,,

Universiteit
Leiden
The Netherlands

Bachelor Data Science & Artificial Intelligence

ClarinetNet:

Enhancing the transcription of wind instruments.

Dominic Tsoi

Supervisors:
Rob Saunders & Peter van der Putten

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS) www.liacs.leidenuniv.nl

26/08/2025

Abstract

Automatic Music Transcription (AMT) has made significant advances through the adoption of neural networks and deep learning, yet wind instruments remain underrepresented in existing datasets, leading to suboptimal transcription performance. This research investigates whether training AMT models on single-instrument datasets can improve transcription accuracy compared to multi-instrument datasets. This paper introduces ClarinetNet, a specialized dataset containing approximately one hour of clarinet music across 21 compositions, designed to address the gap in wind instrument representation. A transcription model was trained and evaluated specifically for clarinet music using this dataset. The experiments used multiple testing strategies including cross-validation and representative set selection. The results show that while ClarinetNet achieved reasonable performance with F1-scores ranging from 0.60 to 0.67, it did not surpass the 0.868 F1-score achieved by multi-instrument datasets for clarinet transcription. Despite not proving the initial hypothesis, this work demonstrates the viability of single-instrument datasets for AMT and establishes a foundation for future research in wind instrument transcription. The study contributes valuable insights into dataset size requirements, representativeness challenges, and the potential for specialized approaches in automatic music transcription.

Contents

1 Introduction

Automatic Music Transcription (AMT) has been around for several decades. As early as the 18th century, people were trying to create AMT systems, such as the mechanical Reproducing Piano [Ins22], where, while playing the instrument, a mechanism writes down markings on a moving strip of paper. In terms of more digital registration, one could say that the oscilloscope is one of the first transcription systems, since it can create a visual representation of sound. Of course, it does not have the capacity to actually transcribe music. One of the first systems to do so is *Traditional Signal Processing*, from the 1980s, quickly followed by *probabilistic modeling* systems and *hidden Markov models* (early 2000s), eventually resulting in what is now state of the art, *neural networks*, and *end-to-end audio-to-notation systems* [BDDE19].

The goal of AMT is to transform raw music audio into a symbolic representation that can be read or analyzed by humans or computers. This symbolic representation of music is commonly presented as sheet music for humans and MIDI-files for computers. Traditionally, transcription is performed by ear, where a human listens to the musical piece and writes the notes down. This often involves replaying the song or specific parts multiple times, maybe even trying to copy the music by playing on a piano or other instrument, to ensure accuracy. A musical piece is often dissected into its different musical components, which are separately notated, meaning that one often starts with identifying the main melody and works their way down from bass lines and harmonics all the way to articulation and dynamics. [BL23] Since this process relies on cognitive functions, it follows that artificial intelligence (AI) might be able to do this, as well.

To effectively train AMT models, suitable datasets are essential. While numerous datasets exist, they often fail to adequately represent wind instruments. For example, data sets such as MusicNet [THK17] and URMP [LLD+16] feature a variety of instruments but lack substantial representation of any single wind instrument. However, specialized datasets, such as GuitarSet [XBP+18] and MAESTRO [HSR+19] focus on singular instruments (guitar and piano respectively), but there is no equivalent dataset for any singular wind instrument. This causes AMT models to perform worse for some other singular instruments than others, as shown by Maman et al. [MB22].

In this paper, we will demonstrate the construction of a singular-instrument dataset for AMT and outline the criteria that it must meet to achieve strong performance. We will make use of the labeling code developed by Maman et al.(2022) [MB22] to prepare the data for the transcription process. Additionally, we will evaluate how a singular instrument dataset compares to a general-purpose dataset like MusicNet and URMP. With this research, we will explore whether we can improve the automatic transcription of singular instruments by using their own specialized dataset. Here we will do this by expanding the clarinet transcription abilities of AMT models, by creating a dataset of clarinet music that previously did not exist. Based on the context described, the research question driving this study is as follows.

To what extent does training on a single-instrument clarinet dataset achieve higher transcription accuracy compared to training on multi-instrument datasets for clarinet music transcription?

The remainder of this thesis is structured as follows. Section 2 provides background information on automatic music transcription and reviews related work in the field, including current AMT models and existing datasets. Section 3 presents the methodology used to create the ClarinetNet dataset, including dataset composition criteria, data collection procedures, labeling methods, and model training and testing approaches. Section 4 presents the experimental results comparing single-instrument and multi-instrument dataset performance. Section 5 discusses the findings, limitations, contributions, and suggestions for future work. Finally, Section 6 concludes the thesis by summarizing the key findings and their implications for automatic music transcription research.

2 Background and Related Work

Automatic music transcription has evolved significantly over the decades, progressing from early signal processing approaches to modern neural network-based systems. This section reviews the development of AMT systems and examines current datasets, with particular focus on the challenges facing wind instrument transcription.

2.1 Historical development of AMT Systems

Trying to create an automatic process for transcribing music has been around for multiple decades. Klopuri and Davy [KD06] identified that one of the first known attempts was made by Moorer in 1975 [Moo75], where he proposed a system that could transcribe two-voice polyphonic compositions. The system consists of different modules that extract different features at three hierarchical levels: low-level signal processing for musical harmony detection and harmonic extraction, intermediate-level processing for note inference and scoring, and high-level processing for musical structure analysis and voice separation. Following Moorer's research, multiple similar systems have been created for different musical purposes, such as Chaf et al. creating a system that identifies notes in polyphonic piano music [CJK+85], Robert's system to specifically isolate specific tunes or melodies from a general ensemble [Mah89] or M. Piszczalski's way of transcription of monophonic music for any instrument [Pis86]. The temporal nature of music initially made AMT resistant to early neural network approaches, but developments in recurrent neural networks in the 2000s and more recent transformer architectures have enabled significant breakthroughs in the field. [Sch15]

2.2 Current AMT models and Datasets

This brings us to the developments in music datasets. Any AI model is heavily dependent on data to allow it to learn to perform its task. This also applies to AMT models, of which MT3 [GSM+21] is an example. A prominent example of such a model is the MT3 model developed by Gardner et al. [GSM+21]. Their research demonstrates that MT3 can not only transcribe solo performances of individual instruments, but also transcribe multi-track music, such as pop/rock bands and wind ensembles. To achieve this versatility, the model was trained on a variety of datasets.

There are a large number of musical datasets, each containing several minutes of music from different instrumentations. Some popular examples are, as highlighted by Ji et al. [JYL23], the Lakh MIDI Dataset (LMD) [Raf16], which incidentally is one of the largest symbolic music corpus. Another is the MAESTRO dataset [HSR+19] that has more than 172 hours of piano performances. Due to it having only piano music in its contents, we call this a single-instrument dataset. Another example of a single-instrument dataset is the GuitarSet dataset [XBP+18], which contains about 3 hours of music. There are also some multi-instrument datasets, such as the MusicNet dataset [THK17], which features 20 different instrumental compositions with 11 different

test instrument	note with instrument F1
Violin	87.3
Viola	61.1
Cello	79.9
Bassoon	78.0
Clarinet	86.8
Horn	75.0

Table 1: Instrument-specific F1 scores (in percentages) reported by Maman et al. [MB22] for their transcription model.

instruments in 34 hours of classical music, or the Slakh2100 dataset [MWSR19], which is derived from the LMD, but is refined to be brought down to 4 core instruments (piano, guitar, bass, and drums) and includes multiple styles of music.

2.3 Challenges in Wind Instrument Transcription

Despite its strengths, MT3 encounters challenges in transcribing music involving wind instruments. Specifically, its performance is suboptimal when tested with the MusicNet [THK17] and URMP [LLD+16] datasets. Gardner et al. [GSM+21] attribute this issue to misalignment between the audio and labels in these datasets, which hampers the model's effectiveness. That is, the timestamp labels assigned to each note in the recording did not match with the audio.

Furthermore, wind instruments present unique challenges for AMT systems compared to keyboard or string instruments. The monophonic nature of wind instruments means that there are no chords, but there are other problems one needs to account for: breath attacks create distinctive onset characteristics, and vibrato and dynamic swells affect pitch stability. Additionally, wind instruments may show timbral variations throughout their range, which basically means the color or tone of an instrument. This can be best understood by listening to and comparing, for example, a classical clarinet solo or a klezmer one. Although they are the same instrument, they sound completely different, even if they play the same notes. It is really just a different style of playing, which is noticeable when listening to it. Timbral differences also appear simply based on which person plays the instrument. This, along with the register breaks, which are jumps quickly between low and high notes, can create discontinuities and confuse some algorithms.

This misalignment issue has been highlighted by other researchers, including Maman et al. [MB22], who developed a method to generate more accurate labels for existing and newly created datasets. Their approach differs from traditional alignment methods by first training a neural network on synthetic data, then using the network's predictions as likelihood terms for Dynamic Time Warping (DTW), rather than relying solely on spectral features. This allows for better handling of inconsistencies between audio recordings and their corresponding MIDI scores, such as repeated cadenzas or subtle timing differences in trill and chord arpeggiation. Their work introduced the MusicNetEM dataset, an improved version of the MusicNet dataset with improved label alignment. By applying their labeling method, they reported significant improvements in AMT performance.

Furthermore, Maman et al. evaluated how well their model transcribed different instruments and reported significant performance variations between instruments. As shown in Table 1, clarinet achieved strong transcription accuracy with an F1-score of 86.8%, while instruments such as the viola showed a lower performance at 61.1%. The specific evaluation metrics and detailed performance analysis are discussed further in Section 3.6.

3 Methodology

To answer the research question, a clarinet-specific dataset[1] was created and labeled using the Maman et al. [MB22] alignment method. The labeled dataset was then used to train a transcription model based on Hawthorne et al. [HSR+19] architecture to test how it performed compared to the multi-instrument baseline reported by Maman et al. [MB22]. The code used for this process is available via a Google Colab page[2].

Training an AMT model on a dataset containing only clarinet music should produce better transcription results for the clarinet than models trained on datasets with many different instruments. To test this idea, this research introduces *ClarinetNet*, a dataset made specifically for clarinet music, and is used to train a transcription model. We followed the same methods used by Maman et al. [MB22] for training and evaluation.

The same labeling method from Maman et al. [MB22] is used to create accurate labels for our dataset, and then adapted their training process to work with the clarinet-only data. The resulting model is tested using different approaches to see how well it performs in clarinet transcription using tests sets derived from the ClarinetNet dataset itself. The model's F1 scores are then compared with the clarinet F1 score reported by Maman et al. [MB22].

Our methodology consists of five main steps: dataset composition, data collection, labeling, model training, and evaluation. For the dataset creation, specific criteria regarding size, content, quality, and structure must be met to function effectively with the model and labeling code.

3.1 Dataset Size

The original dataset used in the research by Maman et al. (2022) [MB22] is the MusicNet dataset [THK17]. The MusicNet dataset consists of 2048 minutes of music, which is about 34 hours. It is composed with the data from multiple recordings of multiple instrumental compositions, such as solo piano, but also string sixtet, and clarinet quintets. In total, there are 20 different ensembles made up of 11 different instruments. Of the entire dataset 173 minutes (2.9 hours) are clarinet tunes, which is about 8%. This amount of data allowed Maman et al. (2022) [MB22] to produce its results.

This also aligns the sizes of some singular instrument datasets. One of these is the GuitarSet [XBP+18], which also has about 3 hours of music. However, a guitar is a multiphonic instrument, which means that it can produce multiple notes at the same time in the form of chords. This can mean that there are multiple combinations of notes, the 3 hours might be a minimum size of a singular multiphonic dataset.

It should also be mentioned that MusicNet does not have an instance of a solo clarinet in its dataset. It is always in ensemble with other clarinets or instruments. Since a clarinet is generally a monophonic instrument, which means that it can only produce one note at a time, a clarinet

[1] Github page: https://github.com/Dominiq7/ClarinetNet

[2] Google Colab link: https://colab.research.google.com/drive/1R9nI2OJNnQUyO8IeFsDxz21HhI50OqIk?usp=sharing

dataset could be of a smaller size than 3 hours. As such, the ClarinetNet dataset will be composed with the aim of being about 1 hour of music.

3.2 Dataset Content and Quality

The dataset focuses on clarinet music, where each piece features a clarinet as the main melodic instrument. Due to limited resources for solo clarinet recordings and MIDI files, non-clarinet accompaniments are permitted, though they should at no point be dominant in the musical composition. The recordings must be of high audio quality, free from static or background noise, and each song should appear only once.

3.3 Data Collection

To collect data, we explored the IMSLP public library [IMS], which offers a wide catalog of clarinet pieces with recordings. I filtered for solo clarinet works, selecting diverse pieces that cover the clarinet's range and are musically grounded. Free improvisations were excluded because they do not have any sheet music on which to base. The total number of pieces from the first selection was 83, but this was dwindled down to 21 pieces.

For each selected piece, we located or created a recording and used Musescore (specifically Musescore 3) notation software [Mus] to generate a corresponding MIDI file. The way this was done was by copying over the notes (note for note) from the original sheet music into the software. Needless to say, this is one of the most time-consuming parts of the entire project. Producing a MIDI file for a piece of 5 minutes takes about 45 minutes, which means that the entire dataset took nearly 10 hours to produce MIDI files. This in addition to the few pieces that were excluded from the datasets, due to there being no audio recording available. The software also assisted in rehearsing the parts and ensured that the music was ready for recording. Of the final 21 pieces in the dataset, 5 had a new recording made and 16 had an online recording already available.

3.4 Data Labeling

After compiling the dataset, we labeled the data using Maman et al.'s [MB22] code. This approach differs from traditional alignment methods by first training a neural network on synthetic data and then using the network's predictions as likelihood terms for Dynamic Time Warping (DTW), rather than relying solely on spectral features.

Traditional spectral-based alignment methods struggle with inconsistencies between audio recordings and their corresponding MIDI scores, such as repeated cadenzas or subtle timing differences in trill and chord arpeggiation. Maman et al.'s method addresses these limitations by using neural network predictions as more informative alignment descriptors. The network generates onset, frame, and offset predictions for each audio recording, which serve as likelihood terms for DTW alignment with the corresponding MIDI score.

The labeling process analyzes each audio file and links MIDI messages with the correct notes based on the network's predicted probabilities. These alignments are saved in TSV (Tab-Separated Values) files, where the onset, offset, and pitch of each note are registered. This is done for every piece of the dataset.

The final ClarinetNet dataset consists of 21 compositions totaling approximately one hour of clarinet music. Table 2 provides detailed information about each piece, including the composer, title, duration, and recording source.

ID	Composer	Title	Duration	Self-made recording
1	Rother	A Tune	0:03:29	Yes
2	Kraussold	Abschied Im Herbst	0:03:10	No
3	De Bleser	Adagio et Allegro	0:03:25	No
4	Gubatz	Barcarole	0:02:22	No
5	Harrington	Blue Miniature	0:01:03	No
6	Higgins	Capriccio on Prokofiev	0:03:52	No
7	Sbraccia	Chiacchierina	0:02:08	Yes
8	St Pierre	Clarifolia	0:03:18	No
9	Vay	Clarinet Solo in C Major	0:04:44	Yes
10	Batista	Danca Espanhola	0:01:18	Yes
11	Tarantola	Diaz	0:02:43	No
12	Pido	Dichotomy	0:01:38	No
13	Gubatz	Dunkle Wolken	0:06:44	No
14	Florczak	Etude for Solo Clarinet	0:02:47	Yes
15	Cibisescu-Duran	Evolutii	0:06:08	No
16	Nichifor	Joke for Andrew Simon	0:01:17	No
17	Bedetti	Music for the King	0:02:14	No
18	De Bleser	Pele Mele	0:04:56	No
19	Gubatz	Renegade	0:01:58	No
20	Sauter	Scherzando	0:01:04	No
21	De Bleser	Uitbreidingen en Verplaatsingen	0:04:29	No
		Total:	1:04:47	

Table 2: ClarinetNet contents; showing the entries ID numbers, composer, title, duration, and whether the recording was made by me.

3.5 Model Training and Testing

For training we use the model used in the research of Maman et al. [MB22], who in turn use the Onset and Frames architecture proposed by Hawthorne et al. [HSR$^+$19]. This architecture was proven to be state-of-the-art for supervised piano transcriptions. It features separate detection heads for note onsets and offsets, and frames of notes (i.e. a group of active/playing notes). The architecture in this research is an enhanced version of Hawthorne et al.'s [HSR$^+$19] architecture, improving its ability to transcribe multi-instrument music. The network processes log-mel spectograms of the audio, through three convolutional layers (64, 64, 128 filters), bidirectional LSTM layers (384 units), fully connected layers (1024 units) and the same three detection heads as the original architecture for onsets, offsets, and frames. The enhanced architecture uses larger layers, which allows it to better handle the complexity of multi-instrumental music.

For evaluation, we use the F1-score, also known as F-measure, which is the harmonic mean of a model's precision (the amount of true positives of all made predictions) and recall (the amount of

Testing strategy	Test set	F1 score model
Entire dataset	All	0.6555
Best split	8, 15, 18	0.6689
Fold 1	1, 11, 14, 20	0.6044
Fold 2	3, 6, 10	0.6670
Fold 3	7, 17, 19, 21	0.5248
Fold 4	4, 12, 13, 15, 18	0.6748
Fold 5	2, 5, 8, 9, 16	0.5351
Average total	—	0.6012
—	Maman et al.	0.868

Table 3: F1 scores calculated during the different testing strategies, with the F1-score result from Maman et al. on clarinet transcriptions.

true positives made on all positive instances). It is a measure that is often used to see how well a model performs in its predictive performance. It's value is between the range of 1 and 0 (though sometimes it is shown as a percentage), where 1 would mean perfect precision and recall, and 0 would mean the worst. Its definition is as follows:

$$\begin{aligned} F_1 &= \tfrac{2 \times Precision \times Recall}{Precision + Recall} \\ Precision &= \tfrac{TruePositives}{TruePositives + FalsePositives} \\ Recall &= \tfrac{TruePositives}{TruePositives + FalseNegatives} \end{aligned}$$

In general, multiple approaches to testing and training splits were used. Based on the training script from the original architecture, we derived a testing script, with which we applied different strategies for dataset splits. Initially, the model was trained and tested with the entire dataset. For the other methods, a training and test set was created in which each dataset was used for its corresponding phase. Here, an 80-20 split was applied. For the first split, the note distributions of each song in the dataset were analyzed and the tracks that have the most representative note distribution of the entire dataset were chosen. For the second split, possible folds were determined and trained and tested accordingly. After calculating all the resulting F1-scores, the average was taken and recorded as the result. The representability of each fold's test set on their train set is shown in Figure 3.

4 Results

Table 3 shows the F1-scores obtained from the models trained and tested using the different data splitting strategies. The model trained and tested on the entire dataset achieved an F1-score of 0.6555. When using the representative test set consisting of pieces 8, 15, and 18, the model achieved an F1-score of 0.6689. Cross-validation results across five folds showed considerable variation, with F1-scores ranging from 0.5248 (Fold 3) to 0.6748 (Fold 4). The average F1-score across all cross-validation folds was 0.6012.

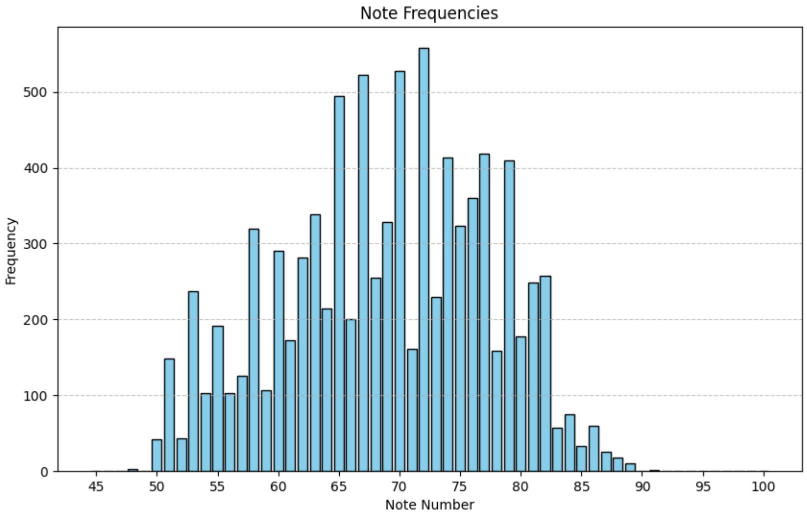

Figure 1: The entire note distribution of the ClarinetNet dataset.

Figure 1 displays the overall note distribution across the entire ClarinetNet dataset. The distribution spans MIDI notes 45 to 95, corresponding approximately to the range from A2 to B6. The distribution shows peak activity concentrated around notes 65-75 (F4 to D#5), with frequency counts reaching over 500 occurrences for the most common notes. Note frequencies generally decrease toward both the lower and upper extremes of the range.

Figure 2 presents individual note distributions for each of the 21 pieces in the dataset. The distributions reveal substantial variation in range usage across pieces. Some compositions, such as pieces 6, 13, and 18, exhibit broader note distributions spanning most of the clarinet's range, while others like pieces 5, 12, 16, and 20 show more concentrated usage in specific registers. Several pieces demonstrate bimodal distributions with activity in both middle and upper registers.

Figure 3 illustrates the note distributions for each training-test split used in the cross-validation analysis. The training sets consistently show broader, more comprehensive note distributions compared to their corresponding test sets. Fold 4, which achieved the highest F1-score (0.6748), used test pieces 4, 12, 13, 15, and 18. Fold 3, which produced the lowest F1-score (0.5248), tested on pieces 7, 17, 19, and 21. The representative test set (pieces 8, 15, 18) demonstrates a note distribution pattern that more closely matches the overall dataset distribution compared to other fold combinations.

5 Discussion

The aim of this research was to see whether a single-instrument dataset could outperform a multi-instrument dataset in automatic music transcription. From the results given by Maman et al. [MB22] we know that a multi-instrument dataset like MusicNet [THK17] can give an F1-score of 0.868 for clarinet transcription. With ClarinetNet, a model with the same architecture produces a result of 0.6555 when trained and tested with the entire dataset, 0.6012 when trained and tested

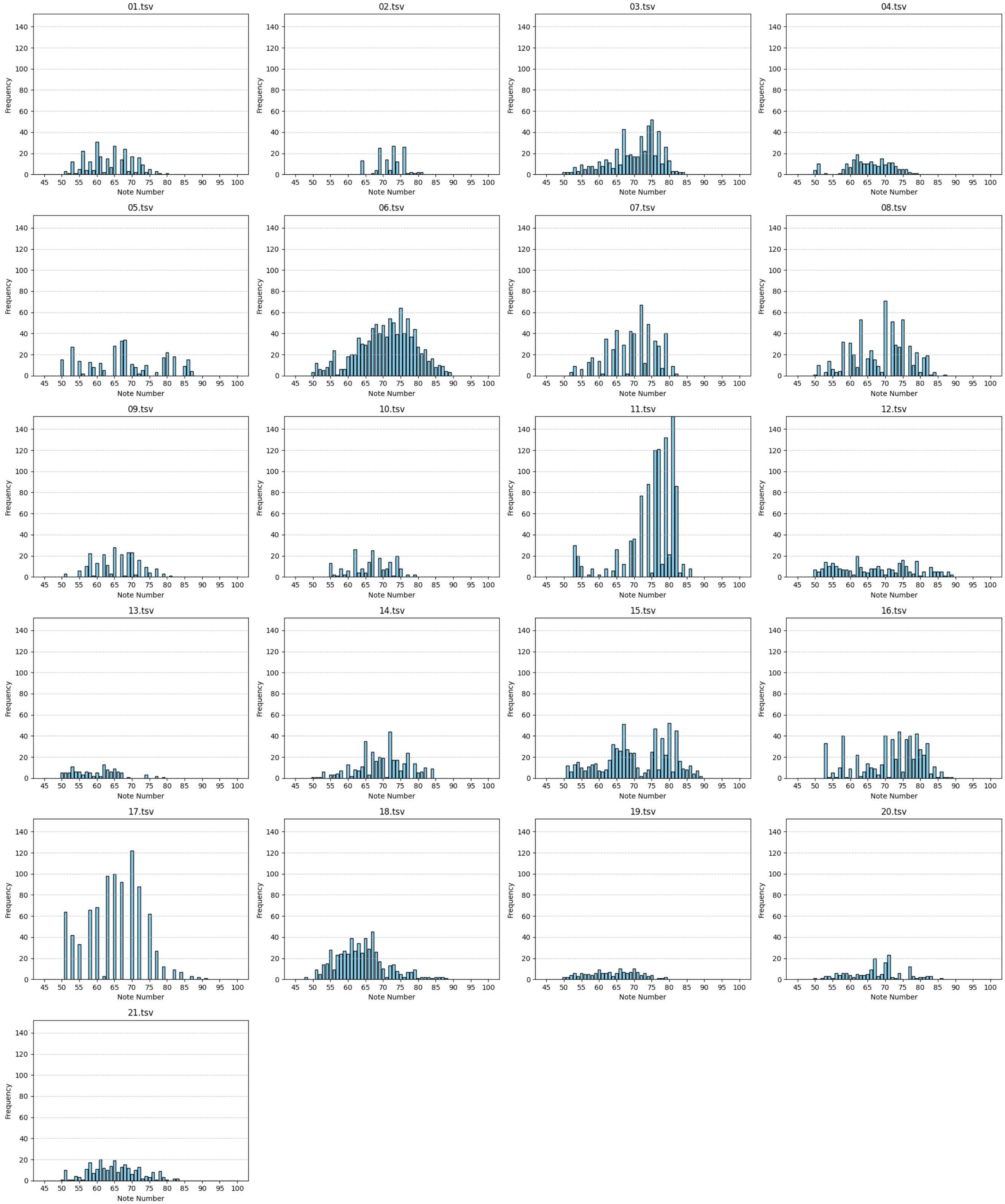

Figure 2: Note distributions per instance in the ClarinetNet dataset.

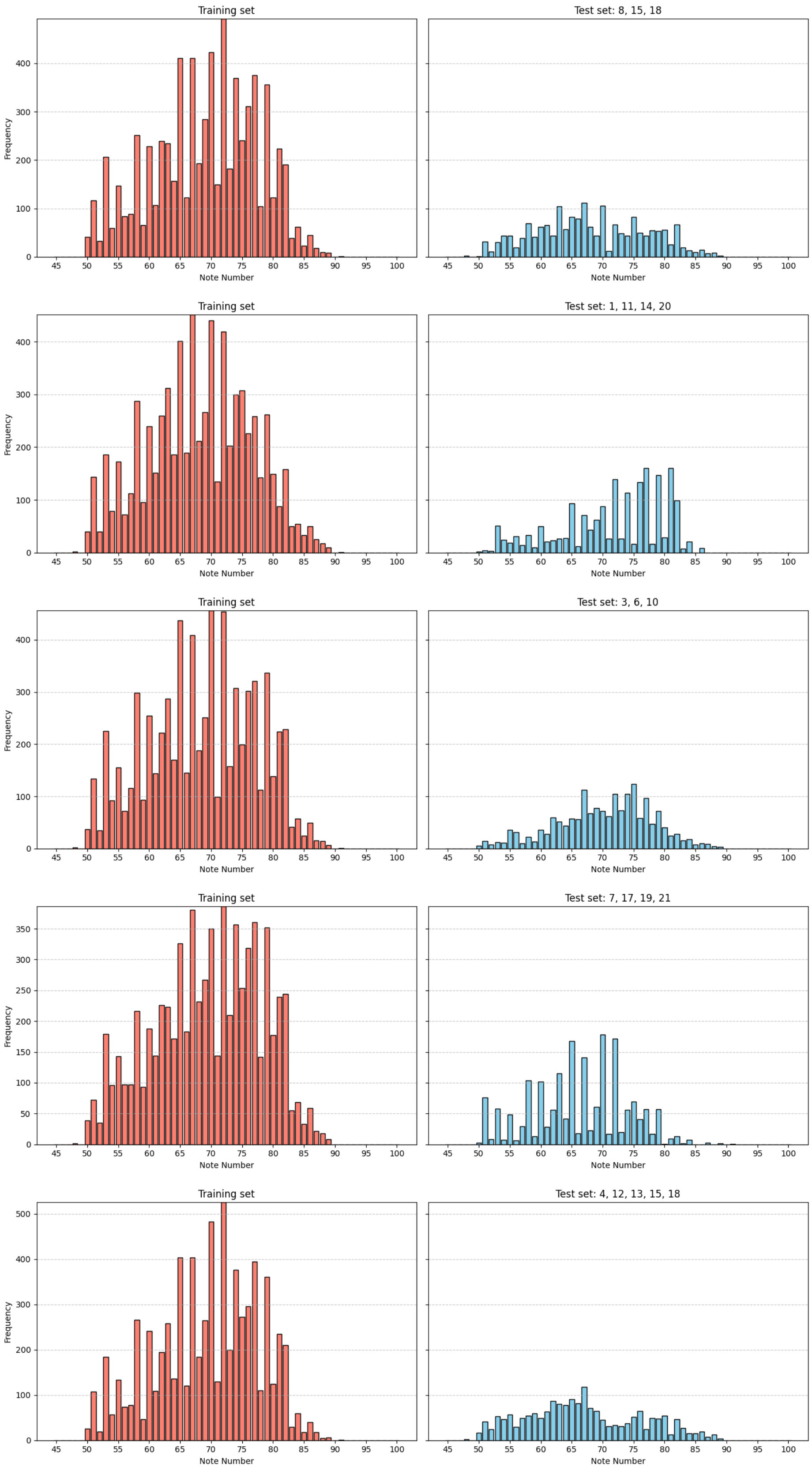

Figure 3: Note distributions of each training-test-split as seen in Section 4.

using cross-validation, and 0.6689 when a most representative test set is composed. All of these F1-scores do not beat the results compared to when the model is trained using a multi-instrument dataset. There are a few possible explanations for why this is the case.

5.1 Limitations

Firstly, the problem may lie with the model itself. Its architecture or training may be poorly suited for the task of learning to record music transcription. However, this is unlikely, as this model has already been proven to work with other datasets and the research has adapted the same training settings as demonstrated by Maman et al. [MB22].

Another possibility could be that the ClarinetNet dataset is not representative enough, meaning that training with this dataset does allow a model to learn enough about clarinet music in order for it to transcribe it accurately. Ultimately, this may also boil down to the quality of the instances in the dataset or to the fact that there are too few instances. It may be possible that the songs that are in the dataset are not representative enough for all possible tracks. However, when instances 8, 15 and 18, or 4, 12, 13, 15 and 18 are used in testing, the model has a more representative test set than with other compositions. In Figures 1, 2, and 3, the note distributions of the dataset and each individual instance are specified. From here it becomes apparent that instances, such as 19, are less representative of the entire dataset than other songs. All this does not have to mean that these tracks are bad. They are complete songs that meet the criteria specified in Section 3.2 and do not have to be removed from the dataset. It just means that for better re-presentability, more instances have to be added that focus on notes and rhythms that the current instances miss.

5.2 Contributions

Looking at the same results produced by Maman et al. [MB22], the Onset and Frames model trained on the ClarinetNet dataset can produce better F1-scores (0.60-0.67) than what the same model architecture can achieve for viola transcription when trained on the MusicNet dataset (0.611). Given that the viola's representation in the MusicNet dataset is 10 hours [THK17], it is a good achievement. Also, being able to produce results higher than 0.60 means that it can work somewhat adequately. It also shows that whilst the MusicNet dataset [THK17] has over 30 hours of music, ClarinetNet can, in fact, produce adequate results while only consisting of 1 hour of music and a fraction of the note count. Knowing this, we can therefore say that this research, even though not proving its hypothesis, shows a new venue that can be taken in AMT.

5.3 Future Work

This research can be seen as a pilot study, demonstrating that a single-instrument approach has potential for AMT research. Several avenues for follow-up studies emerge from these findings.

One could improve upon the ClarinetNet dataset by expanding its size and addressing the representativeness challenges identified in Section 5.1. The analysis of note distributions revealed that some compositions were more representative of the overall dataset than others, suggesting that strategic content selection could significantly improve model performance.

Beyond dataset expansion, this work opens possibilities for instrument-family datasets that could address limitations of both single-instrument and broad multi-instrument approaches. A

WoodwindNet dataset combining clarinet, oboe, flute, and saxophone recordings could leverage shared acoustic characteristics while maintaining specialized focus. Woodwind instruments share the monophonic nature, breath attacks, vibrato, dynamic swells, and timbral variations across registers identified as challenges in Section 2.3. This approach could potentially overcome the representativeness limitations observed in ClarinetNet while avoiding the dilution effects of completely mixed datasets like MusicNet.

This suggests a hierarchical training strategy where separate models are trained for each instrument family, then potentially combined through ensemble methods. Such an approach could exploit the acoustic similarities within instrument families while maintaining the specialized understanding that distinguishes family-grouped datasets from general multi-instrument collections.

Similar family-based datasets could be developed for other instrument groups - BrassNet for trumpet, trombone, and horn, or StringNet for violin, viola, and cello. Each family shares specific acoustic properties and playing techniques that could benefit from targeted training approaches.

Looking further ahead, one could investigate whether strategically combining single-instrument datasets produces superior results to traditional multi-instrument datasets. Rather than mixing all instruments equally as in existing datasets, this approach would maintain the specialized knowledge gained from individual instrument training while building more comprehensive transcription capabilities.

6 Conclusion

This research set out to investigate whether single-instrument datasets could outperform multi-instrument datasets in automatic music transcription, specifically focusing on clarinet music. Through the development of ClarinetNet, a specialized dataset containing one hour of clarinet compositions, we trained and evaluated AMT models using the proven methodology of Maman et al. [MB22].

The results indicate that while ClarinetNet produces competent transcription performance with F1-scores between 0.60 and 0.67, it does not exceed the 0.868 F1-score achieved by multi-instrument datasets like MusicNet [THK17] for clarinet transcription. However, these findings should not be interpreted as a failure of the single-instrument approach. Rather, they highlight several important considerations for future AMT research.

First, the study demonstrates that meaningful transcription results can be achieved with significantly smaller datasets than previously assumed. ClarinetNet's one hour of music represents a fraction of the 34-hour MusicNet dataset [THK17], yet produces results comparable to or better than some instruments in multi-instrument evaluations. This suggests that dataset efficiency may be more important than sheer size for certain applications.

Second, the research reveals the critical importance of dataset representativeness. Our analysis of note distributions across individual pieces showed that some compositions were more representative of the overall dataset than others, directly impacting model performance. This insight provides a roadmap for improving single-instrument datasets through more strategic content selection.

Third, this work establishes the first dedicated clarinet dataset for AMT research, filling a significant gap in wind instrument representation. While the immediate transcription results may not surpass existing approaches, ClarinetNet provides a foundation for future research and demonstrates the feasibility of creating specialized datasets for underrepresented instruments.

The implications extend beyond clarinet transcription. This research opens new avenues for exploring single-instrument approaches across other wind instruments, potentially leading to a collection of specialized datasets that could be combined to create more balanced multi-instrument training sets. Furthermore, the methodology developed here provides a template for creating similar datasets for other monophonic instruments.

In conclusion, while ClarinetNet did not achieve the hypothesized performance improvement over multi-instrument datasets, it successfully demonstrates the potential of single-instrument approaches in AMT. The research contributes valuable insights into dataset construction, evaluation methodologies, and the unique challenges of wind instrument transcription. Future work should focus on expanding dataset size and representativeness, exploring combinations of single-instrument datasets, and investigating whether specialized approaches can complement rather than replace multi-instrument training strategies. This study represents an important step toward more inclusive and effective automatic music transcription systems.

References

[BDDE19] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

[BL23] Bhuwan Bhattarai and Joonwhoan Lee. A comprehensive review on music transcription. *Applied Sciences*, 13(21), 2023.

[CJK$^+$85] Chris Chafe, David A Jaffe, Kyle Kashima, Bernard Mont-Reynaud, and Julius O Smith III. *Techniques for note identification in polyphonic music*. Number 29. CCRMA, Department of Music, Stanford University, 1985.

[GSM$^+$21] Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse H. Engel. MT3: multi-task multitrack music transcription. *CoRR*, abs/2111.03017, 2021.

[HSR$^+$19] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[IMS] IMSLP. IMSLP Petrucci Music Library. https://imslp.org. Accessed: 2024-12-18.

[Ins22] The Pianola Institute. The reproducing piano - early experiments, 2022. https://www.pianola.org/reproducing/reproducing_early.cfm [Accessed: 2025-08-23].

[JYL23] Shulei Ji, Xinyu Yang, and Jing Luo. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Comput. Surv.*, 56(1), August 2023.

[KD06] Anssi Klapuri and Manuel Davy. *Signal Processing Methods for Music Transcription*. 01 2006.

[LLD$^+$16] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating A musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *CoRR*, abs/1612.08727, 2016.

[Mah89] Robert C. Maher. *An approach for the separation of voices in composite musical signals*. PhD thesis, 1989. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2025-05-28.

[MB22] Ben Maman and Amit H. Bermano. Unaligned supervision for automatic music transcription in the wild. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 14918–14934. PMLR, 2022.

[Moo75] James A. Moorer. On the segmentation and analysis of continuous musical sound by digital computer. Master's thesis, Stanford University, Stanford, CA, 1975.

[Mus] Musescore. Musescore free music composition and notation. https://musescore.org/en. Accessed: 2025-01-15.

[MWSR19] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity, 2019.

[Pis86] Martin Piszczalski. *A computational model of music transcription*. PhD thesis, USA, 1986. UMI order no. GAX86-21354.

[Raf16] Colin Raffel. *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, Columbia University, 2016.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[THK17] John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch, 2017.

[XBP+18] Qingyang Xi, Rachel M. Bittner, Johan Pauwels, Xuzhou Ye, and Juan P. Bello. Guitarset: A dataset for guitar transcription. *International Society for Music Information Retrieval Conference, September 2018*, 2018.