Universiteit Leiden

Robustness of Guitar Tablature Transcription under Signal-Level Acoustic Degradation

Márton Bálint Gebe (s3959813)

Supervisors:
Dr. E.M. Bakker & Prof. dr. M.S. Lew

Bachelor Thesis– Data Science and Artificial Intelligence

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

August 11, 2026

Abstract

Guitar tablature transcription estimates both the sounding pitches and the string–fret positions on the guitar used to play them. Previous robustness studies have mainly examined changes in guitar tone, effects, and recording equipment, while environmental background noise and room reverberation have received less attention. This thesis extends that work by evaluating whether TAB-CNN training configurations that include environmentally degraded recordings are more robust under these conditions.

Five training configurations are compared using a fixed TAB-CNN architecture. Training configuration V1 uses clean GuitarSet training, V2 and V3 are two effect-augmentation configurations based on prior work, V4 adds noise- and reverberation-degraded GuitarSet recordings, and V5 is a combined configuration containing acoustic- and electric-guitar data with environmental counterparts. The study contributes two fixed datasets for training and evaluation, GuitarSetENV and GuitarSetENV-Test, and an expanded out-of-domain test-set, EGSet12+.

On the clean GuitarSet, the environmental V4 configuration and the clean V1 TAB-CNN baseline obtains similar F_1–means. Under environmental noise and simulated room reverberation, V4 trained TAB-CNN obtains the highest multi-pitch- and tablature–F_1 means at all three tested SNR levels, while V5 trained TAB-CNN obtains the highest TDR–means. On clean out-of-domain electric-guitar recordings, V3 trained TAB-CNN obtains the highest multi-pitch and tablature F_1–means and the highest TDR–mean. When environmental noise is added to those recordings, V5 trained TAB-CNN obtains the highest multi-pitch and tablature F_1–means and the highest TDR–means at all three SNR levels.

These results extend previous effect-augmentation findings by showing that robustness gains depend on the evaluation domain and degradation type. In our experiments, we found no single overall best-performing TAB-CNN training configuration.

Contents

1 Introduction **1**
- 1.1 Research questions . . . 2
- 1.2 Contributions . . . 3

2 Related work **3**
- 2.1 GuitarSet and TAB-CNN . . . 4
- 2.2 Current methods in guitar tablature transcription . . . 4
- 2.3 Robustness studies . . . 5
- 2.4 Augmentation for acoustic robustness . . . 5

3 TAB-CNN **6**

4 Methodology **8**
- 4.1 Proposed approach . . . 8
- 4.2 Datasets . . . 8
  - 4.2.1 GuitarSet . . . 9
  - 4.2.2 GuitarSetFX . . . 9
  - 4.2.3 GuitarProFX . . . 9
  - 4.2.4 EGSet12 . . . 10
  - 4.2.5 DEMAND . . . 10
  - 4.2.6 EigenScape . . . 10
  - 4.2.7 GuitarSetENV . . . 10
  - 4.2.8 GuitarProFXENV . . . 12
  - 4.2.9 GuitarSetENV-Test . . . 13
  - 4.2.10 EGSet12+ . . . 15
- 4.3 Training configurations V1–V5 . . . 15
  - 4.3.1 V1: clean GuitarSet . . . 16
  - 4.3.2 V2: GuitarSet and GuitarSetFX . . . 16
  - 4.3.3 V3: GuitarSet and GuitarProFX . . . 16
  - 4.3.4 V4: GuitarSet and GuitarSetENV . . . 16
  - 4.3.5 V5: combined acoustic- and electric-guitar training . . . 16
- 4.4 Baseline reproduction and implementation check . . . 17
- 4.5 Evaluation protocol and aggregation . . . 17
- 4.6 Code and data availability . . . 18

5 Results **18**
- 5.1 Clean GuitarSet: in-domain reference . . . 19
- 5.2 GuitarSetENV-Test: environmental degradation within the GuitarSet domain . . . 20
- 5.3 EGSet12: electric-guitar domain shift . . . 20
- 5.4 EGSet12+: combined domain and environmental shift . . . 21

6 Discussion **23**
- 6.1 Limitations . . . 25

7 Conclusion 26

8 Future Work 27

References 27

A Implementation and construction details 30

1 Introduction

Automatic music transcription (AMT) converts musical audio recordings into symbolic representations. Standard Western notation is instrument-agnostic: it represents pitches and durations without specifying how notes are produced on a particular instrument. Tablature is an instrument-specific alternative for fretted string instruments. Its six horizontal lines represent the guitar strings, and a number on a line indicates the fret at which that string is played, with 0 denoting an open string. Figure 1 illustrates the difference between the two representations: staff notation identifies the pitch, whereas tablature must also identify one of the string–fret positions that can produce it.

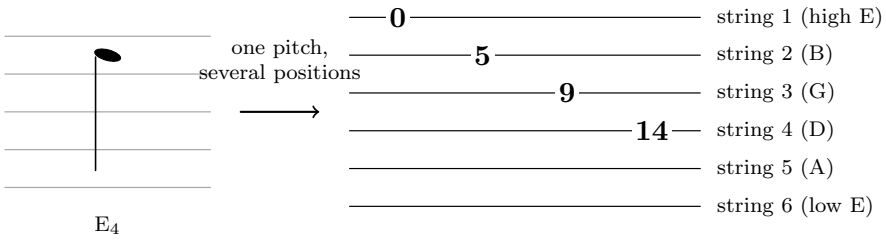

Figure 1: String–fret problem in guitar tablature. The same sounding pitch, E_4, can be played at several positions on a standard-tuned guitar: string 1 open, string 2 fret 5, string 3 fret 9, or string 4 fret 14.

Recovering tablature from audio is therefore more difficult than recovering pitches alone. On a piano, each pitch corresponds to one key, so identifying the pitch also identifies its playing position. On a guitar, the pitch ranges of the six strings overlap, allowing the same pitch to be produced at several string–fret positions. Guitar Tablature Transcription (GTT) is the AMT task that addresses both parts of this problem: it estimates the sounding pitches and the strings and frets used to produce them. This additional string–fret ambiguity distinguishes GTT from pitch-only transcription.

GuitarSet is a central benchmark for this task [21, 19, 9]. It contains 360 acoustic-guitar performances by six guitarists with note-, string-, and fret-level annotations. Wiggins and Kim introduced TAB-CNN in 2019 as a direct framewise audio-to-tablature model for GuitarSet [19]. At the time of publication, it outperformed a state-of-the-art multi-pitch estimation baseline and introduced dedicated metrics for evaluating tablature transcription. The model predicts one tablature state for each of the six strings from a constant-Q spectrogram. Its established GuitarSet protocol, available implementation, and use in the closest prior robustness study make it a suitable architecture for this thesis. Using the same architecture across conditions limits variation due to model choice, although the conditions still differ in their training data and batch size. Later research has developed continuous-pitch formulations [3], larger and synthetic training resources [23, 8, 12], and transfer from piano-transcription representations [11]; Section 2 reviews these developments.

Robustness is important because GuitarSet consists of clean, close-microphone recordings, whereas recordings made outside controlled conditions may contain background noise, room reverberation, different guitar tones, audio effects, and changes in the recording chain. Pedroza et al. studied part of this problem by augmenting TAB-CNN training with electric-guitar tones and effects, obtaining higher reported means on the out-of-domain electric-guitar set EGSet12 [9]. Guitar-TECHS extends this direction with electric-guitar recordings that vary in playing

technique, musical content, and hardware chain [10]. The GTT literature reviewed for this thesis addresses timbre, effects, hardware, and recording-chain variation, but does not provide a systematic evaluation of environmental background noise and room reverberation at controlled signal-to-noise ratios. Work in related transcription and audio-classification tasks indicates that degradation-based augmentation can improve performance under shifted acoustic conditions [4, 6]. Public environmental-noise corpora and room-simulation tools make it possible to study this form of degradation directly [17, 5, 15].

This thesis addresses that gap by comparing five training conditions, V1–V5, for TAB-CNN while keeping the architecture of TAB-CNN fixed. V1 uses the clean GuitarSet. V2 adds GuitarSetFX, and V3 adds GuitarProFX; both provide effect-augmentation reference conditions based on the datasets supplied by Pedroza et al. [9]. V4 pairs the clean GuitarSet training recordings with own GuitarSetENV copies degraded using DEMAND environmental noise and simulated room reverberation [17, 15]. V5 combines GuitarSet, GuitarSetENV, GuitarProFX, and GuitarProFXENV, where GuitarProFXENV is an environmentally degraded version of GuitarProFX.

The comparisons are organised at three levels. The primary comparison is between V1 and V4, the acoustic-guitar-domain conditions, which both use batch size 8, with V4 adding the proposed environmental augmentation to the clean GuitarSet training setup. The secondary comparison is among V2, V3, and V5, which all use batch size 32 and collectively cover the acoustic- and electric-guitar domains. All five configurations are also compared descriptively at a broader level to examine their observed performance across the evaluation conditions. Differences in training-set size and composition arise from the evaluated training recipes and were not independently controlled. Broader comparisons across the two batch-size groups therefore remain informative, but they are not interpreted as isolating the causal effect of a single training factor.

The study uses four evaluation sets: one fold-validation set and three final test sets. Clean GuitarSet provides held-out-player fold validation in the original acoustic-guitar domain. Final testing uses GuitarSetENV-Test, clean EGSet12, and EGSet12+. GuitarSetENV-Test applies EigenScape environmental noise and fixed simulated room reverberation to held-out GuitarSet recordings at 20, 10, and 0 dB SNR. EGSet12 contains twelve out-of-domain electric-guitar performances without added environmental degradation [9], while EGSet12+ adds EigenScape noise to the same source tracks at the same three SNR levels without additional reverberation. Environmental training uses DEMAND recordings, whereas final testing uses EigenScape recordings, so no environmental-noise file is shared between training and evaluation. Multi-pitch and tablature F_1 are the primary performance measures, supported by precision, recall, and the Tablature Disambiguation Rate (TDR).

1.1 Research questions

The following main research question guides this thesis:

> **RQ1: How does a TAB-CNN training configuration that includes environmentally degraded guitar recordings perform on clean and environmentally degraded guitar audio compared with clean-only training?**

This question is examined primarily by comparing the clean GuitarSet baseline V1 with V4, which adds GuitarSetENV recordings containing environmental noise and simulated room reverberation.

A secondary research question examines how environmental augmentation relates to the wider set of training configurations:

> **RQ2: How do clean, effect-augmented, environmentally augmented, and combined training configurations perform across clean and degraded acoustic- and electric-guitar evaluation conditions?**

The main question is addressed using clean GuitarSet fold validation and GuitarSetENV-Test, with V1 and V4 providing the primary comparison. For the secondary question, V2, V3, and V5 provide the main same-batch-size comparison of the other augmentation and training-data strategies, while all five configurations are also considered descriptively across the evaluation conditions. Clean EGSet12 and EGSet12+ additionally examine performance under electric-guitar domain shift, with and without added environmental noise.

1.2 Contributions

This thesis makes four contributions:

1. **GuitarSetENV**: a fixed environmental training dataset containing 360 degraded copies of GuitarSet. In each fold, V4 and V5 use the 300 copies corresponding to the five training guitarists together with the 300 clean recordings [17, 15].
2. **GuitarSetENV-Test**: a fixed final test set containing 1,080 degraded GuitarSet files, created with EigenScape noise and three characterised simulated room impulse responses at 20, 10, and 0 dB SNR [5]. Its environmental-noise recordings are separate from the DEMAND recordings used during training.
3. **EGSet12+**: a fixed set of 36 environmentally degraded versions of the twelve EGSet12 source tracks, created by adding selected EigenScape noise at 20, 10, and 0 dB SNR without simulated room reverberation [9, 5].
4. **A five-condition robustness evaluation**: a comparison of V1–V5 across clean GuitarSet validation, GuitarSetENV-Test, clean EGSet12, and EGSet12+, showing how the leading configuration changes with source domain and degradation setting.

The rest of the paper is organised as follows. Section 2 reviews prior work on GTT, domain robustness, and acoustic augmentation. Section 3 describes TAB-CNN and the performance measures used in this study. Section 4 presents the datasets, the five training configurations, the training setup, and the evaluation protocol. Section 5 reports the results, and Section 6 interprets them and discusses the limitations. Section 7 concludes the thesis, Section 8 presents future work, and Appendix A records additional implementation and construction details.

2 Related work

Automatic guitar tablature transcription is a guitar-specific form of automatic music transcription (AMT), a broader research field surveyed by Benetos et al. [2]. Within this field, early guitar-specific approaches combined multipitch analysis with playability constraints to infer string–fret positions from audio [22].

2.1 GuitarSet and TAB-CNN

GuitarSet, introduced by Xi et al., contains 360 acoustic-guitar performances by six guitarists across several musical styles, with note-, string-, and fret-level annotations [21]. These annotations make it suitable for tablature transcription, where the target is not only the sounding pitch but also the string and fret that produced it. Its player structure also supports the six-fold protocol used in this thesis, in which the recordings of one guitarist form the held-out fold.

TAB-CNN, proposed by Wiggins and Kim, uses the framewise formulation retained in this thesis: for each analysis frame, the model predicts one categorical state per string, with 21 classes per string representing the inactive state, the open string, and frets 1–19 [19]. The study showed that a compact convolutional network using a constant-Q input can estimate string and fret positions directly, and TAB-CNN has since been used as a reference model in later GTT research. This thesis retains GuitarSet, the TAB-CNN architecture, and the held-out-player protocol. All five training conditions use the PyTorch/AMT-Tools implementation associated with Pedroza et al. [9]. A separate rerun of the original Keras implementation is used only as a numerical benchmark check. All five conditions are trained within the same PyTorch/AMT-Tools framework. V1–V3 correspond to the clean and effect-augmentation configurations associated with Pedroza et al., while V4 and V5 introduce the environmental training branches proposed in this thesis.

2.2 Current methods in guitar tablature transcription

After TAB-CNN, GTT research developed along several directions. FretNet reformulates tablature transcription as continuous-valued pitch-contour streaming, addressing a limitation of framewise discrete fret classification: bends and other pitch modulations are represented poorly by fixed pitch classes [3]. A second direction expanded the available training material. SynthTab uses synthesised guitar audio paired with tablature to reduce dependence on limited manually annotated guitar recordings [23]. GOAT provides 5.9 hours of unique direct-input electric-guitar recordings paired with tablature, together with amplifier-processed variants [8].

Evaluation has also broadened beyond GuitarSet. GAPS contributes a larger corpus of classical-guitar recordings with aligned scores and a benchmark transcription model [12]. Its note-level onset metrics are not directly comparable with TAB-CNN's frame-level tablature F_1, so it is relevant here as evidence of dataset expansion rather than as a numerical baseline. Riley et al. additionally adapt a high-resolution piano-transcription model to guitar, showing that guitar transcription can benefit from representations learned from larger non-guitar corpora [11].

Wiggins examines the semi-supervised use of unlabelled guitar audio and reports more consistent improvements on guitars with previously unseen timbres using an autoencoder-based approach [20]. Together, these studies modify the transcription formulation, increase the scale and diversity of training data, or change the learned representation. They do not report a systematic GTT evaluation of additive environmental noise and simulated room reverberation at controlled signal-to-noise ratios. This thesis instead keeps the model architecture fixed and study how different training-data configurations perform under these degradations.

2.3 Robustness studies

Pedroza et al. provide the closest prior study [9]. They investigate whether training with electric-guitar tones and effects improves TAB-CNN's generalisation beyond clean acoustic GuitarSet training. Their experiments compare GuitarSet alone, GuitarSet with GuitarSetFX, and GuitarSet with GuitarProFX. GuitarSetFX contains effect-rendered versions of the GuitarSet performances, while GuitarProFX contains separate solo-guitar material rendered with electric-guitar tones and effects. The models are evaluated on EGSet12, a twelve-performance electric-guitar dataset used as an out-of-domain test set.

Pedroza et al. report higher EGSet12 multi-pitch and tablature F_1–means for both augmented conditions than for clean GuitarSet training, while their clean GuitarSet fold-validation results remain broadly similar. Because the augmented conditions also add training data, the results reflect the complete training configurations rather than isolating guitar-tone and effect diversity as a single causal factor. Pedroza et al. later extended this direction with Guitar-TECHS, an electric-guitar dataset covering techniques, excerpts, chords, and scales recorded through varied hardware chains. They report that adding Guitar-TECHS to the training data increases mean TDR on EGSet12 from 0.695 to 0.776 [10].

Venohr et al. propose cross-version consistency as an annotation-free evaluation method for automatic music transcription and report a close relationship with standard multi-pitch metrics [18]. Their work studies consistency across different recordings of the same musical work rather than signal-level acoustic degradation. It therefore provides a broader robustness-evaluation perspective, but not a training-augmentation baseline for environmental degradation in GTT.

Overall, prior GTT robustness work mainly addresses variation in timbre, effects, hardware, musical material, and recording chains. Pedroza et al. provide the closest reference because they use TAB-CNN, augmented training data, and EGSet12 on out-of-domain evaluation dataset. The present study extends this direction to environmental background noise and simulated room reverberation. The exact V1–V5 training configurations and the use of the author-supplied datasets are described in Section 4.

2.4 Augmentation for acoustic robustness

The methodological basis for the environmental augmentation comes mainly from piano transcription and general audio classification, since these areas contain more direct studies of acoustically degraded recordings. Edwards et al. show that piano-transcription systems can overfit the acoustic properties of their training data and that combining acoustically diverse sources with augmentation improves cross-dataset performance [4]. Kusaka and Maezawa likewise use augmentation to improve polyphonic piano transcription for in-the-wild mobile recordings [7].

In preliminary extended-abstract results, Kim and Lerch investigate white-noise injection at controlled SNR levels. Their results support retaining clean examples alongside degraded examples, since noisy training examples may improve degraded-condition performance while clean examples help preserve clean-condition performance [6]. Data augmentation is also widely used in audio classification, including singing-voice detection [16] and environmental-sound classification [13], and has been reviewed across sound-classification tasks [1].

These findings motivate retaining clean examples while adding controlled degraded examples during training. The proposal in this thesis differs from the cited piano-transcription studies

because GTT must estimate not only the sounding pitch but also the string–fret position used to produce it. A correct pitch can therefore still correspond to an incorrect tablature prediction. The degraded guitar recordings must also remain aligned with their note-, string-, fret-, and timing annotations. This thesis applies the augmentation principle to environmental background noise and simulated room reverberation; the resulting datasets and training configurations are described in Section 4.

3 TAB-CNN

TAB-CNN, introduced by Wiggins and Kim, is a convolutional neural network for framewise guitar tablature transcription [19]. Instead of predicting only the sounding pitches, the model predicts one tablature state for each of the six guitar strings. Each string has 21 possible states: inactive, open string, or frets 1–19. The six string predictions together therefore describe the tablature state of the guitar for one analysis frame.

In the implementation used in this thesis, audio is processed at a sampling rate of 22,050 Hz and converted with a constant-Q transform (CQT) using a hop length of 512 samples, 192 frequency bins, and 24 bins per octave. During training, the data loader samples 200-frame excerpts from the recordings. Nine consecutive CQT frames form the input for one frame prediction, producing an input tensor of $1 \times 192 \times 9$, where the dimensions represent the input channel, frequency bins, and time frames. The 200-frame excerpt length and the nine-frame model input therefore refer to different stages of the training pipeline.

As shown in Figure 2, TAB-CNN contains three 3×3 convolutional layers with 32, 64, and 64 output channels and ReLU activations. These are followed by 2×2 max-pooling, dropout with $p = 0.25$, a 128-unit dense layer, dropout with $p = 0.50$, and 126 output logits. The logits are reshaped into six groups of 21 states, one for each guitar string, and each group is normalised independently with a per-string softmax.

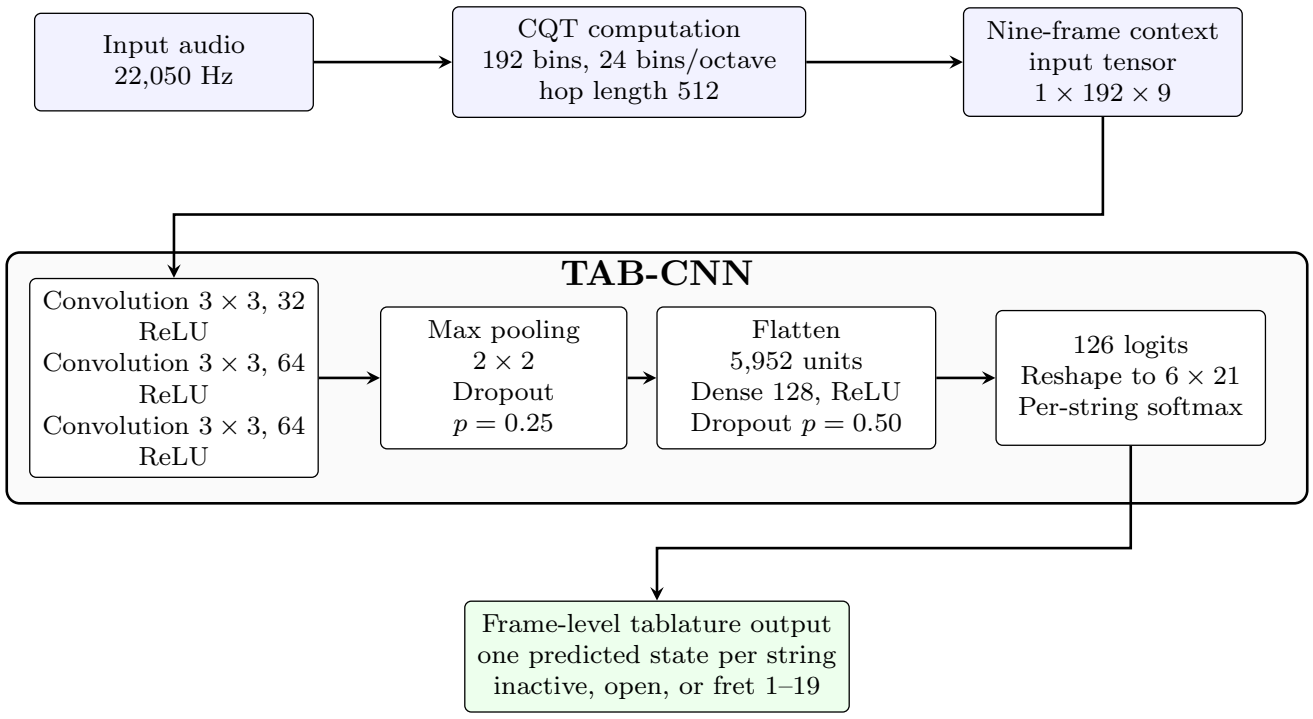

Figure 2: End-to-end TAB-CNN processing flow from input audio to frame-level six-string tablature prediction. The architecture is fixed across training configurations V1–V5.

Performance of TAB-CNN is evaluated for both multi-pitch estimation and tablature transcription. Multi-pitch evaluation measures whether the sounding pitches are detected without requiring the correct string assignment, whereas tablature evaluation also requires the predicted string–fret state to match the reference. For both tasks, precision and recall are defined as

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN},$$

where TP is the number of correctly predicted activations, FP is the number of predicted activations that are not present in the reference, and FN is the number of reference activations that are not predicted. Their harmonic mean gives the F_1 score:

$$F_1 = \frac{2PR}{P + R}.$$

The tablature disambiguation rate (TDR) measures how often a correctly detected pitch is also assigned to the correct tablature position. In the evaluator used in this thesis, it is defined as

$$\mathrm{TDR} = \frac{N_{\text{correct tablature}}}{N_{\text{correct multi-pitch}}},$$

where the numerator counts correctly predicted tablature activations and the denominator counts correctly detected string-collapsed multi-pitch activations. Values above 1 are retained without clipping. Precision, recall, F_1, and TDR are the performance measures reported throughout this thesis.

4 Methodology

4.1 Proposed approach

The proposal of this thesis is to improve the robustness of TAB-CNN to environmental recording conditions by changing the training data while keeping the TAB-CNN architecture fixed. Previous GTT robustness work has mainly introduced variation in guitar tone, effects, and recording equipment. This thesis instead introduces environmental background noise and simulated room reverberation into the training data. Although evaluated here with TAB-CNN, the same dataset-based training approach could in principle be applied to similar GTT models.

The proposed approach retains the original clean labelled recordings and adds fixed environmentally degraded versions of them. Because the degradation changes the audio signal without changing the performed notes, the original note-, string-, fret-, and timing annotations can be retained. Training data and final-test data use different environmental-noise corpora: DEMAND is used to construct the environmental training data, while EigenScape is used for final testing. No environmental-noise recording is therefore shared between environmental training and final testing.

This approach results in the environmental training datasets GuitarSetENV and GuitarProFX-ENV and the fixed final-test datasets GuitarSetENV-Test and EGSet12+. GuitarSetENV, GuitarSetENV-Test, and EGSet12+ are the main dataset contributions of this thesis. These datasets are described first. The five TAB-CNN training configurations V1–V5 are then defined separately using these datasets and the existing datasets from prior work.

4.2 Datasets

The experiments use existing guitar datasets and environmental-noise corpora together with datasets constructed for this thesis. The existing datasets provide the clean and effect-based training material and the original out-of-domain electric-guitar test data. The constructed datasets introduce environmental degradation for training and final testing.

Table 1: Technical specifications of the existing datasets and environmental-noise corpora used in this thesis. Sample rates and bit depths describe the stored source files before resampling by the experimental pipeline.

Resource	Recordings and distribution	Sample rate	Bit depth	Recording duration
GuitarSet	360 recordings; 60 per guitarist from six guitarists	44.1 kHz	16-bit	Close to 30 s per recording
GuitarSetFX	360 effect-rendered recordings; 60 per GuitarSet guitarist	48 kHz	24-bit	Close to 30 s; follows the corresponding GuitarSet recording
GuitarProFX	360 solo-guitar tracks; no player grouping	48 kHz	24-bit	275.1 s on average; 27.5 h in total
EGSet12	12 performances by one guitarist	48 kHz	24-bit	31.65 s on average; 379.8 s in total
DEMAND	15 environments in the corpus; nine used in this thesis	48 kHz	16-bit	300 s per recording
EigenScape	64 recordings in the corpus; 31 from four classes used in this thesis	48 kHz	24-bit	600 s per recording

4.2.1 GuitarSet

GuitarSet contains 360 acoustic-guitar performances by six guitarists, with note-, string-, fret-, and timing annotations [21]. It provides the common clean training data and held-out-player validation data used throughout this thesis. Its six-player structure supports the six-fold protocol used later for V1–V5. The technical specifications of GuitarSet are given in Table 1.

4.2.2 GuitarSetFX

GuitarSetFX contains effect-rendered versions of the original GuitarSet performances and was used in the effect-augmentation study of Pedroza et al. [9]. It introduces changes in guitar tone and audio effects while retaining the original GuitarSet performances. The dataset was supplied directly by Hegel Pedroza and was not created or regenerated in this thesis. The technical specifications of GuitarSetFX are given in Table 1.

4.2.3 GuitarProFX

GuitarProFX contains 360 separate solo-guitar tracks based on material selected from DadaGP and rendered with electric-guitar tones and effects [14, 9]. Unlike GuitarSetFX, its recordings are not effect-rendered versions of the GuitarSet performances and therefore also introduce different musical material. GuitarProFX was supplied directly by Hegel Pedroza and was not created or regenerated in this thesis. The technical specifications of GuitarProFX are given in Table 1.

4.2.4 EGSet12

EGSet12 is the out-of-domain electric-guitar test set used by Pedroza et al. [9]. It contains twelve stereo electric-guitar performances independent of GuitarSet. The original recordings are used without added environmental degradation as the clean electric-guitar final test set. The technical specifications of EGSet12 are given in Table 1.

4.2.5 DEMAND

DEMAND is a collection of environmental-noise recordings from a range of real environments [17]. It supplies the environmental noise used to construct GuitarSetENV and GuitarProFXENV. It is therefore used only for environmental training-data construction. The technical specifications of the DEMAND recordings used in this thesis are given in Table 1.

4.2.6 EigenScape

EigenScape contains spatial environmental recordings from several acoustic scene classes [5]. Recordings from the Busy Street, Park, Shopping Centre, and Train Station classes are used for the degraded final-test datasets. EigenScape is not used to construct the environmental training data. The technical specifications of the EigenScape recordings used in this thesis are given in Table 1.

Table 2: Existing datasets and environmental-noise corpora used in this thesis.

Dataset	Contents used in this thesis	Purpose
GuitarSet	360 annotated acoustic-guitar performances from six guitarists	Clean training data and held-out-player validation data
GuitarSetFX	360 effect-rendered versions of the GuitarSet performances	Effect-augmentation training data
GuitarProFX	360 solo-guitar recordings rendered with electric-guitar tones and effects	Effect-augmentation training data with separate musical material
EGSet12	12 stereo electric-guitar performances	Clean out-of-domain final test data
DEMAND	Environmental-noise recordings	Noise source for environmental training-data construction
EigenScape	Environmental recordings from selected acoustic scene classes	Noise source for environmental final-test construction

4.2.7 GuitarSetENV

GuitarSetENV is the environmental training dataset proposed for the GuitarSet domain. Its purpose is to expose TAB-CNN to environmental background noise and room reverberation while retaining the original clean training recordings. The dataset contains one environmentally degraded copy of each of the 360 GuitarSet recordings. During training, the clean recording and its degraded counterpart can therefore be used together.

Each GuitarSetENV recording is generated offline from its clean GuitarSet source. A deterministic per-track seed fixes the selected noise recording and crop, simulated room, source and microphone

positions, target reverberation time, and SNR. Simulated room reverberation is generated using pyroomacoustics [15]. Room dimensions and source and microphone positions are sampled for each source recording, with a requested RT60 between 0.2 and 0.8 s.

The clean signal is convolved with the generated room impulse response and truncated to its original duration. A segment of DEMAND environmental noise is then mixed into the reverberated signal at an SNR sampled uniformly from 0–24 dB [17]. A clipping guard is applied before the degraded waveform is stored. The original note-, string-, fret-, and timing annotations are copied unchanged. Because construction is performed offline, the same degraded copy is reused throughout training.

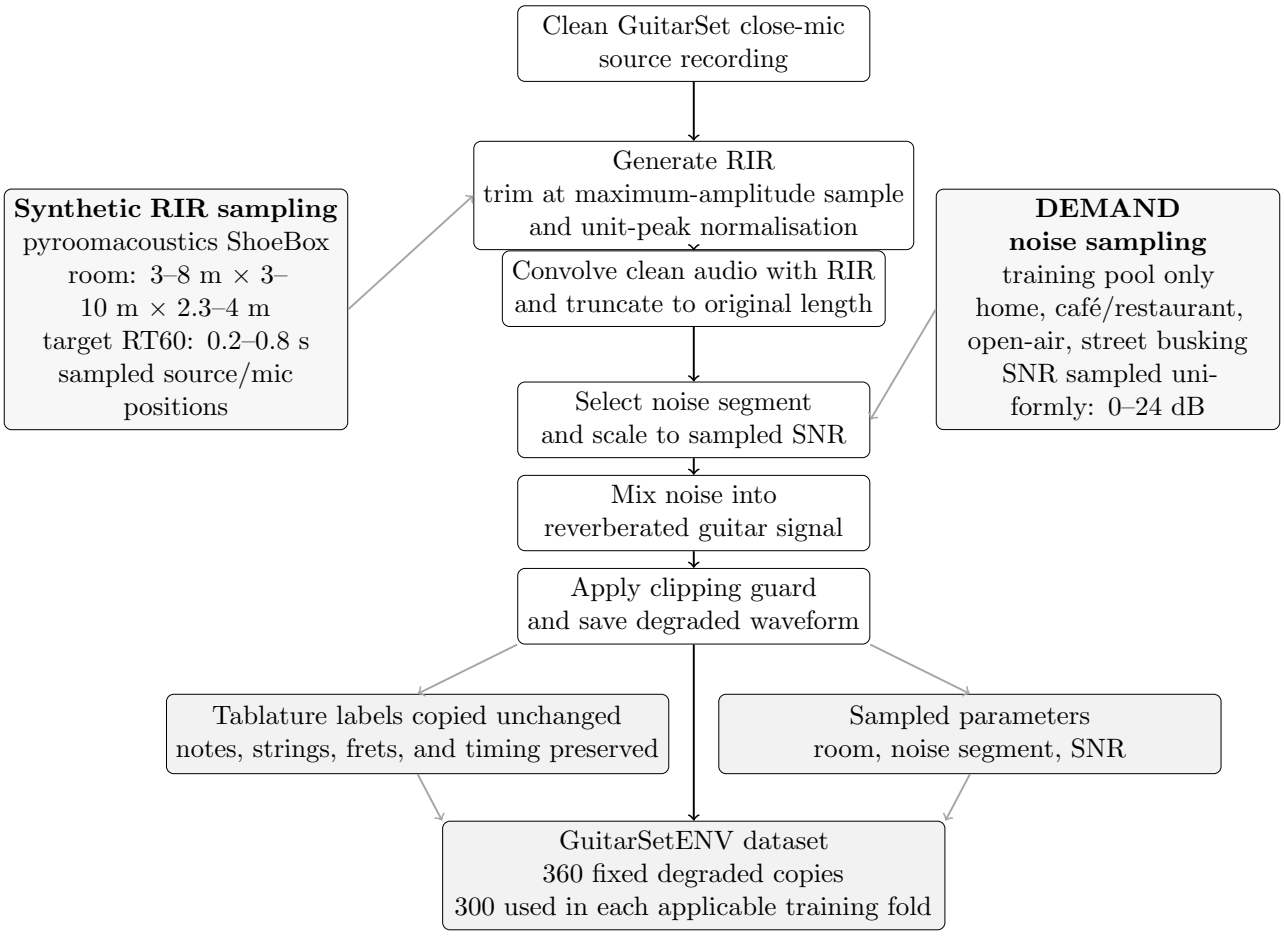

Figure 3: Offline construction of GuitarSetENV. Each GuitarSet recording is paired with one fixed degraded copy containing simulated room reverberation and DEMAND environmental noise at an SNR sampled from 0–24 dB. The original annotations are retained.

Measurements of the generated RIRs produced realised RT60 values from 0.148 to 2.230 s, which is wider than the requested 0.2–0.8 s interval. Additional implementation parameters and fixed generation seeds are reported in Appendix A.

Table 3: Contents and technical specifications of GuitarSetENV.

Property	GuitarSetENV
Source recordings	360 GuitarSet recordings
Generated files	360 fixed degraded copies
Sample rate	44.1 kHz
Bit depth	16-bit
Recording duration	Approximately 30 s per file; unchanged from the corresponding source recording
Noise corpus	DEMAND
Reverberation	Sampled simulated room impulse response
SNR	Sampled uniformly from 0–24 dB
Annotations	Original note-, string-, fret-, and timing annotations retained
Training use	300 copies corresponding to the five training guitarists in each applicable fold
Storage	Generated offline and reused throughout training

4.2.8 GuitarProFXENV

GuitarProFXENV extends the environmental training approach to the electric-guitar material contained in GuitarProFX. Its purpose is to provide V5 with an environmental counterpart to the GuitarProFX branch, so that the combined configuration contains both original and environmentally degraded acoustic- and electric-guitar training data.

GuitarProFXENV was generated offline from the 360 GuitarProFX source recordings using the same general degradation procedure as GuitarSetENV. DEMAND supplies the environmental noise and pyroomacoustics supplies the simulated room reverberation [17, 15]. The target SNR is sampled from 0–24 dB, and the source annotations are copied to the corresponding output.

Of the 360 GuitarProFX recordings, 355 have non-zero signal energy and are environmentally degraded. The remaining five have zero signal energy, so a target SNR cannot be defined for them. These five source files are retained unchanged in GuitarProFXENV.

Table 4: Contents and technical specifications of GuitarProFXENV.

Property	GuitarProFXENV
Source recordings	360 GuitarProFX recordings
Environmentally degraded files	355
Zero-energy sources retained unchanged	5
Total entries	360
Sample rate	48 kHz
Bit depth	16-bit
Recording duration	Variable; unchanged from the corresponding GuitarProFX source recording
Noise corpus	DEMAND
Reverberation	Sampled simulated room impulse response for non-zero sources
SNR	Sampled from 0–24 dB for non-zero sources
Annotations	Source annotations retained
Training use	All 360 entries in every V5 fold

4.2.9 GuitarSetENV-Test

GuitarSetENV-Test is the fixed environmental final test set proposed for the GuitarSet domain. Its purpose is to evaluate whether training with environmental degradation transfers to environmental recordings that were not used to construct the training data. It is therefore constructed separately from GuitarSetENV and uses EigenScape rather than DEMAND noise.

The test set uses three fixed simulated rooms, labelled small, medium, and large. Each room has fixed geometry and fixed source and microphone positions, producing one RIR that is reused throughout testing. Table 5 reports their dimensions, requested and measured reverberation times, and source–microphone distances.

Table 5: Fixed simulated rooms used in GuitarSetENV-Test. Requested RT60 is the construction target; measured RT60 is calculated from the generated RIR.

Room	Dimensions (m)	Requested RT60 (s)	Measured RT60 (s)	Source–mic distance (m)
Small	$3.4 \times 4.0 \times 2.5$	0.3	0.318	0.702
Medium	$5.8 \times 6.5 \times 3.1$	0.5	0.603	1.006
Large	$7.8 \times 9.3 \times 3.8$	0.7	0.937	1.412

Each of the 360 GuitarSet source recordings is assigned one of these rooms together with one EigenScape noise recording and crop. After convolution with the assigned RIR, the recording is mixed with the same noise crop at 20, 10, and 0 dB SNR. For a given source recording, the room, RIR, noise recording, and crop remain fixed across the three SNR versions; only the noise gain changes. The original GuitarSet annotations are reused unchanged, and all assignments are frozen before any model is evaluated.

The three room labels occur equally often within each player. Environmental noise is selected from the Busy Street, Park, Shopping Centre, and Train Station EigenScape scene classes [5]. Final

testing preserves the GuitarSet player split: each fold checkpoint is evaluated only on the degraded versions of the 60 recordings belonging to its held-out guitarist.

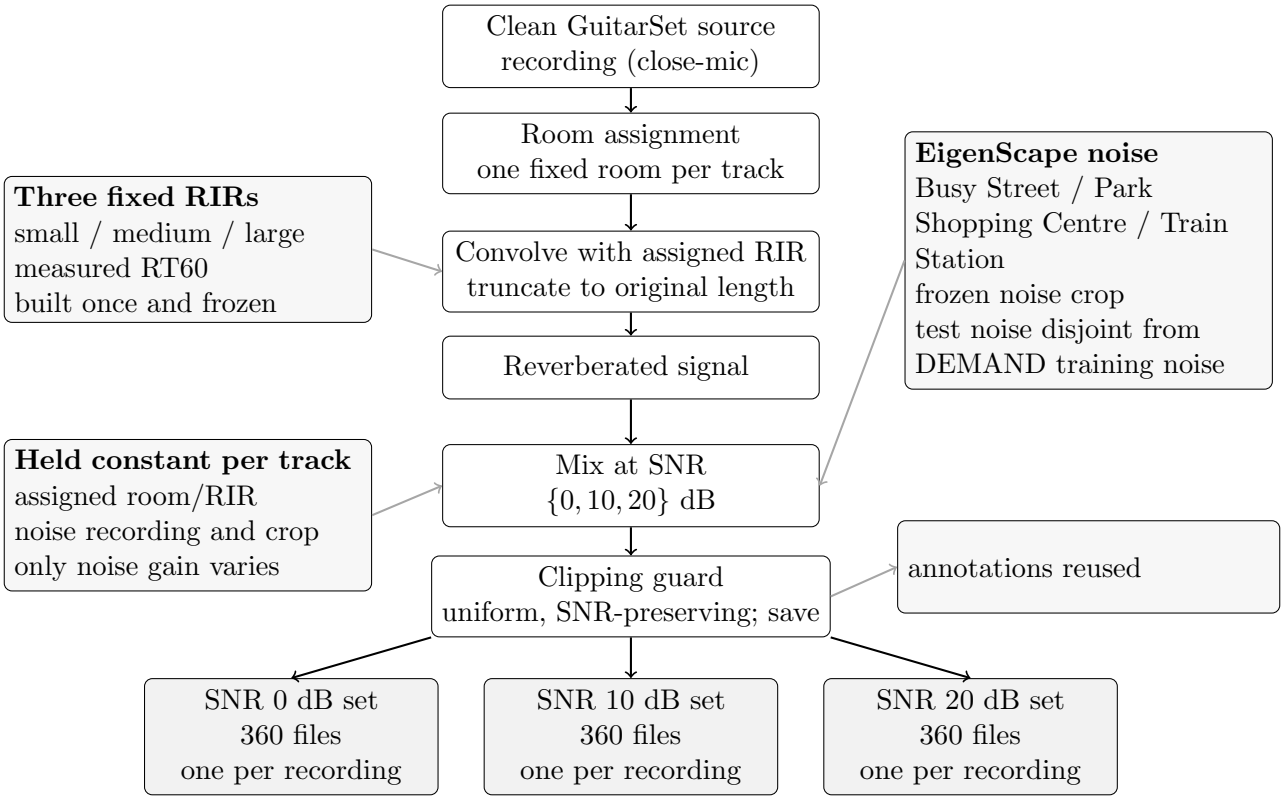

Figure 4: Offline construction of GuitarSetENV-Test. Each GuitarSet source recording is assigned one fixed room and one EigenScape noise recording and crop, frozen before any model testing; its three outputs differ only in the noise scaling used to reach the target SNR.

Table 6: Contents of GuitarSetENV-Test.

Property	GuitarSetENV-Test
Source recordings	360 GuitarSet recordings
Generated files	1,080
Files per SNR level	360
SNR levels	20, 10, and 0 dB
Noise corpus	EigenScape
Noise classes	Busy Street, Park, Shopping Centre, Train Station
Reverberation	Three fixed simulated RIRs
Across-SNR assignment	Same room, RIR, noise recording, and crop; only noise gain changes
Annotations	Original GuitarSet annotations retained
Evaluation use	60 held-out-player files per fold at each SNR

4.2.10 EGSet12+

EGSet12+ is the proposed environmentally degraded extension of EGSet12. Its purpose is to evaluate environmental robustness after the source domain has already shifted from the acoustic-guitar GuitarSet recordings to out-of-domain electric-guitar recordings.

The twelve original EGSet12 recordings remain unchanged and provide the clean electric-guitar test condition. EGSet12+ is constructed by assigning each source track one selected EigenScape recording and one fixed crop. The same noise recording and crop are reused to produce versions at 20, 10, and 0 dB SNR, so only the noise gain changes between the three versions. No simulated room reverberation is applied, and the original stereo format is retained. The original EGSet12 annotations are reused unchanged.

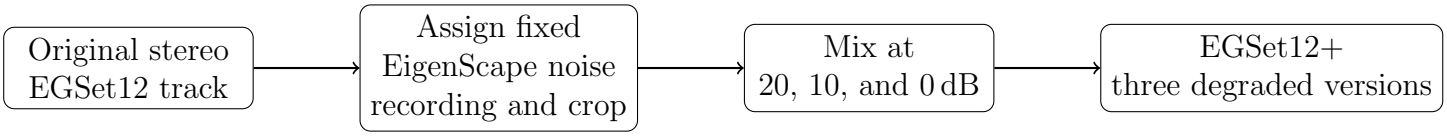

Figure 5: Offline construction of EGSet12+. Each EGSet12 track keeps the same EigenScape recording and crop at all three SNR levels; only the noise gain changes. No simulated RIR is applied.

Table 7: Comparison of original EGSet12 and EGSet12+.

Property	Original EGSet12	EGSet12+
Source tracks	12 electric-guitar performances	Same 12 tracks as sources
Provenance	Pedroza et al.	Constructed in this thesis
Files	12 original files	36 degraded files, 12 per SNR
Audio format	Stereo	Stereo preserved
Environmental noise	None added	Selected EigenScape recordings
Simulated RIR	None	None
SNR	No added environmental noise	20, 10, and 0 dB
Across-SNR crop reuse	Not applicable	Same noise file and crop per track
Annotations	Original annotations	Original annotations retained

4.3 Training configurations V1–V5

All five training configurations use the same six-fold, player-wise GuitarSet split. GuitarSet contains six guitarists and 360 recordings. In each fold, the 300 recordings from five guitarists form the clean training partition, while the 60 recordings from the remaining guitarist form the held-out clean validation fold. Additional GuitarSet-derived datasets follow the same player separation.

All five configurations use the TAB-CNN implementation described in Section 3 through the PyTorch/AMT-Tools pipeline associated with Pedroza et al. [9]. Training uses the Adadelta optimiser with a learning rate of 1.0. Each configuration is trained for 2,500 traversals of its shuffled fold-specific training loader. Checkpoints are stored every 50 traversals, and the checkpoint after traversal 2,500 is used for evaluation. The validation and final-test sets are not used for checkpoint selection.

4.3.1 V1: clean GuitarSet

V1 represents the clean reference configuration. It uses only clean GuitarSet training recordings and therefore provides the reference for configurations that add effect-based or environmental training data.

In each fold, V1 uses the 300 clean GuitarSet recordings from the five training guitarists. The remaining 60 recordings from the held-out guitarist form the validation fold.

Experimental setup: 300 training entries per fold, batch size 8, and 2,500 training traversals.

4.3.2 V2: GuitarSet and GuitarSetFX

V2 represents the GuitarSetFX effect-augmentation configuration associated with Pedroza et al. [9]. It uses the original clean GuitarSet recordings together with effect-rendered versions of the same performances. Its purpose in this thesis is to provide an effect-augmentation reference configuration.

In each fold, V2 uses 300 clean GuitarSet recordings and the corresponding 300 GuitarSetFX recordings from the same five training guitarists. GuitarSet and GuitarSetFX recordings belonging to the held-out guitarist are excluded from training.

Experimental setup: 600 training entries per fold, batch size 32, and 2,500 training traversals.

4.3.3 V3: GuitarSet and GuitarProFX

V3 represents the GuitarProFX effect-augmentation configuration associated with Pedroza et al. [9]. It combines clean GuitarSet with separate guitar material rendered using electric-guitar tones and effects. It therefore provides a second effect-augmentation reference configuration using different musical material.

In each fold, V3 uses the 300 clean GuitarSet recordings from the five training guitarists together with all 360 GuitarProFX recordings. GuitarProFX does not follow the GuitarSet player split and is therefore included in full in every fold.

Experimental setup: 660 training entries per fold, batch size 32, and 2,500 training traversals.

4.3.4 V4: GuitarSet and GuitarSetENV

V4 represents the proposed environmental training configuration in the GuitarSet domain. It retains the clean GuitarSet recordings and adds their corresponding GuitarSetENV copies, so the model is trained on both clean and environmentally degraded versions of the same performances.

In each fold, V4 uses the 300 clean GuitarSet recordings from the five training guitarists together with their 300 corresponding GuitarSetENV copies. Both the clean and degraded versions belonging to the held-out guitarist are excluded from training. The training data therefore have a 1:1 clean-to-environmentally-degraded composition.

Experimental setup: 600 training entries per fold, batch size 8, and 2,500 training traversals.

4.3.5 V5: combined acoustic- and electric-guitar training

V5 represents the combined training configuration. It contains clean and environmentally degraded GuitarSet data together with GuitarProFX and its environmentally degraded counterpart GuitarProFXENV. Its purpose is to expose TAB-CNN to acoustic- and electric-guitar material together with environmental degradation. V5 does not contain GuitarSetFX.

In each fold, V5 uses 300 clean GuitarSet recordings, their 300 corresponding GuitarSetENV copies, all 360 GuitarProFX recordings, and all 360 GuitarProFXENV entries. The GuitarSet-derived branches preserve the player-wise split, while GuitarProFX and GuitarProFXENV are included in full.

Experimental setup: 1,320 training entries per fold, batch size 32, and 2,500 training traversals.

Table 8: Summary of the five TAB-CNN training configurations. Counts are per cross-validation fold.

Condition	Training data per fold	Entries	Batch
V1	300 GuitarSet	300	8
V2	300 GuitarSet + 300 corresponding GuitarSetFX	600	32
V3	300 GuitarSet + 360 GuitarProFX	660	32
V4	300 GuitarSet + 300 corresponding GuitarSetENV	600	8
V5	300 GuitarSet + 300 GuitarSetENV + 360 GuitarProFX + 360 GuitarProFXENV	1,320	32

4.4 Baseline reproduction and implementation check

Before training V1–V5, the released Keras implementation of TAB-CNN by Wiggins and Kim was rerun on clean GuitarSet [19]. This separate run was used to check whether the published clean-condition benchmark could be approximately recovered in the local Keras setup.

The Keras rerun is not one of V1–V5. The five experimental configurations are trained, validated, and finally tested using the PyTorch/AMT-Tools pipeline associated with Pedroza et al. [9]. The Keras result therefore serves only as a numerical benchmark check against the original TAB-CNN result.

The locally trained V2 and V3 models use the GuitarSetFX and GuitarProFX datasets supplied by Hegel Pedroza together with the PyTorch/AMT-Tools implementation. Their scores are close to, but not identical to, the corresponding values reported by Pedroza et al. They are therefore treated as approximate reproductions of the two effect-augmentation configurations.

4.5 Evaluation protocol and aggregation

All five configurations use the same fold assignments, scoring implementation, and performance measures. The performance measures are defined in Section 3. Clean GuitarSet is used for fold validation, while GuitarSetENV-Test, EGSet12, and EGSet12+ are final test sets.

For clean GuitarSet, the final checkpoint from fold k is evaluated only on the 60 clean recordings of guitarist k, who was excluded from training in that fold. Each metric is first averaged across those 60 recordings. The reported clean GuitarSet result is then the mean and sample standard deviation across the six fold-level means.

GuitarSetENV-Test follows the same out-of-fold player restriction. At each SNR level, the final checkpoint from fold k is evaluated only on the 60 degraded recordings belonging to guitarist k. Each metric is first averaged across these 60 recordings, after which the mean and sample standard deviation are calculated across the six fold-level means.

For EGSet12 and EGSet12+, all six fold checkpoints evaluate all twelve source tracks. For each source track and evaluation condition, each metric is first averaged across the six checkpoints. The reported mean and sample standard deviation are then calculated across the twelve source-track means. EGSet12+ is evaluated separately at 20, 10, and 0 dB SNR.

The GuitarSet-based and EGSet12-based standard deviations therefore describe variation across different observational units: six held-out-player fold means for GuitarSet and GuitarSetENV-Test, and twelve source-track means for EGSet12 and EGSet12+. Stored F_1 values are averaged directly rather than reconstructed from aggregated precision and recall. No inferential statistical tests are performed.

4.6 Code and data availability

The public repository associated with this thesis is available at https://github.com/MarciTheOfficial/tabcnn-environmental-robustness.git. It contains the code and configuration files used for dataset construction, model training, and evaluation. GuitarSetFX and GuitarProFX were supplied privately by Hegel Pedroza and are therefore not redistributed. Because GuitarProFXENV is derived from GuitarProFX, it is subject to the same restriction. The repository instead provides the code required to use these datasets when they have been obtained from their original authors.

5 Results

The results are organised into four evaluation families that examine the training conditions along two complementary dimensions: recording domain and added environmental degradation. Clean GuitarSet provides the in-domain acoustic-guitar reference. GuitarSetENV-Test retains the GuitarSet performance domain but adds environmental noise and simulated room reverberation. Clean EGSet12 introduces an out-of-domain electric-guitar setting without added environmental degradation, while EGSet12+ combines this electric-guitar domain shift with additive environmental noise.

This structure provides relative views of robustness across the different training conditions. GuitarSetENV-Test includes simulated reverberation, whereas EGSet12+ uses additive noise only. The GuitarSet and EGSet12 families also differ in source recordings, dataset size, and aggregation procedure. Results are therefore interpreted within each evaluation family while taking these differences into account.

The primary comparison is between V1 and V4, which both use a batch size of 8. V1 uses clean GuitarSet training, while V4 extends this training recipe with the proposed GuitarSetENV environmental augmentation. V2, V3, and V5 form the secondary comparison group because all three use a batch size of 32 and represent different augmentation and training-data strategies. All five conditions are also compared descriptively at a broader level. Differences in training-data size and composition are part of the evaluated training recipes and were not independently controlled.

All comparisons are descriptive because no inferential statistical tests were performed. For clean GuitarSet and GuitarSetENV-Test, values are means and sample standard deviations across the six held-out-player fold means. For EGSet12 and EGSet12+, each metric is first averaged across the six fold checkpoints for each source track, after which the mean and sample standard deviation are

calculated across the twelve source-track means. The complete aggregation procedures are defined in Section 4.5.

5.1 Clean GuitarSet: in-domain reference

Table 9 reports performance on the clean GuitarSet held-out-player splits. This evaluation provides the clean in-domain reference and shows how the complete training configurations perform without added environmental degradation or electric-guitar domain shift.

Table 9: Performance on clean GuitarSet. Values computed in this thesis are means ± sample standard deviation across the six held-out-player fold means. The W&K row is quoted from Wiggins and Kim [19], while the reported GuitarSetFX and GuitarProFX rows are quoted from Pedroza et al. [9]. Their complete aggregation conventions are not assumed to be identical to those used in this thesis. The Keras rerun is a separate benchmark check and is not one of V1–V5. Conditions V1–V5 are defined in Table 8. Bold marks the highest mean among V1–V5 only.

	Multi-pitch estimation			**Tablature estimation**			
Setup	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**TDR**
W&K reported	0.826±0.025	0.900±0.016	0.764±0.043	0.748±0.047	0.809±0.029	0.696±0.061	0.899±0.033
Keras rerun	0.821±0.029	0.905±0.015	0.754±0.052	0.744±0.047	0.814±0.023	0.687±0.065	0.900±0.029
V1	0.831±0.021	0.907±0.022	0.774±0.038	0.743±0.032	0.799±0.023	0.699±0.045	0.898±0.020
GuitarSetFX reported	0.837±0.019	0.904±0.019	0.785±0.038	0.746±0.030	0.795±0.022	0.708±0.044	0.896±0.021
V2	0.829±0.023	0.911±0.021	0.767±0.042	0.740±0.035	0.803±0.019	0.691±0.050	0.895±0.022
GuitarProFX reported	0.830±0.018	0.908±0.014	0.769±0.027	0.743±0.029	0.802±0.029	0.696±0.035	0.900±0.021
V3	**0.836**±0.018	0.903±0.022	**0.784**±0.033	**0.748**±0.030	0.797±0.021	**0.710**±0.044	**0.902**±0.020
V4	0.832±0.018	0.904±0.023	0.777±0.033	0.744±0.031	0.798±0.030	0.702±0.039	0.898±0.022
V5	0.823±0.015	**0.912**±0.020	0.756±0.022	0.737±0.026	**0.808**±0.022	0.683±0.032	0.899±0.018

The separate Keras rerun differs from the published Wiggins and Kim multi-pitch and tablature F_1–means by 0.005 and 0.004, respectively. This provides a numerical check against the original benchmark but does not validate the later PyTorch pipeline or establish equivalence between the two implementations. The locally trained V2 and V3 results are also close to the corresponding values reported by Pedroza et al., supporting the approximate reproduction described in Section 4.4.

The primary V1–V4 comparison shows very similar performance on clean GuitarSet. Their multi-pitch F_1–means are 0.831 and 0.832, respectively, while their tablature F_1–means are 0.743 and 0.744. The main separation between these two conditions therefore appears under environmental degradation rather than on clean GuitarSet.

Within the secondary V2–V3–V5 comparison group, V3 has the highest observed multi-pitch and tablature F_1–means and the highest recall and TDR–means, while V5 has the highest precision means. At the broader V1–V5 level, V3 also has the highest observed multi-pitch and tablature F_1–means, at 0.836 and 0.748.

5.2 GuitarSetENV-Test: environmental degradation within the GuitarSet domain

Table 10 reports performance on the held-out GuitarSet recordings after the addition of EigenScape environmental noise and simulated room reverberation. This evaluation retains the GuitarSet performance domain while varying the environmental-degradation level.

Table 10: Performance on GuitarSetENV-Test at 20, 10, and 0 dB SNR. At each SNR level, values are means ± sample standard deviation across the six held-out-player fold means. GuitarSetENV-Test uses EigenScape environmental noise and simulated room reverberation. The V1–V5 condition labels are defined in the Methodology. Bold marks the highest observed mean among V1–V5 within each SNR level and metric.

		Multi-pitch estimation			**Tablature estimation**			
SNR	Condition	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	**TDR**
20 dB	V1	0.673±0.026	0.851±0.042	0.570±0.040	0.568±0.033	0.702±0.054	0.488±0.037	0.838±0.032
	V2	0.694±0.024	0.858±0.035	0.595±0.039	0.590±0.026	0.719±0.032	0.510±0.036	0.849±0.020
	V3	0.711±0.018	0.848±0.037	0.622±0.028	0.605±0.027	0.708±0.040	0.537±0.031	0.856±0.023
	V4	**0.750**±0.016	0.882±0.027	**0.664**±0.028	**0.655**±0.027	0.759±0.035	**0.585**±0.031	0.875±0.022
	V5	0.739±0.023	**0.900**±0.022	0.638±0.031	0.648±0.027	**0.777**±0.023	0.564±0.034	**0.876**±0.020
10 dB	V1	0.546±0.034	0.793±0.065	0.432±0.044	0.432±0.040	0.609±0.082	0.346±0.037	0.769±0.053
	V2	0.577±0.025	0.842±0.036	0.451±0.030	0.476±0.029	0.682±0.037	0.375±0.029	0.816±0.026
	V3	0.608±0.019	0.829±0.042	0.492±0.024	0.499±0.028	0.663±0.047	0.410±0.026	0.816±0.030
	V4	**0.698**±0.023	0.881±0.025	**0.591**±0.032	**0.599**±0.031	0.744±0.036	**0.512**±0.034	0.859±0.024
	V5	0.681±0.032	**0.902**±0.021	0.560±0.041	0.588±0.031	**0.767**±0.024	0.488±0.039	**0.861**±0.019
0 dB	V1	0.328±0.020	0.501±0.103	0.265±0.030	0.220±0.032	0.335±0.100	0.176±0.022	0.606±0.083
	V2	0.345±0.015	0.608±0.060	0.256±0.014	0.260±0.021	0.453±0.057	0.193±0.015	0.715±0.050
	V3	0.362±0.011	0.581±0.085	0.283±0.016	0.274±0.017	0.432±0.074	0.215±0.011	0.721±0.042
	V4	**0.518**±0.035	0.817±0.033	**0.395**±0.035	**0.430**±0.037	0.665±0.044	**0.330**±0.034	0.822±0.030
	V5	0.496±0.044	**0.850**±0.022	0.364±0.043	0.418±0.040	**0.702**±0.041	0.309±0.038	**0.833**±0.033

For every training condition, multi-pitch and tablature F_1 decrease as the SNR is reduced from 20 to 0 dB. In the primary V1–V4 comparison, V4 has substantially higher multi-pitch and tablature F_1–means than V1 at all three SNR levels. Its tablature F_1–means are 0.655, 0.599, and 0.430 at 20, 10, and 0 dB, respectively, compared with 0.568, 0.432, and 0.220 for V1. V4 also has higher recall means than V1 at every SNR level.

Within the secondary V2–V3–V5 comparison group, V5 has the highest multi-pitch and tablature F_1, precision, and TDR–means at all three SNR levels, while V3 has higher recall than V2. At the broader V1–V5 level, V4 has the highest observed multi-pitch and tablature F_1 and recall means at every SNR level, while V5 has the highest precision and TDR–means. Recall is lower than the corresponding precision for every condition and SNR level.

5.3 EGSet12: electric-guitar domain shift

Table 11 reports performance on the twelve clean EGSet12 source tracks. Unlike clean GuitarSet, EGSet12 represents an out-of-domain electric-guitar setting with different performance material, guitar tones, and recording characteristics. No environmental noise was added to these recordings as part of this thesis.

Table 11: Performance on clean EGSet12. For the thesis conditions V1–V5, each metric is first averaged across the six fold checkpoints for each source track. Values are then reported as means ± sample standard deviation across the twelve source-track means. The three Pedroza rows are quoted from Pedroza et al. [9] and are included as external reference values; their complete aggregation convention is not assumed to be identical to that used in this thesis. Bold marks the highest observed mean among V1–V5 only.

	Multi-pitch estimation			**Tablature estimation**			
Condition	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	**TDR**
Pedroza: clean	0.638±0.060	0.819±0.080	0.530±0.067	0.447±0.071	0.565±0.089	0.375±0.067	0.695±0.075
V1	0.649±0.117	0.829±0.117	0.543±0.126	0.460±0.160	0.573±0.169	0.390±0.154	0.702±0.156
Pedroza: GuitarSetFX	0.740±0.055	0.835±0.085	0.679±0.052	0.557±0.088	0.619±0.100	0.518±0.084	0.755±0.106
V2	0.719±0.128	0.860±0.109	0.631±0.154	0.548±0.181	0.643±0.177	0.487±0.188	0.758±0.162
Pedroza: GuitarProFX	0.719±0.061	0.839±0.082	0.647±0.068	0.585±0.084	0.658±0.073	0.541±0.087	0.819±0.075
V3	**0.731**±0.143	0.837±0.108	**0.667**±0.179	**0.599**±0.168	0.663±0.137	**0.559**±0.191	**0.827**±0.114
V4	0.645±0.108	0.833±0.108	0.534±0.118	0.446±0.146	0.564±0.154	0.376±0.144	0.692±0.161
V5	0.722±0.141	**0.874**±0.110	0.630±0.168	0.594±0.179	**0.697**±0.148	0.529±0.193	0.825±0.138

The primary V1–V4 comparison shows similar performance on clean EGSet12, with V4 slightly below V1. Their multi-pitch F_1–means are 0.649 and 0.645, respectively, while their tablature F_1–means are 0.460 and 0.446. Environmental augmentation therefore does not provide an advantage over clean training on this clean out-of-domain electric-guitar evaluation.

Within the secondary V2–V3–V5 comparison group, V3 has the highest observed multi-pitch and tablature F_1–means, recall means, and TDR–mean, while V5 has the highest precision means. At the broader V1–V5 level, V3 also has the highest observed multi-pitch and tablature F_1–means, at 0.731 and 0.599. The locally computed V2 and V3 values are broadly similar to the corresponding results reported by Pedroza et al. and provide an approximate reproduction of their effect-augmentation conditions. The remaining numerical differences may partly reflect checkpoint and aggregation details that are not fully documented in the published description.

5.4 EGSet12+: combined domain and environmental shift

Table 12 reports performance on EGSet12+ at 20, 10, and 0 dB SNR. EGSet12+ retains the out-of-domain electric-guitar source recordings of EGSet12 and adds EigenScape environmental noise. It therefore evaluates the combined presence of electric-guitar domain shift and additive environmental degradation. Unlike GuitarSetENV-Test, it does not include additional simulated room reverberation.

Table 12: Performance on EGSet12+ at 20, 10, and 0 dB SNR. Each metric is first averaged across the six fold checkpoints for each source track. Values are then reported as means ± sample standard deviation across the twelve source-track means. EGSet12+ adds EigenScape environmental noise without simulated room reverberation. Conditions V1–V5 are defined in the Methodology. Bold marks the highest observed mean among V1–V5 within each SNR level and metric.

		Multi-pitch estimation			**Tablature estimation**			
SNR	Condition	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	**TDR**
20 dB	V1	0.578±0.135	0.810±0.121	0.462±0.135	0.402±0.173	0.549±0.187	0.325±0.154	0.676±0.178
	V2	0.664±0.145	0.858±0.118	0.558±0.159	0.495±0.198	0.624±0.196	0.420±0.190	0.730±0.178
	V3	0.678±0.156	0.833±0.112	**0.592**±0.182	0.542±0.191	0.640±0.161	0.483±0.199	0.796±0.144
	V4	0.614±0.119	0.824±0.113	0.499±0.127	0.415±0.144	0.544±0.153	0.342±0.139	0.673±0.151
	V5	**0.691**±0.150	**0.873**±0.108	0.588±0.172	**0.560**±0.189	**0.685**±0.167	**0.485**±0.194	**0.807**±0.149
10 dB	V1	0.478±0.131	0.715±0.124	0.374±0.125	0.315±0.171	0.455±0.205	0.248±0.142	0.622±0.218
	V2	0.539±0.163	0.806±0.129	0.425±0.155	0.401±0.181	0.583±0.188	0.316±0.155	0.720±0.183
	V3	0.580±0.172	0.770±0.131	0.488±0.182	0.449±0.194	0.571±0.178	0.385±0.187	0.760±0.176
	V4	0.533±0.129	0.783±0.128	0.418±0.130	0.341±0.134	0.487±0.142	0.271±0.125	0.630±0.138
	V5	**0.612**±0.162	**0.847**±0.120	**0.501**±0.173	**0.478**±0.184	**0.635**±0.167	**0.398**±0.182	**0.772**±0.162
0 dB	V1	0.320±0.089	0.475±0.143	0.258±0.087	0.195±0.123	0.287±0.182	0.156±0.099	0.571±0.259
	V2	0.339±0.125	0.589±0.197	0.252±0.099	0.244±0.131	0.411±0.210	0.181±0.096	0.685±0.211
	V3	0.386±0.146	0.560±0.202	**0.314**±0.126	0.283±0.150	0.392±0.210	**0.232**±0.122	0.705±0.215
	V4	0.366±0.132	0.700±0.160	0.267±0.121	0.207±0.109	0.375±0.132	0.153±0.095	0.548±0.148
	V5	**0.410**±0.145	**0.750**±0.180	0.304±0.133	**0.300**±0.142	**0.516**±0.175	0.227±0.124	**0.713**±0.188

All five conditions show lower multi-pitch and tablature F_1–means as the SNR decreases. In the primary V1–V4 comparison, V4 has higher multi-pitch and tablature F_1–means than V1 at all three SNR levels. This shows that the environmental V4 training recipe also improves robustness relative to clean V1 training when environmental noise is added to the out-of-domain electric-guitar recordings, although V4 does not lead this evaluation overall.

Within the secondary V2–V3–V5 comparison group, V5 has the highest multi-pitch and tablature F_1–means at all three SNR levels. At 20 dB, V3 has the highest multi-pitch recall mean, while V5 has the highest means for the remaining metrics. At 10 dB, V5 has the highest observed mean for all seven reported metrics. At 0 dB, V3 has the highest multi-pitch and tablature recall means, while V5 has the highest means for the remaining metrics.

At the broader V1–V5 level, V5 therefore has the highest observed multi-pitch and tablature F_1–means at all three SNR levels. This differs from GuitarSetENV-Test, where V4 has the highest observed F_1–means, indicating that the leading complete training recipe depends on the evaluation condition.

Table 13 compares relative performance loss between clean EGSet12 and EGSet12+ at 0 dB. The comparison uses the clean values reported in Table 11 and the 0 dB values reported in Table 12.

Table 13: Relative decline from clean EGSet12 to EGSet12+ at 0 dB SNR. Decline is calculated for each condition as $(\overline{x}_{\text{clean}} - \overline{x}_{0\,\text{dB}})/\overline{x}_{\text{clean}} \times 100$ using the underlying unrounded means. Bold marks the smallest displayed decline per metric. Equal displayed values are treated as ties.

	Multi-pitch F_1			**Tablature** F_1		
Condition	Clean	0 dB	Decline (%)	Clean	0 dB	Decline (%)
V1	0.649	0.320	50.8	0.460	0.195	57.5
V2	0.719	0.339	52.9	0.548	0.244	55.5
V3	0.731	0.386	47.2	0.599	0.283	52.8
V4	0.645	0.366	**43.2**	0.446	0.207	53.7
V5	0.722	0.410	**43.2**	0.594	0.300	**49.5**

In the primary V1–V4 comparison, V4 shows a smaller multi-pitch F_1 decline than V1, at 43.2% compared with 50.8%, while their tablature F_1 declines are 53.7% and 57.5%, respectively. Within the secondary V2–V3–V5 group, V5 has the smallest displayed decline for both metrics, with 43.2% for multi-pitch F_1 and 49.5% for tablature F_1. At the broader V1–V5 level, V4 and V5 are tied for the smallest displayed multi-pitch F_1 decline, while V5 has the smallest tablature F_1 decline. These relative changes complement the absolute scores and are interpreted descriptively.

6 Discussion

The results show that robustness in this study depends on the evaluation condition. The four evaluations examine two related dimensions: the source domain, represented by GuitarSet and EGSet12, and the presence or absence of added environmental degradation. This organisation is useful for interpreting the results, but it is not a controlled factorial design because the GuitarSet and EGSet12 evaluations use different source recordings, degradation procedures, sample sizes, and aggregation units.

The main research question concerns V1 and V4 within the GuitarSet domain. On clean GuitarSet, the displayed multi-pitch and tablature F_1–means of the two conditions differ by only 0.001. This does not establish equivalence, but it shows that the complete V4 training configuration did not produce a clear reduction in the displayed clean-condition means. The difference between the conditions becomes larger on GuitarSetENV-Test. V4 has the highest observed multi-pitch and tablature F_1 and recall means at every SNR level, and its separation from V1 increases as the SNR decreases. Under the applied protocol, the complete V4 configuration therefore performs better than the clean baseline when the GuitarSet recordings contain environmental noise and simulated room reverberation, while their clean results remain similar.

A plausible interpretation is that the GuitarSetENV branch gives V4 experience with the same broad degradation family used in GuitarSetENV-Test: additive environmental noise combined with simulated room reverberation. The environmental-noise recordings are not shared between training and evaluation, because training uses DEMAND and testing uses EigenScape. The result is therefore consistent with transfer to unseen noise recordings within the selected degradation family. It does not, however, demonstrate robustness to environmental recordings or room conditions in general. This clean-versus-degraded pattern is broadly consistent with earlier robustness-oriented

augmentation studies, in which the benefit of augmented training becomes more visible under shifted or degraded evaluation conditions [4, 6, 13].

The clean EGSet12 results show a different form of robustness. Within the secondary V2–V3–V5 comparison group, V3 has the highest observed multi-pitch and tablature F_1–means, recall means, and TDR–mean, while V5 has the highest precision means. This direction is consistent with the effect-augmentation results reported by Pedroza et al. [9]. GuitarSetFX and GuitarProFX introduce guitar-tone and effect diversity, and GuitarProFX additionally introduces separate musical material. These factors make V2 and V3 more closely related to the type of domain shift represented by EGSet12 than V4, whose additional data change the acoustic environment of GuitarSet recordings. At the broader V1–V5 level, V3 also has the highest observed F_1–means on this evaluation. The present experiment cannot determine whether V3's result is explained by electric-guitar tones, effects, separate musical material, additional data, or their combination.

EGSet12+ combines the electric-guitar domain shift of EGSet12 with additive environmental noise. In the primary V1–V4 comparison, V4 has higher multi-pitch and tablature F_1–means than V1 at all three SNR levels. Within the secondary V2–V3–V5 comparison group, V5 has the highest observed multi-pitch and tablature F_1–means at all three SNR levels. Its training data include both GuitarProFX and GuitarProFXENV in addition to the clean and environmentally degraded GuitarSet branches. The result is therefore consistent with the broader V5 training recipe being useful when both electric-guitar domain shift and environmental degradation are present. At the broader V1–V5 level, V5 also has the highest observed F_1–means on EGSet12+.

V5 should nevertheless be interpreted as the larger combined training configuration constructed for this thesis, rather than as evidence that effect-based and environmental augmentation have a complementary or synergistic causal effect. Within the V2–V3–V5 comparison group, the batch size is shared, but the conditions still differ in training-data composition and training-set size. Comparisons between V5 and the batch-size-8 conditions additionally include a batch-size difference. Its results therefore describe the complete expanded recipe. A controlled comparison would require the relevant branches to be varied independently under matched training settings.

The contrast between V4 and V5 also shows why robustness should not be summarised using one overall ranking. V4 has the highest observed F_1 means when GuitarSet recordings are degraded by noise and simulated reverberation. V5 has the highest observed F_1–means when additive noise is applied to the out-of-domain EGSet12 recordings. These evaluations differ in source domain and in the use of simulated reverberation, so the ranking change cannot be assigned to one factor. The supported conclusion is that V4 and V5 provide different advantages under the tested evaluation conditions, and that no training condition is the strongest across all domains and metrics.

The precision and recall results provide additional descriptive information. On GuitarSetENV-Test, recall is lower than precision for every condition and SNR level. At the aggregate metric level, this means that the models recover a smaller proportion of the reference activations than the proportion of their predicted activations that are correct. This is compatible with missed activations contributing strongly to the degraded-condition performance loss, but a prediction-level error analysis would be required to determine the exact error patterns.

Absolute performance and relative retention also answer different questions. V3 has the highest absolute F_1–means on clean EGSet12, while V5 has the highest means on the noisy EGSet12+ conditions. From clean EGSet12 to EGSet12+ at 0 dB, V4 and V5 both show a 43.2% multi-pitch F_1 decline at the reported precision, while V5 has the smallest displayed tablature F_1 decline.

Relative retention should therefore be considered together with the absolute clean and degraded scores rather than used as a separate overall model ranking.

6.1 Limitations

The conclusions are limited to the experimental setup used in this thesis. All five conditions use one TAB-CNN architecture and the same GuitarSet-based fold protocol. V1 and V4 both use batch size 8, while V2, V3, and V5 use batch size 32. Within these groups, the conditions differ in training-data composition and number of entries as part of the evaluated augmentation and training recipes. The observed differences therefore compare complete training configurations and do not independently isolate the effects of augmentation type, training-data amount, or other training factors. Broader comparisons across V1–V5 additionally include the batch-size difference between the two groups. V5 is also a nested expansion that reuses the GuitarSetENV and GuitarProFX branches rather than an independently balanced condition.

V2 and V3 are approximate reproductions of the effect-augmentation conditions studied by Pedroza et al. The locally trained models use the datasets and PyTorch/AMT-Tools code supplied by the authors, and their results are close to the corresponding published values. Exact numerical replication is not claimed because some checkpoint and aggregation details could not be verified from the published description.

The four evaluation families provide a useful conceptual comparison, but they do not form a balanced two-by-two experiment. GuitarSetENV-Test adds EigenScape noise and simulated room reverberation to GuitarSet, whereas EGSet12+ adds EigenScape noise without additional reverberation to EGSet12. The GuitarSet and EGSet12 families also contain different performances and use different aggregation units. The separate effects of source-domain shift, environmental noise, and reverberation therefore cannot be obtained by directly subtracting results across the four evaluations.

All degraded evaluation files were generated offline from recordings that were originally clean. GuitarSetENV-Test uses environmental recordings from EigenScape and simulated rooms produced using the same RIR simulator family used for GuitarSetENV training. Training and evaluation use different noise corpora, but they still share the broader additive-noise-and-simulation design. EGSet12+ uses additive noise only. The results therefore support claims about transfer within these selected degradation procedures, not about guitar performances recorded directly in arbitrary noisy or reverberant environments. The realised GuitarSetENV reverberation times ranged from 0.148 to 2.230 s, exceeding the nominal target interval of 0.2–0.8 s; the stored training data therefore contain broader reverberation than originally requested.

The EGSet12-based evaluation is also small. Original EGSet12 contains twelve performances from one guitarist and recording chain [9]. EGSet12+ reuses these twelve sources at three SNR levels, and each source track is paired with one fixed environmental scene. Track and scene effects are consequently confounded, and the degraded conditions are repeated derivatives rather than independent performances. The six checkpoint predictions for each track are also related because they come from the six folds of the same architecture and training procedure.

Finally, the reported spreads are sample standard deviations rather than confidence intervals. The GuitarSet-based standard deviations describe variation across six held-out-player fold means, whereas the EGSet12-based standard deviations describe variation across twelve source-track means. Their magnitudes should therefore not be compared directly. No inferential statistical tests

were performed, and overlap between mean ± standard deviation intervals does not demonstrate equivalence or the absence of a difference. All conclusions in this thesis are consequently descriptive and limited to the evaluated recordings and training conditions.

7 Conclusion

This thesis investigated how environmental-degradation training affects the robustness of TAB-CNN. The primary comparison was between V1 and V4, which both used a batch size of 8. Their multi-pitch and tablature F_1 means are similar on clean GuitarSet. On GuitarSetENV-Test, however, V4 has higher observed means than V1 for every reported metric at 20, 10, and 0 dB SNR. It also obtains the highest multi-pitch and tablature F_1 and recall means among V1–V5 at all three SNR levels. Under the applied protocol, the complete V4 configuration therefore provides greater robustness to the tested environmental noise and simulated room reverberation while retaining similar reported performance on clean GuitarSet.

Across GuitarSetENV-Test, recall is lower than precision for all five training conditions at every SNR level. For V4, this difference becomes larger as the SNR is reduced: precision remains comparatively high, while recall falls substantially. This pattern indicates that missed reference activations contribute substantially to the degradation-related performance loss, while a comparatively high proportion of the predicted activations remains correct. The secondary comparison was among V2, V3, and V5, which all used a batch size of 32 and together included acoustic- and electric-guitar training data. V3 has the highest observed multi-pitch and tablature F_1–means on clean GuitarSet and clean EGSet12. On EGSet12+, however, V5 has the highest observed means for both F_1 metrics at every tested SNR level. Under the applied protocol, V3 therefore provides the strongest clean-condition F_1 results in this group, whereas V5 provides the strongest F_1 results when the electric-guitar domain shift is combined with additive environmental noise. In the broader comparison of V1–V5, no training condition is strongest across all evaluation families.

These comparisons concern complete training configurations rather than controlled ablations. Batch size is fixed within each comparison group, but the conditions differ in training-data composition and training-set size. The two groups also use different batch sizes. The evaluation is limited to one model architecture, offline-generated degradations, and twelve out-of-domain source tracks. The reported differences are descriptive because no inferential statistical tests were applied. The main contributions of this thesis are:

- the construction of GuitarSetENV as an environmentally degraded training dataset;
- the construction of the fixed 1,080-file GuitarSetENV-Test and 36-file EGSet12+ evaluation datasets;
- a five-condition comparison across clean and degraded acoustic- and electric-guitar evaluation settings, using DEMAND noise for training-data construction and EigenScape noise for final testing.

Overall, the results extend previous GTT robustness work beyond guitar-tone and effect variation. They show that stronger environmental degradation is associated with a reduced ability to recover reference activations and that environmentally degraded training data can improve robustness

under the tested conditions. They also show that the strongest observed training configuration depends on the evaluation domain and degradation type.

8 Future Work

Future work should first test whether the greater robustness observed for V4 on GuitarSetENV-Test transfers to guitar performances recorded directly in noisy and reverberant environments. Controlled experiments are also needed to isolate the contributions of the training-data components. The main priorities are:

1. Evaluate guitar performances recorded directly in noisy and reverberant environments. The recordings should cover different spaces, environmental conditions, performers, and recording chains and should be independent of the offline augmentation procedures used in this thesis.

2. Construct controlled ablations of V4 and V5. For V4, environmental noise and simulated room reverberation should be varied separately. For V5, the inclusion of GuitarSetENV, GuitarProFX, and GuitarProFXENV should be varied separately. The variants should use the same batch size and matched training-data and optimisation budgets so that the individual and combined contributions can be distinguished.

3. Expand the EGSet12-based evaluation with more performers, source tracks, guitar tones, and recording chains. Environmental scenes should be balanced across the source tracks, and both additive-noise and reverberant conditions should be included.

4. Apply paired statistical analyses to the retained per-track results and conduct prediction-level error analysis. The analyses should account for repeated source tracks across conditions and SNR levels and for the related predictions produced by the six cross-validation checkpoints. The error analysis should examine false-positive, false-negative, and predicted-activation counts directly.

5. Test environmental-degradation training with additional tablature-transcription architectures and larger training datasets. This would determine whether the observed patterns extend beyond TAB-CNN and GuitarSet-based training.

References

[1] Olusola O Abayomi-Alli, Robertas Damaševičius, Atika Qazi, Mariam Adedoyin-Olowe, and Sanjay Misra. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics*, 11(22):3795, 2022.

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

[3] Frank Cwitkowitz, Toni Hirvonen, and Anssi Klapuri. FretNet: Continuous-valued pitch contour streaming for polyphonic guitar tablature transcription. In *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023.

[4] Drew Edwards, Simon Dixon, Emmanouil Benetos, Akira Maezawa, and Yuta Kusaka. A data-driven analysis of robust automatic piano transcription. *IEEE Signal Processing Letters*, 31:681–685, 2024.

[5] Marc Ciufo Green and Damian Thomas Murphy. Eigenscape: A database of spatial acoustic scene recordings. *Applied sciences*, 7(11):1204, 2017.

[6] Yonghyun Kim and Alexander Lerch. Towards robust transcription: Exploring noise injection strategies for training data augmentation. In *Late-Breaking Demo Session, International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[7] Yuta Kusaka and Akira Maezawa. Mobile-AMT: Real-time polyphonic piano transcription for in-the-wild recordings. In *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO)*, pages 36–40, 2024.

[8] Jackson Loth, Pedro Sarmento, Saurjya Sarkar, Zixun Guo, Mathieu Barthet, and Mark Sandler. GOAT: A large dataset of paired guitar audio recordings and tablatures. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, pages 655–662, 2025.

[9] Hegel Pedroza, Wallace Abreu, Ryan M Corey, and Iran R Roman. Leveraging real electric guitar tones and effects to improve robustness in guitar tablature transcription modeling. In *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24)*, pages 143–146, 2024.

[10] Hegel Pedroza, Wallace Abreu, Ryan M Corey, and Iran R Roman. Guitar-TECHS: An electric guitar dataset covering techniques, musical excerpts, chords and scales using a diverse array of hardware. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[11] Xavier Riley, Drew Edwards, and Simon Dixon. High resolution guitar transcription via domain adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1051–1055, 2024.

[12] Xavier Riley, Zixun Guo, Andrew C. Edwards, and Simon Dixon. GAPS: A large and diverse classical guitar dataset and benchmark transcription model. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pages 611–617, 2024.

[13] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[14] Pedro Sarmento, Adarsh Kumar, CJ Carr, Zack Zukowski, Mathieu Barthet, and Yi-Hsuan Yang. Dadagp: A dataset of tokenized guitarpro songs for sequence models. *arXiv preprint arXiv:2107.14653*, 2021.

[15] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355, 2018.

[16] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 121–126, 2015.

[17] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19, page 035081. Acoustical Society of America, 2013.

[18] Yannik Venohr, Yiwei Ding, and Christof Weiss. Towards robust music transcription by measuring cross-version consistency in western classical music. In *Ismir 2025 Hybrid Conference*, 2025.

[19] Andrew Wiggins and Youngmoo E Kim. Guitar tablature estimation with a convolutional neural network. In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, pages 284–291, 2019.

[20] Andrew Franklin Wiggins. *Leveraging unlabeled data to improve guitar tablature transcription.* PhD thesis, Drexel University, 2024.

[21] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. GuitarSet: A dataset for guitar transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pages 453–460, 2018.

[22] Kazuki Yazawa, Daichi Sakaue, Kohei Nagira, Katsutoshi Itoyama, and Hiroshi G Okuno. Audio-based guitar tablature transcription using multipitch analysis and playability constraints. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 196–200. IEEE, 2013.

[23] Yongyi Zang, Yi Zhong, Frank Cwitkowitz, and Zhiyao Duan. SynthTab: Leveraging synthesized data for guitar tablature transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1286–1290, 2024.

A Implementation and construction details

This appendix collects exact implementation values referenced from Section 4.

Seeds and determinism. Table 14 lists the fixed seeds behind the constructed datasets and training runs. Checkpoints are saved every 50 traversals, and the final checkpoint at traversal 2,500 is used for evaluation.

Table 14: Fixed seeds used in this thesis.

Component	Seed	Purpose
GuitarSetENV builder	42	Global seed; deterministic per-track states and room seeds
GuitarProFXENV builder	42	Same global seed; per-track streams namespaced to `GuitarProFXENV.v1`
GuitarSetENV-Test plan	20260619	Master seed for room, noise, and crop assignments
EGSet12+ assignment	20260708	Global seed for noise and crop assignments
Model training	0	Per-fold training seed for every condition

GuitarSetENV (training data). Rooms are sampled as pyroomacoustics ShoeBox geometries with dimensions 3–8 m × 3–10 m × 2.3–4 m. The DEMAND training environments are DKITCHEN, DLIVING, NFIELD, NPARK, PCAFETER, PRESTO, PSTATION, SCAFE, and SPSQUARE. Across the 360 degraded files, the achieved SNR values span 0.11–23.96 dB (mean 12.40 dB) and the achieved RT60 values span 0.148–2.230 s (mean 0.711 s).

GuitarProFXENV (training data). GuitarProFXENV applies DEMAND noise and simulated reverberation to the 355 non-silent GuitarProFX tracks. The five silent source stems are retained unchanged. Their source annotations were also retained, but the correspondence between the zero-energy audio and these annotations was not independently verified. The original annotations are reused because the augmentation changes the audio but not the target notes or string–fret labels.

GuitarSetENV-Test. The 31 eligible EigenScape recordings span the four scene groups; each SNR level contains 360 outputs, 60 per player and 120 per room, and the scene totals are not balanced. A clipping guard rescales a mixture without changing its SNR where needed. Table 15 lists the fixed room identifiers used to reproduce this test set.

Table 15: Exact source and microphone positions and generation identifiers for the three fixed GuitarSetENV-Test rooms. Coordinates are in metres as (x, y, z).

Room	Source position	Microphone position	Room seed	RIR filename
Small	$(2.15, 1.95, 1.20)$	$(1.45, 1.95, 1.25)$	1,300,301	`room_small.wav`
Medium	$(3.65, 3.20, 1.30)$	$(2.65, 3.10, 1.35)$	1,500,503	`room_medium.wav`
Large	$(4.85, 4.65, 1.35)$	$(3.45, 4.50, 1.45)$	1,700,707	`room_large.wav`

EGSet12+. The twelve original EGSet12 recordings remain untouched; EGSet12+ contains 36 generated derivatives, twelve at each of 20, 10, and 0 dB SNR. EGSet12+ uses four EigenScape scene categories with three selected recordings from each category. Each of the twelve EGSet12 source tracks receives one noise recording and crop, and the same assignment is reused at 20, 10, and 0 dB, so only the noise gain changes. The original stereo format is preserved, and no RIR is applied. Table 16 gives the frozen assignment for every source track; crop starts are sample indices.

Table 16: The EigenScape recording and crop assigned to each EGSet12 track. Each track keeps the same recording and the same crop at all three SNR levels, so only the noise gain changes between conditions. Every crop begins on a whole second; the sample index at 48 kHz is the start time in seconds multiplied by 48,000.

Track	Scene	EigenScape recording	Crop start
01	Train Station	`TrainStation.4.flac`	6:25
02	Busy Street	`BusyStreet.7.flac`	7:35
03	Shopping Centre	`ShoppingCentre.8.flac`	6:55
04	Train Station	`TrainStation.7.flac`	1:05
05	Busy Street	`BusyStreet.1.flac`	4:25
06	Shopping Centre	`ShoppingCentre.1.flac`	9:15
07	Park	`Park.5.flac`	2:15
08	Train Station	`TrainStation.3.flac`	7:00
09	Busy Street	`BusyStreet.8.flac`	5:40
10	Shopping Centre	`ShoppingCentre.3.flac`	3:30
11	Park	`Park.3.flac`	4:20
12	Park	`Park.8.flac`	4:30

Evaluation protocol summary. Every training condition is evaluated with the same protocol. For clean GuitarSet and GuitarSetENV-Test, each fold's final checkpoint scores the held-out player's recordings, with GuitarSetENV-Test scored separately at each SNR level; for EGSet12 and EGSet12+, all six fold checkpoints score the twelve source tracks and their degraded derivatives.

Source access and permissions. GuitarSet, DEMAND, EigenScape, and EGSet12 are publicly available. GuitarSetFX and GuitarProFX were supplied privately by Pedroza et al. and are not publicly redistributed. GuitarProFXENV derives from GuitarProFX and is subject to the same restriction.