Performance Analysis of a Convolutional Transformer-Based Automatic Chord Recognition Model for Extended Chord Qualities

Daniel Antonio de Jesus Melo[1,*], **Pedro Augusto Gomes Medeiros**[1], **Luigi Emanuel Martins Schmitt**[1], **Telmo M. Silva Filho**[2], **Yuri de Almeida Malheiros Barbosa**[1], **Thaís Gaudencio do Rêgo**[1]

[1]ARIA – Universidade Federal da Paraíba
Rua dos Escoteiros, s/n - Campus de Mangabeira, João Pessoa – PB

[2] School of Engineering Mathematics and Technology – University of Bristol
University of Bristol, Beacon House, Queens Road, Bristol, BS8 1QU, United Kingdom

danjesusmelo@gmail.com, gaudenciothais@gmail.com

Abstract. *Automatic Chord Recognition (ACR) has seen significant advancements with the advent of deep learning, particularly with convolutional transformer architectures. However, a detailed performance analysis, especially concerning their ability to recognize chord qualities, remains an area for investigation. This paper presents an analysis of a convolutional transformer-based automatic chord recognition model for extended chord qualities. We evaluate its accuracy on three benchmark datasets—McGill-Billboard, RWC-Pop, and JAAH—, focusing on root notes and specific chord qualities (maj, min, maj7, min7, 7, hdim7, and 7(b9)). Using annotations from the original ground truth files and the CHOCO (Chord Corpus) as a reference, with a duration-weighted evaluation methodology, we conduct an assessment of the model's capabilities. Our results reveal that while the model achieves high accuracy on basic triads, its ability to correctly identify extended chords drops in comparison. We conclude that the model's limitations stem from the underrepresentation of extended chords in common training corpora, underscoring the critical need for more balanced and harmonically diverse datasets to potentially mitigate the reported issues.*

1. Introduction

The Automatic Chord Recognition (ACR) field matured considerably, developing a repertoire of techniques based on probabilistic models, such as Hidden Markov Models (HMMs), and more classical machine learning approaches. This period of evolution was extensively documented in literature reviews, such as that of McVicar et al. [1], which captured the state of the art before the most recent revolution in the field. In the following years, ACR, like many other areas of artificial intelligence, underwent a profound transformation driven by deep learning. Models that treat chord sequences as a language, such as Recurrent Neural Networks (RNNs) [2], gave way to architectures based on Transformers with convolutional layers embedded within the network [3] [4], which are capable of capturing long-term dependencies in musical context and drastically elevated the level of precision.

Progress in ACR has always been intrinsically linked to the curation of high-quality datasets. Some reference datasets, such as the RWC Popular Music Database [5] and the Billboard dataset [6], was fundamental in providing robust ground truth annotations for training and evaluating ACR models. Nevertheless, dataset development brought its own challenges, notably the lack of standardized syntax for chord annotation regarding several datasets. Efforts such as the JAMS specification [7] and, more recently, the CHOCO corpus [8], sought to mitigate this problem by proposing unified annotation formats and normalizing inconsistencies between different data sources, facilitating reproducible research and comparative evaluation.

This article presents a detailed analysis of a convolutional transformer-based automatic chord recognition model performance, focusing on extended chord qualities. This model was chosen for analysis based on its extensive worldwide usage and accessibility for research purposes through the Music AI platform, which charges $0.04 per minute of processed audio content. We evaluate its performance on three reference datasets: Billboard, RWC-Pop, and the JAAH [9]; which together cover a diverse range of musical styles. Adopting an evaluation methodology that considers the chord duration as segment comparison presented in Pauwels et al. [10], with the understanding that extended chord qualities (tetrads) derive from base chord qualities (triads), and using the normalized annotations of the CHOCO corpus as an additional reference, we investigate not only the model's accuracy but also the impact of data standardization on evaluation. The objective is to provide a realistic benchmark for the performance and investigate possible improvements to it.

2. Related Works

In Korzeniowski et al. [2], the author performs a large-scale evaluation of natural language models, focusing primarily on RNN-based architectures, observing that they achieve better results regarding musical context than traditional n-gram models for chord recognition tasks. The results found in this paper were important to validate that the ACR field needed to evolve in understanding not only the temporal segmentation of chord notes, but also to consider harmonic aspects of music as a whole.

In Park et al. [3], a bi-directional Transformer architecture is applied to chord recognition, a framework that proved competitive compared to other models in some

*Supported by Music AI.

evaluation metrics, precisely because capture long-term dependencies, making use of essential information regardless of distance. With the advent of self-attention, it is also possible to extract peripheral harmonic features over a longer time window. One of the main challenges in the field is balancing the fact that complex chords, e.g., **C:minmaj7(b9,b13)**, will often not have the same representation in the data as simpler chords, e.g., **C:maj**. Addressing this issue, Rowe et al. [11] presents a convolutional transformer model trained on a large set of chords with more than 170 distinct types, and to solve the balancing problem, they implement a curriculum learning procedure, first presented by Bengio et al. [12], training scheme that proved superior to existing methods at the time of publication.

In Akram et al. [4], a conformer architecture is proposed, which integrates convolutional layers with self-attention mechanisms, effectively capturing both local spectral features and long-range harmonic dependencies. The proposed model significantly outperforms state-of-the-art methods, demonstrating superior performance in both framewise accuracy and class-wise accuracy, particularly in recognizing rare and complex chord types. To balance the relationship between chords, a re-weighted loss function was employed, ensuring balanced learning across frequent and underrepresented chord classes.

Currently, Korzeniowski et al. [2] serves as the primary reference, as it fundamentally addresses the analysis of model performance for chord recognition, whereas other studies propose new architectures to achieve performance improvements in this task. Our research initiates with a theoretical performance analysis of a specific model on specific chords, targeting their improvement as an initial focus, with the long-term goal of enhancing detection accuracy for chord qualities, that are underrepresented in training datasets commonly employed by the aforementioned works and others in the field.

3. Methodology

To develop the analysis, the first step was to find datasets with relevant chord annotations and with an interesting number of complex chords. Thus, three datasets were selected: Billboard [6], RWC-POP [5], and JAAH [9]. The evaluation of songs from these three datasets was performed using the Music AI API (https://music.ai/platform/api/) with the "Transcribe Chords" workflow. With the model inference results in hand, we performed normalization of the annotation formats from the datasets, which had been annotated differently. For this purpose, we decided to use a new annotation reference, namely the annotations of the same three previously named datasets, also found in CHOCO (Chord Corpus) [8], where the chord annotations are in .jams format (JSON Annotated Music Specification) [7].

Once the dataset annotations were converted, we could now perform a comparison between the ground truth annotations, the CHOCO annotations, and those inferred by the chord recognition model. The comparison was performed based on the evaluation segments presented on Pauwels et al. [10], with some modifications for the JAAH dataset, which did not have precise chord duration annotations, but rather a symbolic notation representing how many chords exist in each measure, the time at which beats occur, and the music's time signature. Our analysis focused on specific chord qualities: **maj**, **min**, **maj7**, **min7**, **7**, **hdim7**, and **7(b9)**, decision made regarding their harmonic functions, the II - V - I chord progression and how commonly some of these chords are found in the datasets. The corresponding intervals of each of the chords can be found in Table 1.

Chord	Intervals
maj	1, 3, 5
min	1, b3, 5
maj7	1, 3, 5, 7
min7	1, b3, 5, b7
7	1, 3, 5, b7
hdim7	1, b3, b5, b7
7(b9)	1, 3, 5, b7, b9

Table 1: Chord Qualities considered in this work and their corresponding intervals.

3.1. Chord Notation

Within the chord recognition literature, Harte et al. [13] plays a fundamental role in implementing a universal way to represent chords, their intervals, and providing a suitable format for computational tasks. The Harte library contains, by default, 48 distinct chord qualities, which does not encompass the complete vocabulary found in the selected datasets for this work. For this reason, we expanded the Harte standard dictionary, incorporating within the library's framework the relevant extended chords that were not mapped by default, making it possible to analyze more complex chords while maintaining the library's base structure. We added 41 additional chord qualities, totaling 89 distinct chord qualities, not counting the inversions that are found in the datasets analyzed in this work.

3.2. Datasets

All three datasets are available for research purposes by contacting the authors via the email addresses provided in their publications,Table 2 shows general information about each dataset used. It is worth noting that the Billboard analysis was performed using only the CHOCO annotations, the JAAH using only the original ground-truth annotations, and the RWC-Pop using both CHOCO and ground-truth. This decision was made because of the lack of precise and consistent annotations in the Billboard ground-truth and in the files available in CHOCO for JAAH. The decision to use these datasets is based on the variety of distinct chords present in the annotations, as well as the fact that we have datasets with defined musical genres in the case of RWC-Pop and JAAH, and one that possesses a variety of musical genres and also relevant density in the case of Billboard.

Dataset	Total time (h:m:s)	Number of Songs	Distinct Chords
RWC-POP	06:46:45	100	57
JAAH	06:50:24	111	56
Billboard	42:25:11	740	89

Table 2: Details about the datasets used in this work.

3.3. Data Normalization

In this work, we propose and detail a systematic methodology for converting musical annotations from .json files to the .jams format (JSON Annotated Music Specification for Reproducible MIR Research), with the aim of ensuring interoperability, reproducibility, and validation in the context of Music Information Retrieval research. The .jams format has consolidated as the standard for storage and exchange of musical annotations due to its provision of a formal JSON schema for generic annotations, support for multiple annotation layers within a single file, schema definitions for various annotation types—including beats, measures, chords, segments, and classifications—validation mechanisms and error detection capabilities. Given these advantages, converting legacy databases and custom annotations to the .jams format maximizes data reusability, facilitates fair algorithm comparisons, and simplifies automatic evaluation. As a fundamental component of this process, we used the Harte library [13] to ensure the standardization of the chord nomenclature. This library is based on Harte notation, ensuring uniformity and reducing ambiguities in chord symbols. For enhanced robustness, a chord extension dictionary was employed, enabling automatic validation of symbols found in the original files.

The conversion procedure can be subdivided into four main stages. Initially, .json files containing metadata information (title, artist, duration) and structured sequences of chords and beats are loaded. Each chord sequence is compared against the Harte library extension dictionary, identifying and registering possible invalid, or inconsistent chords, and promoting data cleaning for subsequent analysis or correction. For each section/annotation, beats serve as temporal anchor points. Thus, each chord symbol is associated with a time interval defined by the span between subsequent beats. This approach enables the precise alignment of the duration windows for each chord, which is fundamental for future quantitative analyses. Following validation and temporal mapping, a new object is created respecting the official .jams schema. This includes file metadata (title, artist, duration, etc.), annotation metadata (annotator information, source corpus, annotation type, version, etc.), inclusion of all chord sequences with start time information, duration, standardized symbol and confidence level, and registration of any unrecognized chords for posterior analysis. The generated object is serialized in .jams format through Python's standard JSON library, maintaining readable indentation and complete structural conformity. The resulting files are saved in organized directories, ready for consumption in automatic evaluation pipelines.

3.4. Evaluation Method

The performance evaluation of the model was conducted using a temporal segment comparison methodology, which can be found in Pauwels et al. [10]. The fundamental premise of this method is that evaluation should be continuous over time, rather than based on a discrete count of chord events. To implement this, the first step consists of decomposing the song timeline into a series of micro-segments. This decomposition is achieved by collecting all start and end timestamps of all chords from both the model-generated transcription and the reference annotation (ground-truth or CHOCO). The unified and ordered set of these temporal markers serves as division points, slicing the total duration of the music into contiguous intervals. The result is a new temporal representation where, within each micro-segment, it is guaranteed that no chord change occurs in either source. For example, if the model predicts a C:maj chord from 0.0s to 2.0s and the reference annotates a C:maj from 0.0s to 1.8s, the system would create two segments for analysis: one from 0.0s to 1.8s and another from 1.8s to 2.0s, allowing for granular analysis of agreement and disagreement. The purpose of this is to ensure that correct predictions and errors have appropriate weight in the performance analysis. Figure 1 illustrates the method's operation.

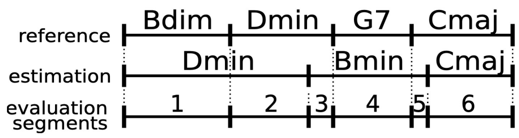

Figure 1: Example of Chord Sequence with evaluation segments.

Once the song is partitioned, comparison and score calculation occur. For each segment, the system identifies the active chord from each source (model and reference), sampling the midpoint of the interval to determine which chord represents it. From this, we evaluate the similarity between the two. The final accuracy score is calculated through a duration-weighted average. Specifically, each segment's duration is added to a "correct time" accumulator if the chords match, or to an "incorrect time" accumulator, otherwise. The final metric is the ratio between "total correct time", and the total duration of the music.

The main advantage of this approach is its ability to mitigate the impact of small discrepancies in chord boundaries (segmentation boundaries). A model that correctly identifies a chord but with slight imprecision in its start or end is penalized only for the small duration of the asynchrony, rather than having the entire chord counted as an error. This whole process results in an evaluation that more faithfully reflects the perceptual quality and practical utility of the transcription. A key consideration is that the model used in this analysis performs beat-aligned chord detection during the recognition process, which makes the implementation of evaluation segments critical for maintaining comparison fidelity.

4. Results

All chord quality accuracy rates reported in this analysis are calculated exclusively for instances where the model correctly identified the root note. Chord qualities omitted from the matrices correspond to chord types for which the model achieved zero detection accuracy in the respective datasets.

Results are presented in the following structure: comparative analysis of three datasets (RWC-POP, JAAH, Billboard) using metrics including Total Time (in seconds), chord occurrence Frequency within each dataset, and model inference Accuracy. Additionally, confusion matrices are provided for both root note and chord quality evaluation, with the latter filtered to show only the target chord qualities (maj, min, maj7, min7, 7, hdim7, and 7(b9)) emphasized in this study for improved readability.

4.1. RWC-POP

Upon analyzing the model's inferences, the results contained in Table 3 were obtained, which demonstrates a large representativeness of **maj** chords in the dataset, and a very small one in relation to **hdim7** and **7(b9)**, that are part of the focus of our performance analysis. Observing the accuracy rate, we can observe some interesting cases: the **7** chord type obtained *0.69* of accuracy, despite having many appearances throughout the dataset, while at the same time, the few times that the **hdim7** chord appeared, it obtained *0.88* of accuracy, which could indicate that within the songs used to train this model, the **hdim7** chord appeared more than the **7** chord or that it is so unique that does not need many examples to be correctly detected.

An analysis of the confusion matrix presented in Figure 2 reveals that the model systematically chooses sharp (#) over flat (b) notation for enharmonic representations. This preference does not constitute a fundamental limitation, as the model operates without the constraint of tonal center-based chord detection. Analysis of the chord quality confusion matrix presented in Figure 3 was restricted to our selected target chords, as previously explained. The results reveal that the **7(b9)** chord was predominantly inferred as simply **7**, demonstrating an additional characteristic of the model: when unable to accurately determine chord extensions, it defaults to the simpler nomenclature with higher confidence - in this instance, the basic **7** chord.

Chord	Total Time	Frequency	Accuracy
maj	10810.61	12198	0.87
min	3518.80	3756	0.74
maj7	1787.81	1685	0.76
min7	3346.13	3890	0.81
7	1675.22	1879	0.69
hdim7	94.68	97	0.88
7(b9)	13.72	18	0.03
maj6	569.07	722	0.04
sus4	344.90	429	0.40
dim	197.79	289	0.47
aug	99.47	127	0.20
min6	70.51	64	0.03
7(13)	31.62	30	0.25
maj7(9)	6.65	8	0.21

Table 3: RWC-POP Results

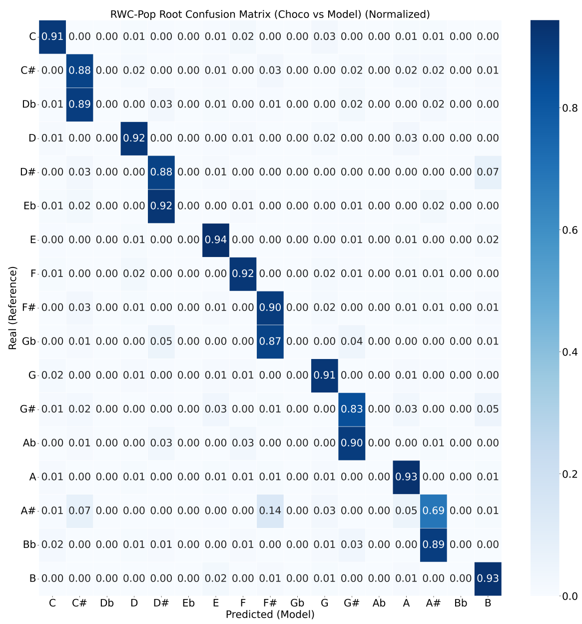

Figure 2: RWC Chord Root Matrix - CHOCO vs Model Inference.

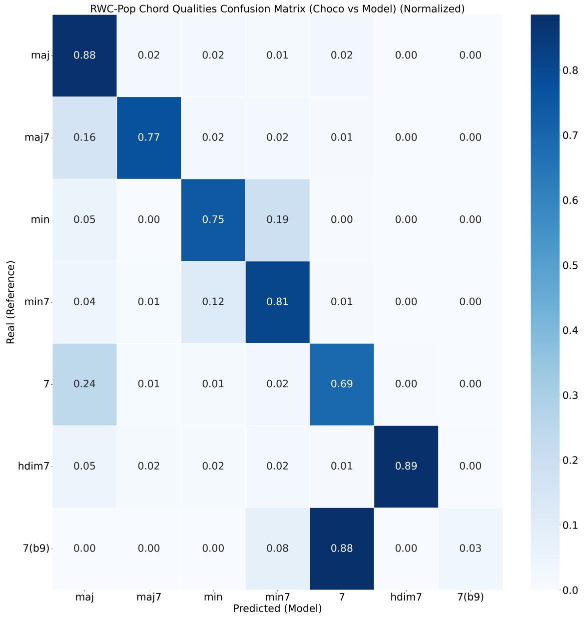

Figure 3: RWC Chord Quality Matrix - CHOCO vs Model Inference.

4.2. JAAH

Being a dataset with songs restricted to a single genre, jazz, this enables evaluation of the model's performance on chords with **7**, primarily, since it is the most representative chord extension in the dataset. The overall accuracy for this dataset decreased considerably, even for major and minor triads, a situation that can be evaluated as being the result of few jazz examples used for model training. Even though the model detects chords, the musical genre influences the training process, and in this specific case, the results shown in Table 4 strongly indicate this.

Among our most significant findings is the result presented in Figure 4, which shows considerable variation in chord root note detection accuracy, generally reflecting low performance levels. This phenomenon stems from JAAH's extensive use of inverted chords, including not only conventional first, second, and third inversions, but also inversions based on extension tones such as the **6th**, **9th**, and **11th** degrees. This characteristic, combined with the limited representation of inverted chords in standard training datasets within the field, directly impacts the performance metrics observed in this confusion matrix.

Another analysis that can be made by observing Figure 5 is the model's strong tendency to classify chords as being **7** chords, where in this specific dataset many chords do indeed include the minor seventh, but with other notes forming the chord extensions.

Chord	Total Time	Frequency	Accuracy
maj	3997.67	4354	0.42
min	1139.48	1247	0.31
maj7	1029.77	1339	0.21
min7	2917.54	3915	0.19
7	8679.20	10326	0.54
hdim7	450.52	512	0.07
7(b9)	446.26	524	0.03
maj6	945.75	1296	0.08
dim7	506.82	780	0.05
min6	588.46	623	0.18
7(13)	171.85	254	0.04
7(9)	117.53	119	0.04
dim	54.73	88	0.27
minmaj7	59.87	68	0.01
aug	25.64	35	0.08

Table 4: JAAH Results

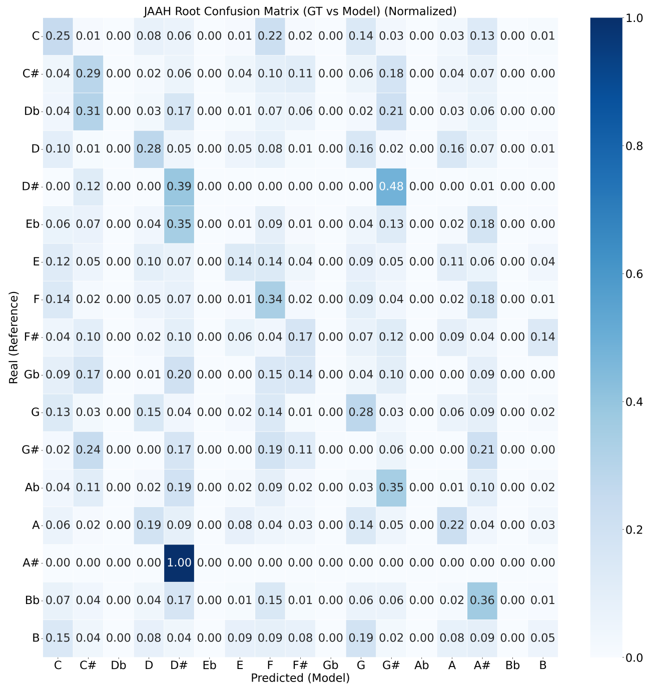

Figure 4: JAAH Chord Root Matrix - Ground Truth vs Model Inference

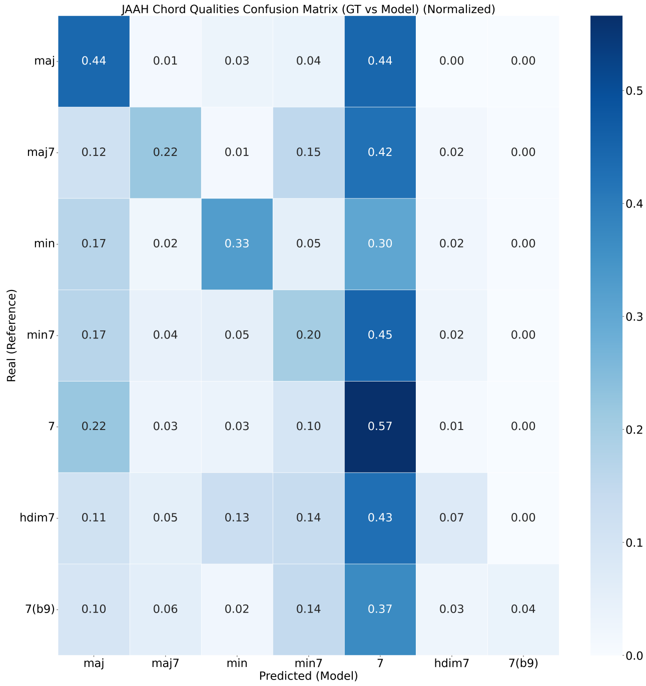

Figure 5: JAAH Chord Quality Matrix - Ground Truth vs Model Inference

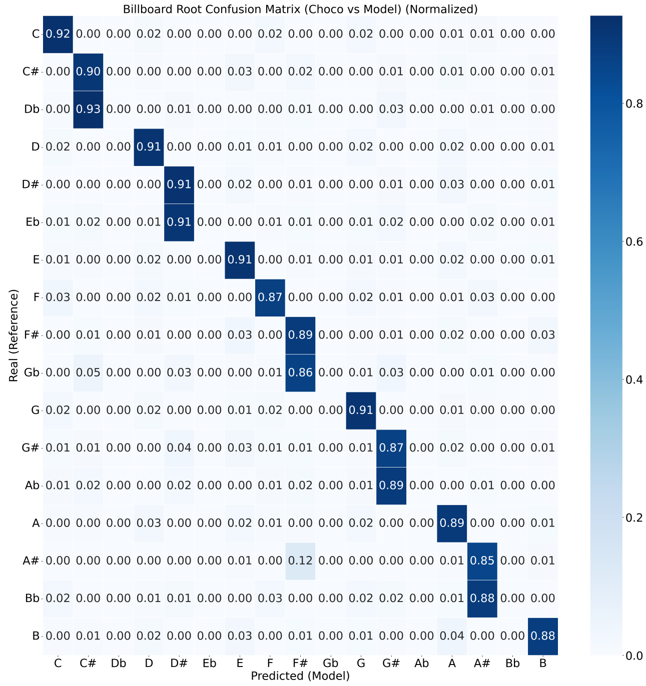

Figure 6: Billboard Chord Root Matrix - Choco vs Model Inference

4.3. Billboard

This dataset contains by far the largest number of songs among the three analyzed, once again producing results consistent with our initial hypotheses. The difference in precision between **hdim7** and **7(b9)** reveals that the model demonstrates significantly greater proficiency in identifying the former compared to the latter, despite their similar representation levels in the evaluated dataset. Notably, additional chord types beyond our primary focus were included to provide context for overall performance, as shown in Table 5. Metrics are presented in Figure 6 and Figure 7.

Chord	Total Time	Frequency	Accuracy
maj	74281.44	92574	0.92
min	18150.18	20782	0.79
maj7	3972.14	4097	0.66
min7	11867.88	14051	0.71
7	14374.99	16532	0.66
hdim7	238.05	308	0.62
7(b9)	150.27	229	0.05
5	2440.33	3521	0.15
maj7(9)	1946.87	2145	0.12
sus4	1421.66	2126	0.19
maj6	1504.06	1994	0.25
7(9)	1183.60	1099	0.46
dim	280.52	458	0.65
min6	343.43	432	0.43
aug	69.66	90	0.54
minmaj7	48.98	64	0.36

Table 5: Billboard Results

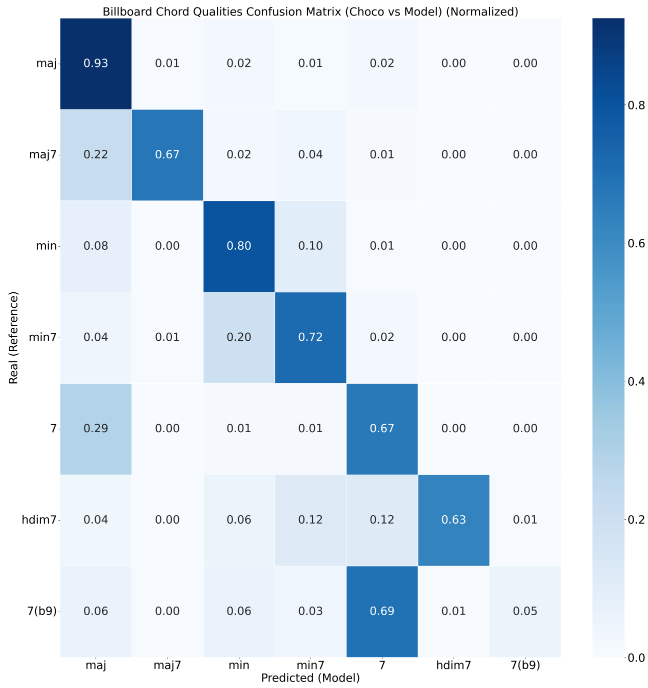

Figure 7: Billboard Chord Quality Matrix - Choco vs Model Inference

5. Discussion

To provide a comprehensive understanding of the model's performance across different musical contexts, we present a comparative analysis of the three evaluated datasets. Table 6 summarizes the average performance metrics across all datasets, revealing significant patterns in the model's behavior.

Chord	Avg Total Time	Avg Frequency	Avg Accuracy
maj	29696.57	36375.33	0.74
min	7602.82	8595.00	0.61
maj7	2263.24	2373.67	0.54
min7	6043.85	7285.33	0.57
7	8243.14	9579.00	0.63
hdim7	261.08	305.67	0.52
7(b9)	203.42	257.00	0.04
maj6	1006.29	1304.00	0.12
sus4	883.28	1277.50	0.30
5	813.44	1173.67	0.05
maj7(9)	976.76	1076.50	0.17
7(9)	650.57	609.00	0.25
min6	334.14	373.00	0.21
dim	177.68	278.33	0.46
aug	64.92	84.00	0.27
minmaj7	36.28	44.00	0.12

Table 6: Average Results Across Datasets

The comparative analysis reveals several critical insights that extend beyond individual dataset performance. First, the model demonstrates a clear hierarchy in chord recognition accuracy that correlates inversely with harmonic complexity. Basic triads (**maj**: 0.74, **min**: 0.61) significantly outperform extended chords, with **7(b9)** achieving only 0.04 average accuracy across all datasets.

Second, the substantial variation in performance across datasets highlights genre-specific challenges. The JAAH dataset's jazz context consistently produced the lowest accuracy rates, even for fundamental chord types, suggesting that the model's training corpus inadequately represents jazz harmonic language. This finding is particularly significant given that jazz extensively employs the extended chords that show poor recognition performance.

Third, the frequency-accuracy relationship reveals effects of the imbalance in training data. While **maj** chords appear 36,375 times on average with 0.74 accuracy, **7(b9)** chords appear only 257 times with 0.04 accuracy. However, **hdim7** chords, despite having similar low frequency (305.67 average), achieve substantially better accuracy (0.52), suggesting that harmonic distinctiveness may compensate for limited training examples in some cases.

6. Conclusion and Future Work

The model's overall performance reveals a strong bias toward correctly identifying the fundamental components of base chords (triads), while exhibiting diminished performance on extended chords. This finding was expected, considering that the datasets available for training in this field mostly lack sufficiently extended chords to enable high-complexity models, such as those cited in this work, to effectively learn to detect these types of chords. Considering that the tests were conducted on a model to which we do not have detailed access, only the API to use it in its current state, the research on this topic will proceed by training, validating, and testing a model with architecture similar to the one evaluated here using different datasets and trying to find a balance in recognition performance between base chords and extended chords. This approach is expected to yield more accurate results, as our analysis clearly demonstrated that the predominance of base chord qualities in training data leads to significantly reduced accuracy for extended chords, which are underrepresented in standard training datasets.

References

[1] Matt McVicar, Raul Santos-Rodriguez, Yizhao Ni, and Tijl De Bie. Automatic chord estimation from audio: a review of the state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):556–575, 2014.

[2] Filip Korzeniowski, David R. W. Sears, and Gerhard Widmer. A large-scale study of language models for chord prediction. *arXiv preprint arXiv:1804.01849*, 2018.

[3] Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim, and Jonghun Park. A bi-directional transformer for musical chord recognition. *arXiv preprint arXiv:1907.02698*, 2019.

[4] Muhammad Waseem Akram, Stefano Dettori, Valentina Colla, and Giorgio Carlo Buttazzo. Chordformer: a conformer-based architecture for large-vocabulary audio chord recognition. *arXiv preprint arXiv:2502.11840*, 2025.

[5] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, 2002.

[6] John Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 633–638, 2011.

[7] Eric Humphrey, Justin Salamon, Oriol Nieto, Jon Forsyth, Rachel Bittner, and Juan Bello. Jams: a json annotated music specification for reproducible mir research. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 2014.

[8] Jacopo De Berardinis, Albert Meroño-Peñuela, Andrea Poltronieri, and Valentina Presutti. Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. *Scientific Data*, 10(1):641, 2023.

[9] Vsevolod Eremenko, Emir Demirel, Baris Bozkurt, and Xavier Serra. Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. *Zenodo*, 2018.

[10] Johan Pauwels and Geoffroy Peeters. Evaluating automatically estimated chord sequences. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 749–753, 2013.

[11] Luke Rowe and George Tzanetakis. Curriculum learning for imbalanced classification in large vocabulary automatic chord recognition. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[12] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48. ACM, 2009.

[13] Christopher Harte, Mark Sandler, Samer Abdallah, and Emilia Gómez. Symbolic representation of musical chords: a proposed syntax for text annotations. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 66–71, 2005.