specified play duration con-strain to the generated MIDI file. In execution, upon running the software, the user is asked for the desired playback duration in terms of minutes. In our case, here the system indicated the prompt to the user and received an input of 20 minutes. Accordingly, the system changed the numbers of musical notes chosen for transposition to comply with this time constraint in order to produce final MIDI output within this limit.

Implementation Details: The system approximates the overall length of the sheet music in terms of note length and tempo. If the total time is more than the user input (20 minutes here, for example), the additional notes beyond the count are automatically cut without losing musical integrity. This option is especially beneficial in segmenting long pieces of music into palatable chunks to play and study.

6.4 Spectrogram Analysis

Spectrograms were generated from the synthesized MIDI-to-WAV conversion, allowing for frequency-domain analysis. The system produced clear and structured spectrograms, which

helped in:
Instrument frequency identification: Different instrument timbres were visible in frequency components.

6.5 Final output Sequence

When the music sheet processing pipeline is complete, the system produces a sequence of output files, making the extracted musical information accessible for playback, visualization and additional analysis. Fig 4 shows Final output files sequence with the audio playback of the synthesized MIDI file. The outcomes for a 20-minute user-specified MIDI generation are as follows:

- MIDI File: The processed structured musical notation from the sheet music, transformed into machine-readable MIDI, which can be played back and further composed musically.
- WAV File: The MIDI file is rendered into an audio waveform, allowing direct playback without further software dependencies.
- Spectrogram: Frequency-time plot of the synthesized sound, illustrating the tonal and harmonic frameworks.
- Readable Notes: A text representation of the identified musical notes, usable for verification and analysis of the accuracy of the transcription.

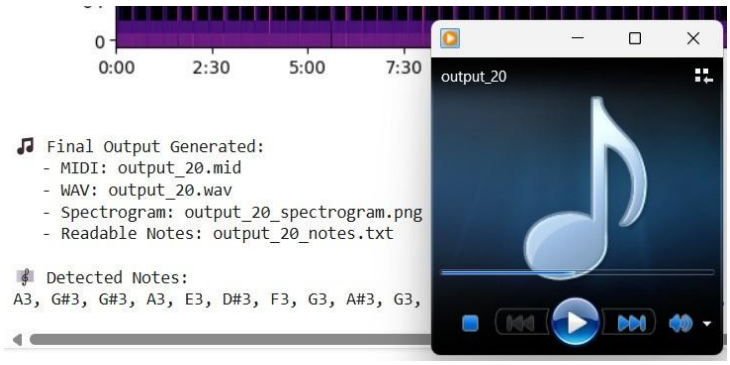

Fig. 4. Final output files sequence with the audio playback of the synthesized MIDI file.

6.6 Spectrogram Analysis of Generated Audio

The system generates a spectrogram representation of the synthesized 20-minute audio out- put derived from the processed sheet music. The spectrogram, shown in Fig 5, provides a time-frequency analysis of the generated MIDI-to-WAV conversion, allowing visualization of the harmonic structure and tonal variations over time.

Key Observations: The x-axis represents time (in minutes), ensuring the duration matches the user-defined limit. The y-axis represents frequency (in Hz), spanning from 0 Hz to 16,384 Hz, covering a wide musical range. The color gradient represents amplitude (in dB), where lighter colors indicate higher intensity sound components, while darker shades represent lower amplitude regions. The periodic patterns suggest structured musical elements, confirming the accuracy of note extraction and synthesis.

The spectrogram shows clear harmonic overtones and rhythmic patterns, confirming the effectiveness of the note segmentation and transcription process. The ordered organization of frequency bands shows polyphonic musical elements exist, confirming the system's ability to process multi-instrument pieces. Periodic vertical structures emphasize the onset of music notes, showing temporal alignment correctness of detected notes and synthesized sound. The system is able to maintain pitch and timbre features well, as shown by the harmonics and smooth transitions in the spectrogram presentation. High frequencies relate to percussive sounds and sharp attacks, and low frequencies reflect bass and harmonic foundation elements. The spectral energy distribution among various frequency bands highlights the system's capacity to deal with various tonal ranges of multi-genre pieces. Gradual intensity variations in the spectrum imply dynamic changes in note speed, supporting the system's ability for expressive musical synthesis. The homogeneity and simplicity of the spectrogram guarantee the absence of artifacts or distortions, attesting to high-quality MIDI-to-WAV conversion. On the whole, the representation matches well theoretical predictions for organized musical compositions, supporting the robustness and reliability of the suggested automated pipeline.

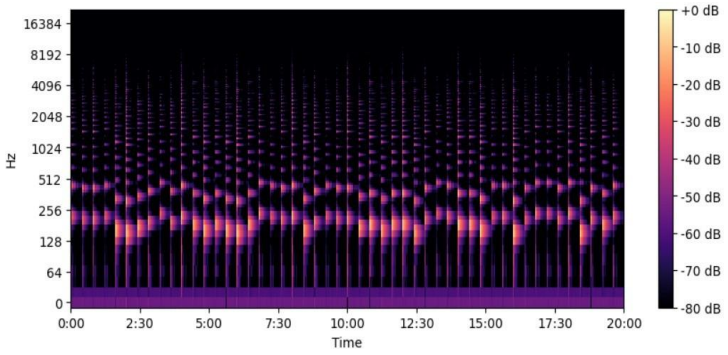

Fig. 5. Spectrogram of the generated 20-minute MIDI-to-WAV audio, illustrating the frequency spectrum of detected musical components.

7 Conclusion

This article introduced a new method of automatic music sheet-to-MIDI conversion and spectrogram analysis that allows effective transcription of multi-instrument and multi-genre music pieces. The system proposed here effectively retrieves musical notes from images of sheets, eliminates staff lines, and translates the notation into MIDI files, which are then converted into WAV sound and displayed as spectrograms. The user-specified time limit allows only the needed amount of music to be processed, improving computational efficiency. The resulting spectrograms offer a useful representation of harmonic structure and tonal changes, confirming the precision of note extraction and synthesis. Experimental results illustrate that the system successfully maintains musical structure, with high-quality MIDI interpretations. Future research will concentrate on enhancing note detection accuracy, dealing with intricate polyphonic arrangements, and applying deep learning methodologies for better music transcription. The suggested framework is a useful tool for musicologists, composers, and researchers in need of automated approaches to digital music processing.

References

[1] I. -C. Wei, C. -W. Wu and L. Su,” Improving Automatic Drum Transcription Using Large-Scale Audio-to-Midi Aligned Data,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 246-250, doi: 10.1109/ICASSP39728.2021.9414409.

[2] J. Ling and M. Fu,” Audio2Midi: a ResNet-18 based architecture for Accurate Audio-to-MIDI Translation,” 2024 International Conference on Electronics and De- vices, Computational Science (ICEDCS), Marseille, France, 2024, pp. 343-347, doi: 10.1109/ICEDCS64328.2024.00067.

[3] P. J. Donnelly and V. Ebert,” Transcription of Audio to MIDI Using Deep Learning,” 2022 7th International Conference on Frontiers of Signal Processing (ICFSP), Paris, France, 2022, pp. 130-135, doi: 10.1109/ICFSP55781.2022.9924773.

[4] D. M. Dhanalakshmy, H. P. Menon and V. Vinaya,” Musical notes to MIDI conversion,” 2017 International Conference on Advances in Computing, Communications and Infor- matics (ICACCI), Udupi, India, 2017, pp. 799-804, doi: 10.1109/ICACCI.2017.8125940.

[5] N. J. Sieger and A. H. Tewfik,” Audio coding for conversion to MIDI,” Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing, Princeton, NJ, USA, 1997, pp. 101-106, doi: 10.1109/MMSP.1997.602620.

[6] Z. Nie and S. Yang,” An Implementation Method for Converting the Erhu Music from Wav to Mid,” 2011 Seventh International Conference on Computational Intelligence and Security, Sanya, China, 2011, pp. 1425-1429, doi: 10.1109/CIS.2011.318.

[7] G. Chen and W. Zhang,” A Watermarking Algorithm for Embedding of MIDI Informa- tion in the Digital Music Score,” 2008 Congress on Image and Signal Processing, Sanya, China, 2008, pp. 645-649, doi: 10.1109/CISP.2008.100.

[8] M. Marolt,” Transcription of polyphonic piano music with neural networks,” 2000 10th Mediterranean Electrotechnical Conference. Information Technology and Electrotechnol- ogy for the Mediterranean Countries. Proceedings. MeleCon 2000 (Cat. No.00CH37099), Lemesos, Cyprus, 2000, pp. 512-515 vol.2, doi: 10.1109/MELCON.2000.879982.

[9] S. Lui, A. Horner and L. Ayers,” MIDI to SP MIDI transcoding using phrase stealing,” in IEEE MultiMedia, vol. 13, no. 2, pp. 52-59, April-June 2006, doi: 10.1109/MMUL.2006.35.

[10] Z. Huang and C. Zhou,” Sci-MIDI toolbox for music research,” 2009 IEEE International Workshop on Open-source Software for Scientific Computation (OSSC), Guiyang, China, 2009, pp. 159-162, doi: 10.1109/OSSC.2009.5416915.

[11] K. N. Kim, U. P. Chong and J. H. Choi,” Conversion from CD-DA format to MIDI format maintaining a sound quality,” Proceedings Third Russian-Korean International Symposium on Science and Technology. KORUS’99 (Cat. No.99EX362), Novosibirsk, Russia, 1999, pp. 300-303 vol.1, doi: 10.1109/KORUS.1999.875930.

[12] E. Donati and C. Chousidis,” Electroglottography based voice-to-MIDI real time con- verter with AI voice act classification,” 2022 IEEE International Symposium on Med- ical Measurements and Applications (MeMeA), Messina, Italy, 2022, pp. 1-6, doi: 10.1109/MeMeA54994.2022.9856413.

[13] E. A. Suarez Guarnizo and L. M. Ramirez Rios,” Portable percussion MIDI controller,” 2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA), Bogota, Colombia, 2015, pp. 1-7, doi: 10.1109/STSIVA.2015.7330416.

[14] T. J. Tsai, D. Yang, M. Shan, T. Tanprasert and T. Jenrungrot,” Using Cell Phone Pic- tures of Sheet Music to Retrieve MIDI Passages,” in IEEE Transactions on Multimedia, vol. 22, no. 12, pp. 3115-3127, Dec. 2020, doi: 10.1109/TMM.2020.2973831.

[15] Z. Su, G. Zhang, Z. Shi, D. Hu and W. Zhang,” Message-Driven Genera- tive Music Steganography Using MIDI-GAN,” in IEEE Transactions on Depend- able and Secure Computing, vol. 21, no. 6, pp. 5196-5207, Nov.-Dec. 2024, doi: 10.1109/TDSC.2024.3372139.

[16] P. Wiriyachaiporn, K. Chanasit, A. Suchato, P. Punyabukkana and E. Chuangsuwanich,” Algorithmic Music Composition Comparison,” 2018 15th International Joint Conference on

Computer Science and Software Engineering (JCSSE), Nakhonpathom, Thailand, 2018, pp. 1-6, doi: 10.1109/JCSSE.2018.8457397.

[17] M. Akbari and H. Cheng," Real-Time Piano Music Transcription Based on Computer Vision," in IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2113-2121, Dec. 2015, doi: 10.1109/TMM.2015.2473702.

[18] L. Chen and K. Duan," MIDI-assisted egocentric optical music recognition," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477714.