ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

Estuscience – Se, 2024, 25 [3] pp. 442-455, DOI: 10.18038/estubtda.1467350

RESEARCH ARTICLE

# END-TO-END AUTOMATIC MUSIC TRANSCRIPTION OF POLYPHONIC QANUN AND OUD MUSIC USING DEEP NEURAL NETWORK

**Emin GERMEN** [1] *, **Can KARADOĞAN** [2]

[1] Electrical & Electronics Engineering Dept. Engineering Faculty Eskisehir Technical University, Eskişehir, Turkey
egermen@eskisehir.edu.tr- 0000-0003-1301-3786

[2] MIAM Center of Advanced Studies in Music, State Conservatory, Istanbul Technical University, İstanbul, Turkey
karadoganc@itu.edu.tr - 0000-0003-3611-6980

## Abstract

This paper introduces an automatic music transcription model using Deep Neural Networks (DNNs), focusing on simulating the "trained ear" in music. It advances the field of signal processing and music technology, particularly in multi-instrument transcription involving traditional Turkish instruments, Qanun and Oud. Those instruments have unique timbral characteristics with early decay periods. The study involves generating basic combinations of multi-pitch datasets, training the DNN model on this data, and demonstrating its effectiveness in transcribing two-part compositions with high accuracy and F1 measures. The model's training involves understanding the fundamental characteristics of individual instruments, enabling it to identify and isolate complex patterns in mixed compositions. The primary goal is to empower the model to distinguish and analyze individual musical components, thereby enhancing applications in music production, audio engineering, and education.

**Keywords**

Music transcription,
Signal processing,
Deep learning,
Constant Q transform

**Time Scale of Article**

Received :11 April 2024
Accepted : 18 Septembert 2024
Online date :30 September 2024

## 1. INTRODUCTION

End-to-end music transcription is a highly challenging subject, captivating those in signal processing research, musicians, and the people involved in music technology[1]. This field initially took shape with advancements in multi-instrument music transcription, rooted in Blind Source Separation (BSS) [2], [3] regarding signal processing and linear system methods like Non-Negative Matrix Factorization[4]. However, the landscape evolved[5] significantly with the advent of accessible processing capabilities and breakthroughs in Deep Neural Networks (DNN)[6]–[8]. These developments have expanded the interest in BSS to researchers focused on Data Learning. Beyond mere source separation, the domain has diversified to include various methodologies, such as extracting instruments from music compositions[5], segregating music and voices through 2D Fourier Transform techniques[9], and isolating monaural speech[10] further enriching the research literature.

The human auditory system can discern distinct patterns and separate acoustic sources, such as different musical instruments or vocals. However, a crucial aspect of this capability is that it often requires some degree of familiarization or training[11]. Individuals lacking exposure to a particular musical genre or

*Corresponding Author: egermen@eskisehir.edu.tr unfamiliar with specific types of instruments may find it challenging to isolate and recognize distinct components within the music.

For instance, someone with no background in classical Indian music might struggle to identify patterns in Indian Raga or Sargam music. These compositions, rich in their unique structure and rhythm, could seem intricate and elusive to an untrained ear. Similarly, Western classical music, known for its complex harmonies and arrangements, might appear as a series of perplexing patterns to an adolescent unfamiliar with this style. In another scenario, a person who has spent considerable time immersed in Western Classical music might find metal music overwhelming or even disturbing. Often characterized by its intense and heavily distorted guitar riffs, metal music starkly contrasts Western Classical music's more structured and melodic nature. This divergence can make it difficult for someone accustomed to the latter to comprehend and appreciate the former's musical themes, nuances, and even instruments.

Overall, the ease with which a person can recognize and separate elements within a musical piece significantly depends on their prior exposure and understanding of the specific musical style. This highlights the importance of cultural and experiential factors in shaping our auditory perception and appreciation of music. The main focus in this context is the trained ear and how this phenomenon is modeled.

The concept of a "trained ear" in music pertains to the ability to discern and identify specific instruments and the music played by them. This skill, often developed over time through exposure and practice, involves a deep understanding of the distinct characteristics of various instruments[12]. The process of acquiring such a nuanced auditory ability is both elaborate and complex, reflecting the intricate nature of musical perception and appreciation[13]. The concept of "training," particularly in the context of music separation, represents a significant milestone in the field of music analysis. This training refers to the process of developing systems that can identify specific patterns within a mixed musical piece, which is effectively treated as a signal in this context. Various methodologies have been explored to achieve this, as noted in the works of [14], [15]

In the context of using models, such as Deep Neural Networks (DNNs)[16], to replicate or assist in this process of music separation and identification, the starting point often involves understanding the basic characteristics of individual instruments. By training the model with these fundamental traits, it can begin to recognize and isolate the complex patterns specific to each instrument. This approach involves feeding the model with data that encapsulates the unique acoustic signatures of different instruments. As the model learns these characteristics, it becomes increasingly adept at identifying these instruments within a mixed musical composition.

The goal is to enable the model to not only recognize an instrument but also to isolate the music played by that instrument from a composition featuring multiple instruments. This ability to discriminate and analyze individual components of a musical piece is valuable for various applications, including music production, audio engineering, and even in enhancing music education and research. It is a testament to how combining human auditory skills and advanced technological models can lead to a deeper and more precise understanding of music.

This study aims to explore the differentiation of instruments and their music within composite audio data, focusing on recognizing distinct, well-defined signatures of individual instruments. Specifically, the Turkish plucked instruments, Qanun and Oud, have been selected for analysis in this model. A DNN model has been developed and trained to utilize the fundamental characteristics of the amalgamated sounds from these instruments. The trained model is then employed to distinguish polyphonic music compositions involving these instruments. The outcomes of this approach have been notably satisfactory.

One of the most significant contributions of this study is the development of a system that enables the separation of complex musical data consisting of two or more sounds, using only the fundamental composite data characteristics of the acoustic instruments to be separated, independently of the instrument databases used in previously trained models. Moreover, it is noteworthy that this study successfully applies such a separation to Turkish music instruments, which have not been studied in the world literature before. The successful results obtained demonstrate that this approach can also be applied to Turkish music.

The paper's organization is structured as follows: Chapter Two presents a comprehensive description of the model utilized for transcribing two-part music. This chapter will detail the preparation of the dataset used for training, as well as the features involved. Chapter Three is devoted to an in-depth discussion of the results derived from this study. Conclusively, Chapter Four offers a summation and final thoughts of the paper.

## 2. AUTOMATIC MUSIC TRANSCRIPTION MODEL

In this detailed section, we delve into the complexities of the model architecture, the generation of training data, and the signal processing techniques employed for feature extraction, all of which are integral to our research study. The framework is given in  Figure 1.

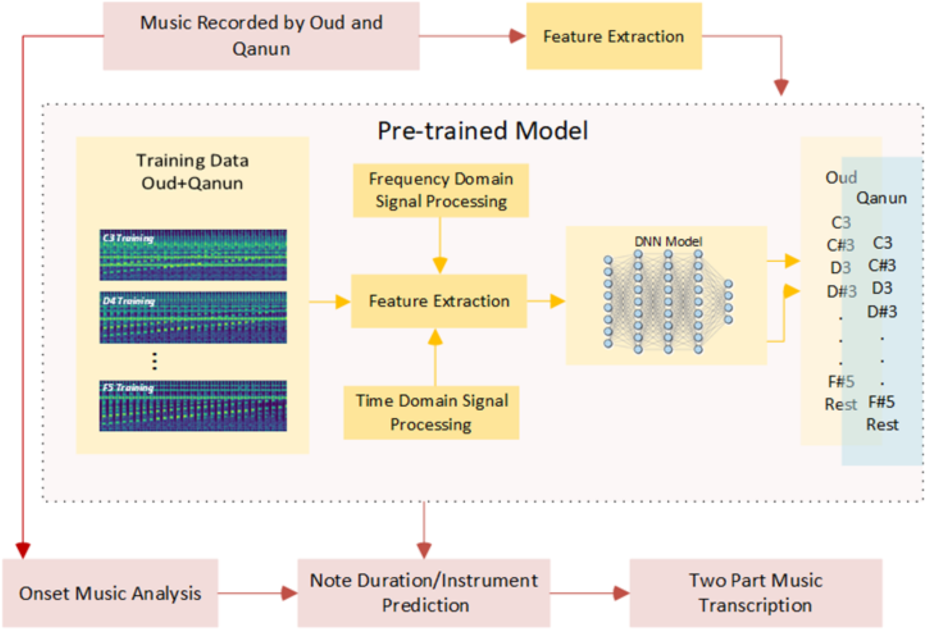


**Figure 1.** The framework of the two-part music transcription

Training Data Generation:

The primary objective of constructing this model is to facilitate the understanding of the mechanism behind the ear training phenomena by scrutinizing the fundamental instrumental combinations in music creation. This research employs a blend of music incorporating Turkish Qanun and Oud instruments, aiming to generate a multi-instrument, multi-pitch dataset. A critical step in this process is the disentanglement of the musical pieces, which necessitates capturing the distinct multi-timbral characteristics of these two instruments playing concurrently.

The concept of a "well-trained ear" in this context signifies proficiency in precisely discerning, identifying, and interpreting sounds, especially within music. This ability is greatly esteemed among musicians, audio engineers, and music lovers. The most crucial element of such a trained ear is the recognition of pitches in polyphonic music. This study investigates whether such training can be mimicked through the note combinations of two simultaneously played instruments. The methodology adopted is straightforward: while the Oud plays a single note (e.g., C3), the Qanun explores various notes spanning three octaves, and all possible combinations are systematically recorded.

Figure 2 exemplifies the training data for the A4 note on the Qanun. In this illustration, the Qanun plays the A4 note while the Oud progresses in chromatic from C3 to A5. Notably, this method also can accommodate training for microtonal intervals. In this work, they are omitted. For each note, a 56-second sample of mixed music is captured for training purposes. The predicted output of 56 seconds data truly considered as A4.  In total, thirty-seven distinct datasets were compiled for the Qanun, covering chromatic notes from C3 to A5. Therefore, the corpus for Qanun consists of total 2072 seconds long data with 37 different predicted outputs to span the notes C3 to A5. An additional dataset was added to the Qanun corpus to represent musical pauses, where only the Oud is audible or there is silence. This resulted in 37 diverse datasets dedicated to identifying Qanun phrases within the mixed music. Here 37 different outputs are C3, C3#, D3, D3#, … A5, and Rest for Qanun, which are equally probably distributed. The same methodology was replicated for compiling the Oud data. This systematic approach ensures a thorough coverage of the musical spectrum, enabling the study to capture the intricacies of pitch variation and harmonic relationships between the two instruments. By focusing on the combination of notes from two distinct instruments, the study paves the way for advanced understanding and modeling of complex musical interactions.

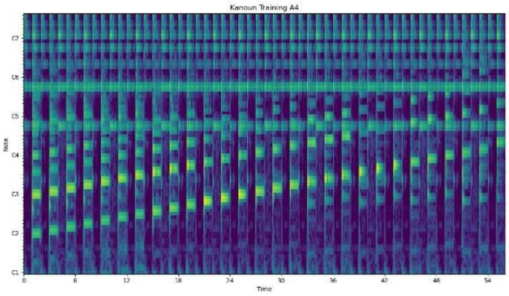


**Figure 2**. A4 training data for Qanun with Oud spanning the scale in 3 octaves.

## Feature Extraction

The features are generated using Short Time Fourier Transform (STFT), Constant Q Transform (CQT), Spectral Centroid (ST) and Band Energy Ratio (BER) combinations of music signal. In music analysis, the STFT [17] is a key tool for tasks like pitch detection, note mapping, and timing. It is effective because it analyzes music in both time and frequency, providing a detailed view of musical pieces. STFT helps break down complex polyphonic compositions into simpler parts, making it easier to understand the structure of the music. However, it is important to mention that while the STFT is good for understanding the sound qualities of instruments, it is not the best at picking up the finer details of musical notes, especially in the 64 Hz to 1.5 kHz frequency range. The CQT [18] is a better tool for this purpose. Its bins are arranged in line with musical notes, especially useful in Western music where an octave is divided into 12 semitones, as denoted in Equation (1).

$$f_m = f_{min} 2^{\frac{m}{12*b}} \tag{1}$$

Here $f_{min}$ is the minimal frequency in analysis, and $b$ is the number of bins per semi-tone. The CQT can be derived with a quality factor $Q = \frac{f_m}{f_{m+1} - f_m} = 1/(2^{\frac{1}{12b}} - 1)$

$$F_{CQT}[m, \Lambda] = \frac{1}{N_m} \sum_{k=0}^{N_m - 1} x[k + \Lambda R] w_m[k] e^{\left(-i\frac{2\pi Qk}{N_m}\right)} \tag{2}$$

To accurately identify both the exact frequencies of musical notes and the unique sound qualities of instruments, a feature set combining the STFT and the CQT has been developed. For the STFT, a 2048-point Fast Fourier Transform (FFT) is used, focusing on the first 500 frequency components from 0 Hz to 9 kHz. At the same time, the CQT uses 80 bins to cover the range of musical notes from the C in the first octave to the C in the eighth octave, between 32.7 Hz and 4186 Hz. This combined use of STFT and CQT provides a thorough and detailed representation of the music, capturing both the frequencies of notes and the distinct sound characteristics of various instruments across a wide frequency range.

The Spectral Centroid [19] shown in Equation (3) is a pivotal feature in music analysis, serving as a quantitative measure that denotes the 'center of gravity' of the spectrum for each analyzed bin. This metric is crucial in discerning the musical texture and timbral qualities of different instruments within a composition. It effectively indicates the concentration of energy across the spectrum, providing insights into the brightness or dullness of the sounds.

$$SC_k = \frac{\sum_{n=1}^{N} m_k(n)n}{\sum_{n=1}^{N} m_k(n)} \tag{3}$$

Like the Spectral Centroid, the BER is another critical audio feature that can be extracted for in-depth music analysis. This parameter is calculated for each window of the audio signal, as detailed in Equation (4). The BER quantifies the energy distribution within specific frequency bands of the audio spectrum, thereby providing a nuanced understanding of the spectral characteristics of the sound. This measurement is essential for tasks such as identifying the dominant frequency bands in a piece of music, understanding the textural components of sound, and distinguishing between different types of sounds or instruments based on their energy distribution across the spectrum.

$$BER_k = \frac{\sum_{n=1}^{F-1} m_k(n)^2}{\sum_{n=F}^{N} m_k(n)^2} = \frac{mean\ square\ of\ lower\ frequency\ components}{mean\ square\ of\ higher\ frequency\ components} \tag{4}$$

### Training the DNN Model

The study employs a comprehensive dataset derived from the multipitch characteristics of the Qanun and Oud instruments, covering octaves ranging from C3 to B5. From this dataset, the features are extracted and utilized as the foundational data for training a sophisticated 7-layer Deep Neural Network (DNN). The architecture of this DNN is thoughtfully designed, with the input layer comprising 586 nodes for the FFT components, CQT bins, Spectral Centroids, and BER features to identify the mixture components of two instruments.

The model is a deep neural network built using the Keras Sequential API, designed for multi-class classification. It consists of five hidden layers with 586, 400, 200, 400, and 200 neurons, each utilizing the ReLU activation function to introduce non-linearity and enhance the network's ability to learn complex patterns. The output layer comprises 37 neurons with softmax activation, producing a probability distribution across 37 different musical notes that span the 3 octaves. Compiled with the categorical cross-entropy loss function and the Adam optimizer, the model is optimized for efficient training and effective convergence. The accuracy metric evaluates performance throughout the training and testing phases, ensuring the model's effectiveness in classifying data into multiple categories. The model is trained using 67% of the data and tested on the remaining 33%, providing a more thorough evaluation of its classification capabilities.

### Transient Detection and Onset Analysis

Music transcription fundamentally involves recognizing temporal changes within a musical piece, such as variations in notes played by instruments or their intermittent silences. These fluctuations, termed as transients, are particularly pronounced during the attack phase of an instrument's note. In the context of this study, which focuses on plucked instruments like the Oud and Qanun, these transients are distinctly identifiable. To accurately track these transients and thereby detect structural shifts within the time series of the music, both the energy envelope and zero crossings are utilized. Concurrently, frequency-domain features, including the spectral centroid, spectral skewness, and spectral kurtosis, are employed to provide a more granular understanding of potential onset points. Figure 3 exemplifies this process, showcasing onset detection in a recording featuring both the Oud and Qanun, with illustrations in both time and frequency domains. This comprehensive approach allows for a precise music transcription, capturing the dynamic and nuanced interplay of these instruments.

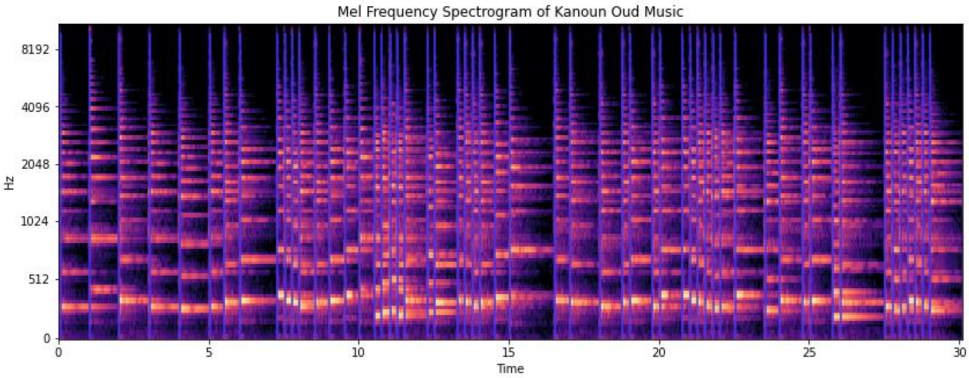


**Figure 3.** Onset times of music played by Qanun and Oud

## The Music Transcription

After completing the model's training process, a two-part musical composition is processed through the trained DNN. The output from this DNN comprises potential notes that form the basis of the musical piece. Concurrently, onset analysis is employed to identify and track the potential transients of notes and the intervals of silence (rests) within the music.

In conducting a comparative analysis of original melodies and their transcriptions, a methodology was utilized to reduce them into smaller constituent elements. This was achieved by transforming the duration of each note and rest within the compositions into an equivalent duration represented by sixteenth notes. The decomposition of these melodies into increments of sixteenth notes allowed for a uniform and exact measure of temporal divisions. This standardization was crucial in representing rhythmic patterns consistently, and it facilitated an accurate examination of nuanced differences, including syncopations and articulations, within the musical pieces.

## 3. RESULTS

The trained model with data of simple combinations of Oud and Qanun is used to separate and transcribe two-part music. We evaluated the model's efficiency using an originally composed piece, as detailed in Figure 4. This composition, featuring both instruments, was recorded in a CD-quality at a tempo of 60 beats per minute (bpm) and endured 1 minute and 38 seconds.

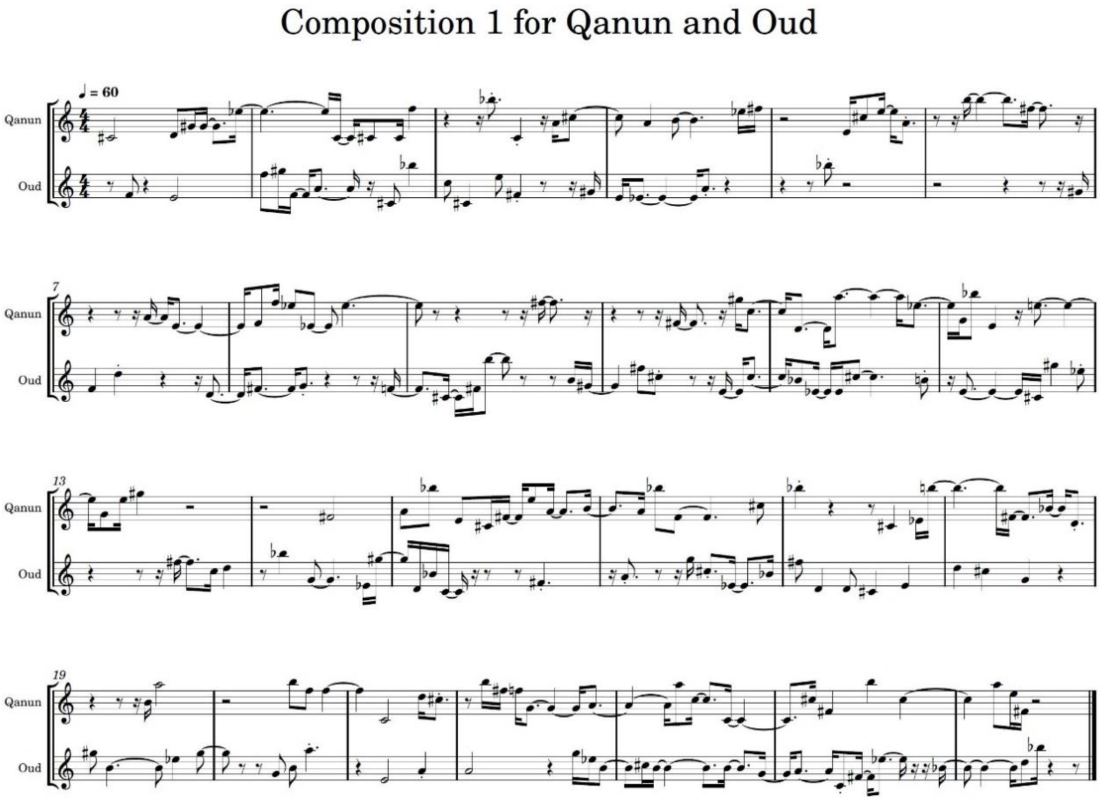


**Figure 4**. The Composition 1 for Qanun and Oud Music

To evaluate the model’s performance, the entire musical composition was segmented into precise segments, each corresponding to a duration of sixteenth notes. This structured division allowed for a detailed analysis of the model's accuracy and F1 measure in recognizing and transcribing each note segment. The F1 measure is a special metric that combines precision (how often we were right when we thought we were) and recall (how many of the right things we caught) into one score. Those are calculated as :

$$Precision = \sum_{n=1}^{N} \frac{True\ Positive[n]}{True\ Positive\ [n] + False\ Positive\ [n]} \tag{5}$$

$$Recall = \sum_{n=1}^{N} \frac{True\ Positive[n]}{True\ Positive\ [n] + False\ Negative\ [n]} \tag{6}$$

$$Accuracy = \sum_{n=1}^{N} \frac{True\ Positive[n]}{True\ Positive\ [n] + False\ Positive\ [n] + False\ Negative\ [n]} \tag{7}$$

$$F1\ measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

The Qanun Transcription of the music is given in Figure 5. Figure 6 shows the Confusion Matrix corresponding to Qanun's transcription. Here, it is noted that NO represents the rest.

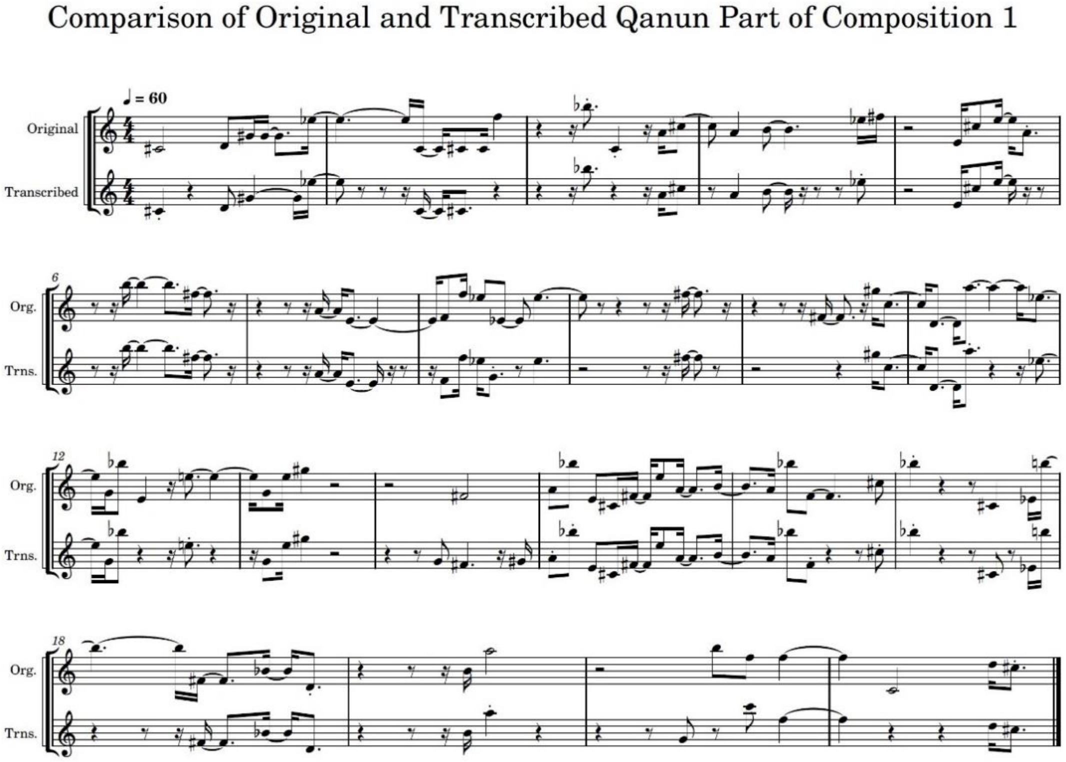
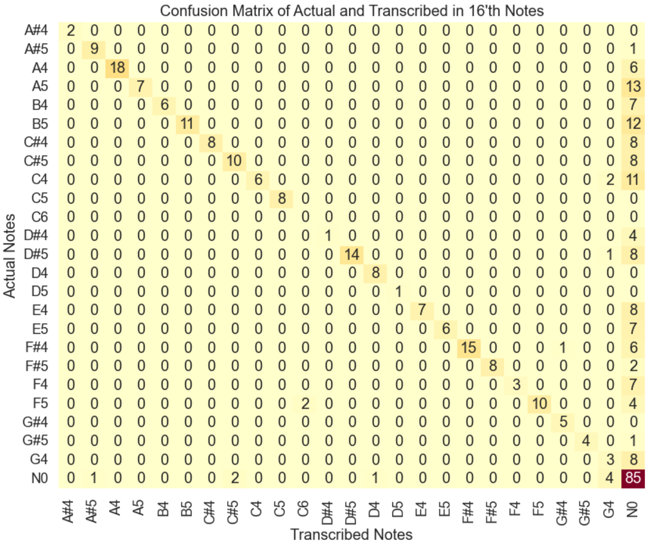

Comparison of Original and Transcribed Qanun Part of Composition 1
♩ = 60
Original
Transcribed
6
Org.
Trns.
12
Org.
Trns.
18
Org.
Trns.

**Figure 5.** Transcription of Extracted Qanun from Composition 1

In this confusion matrix of Qanun of Composition 1:

The Accuracy: 0.6538461538461539, F1 Score: 0.6636862141708085 are calculated.



| | A#4 | A#5 | A4 | A5 | B4 | B5 | C#4 | C#5 | C4 | C5 | C6 | D#4 | D#5 | D4 | D5 | E4 | E5 | F#4 | F#5 | F4 | F5 | G#4 | G#5 | G4 | N0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A#4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A#5 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| A4 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| A5 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |
| B4 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| B5 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| C#4 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| C#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| C4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 11 |
| C5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| D#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 8 |
| D4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| E5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| F#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 1 | 0 | 0 | 6 |
| F#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 2 |
| F4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 7 |
| F5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 4 |
| G#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| G#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 1 |
| G4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 8 |
| N0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 85 |

**Figure 6.** Confusion matrix of Transcribed Qanun part

The experiment conducted to transcribe qanun music yielded an accuracy of approximately 0.65, and the F1 score is 0.66, which, upon examination of the confusion matrix and results, was found to be implausible. Notably, a significant number of rests were transcribed instead of notes, primarily due to the distinctive pitch characteristics of the qanun instrument. The limited sustain inherent to the qanun often renders the related sound imperceptible, leading to its transcription as a rest. Nevertheless, it is crucial to highlight that despite this limitation, the qanun music's transcribed sections successfully capture the melody's progression, which remains the most critical aspect in this context. The transcriptions effectively preserve the descriptive elements associated with the melody's development and trajectory. However, it is essential to acknowledge that the accuracy metric of 0.65 may not fully encompass the transcription's overall quality, particularly when a specific instrument or music-specific attributes have not been adequately accounted for.

The primary objective of the transcription process is to accurately track the progression of the melody and capture the temporal changes in notes. Figure 7 illustrates that, despite successfully identifying the overall melody progression in both measure 1 and measure 2, there are discrepancies in specific note durations.

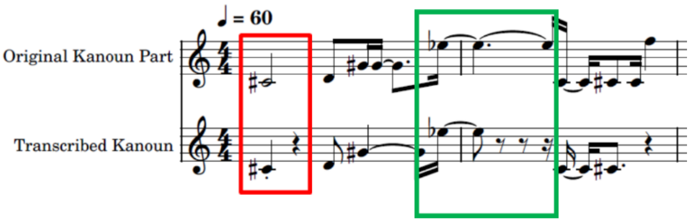


**Figure 7.** Extract from the Qanun music.

In measure 1, for instance, the sustained C#4, which lasts for 16 sixteenth notes, is transcribed as 8 sixteenth notes followed by 8 rests. This transcription fails to represent the continuous nature of the note's duration accurately. Similarly, at the beginning of the second measure, the sustained E♭5, which endures 8 sixteenth notes, is transcribed in the correct time position. It is shown in the second box in Figure 7. However, it is represented as 3 sixteenth notes instead of the intended 8. This discrepancy disregards the sustained quality of the note and fails to capture its full duration.

As mentioned before, in the case of qanun music, the characteristic pitch of the instrument may lead to the transcriptions incorrectly identifying rests due to the lack of sustain in the instrument. Since the sustain is insufficient to sustain the sound, it is transcribed as a rest.

The primary focus of melody transcription lies in accurately capturing the progression of the melody. The events presented in Figure 7**,** while not impacting the recognition of the melody itself, significantly affect metrics such as the confusion matrix, as well as satisfaction measures like Accuracy and F1 scores. A specialized heuristic has been developed to solve this issue. At the core of this heuristic is a decision-making process based on probability metrics, particularly when determining the musical notes at the output layer in DNN. Here, the algorithm meticulously evaluates the likelihood of each possible note. The one with the highest probability score is typically selected as the most appropriate note value. However, the heuristic introduces a nuanced twist in scenarios where the output indicates a rest, which, in musical terms, means a pause or silence. Instead of automatically accepting this rest, the algorithm delves deeper, examining the probability value associated with the note immediately preceding the supposed rest. If this preceding note's probability value is found to be higher than that of any other outputs except the rest, the heuristic dynamically adjusts its decision. In this case, it opts to consider the preceding note as being more representative of the intended musical expression rather than the rest. The example illustrated in Figure 7 within the red box shows a probability value of 0.82 at the end of the duration of the quarter note for C3. Subsequently, the rest probabilities are 0.47, 0.52, 0.57, and 0.48, which are higher than all other note possibilities. Meanwhile, C3 exhibits probability values of 0.26, 0.26, 0.21, and 0.20 during this time, making them the second-highest probabilistic values identified by the algorithm. Since the previous note value aligns with these second most probable notes over time, the algorithm decides to continue the note instead of replacing it with a rest value.

This heuristic evaluates whether rests are present in the music, considering the rapid decay attributes of the qanun instrument. This assessment is done without compromising the transcription accuracy of individual notes. After implementing this heuristic, the transcribed music played on the qanun has been regenerated and is displayed in Figure 8. Although the musical characteristics remain unchanged, there has been a notable improvement in both the F1 score and overall accuracy.

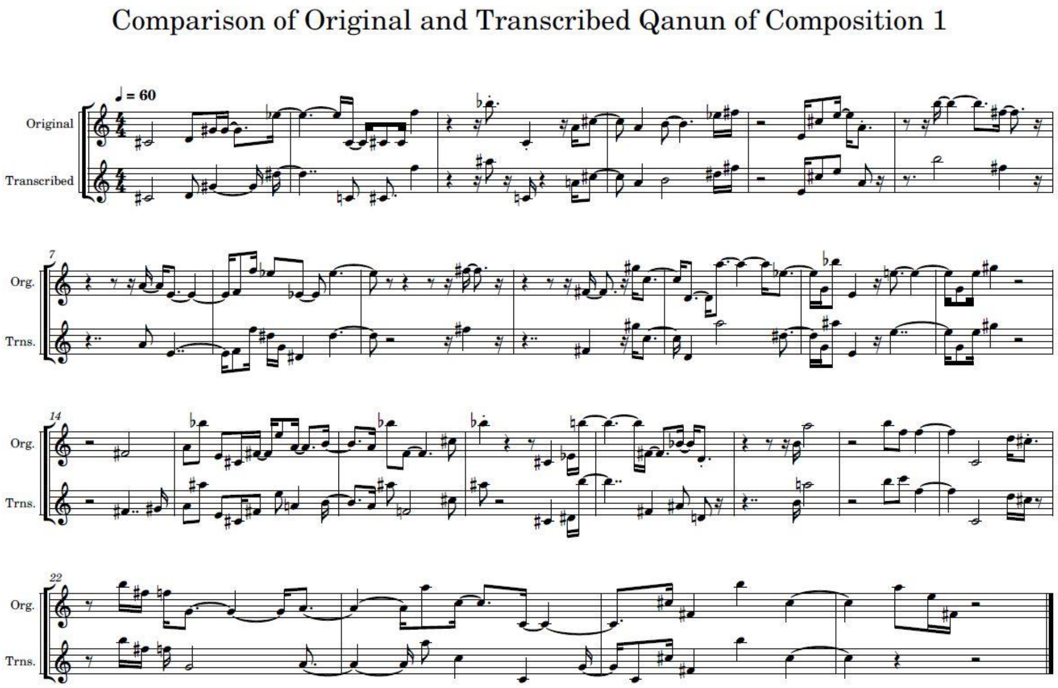
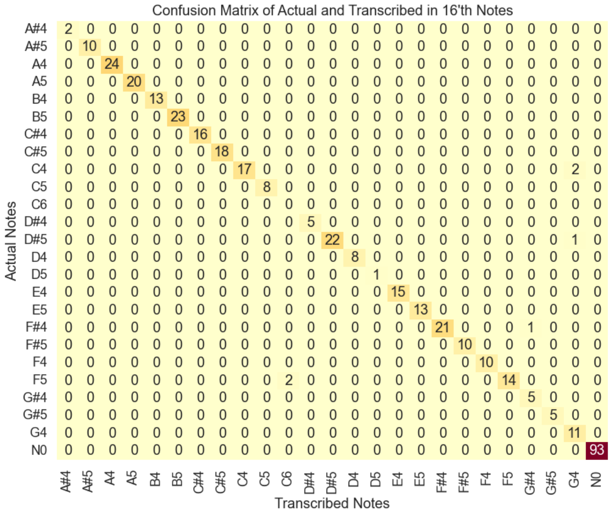

Comparison of Original and Transcribed Qanun of Composition 1
♩ = 60
Original
Transcribed
7
Org.
Trns.
14
Org.
Trns.
22
Org.
Trns.

**Figure 8.** Transcription of Extracted Qanun from Composition 1

In this transcription:
The Accuracy : 0.9846153846153847 F1 Score : 0.9873858802696013 are calculated.
The confusion matrix is given in Figure 9



| | A#4 | A#5 | A4 | A5 | B4 | B5 | C#4 | C#5 | C4 | C5 | C6 | D#4 | D#5 | D4 | D5 | E4 | E5 | F#4 | F#5 | F4 | F5 | G#4 | G#5 | G4 | N0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A#4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A#5 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A4 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A5 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B4 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B5 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C#4 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| C5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| D4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| F#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| F4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 |
| F5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 |
| G#4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| G#5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 |
| G4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 |
| N0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 93 |

**Figure 9**. The Confusion matrix recalculated using heuristics.

The Oud part's transcription results exhibit a similar quality level as depicted in Figure 10. The evaluation of these results includes the calculation of Accuracy and F1 score metrics, which are considered impressive for assessing the performance of the transcription model. In this transcription: The Accuracy : 0.9894736842105263 F1 Score : 0.9918660287081339 are calculated.

The same experiments have been repeated randomly generated compositions and the results are shown in Table 1

**Table 1.** Quality Metrics of Transcription of Qanun and Oud in different compositions

| | Qanun | | Oud | |
|---|---|---|---|---|
| | Accuracy | F1 Measure | Accuracy | F1 Measure |
| **Composition 1** | 0.984 | 0.987 | 0.989 | 0.991 |
| **Composition 2** | 0.991 | 0.990 | 0.987 | 0.989 |
| **Composition 3** | 0.989 | 0.991 | 0.994 | 0.997 |

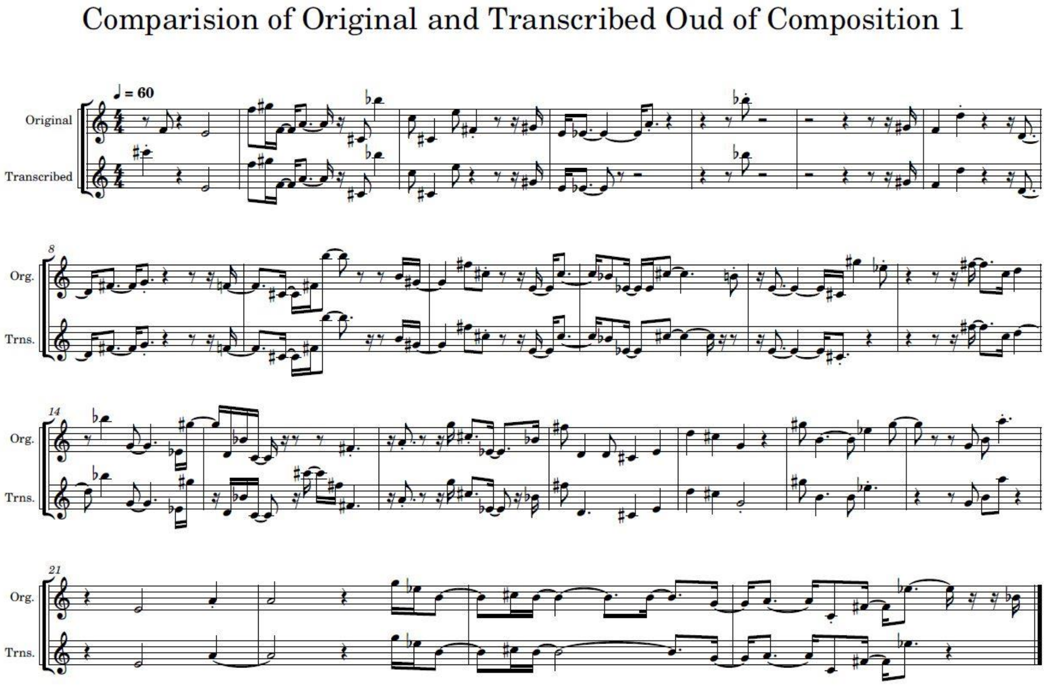

Comparision of Original and Transcribed Oud of Composition 1
♩ = 60
Original
Transcribed
8
Org.
Trns.
14
Org.
Trns.
21
Org.
Trns.

**Figure 10.** Comparison of Original and Transcribed Oud of Melody 1

To evaluate the effectiveness of the proposed model, we examined and trained the current state-of-the-art automatic music transcription models. One of the most significant challenges is that these models heavily rely on corpora based on Western music and Western instruments. For example, the commercially available Automatic Music Transcription software, AnthemScore 5 [20] attempted to transcribe the music shown in Figure 4, resulting in a highly inaccurate transcription, as depicted in Figure 11. In this software, the guitar is the closest instrument to the qanun and oud.

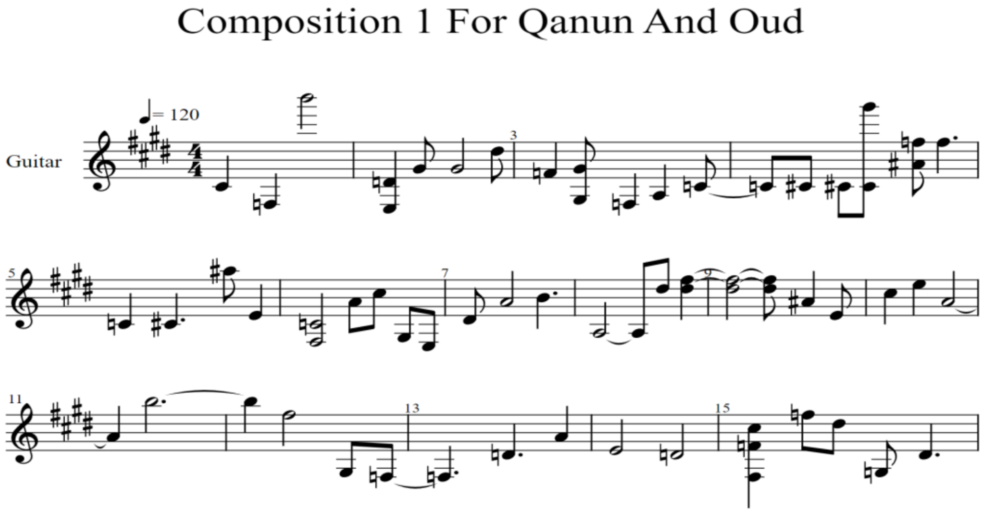


**Figure 11.** AnthemScore 5 output of Composition 1 For Qanun and Oud

Another well-known Automatic Music Transcription model, based on probabilistic latent component analysis using matrix factorization, developed by [21], was tested on the music shown in Figure 4. The output was a time series of probabilities for 80 different notes in matrix form. This matrix was then resolved, and the notes were configured. However, the model failed to separate the two instruments, and for each time sample, the most probable two note signatures were considered. The accuracy of this approach for the qanun was calculated as 0.342, and the F1 measure was calculated as 0.57, both of which are significantly lower than the proposed model.

## 4. CONCLUSION

In this research, an end-to-end automatic music transcription solution based on learning basic trained patterns has been formulated. The underlying motivation is to simulate the trained ear on specific instruments or combinations of instruments. Here, the test bench has been developed using the traditional Turkish instruments Qanun and Oud. The DNN has been trained by the basic combinations of those two instruments. Those multipitched and plucked instruments are highly susceptible to the early decay periods; however, an algorithm has also been devised to grasp the true melodic formations.

Recent developments in machine learning-based algorithms for automatic music transcription predominantly involve models trained on specific data structures. These models have been effectively applied to the analysis of Western music and the separation of Western instruments. However, the literature lacks a comprehensive corpus for authentic instruments such as the Qanun, Oud, Tambour, etc, The proposed work aims to address this gap by creating a dedicated corpus for these instruments and utilizing it for music transcription. Nonetheless, a notable limitation of the model is its dependency on the corpus for introducing and separating the instruments. Consequently, a model trained on a corpus created for instruments like the Qanun and Oud would not be applicable to the separation of other instruments.

In this research, different compositions with very complicated combinations of those instruments have been generated in polyphonic scope, and the transcription of two-part music related to two instruments has been obtained in a very satisfactory manner. Here, this approach is the fundamental step to identify a much more complex pattern regarding the Maqam, which depends on microtonal structures. In Maqam music, the whole interval has been separated into 5 to 9 microtones, and a half interval is divided into 4 microtonal intervals. This proposed approach will be improved to transcribe classical Turkish music.

The proposed approach, currently successful in transcribing complex Western music patterns, is expected to evolve and become capable of handling the sophisticated structures of classical Turkish music. This progression is not just a technical achievement but also an important cultural one. By improving the transcription of Maqam and other traditional Turkish music forms, this technology plays a vital role in preserving and promoting the understanding of Turkey's rich musical heritage. The ability to accurately transcribe and study these complex musical forms could open new avenues for musicological research and appreciation, bridging the gap between traditional musical art forms and modern technological advancements.

## CONFLICT OF INTEREST

The authors declare that they have no conflict of interest regarding the publication of this article.

## CRediT AUTHOR STATEMENT:

**Emin Germen:** Conceptualization, Methodology, Software, Formal Analyses, Writing, Validation, Investigation. **Can Karadoğan:** Conceptualization, Validation, Supervision, Resources.

## REFERENCES

[1] Benetos E, Dixon S, Duan Z, Ewert S. Automatic Music Transcription: An Overview. IEEE Signal Process Mag. 2019;36(1):20-30. doi:10.1109/MSP.2018.2869928

[2] Bertin N, Badeau R, Richard G. Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. IEEE; 2007:I-65-I-68. doi:10.1109/ICASSP.2007.366617

[3] Ansari S, Alatrany AS, Alnajjar KA, et al. A survey of artificial intelligence approaches in blind source separation. Neurocomputing. 2023;561:126895. doi:https://doi.org/10.1016/j.neucom.2023.126895

[4] Munoz-Montoro AJ, Carabias-Orti JJ, Cabanas-Molero P, Canadas-Quesada FJ, Ruiz-Reyes N. Multichannel Blind Music Source Separation Using Directivity-Aware MNMF With Harmonicity Constraints. IEEE Access. 2022;10:17781-17795. doi:10.1109/ACCESS.2022.3150248

[5] Uhlich S, Giron F, Mitsufuji Y. Deep neural network based instrument extraction from music. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. ; 2015. doi:10.1109/ICASSP.2015.7178348

[6] Nishikimi R, Nakamura E, Goto M, Yoshii K. Audio-to-score singing transcription based on a CRNN-HSMM hybrid model. APSIPA Trans Signal Inf Process. 2021;10(1):e7. doi:10.1017/ATSIP.2021.4

[7] Sigtia S, Benetos E, Boulanger-Lewandowski N, Weyde T, d'Avila Garcez AS, Dixon S. A hybrid recurrent neural network for music transcription. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2015:2061-2065. doi:10.1109/ICASSP.2015.7178333

[8] Sigtia S, Benetos E, DIxon S. An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Trans Audio Speech Lang Process. 2016;24(5):927-939. doi:10.1109/TASLP.2016.2533858

[9] Seetharaman P, Pishdadian F, Pardo B. Music/Voice separation using the 2D fourier transform. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. ; 2017. doi:10.1109/WASPAA.2017.8169990

[10] Huang P Sen, Chen SD, Smaragdis P, Hasegawa-Johnson M. Singing-voice separation from monaural recordings using robust principal component analysis. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. ; 2012. doi:10.1109/ICASSP.2012.6287816

[11] Tervaniemi M. Musicians - Same or different. In: Annals of the New York Academy of Sciences. Vol 1169. ; 2009. doi:10.1111/j.1749-6632.2009.04591.x

[12] Andrianopoulou M. Aural Education: Reconceptualising Ear Training in Higher Music Learning. Taylor & Francis; 2019. https://books.google.com.tr/books?id=p_S2DwAAQBAJ

[13] Corey J. Technical ear training: Tools and practical methods. In: Proceedings of Meetings on Acoustics. Vol 19. AIP Publishing; 2013:025016-025016. doi:10.1121/1.4795853

[14] Chabriel G, Kleinsteuber M, Moreau E, Shen H, Tichavsky P, Yeredor A. Joint Matrices Decompositions and Blind Source Separation. A Survey of Methods, Identification and Applications. Signal Processing Magazine, IEEE. 2014;31:34-43. doi:10.1109/MSP.2014.2298045

[15] Luo Z, Li C, Zhu L. A comprehensive survey on blind source separation for wireless adaptive processing: Principles, perspectives, challenges and new research directions. IEEE Access. Published online 2018. doi:10.1109/ACCESS.2018.2879380

[16] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. doi:10.1038/nature14539

[17] Sigtia S, Benetos E, Boulanger-Lewandowski N, Weyde T, D'Avila Garcez AS, Dixon S. A hybrid recurrent neural network for music transcription. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Vol 2015-August. Institute of Electrical and Electronics Engineers Inc.; 2015:2061-2065. doi:10.1109/ICASSP.2015.7178333

[18] Brown J. Calculation of a Constant Q Spectral Transform. Journal of the Acoustical Society of America. 1991;89:425. doi:10.1121/1.400476

[19] Giannakopoulos T, Pikrakis A. Introduction to Audio Analysis: A MATLAB Approach. Introduction to Audio Analysis: A MATLAB Approach. Published online 2014:1-266. doi:10.1016/C2012-0-03524-7

[20] AnthemScore 5 Music AI 2024. https://www.lunaverus.com/

[21] Benetos E, Cherla S, Weyde T. An Efficient Shift-Invariant Model for Polyphonic Music Transcription. https://code.soundsoftware.ac.uk/projects/amt_mssiplca_fast