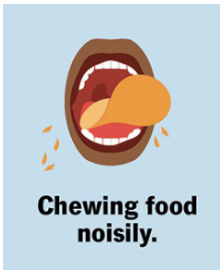
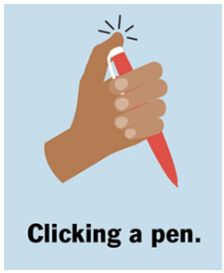
Universiteit Leiden
The Netherlands

Media Technology

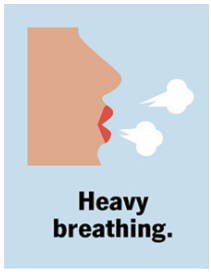


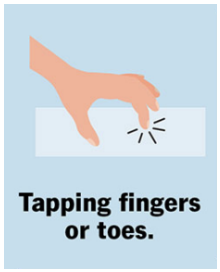
Misophonia Sound Recognition Using PaSST

Charlotte Marosvölgyi



Supervisors:
Rob Saunders & Tessa Verhoef





MASTER THESIS

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

12/09/2025

Abstract

Misophonia is a relatively underexplored condition in which specific everyday sounds, often produced by humans, evoke strong negative emotional reactions such as anger, anxiety, or disgust. Despite increasing interest in this condition, currently offered treatments and the availability of relevant datasets for research are limited. In this study, a transformer-based audio classification approach is explored using the PaSST (Patchout faSt Spectrogram Transformer) model to identify misophonia-related trigger sounds. We explore PaSST's performance and compare it to a Vision Transformer (ViT) baseline. We further extend the dataset with nine misophonia-related classes and analyze performance as audio duration decreases. PaSST reaches higher validation accuracy (97.1%) but slightly higher validation loss (0.224) than ViT (92.29%, 0.1956). Across 59 classes (16 misophonia), PaSST achieves a recall of 91% on five-second clips when testing on the misophonia trigger sounds. Performance decreases with shorter clip lengths. Recall remains above 80% down to 1.25 s, but falls below 80% at 1.0 s and shorter. Based on these findings, we can state that the PaSST model achieves state-of-the-art performance for misophonia trigger detection on the ESC-50 dataset, extended with the nine misophonia categories, offering a strong foundation for future applications in adaptive noise cancellation or other therapeutic support tools for individuals with misophonia.

Contents

1 Introduction

Misophonia ('hatred of sound') is a condition where specific sounds trigger intense emotional reactions in individuals such as irritation, anger, or even rage. Common examples of these sounds include chewing, breathing, sniffing, or tapping. Although first named and described in the early 2000s, misophonia remains under-recognized in clinical settings, and its exact workings and diagnostic criteria continue to be subjects of debate. What is clear, however, is that misophonia can severely impact daily functioning, social interactions, and mental well-being. Some individuals may avoid environments in which misophonic sounds will occur, and others might use coping mechanisms, such as noise-cancelling headphones, to help suppress sound and manage their symptoms. However, there are currently no standardized tools to automatically detect or respond to trigger sounds in real time. Advancements in machine learning and audio classification may help to address these shortcomings. This research builds on recent developments in transformer-based models for audio classification, in particular the Patchout faSt Spectrogram Transformer (PaSST).

The study aims to evaluate how well such models can recognize misophonia-related trigger sounds. First, we use a subset of the ESC-50 dataset focusing on eight sound categories identified in previous literature as common misophonia triggers [BBA23]. Considering the limited availability of existing datasets for misophonia trigger sounds, we also expand this subset with eight new self-recorded categories to create a more comprehensive dataset.

This thesis addresses three key objectives:

1. Exploring whether the PaSST model can match the results of vision transformers in classifying known misophonia trigger sounds.

2. Assessing the model's performance when extending the sound dataset with more misophonia trigger sounds, as well as non-misophonia sounds.

3. Investigating the performance of the PaSST model when trimming the audio clips to shorter durations to simulate real-time sound recognition for potential noise cancellation purposes.

By addressing these objectives, we aim to answer the overarching **research question**:

"Can an audio transformer model like PaSST be used to produce a state-of-the-art misophonia trigger sound detection system?"

By evaluating the PaSST model, expanding the ESC-50 dataset and running experiments on different audio fragment durations, this thesis contributes to the development of potential misophonia support tools. The thesis is structured as follows: Section 2 discusses the background on misophonia and Noise Cancellation; Section 3 discusses Vision Transformers and PaSST; Section 4 covers data and training; Section 5.1 describes the experiments; Section 5.2 reports the results; Section 6 discusses the results of the experiment and the conclusions drawn from them.

2 Background

2.1 Misophonia

Misophonia was first identified and labeled by Jastreboff and Jastreboff in 2001 [JJ01]. The term described a distinct form of sound tolerance that did not fit the criteria of existing conditions on sound sensitivity, such as *hyperacusis* (physical sensitivity to sound) and *phonophobia* (fear response to sounds). As opposed to general hypersensitivity to loudness, misophonia involves strong negative emotional responses such as anger and disgust when exposed to specific sounds (e.g., chewing, breathing, or tapping) [JJ01]. Although the term was introduced in 2001, misophonia did not attract the attention of psychiatric research

until 2013, when Schröder et al. presented a clinical case study to define the condition as a distinct psychiatric condition [SVD13]. Since then, the condition has gained significant research attention, and a consensus definition was not published until recently in 2022 by researchers Swedo et al. [SBD+22]. Despite clinical recognition, misophonia is not yet formally recognized in diagnostic manuals such as DSM-5-TR (Diagnostic and Statistical Manual of Mental Disorders) or ICD-11 (International Classification of Diseases), though clinical demand and growing empirical evidence strive for its inclusion [MHI+23].

Studies suggest that misophonia sufferers experience involuntary emotional and physiological responses that significantly impact their daily functioning, relationships, and mental health. However, treatment options are limited. Approaches such as Cognitive Behavioral Therapy (CBT), sound therapy, and mindfulness-based interventions have been explored with some success, though evidence remains sparse [MDW+23]. Research is ongoing to better understand the condition, its underlying mechanisms, and effective treatment options.

From a neurological perspective, neuroimaging research indicates that misophonia may not come from issues in the auditory cortex itself. Instead, it seems to stem from unusual connectivity between auditory processing and motor systems. Kumar et al. [KDE+21] showed that people with misophonia do not exhibit different responses in the auditory cortex to trigger sounds. However, they demonstrate stronger connections between auditory areas and motor regions used for actions like chewing. These motor areas are especially active when exposed to trigger sounds. This supports a hyper-mirroring model, where harmless sounds like chewing or breathing take on excessive significance due to their strong links to the neural systems that control those actions. This shows, misophonia relates less to sound volume or acoustic features and more to how specific sounds become overly connected to meaning and social context. Savard and Coffey [SC25] expanded on this by proposing a broader cognitive framework. They emphasized that misophonia cannot be simplified to a single abnormal circuit. Instead, it emphasizes the importance of deviating interactions across larger brain networks, including auditory pathways, the salience and attention systems, and executive control networks. They suggest that these networks together assign too much importance to a narrow set of everyday sounds. This framework aligns with the challenge in machine learning, which involves teaching models to differentiate a small group of trigger sounds from a much larger collection of neutral sounds, even when the acoustic differences are slight.

2.2 Noise Cancellation Technology

Noise cancellation technology, particularly active noise cancellation (ANC), plays a vital role in auditory interventions [BFP+25]. ANC uses destructive interference to cancel out sound by generating sound waves of the opposite phase. This technology is implemented in headphones to suppress sounds such as air conditioning or passing traffic. A recent study by Wunrow [Wun24] explored the potential of selective noise cancellation as a treatment approach for misophonia. In this work, a convolutional neural network (CNN) was developed to recognize common trigger sounds and selectively cancel them from audio clips. The system was evaluated in a user study, where participants rated their reactions to original trigger sounds, non-trigger sounds, and selectively cancelled trigger sounds. The findings showed that the participants reported significantly reduced misophonic reactions when exposed to selectively cancelled audio. This shows the promise of this technique as a foundation for future audiological interventions. Although deep learning algorithms such as Convolutional Neural Networks (CNNs) have achieved high accuracy performance, recent studies show that transformer models have surpassed CNNs as the dominant technology in the field of audio classification [BBA23].

3 Related Work

3.1 Vision Transformer

Bahmei et al. [BBA23] conducted one of the first studies to recognize misophonia sounds using a machine-learning-based approach called Vision Transformer (ViT). The transformer, first introduced by Vaswani et al. [Vas17], uses an attention mechanism that selects the most important parts of the input. Although the original transformer consists of an encoder and a decoder, ViT primarily uses the encoder for classification. In the work by Bahmei et al., a 2D spectrogram (derived from audio) is split into fixed-size patches. Then, positional embeddings (encoding the order of patches) and a special classification token are added to the transformer encoder input. The encoder learns global and relative positional information of the patches. The transformer encoder block itself includes three components: the Multi-Head Self-Attention (MSA) to learn dependencies, a Multi-Layer Perceptron (MLP) to capture complex patterns, and Layer Normalization (LN) to improve training time and performance. Finally, a classification head consisting of four fully connected layers with ReLU activations and Softmax (to produce probabilities per class) is used. The Adam optimizer is used for training, and predicted classes are then represented as one-hot encoded vectors. This approach effectively classified misophonia triggers on ESC-50 (+ chewing). Their work demonstrates that transformer architectures can effectively classify trigger sounds, motivating further research into deep learning approaches for misophonia-related audio tasks.

3.2 PaSST

Koutini et al. [KEzW21] propose Patchout faSt Spectrogram Transformers (PaSST) as an efficient and high-performing architecture for audio classification. The model improves upon previous transformer-based approaches like AST (Audio Spectrogram Transformer) by introducing a Patchout mechanism, which randomly removes parts of the transformer's input sequence during training. This encourages the transformer to perform classification while using a sequence that is incomplete. This acts both as a regularizer and as a complexity reducer, allowing state-of-the-art results to be achieved on large datasets such as Audioset [1] while significantly lowering memory usage and training time [KEzW21]. The model has also been shown to generalize performance in tasks such as ESC-50, OpenMIC, and FSD50K [2]. In this thesis, the PaSST model is selected over CNN or traditional transformer models because it is designed specifically for audio training and has shown strong performance on short audio samples [KEzW21]. With Patchout, training time is reduced and the model fits on a single consumer-grade GPU [KEzW21].

3.3 Datasets in Misophonia Research

A significant barrier in the field of misophonia research has been the scarcity of dedicated sound datasets. Researchers often rely on existing datasets that are limited for misophonia trigger sounds and lack coverage of individual trigger sounds reported by sufferers. One recent attempt to address this gap is the Free Open-Access Misophonia Stimuli (FOAMS) database, introduced by Orloff, Benesch and Hansen [OBH23]. They offer a publicly available sound bank consisting of 32 audio examples (8 categories with 4 audio fragments each), complete with pilot discomfort ratings. However, the dataset only covers a fraction of recognized misophonia triggers and relies on user-uploaded recordings from various sources, resulting in variability in audio quality and limited diversity. Another dataset-related study utilized the International Affective Digitized Sounds

[1] https://research.google.com/audioset//index.html
[2] https://github.com/kkoutini/PaSST?tab=readme-ov-file

(IADS-2) library [BL07], which is a dataset widely used in emotion research. Researchers Trumbull et al. [TLM+24] used the dataset to explore affective responses to the audio clips, including trigger sounds such as chewing, sneezing, and paper rustling. Although the study contributes useful information, IADS-2 is not tailored to misophonia and is not fully open-access, which makes it difficult for researchers to use the sounds in their studies. In addition to the audio datasets, Samermit et al. [SYA+22] developed the Sound-Swapped Video (SSV) database, pairing trigger sounds with context-aligned, affect-neutral video sources, to study how visual context modulates emotional responses. Although this research is valuable for psychological experiments, the focus of SSV is not on standardized audio stimulus sets that are usable for machine learning models.

In contrast to these limited resources, this thesis leverages and extends the ESC-50 dataset, a widely used benchmark for environmental sound classification introduced by Piczak [Pic15]. ESC-50 includes 2,000 labeled audio samples evenly distributed in 50 environmental categories (40 examples per class), with a standardized format (five-second `WAV` files (mono), sample rate 44.1 kHz). Several sounds within ESC-50 are relevant for misophonia research (e.g., breathing, coughing, snoring, keyboard typing), making it a valuable foundation.

To build on this, our study introduces an expansion of ESC-50 by adding eight newly recorded misophonia-related classes, such as joint cracking, sniffing, finger tapping, and humming, plus an additional category *chewing*, which was collected from the Kaggle dataset [3]. Each new class consists of 40 recordings in a format consistent with ESC-50 [4]. Hereafter, we refer to the extended dataset (ESC-50 plus nine additional classes) as **ESC-59**, containing 16 misophonia-related classes in total.

4 Methodology

4.1 Research Design

This study uses supervised learning for audio classification using the PaSST model [KEzW21]. The goal is to investigate the effectiveness of this transformer-based architecture in classifying misophonia-related trigger sounds. The study follows a comparative design: It first mirrors the structure of an existing study that used a *Vision Transformer* for audio classification [BBA23], and then extends it with additional misophonia-related categories to evaluate the model's scalability and performance under variable input lengths.

4.2 Dataset

4.2.1 Initial Dataset Selection

The initial dataset used for this study was the **ESC-50 dataset** and can be found on Gitub[5]. The dataset contains 2000 labeled environmental audio recordings of five seconds each across 50 classes (40 recordings per class). Seven categories from ESC-50 were selected that are commonly associated with misophonia trigger sounds (e.g., breathing, coughing, snoring, drinking, mouse clicking, keyboard typing, and clock ticking). Since chewing is one of the most commonly reported misophonia triggers but is not included in ESC-50, we added this category separately using recordings from Ma's 2020 eating sound dataset [MGMVR20]. Ma conducted 11,141 recordings of 20 different food types [MGMVR20]. For our curated dataset, two recordings from each of these food types were extracted to mimic the most variable chewing sounds and to remain consistent with 40 recordings per category. Together, these eight categories (seven from ESC-50 plus chewing from

[3] https://www.kaggle.com/datasets/mashijie/eating-sound-collection

[4] The original ESC-50 dataset consists of recordings of five seconds with a sample rate of 44.1 kHz, however, all recordings are resampled to a sample rate of 32 kHz during the PaSST model training process.

[5] https://github.com/karolpiczak/ESC-50

Ma's dataset) match the set chosen in Bahmei et al. [BBA23].

4.2.2 Extended Dataset

After confirming that the PaSST model produces comparable results to the baseline study, the dataset was extended with eight additional misophonia-relevant categories. All sounds were recorded manually, where again each category contains 40 audio recordings.
The new categories that were chosen are: rustling, utensils, pen clicking, throat clearing, sniffing, joint cracking, finger tapping, and humming. The categories were chosen based on a variety of identified trigger sounds by health sources such as the Cleveland Clinic [Clend] and the Misophonia Institute [Misnd]. The spectrum of potential misophonia triggers is virtually infinite, so these eight categories were merely chosen as a starting point for extending the dataset with potential trigger sounds. They include a mix of oral sounds (throat clearing, sniffing, humming), manual activity sounds (pen clicking, finger tapping, utensils), and body-related sounds (joint cracking, rustling). Other triggers mentioned in these sources, such as environmental trigger sounds (e.g., dripping taps, household appliances), were not chosen to extend the dataset since these can often be eliminated or avoided. In contrast, human-generated sounds, such as pen clicking during a meeting or throat clearing during a lecture, are harder to control.

4.3 Recording Procedure

For each of the eight categories, 40 audio samples of exactly five seconds were recorded, matching the sample size and duration used in the ESC-50 dataset to maintain a balanced class distribution and enable fair comparison with prior work. Since eight new categories were manually added, a total of 320 five-second audio fragments were required. There were several considerations when choosing a recording method:

- **Recording setup** All recordings were captured using the built-in microphone of an iPhone via a dedicated recording application called Voice Record Pro [Day12]. This choice ensured consistency and clear audio quality. Although a headphone microphone was considered, given its relevance to potential future integration with active noise-cancelling headphones, research shows the audio quality of an iPhone microphone is comparable to that of a headset microphone [FMH+23].

- **Environments** To stay consistent with real-life conditions in which misophonia trigger sounds occur and the ESC-50 dataset, recordings were made in both quiet, controlled rooms and mildly noisy public environments such as cafés and trains. Background noise levels were kept below conversational speech to ensure the target sound remained clearly identifiable while introducing some environmental variability.

- **Technical parameters** Recordings were sampled at 32 kHz, 16-bit depth, mono channel, and saved in `WAV` format. These parameters match those of the ESC-50 dataset and pretrained PaSST models, ensuring compatibility and reducing preprocessing complexity.

- **Quality control** All recordings were manually reviewed to confirm that the trigger sound was clearly present and recognizable. Samples with unrelated or obstructive sounds were discarded and re-recorded.

By using human-made trigger sounds that were recorded in various environments and kept close to the way ESC-50 recordings were made, the extended dataset provides a strong basis for testing how well the PaSST model can detect misophonia triggers.

4.4 Audio Preprocessing

All audio samples were standardized to five-second fragments, consistent with the audio fragments of the ESC-50 dataset. For further experiments, we also wanted to decrease the duration of the audio clip to see if the model classifies the sounds equally compared to the five-second fragments.
For this purpose, we have written a program that takes in an audio file and creates a new audio file with a specified duration (i.e., 2 s, 1.5 s, 1.25 s, 1 s, and 0.5 s). The new file consists of a trimmed segment of the original file and contains the moment where the amplitude was maximal. We also tested different positions of this maximum amplitude to see if this would make a difference for the performance of the model. For audio clips of length 1 s up until 1.5 s, two different trimming strategies were applied: positioning the maximum amplitude in the audio clip at 20% of the trimmed audio segment and at 50%. For example, when we take audio clips of 1.5 s and a trimming method of 20%, the maximum amplitude of the trimmed audio file will be at 0.3 s, since $0.3/1.5 = 20\%$. Vice versa, for the 50% trimming method, the maximum amplitude of a 1.5 s audio file would be at 0.75 s. If this is not possible because the maximum amplitude is found at the edge of the original audio clip, the trimmed segment will be positioned at that edge.
After quickly reviewing all new audio segments manually, we still had to delete some audio files where the highest amplitude in the fragment was not recognizable as an actual trigger sound, and re-record this sound. After doing so, the new testing dataset was complete.

4.5 Training Configuration

As mentioned in Section 3, the PaSST model is used due to its strong performance and reduced training time. Moreover, pretrained weight files are available, allowing for initial testing of the setup and providing a baseline for training on the expanded dataset [KEzW21]. The pretrained model that was used for our experiments was the `esc50-passt-s-n-f128-p16-s10-fold1-acc.967.pt`, and can be found on the PaSST github[6].
Further, the training process went as follows. Each class in the ESC-50 and the additional misophonia trigger set contains 40 audio clips of 5 seconds. We applied a stratified 60/40 split per class, where 24 clips were assigned to the training set, and 16 clips were used for the test set.
The 24 training clips were used to fit the PaSST model. During training, the PaSST implementation automatically reserves a portion of the training set for internal validation (used for early stopping and hyperparameter tuning). Thus, the validation set is a random subset of the training data and is not manually defined. The 16 held-out test clips per class were never seen by the model during either training or validation and were used exclusively for final evaluation and reporting of performance.
Training Details:

- Optimizer: AdamW
- Learning Rate: 0.00001
- Loss Function: Cross-entropy
- Epochs: 10
- Batch Size: 12 (training) + 20 (validation) = 32
- Data Augmentation: Mixup (alpha = 0.3) and SpecAugment (time mask = 80, frequency mask = 48)
- Training/Validation/Test Split: Training on 24 samples per category, testing on 16 samples per category of interest
- Training was performed using PyTorch Lightning on an NVIDIA T4 GPU.

4.6 Experimental setup

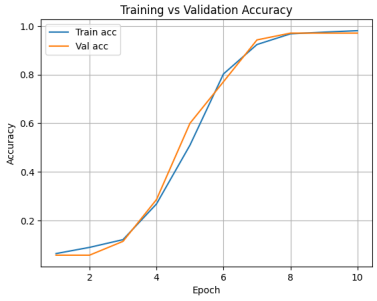
The PaSST model was trained over 10 epochs using a batch size of 32. The PaSST model was

[6] https://github.com/kkoutini/PaSST/releases/tag/v.0.0.6

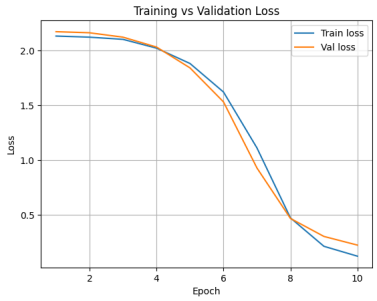
trained on the 24 training clips per class, with an internal validation split automatically generated by the framework. Model checkpoints with the lowest validation loss were retained. After training, all reported results are based on the 16 held-out test clips per class, which provide an unbiased estimate of generalization to unseen data.

4.7 Evaluating the model

To see whether the PaSST model was suited for further experimentation, we first trained the model on the seven chosen misophonia trigger sounds from the original ESC-50 dataset plus the added chewing sounds category, and computed the validation accuracy and validation loss to compare it to the research by Bahmei et al. [BBA23].



(a) Accuracy



(b) Loss

Figure 1: Graphs showing the training and validation accuracy and loss during 10 epochs based on our results.

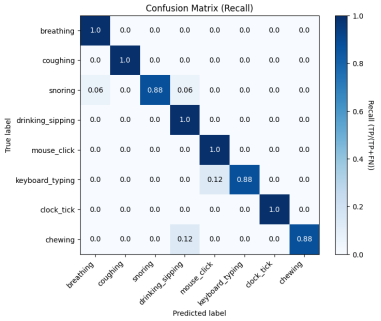
Their model shows a validation accuracy of 92.29% and a validation loss of 0.1956, see Figures 14 and 15 in the appendix. This indicates that their model is capable of accurately classifying trigger sounds. After training our PaSST model, we observed similar results with a validation accuracy that reaches 97.1% and a validation loss that reaches 0.224, see Figure 1.
We also compared the precision, recall, and F1 scores as well as the confusion matrix from the Bahmei paper. Table 1 shows the performance metrics of our PaSST model and the Bahmei et al. paper, respectively. The numbers that are marked show when either the PaSST model or the ViT model outperforms one another.[7] Lastly, the confusion matrices from both the PaSST model and Bahmei et al. can be observed; see Figure 2.

Categories	Precision		Recall		F1-score	
	PaSST	ViT	PaSST	ViT	PaSST	ViT
Breathing	0.94	**1.00**	**1.00**	0.97	0.97	**0.98**
Coughing	**1.00**	0.92	**1.00**	**1.00**	**1.00**	0.95
Snoring	**1.00**	**1.00**	0.88	**0.98**	0.93	**0.99**
Drink sipping	0.84	**1.00**	**1.00**	0.96	0.91	**0.98**
Mouse clicking	0.89	**1.00**	**1.00**	0.97	0.94	**0.98**
Keyboard typing	**1.00**	**1.00**	0.88	**0.98**	0.93	**0.99**
Clock ticking	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Chewing	**1.00**	-	**0.88**	-	**0.93**	-
Average	0.96	**0.99**	0.95	**0.98**	0.95	**0.98**

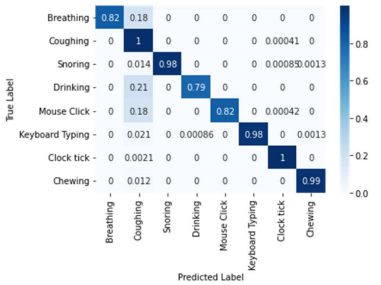
Table 1: Performance metrics for the PaSST and ViT model on **five-second** audio clips, showing precision, recall, and F1-score for each class.

These results show that the performance of our PaSST model is sufficient for further training and running experiments. For further performance evaluation, the PaSST model was trained on the aforementioned ESC-59 dataset. When measuring performance metrics on the test set, we used *recall* as the determining factor, as *precision* would only be accurate if we tested on all categories; a non-misophonia sound classified as a misophonia trigger sound would decrease the precision as the number of false positives (FP) increases, while the *recall* is fixed.

[7] Bahmei et al. [BBA23] did not include the "chewing" category in their reported metrics table, and it remains unclear how the presented metrics were derived from the accompanying confusion matrix, see Figure 2.



(a) PaSST



(b) Vision Transformer

Figure 2: Confusion matrices from both the PaSST model and the Bahmei et al. study.

Furthermore, confusion matrices were used not only to examine *recall* but also to identify whether a misclassified sound was assigned to another misophonia trigger category.
After training on five-second fragments, a series of experiments was conducted to determine the minimum effective audio length required for reasonable classification performance. The trained model was then tested on progressively shortened audio fragments to assess performance degradation at different timescales.

Important note: The confusion matrices are row-normalized and restricted to the 16 misophonia classes; misophonia sounds wrongly classified as non-misophonia sounds (false negatives) are therefore not displayed. For example, out of the 16 test samples for *chewing*, 15 were classified correctly (94%), and one sample was misclassified as a non-misophonia sound. Training and testing exclusively on misophonia trigger sounds would reveal exactly how each sound is classified, as in the study by Bahmei et al. [BBA23]. However, we chose not to omit the other categories, since the intended application of our model should function in real-life settings where various sounds are continuously present. Even with all categories included, our model achieved similar recall, precision, and F1-scores, as well as comparable validation accuracy and loss.

5 Experiments & Results

In this section, the conducted experiments and the results of the experiments will be discussed.

5.1 Experiments

The experiments that were conducted to evaluate the effectiveness of the PaSST model on different lengths of audio clips were done in the following manner. First, a baseline was established with the audio clips of five seconds in duration. After this, the model was run on the processed audio clips of 2 s, 1.5 s, 1.25 s, 1 s, and 0.5 s. Finally, confusion matrices and reports containing precision, recall, and F1-scores were made for the performance comparison of the models.

5.2 Results

In this section, the results of the experiments will be presented. The results will then be interpreted in Section 6. For every combination of audio clip duration and trimming method that the model was tested on, a table with the metrics precision, recall, and F1-score was made, as well as the corresponding confusion matrix.

5.2.1 Baseline results

The results from the five-second audio clips are used as a baseline for comparing the results of the shortened audio clips. Table 2 shows the precision, recall, and F1-scores for each category. Figure 3 shows the associated confusion matrix. Next, the results of two-second audio clips are

shown, as they were the shortest audio clips that still had similar results to the five-second audio clips. Table 3 shows the recall, precision, and F1-scores for the two-second audio segments along with their confusion matrix in Figure 4.

Categories	Precision	Recall	F1-score
Breathing	0.92	0.75	0.83
Coughing	1.00	0.88	0.93
Snoring	1.00	0.94	0.97
Drink sipping	1.00	0.75	0.86
Mouse clicking	0.93	0.88	0.90
Keyboard typing	1.00	1.00	1.00
Clock ticking	1.00	0.94	0.97
Chewing	0.89	1.00	0.94
Rustling	1.00	1.00	1.00
Utensils	1.00	1.00	1.00
Pen clicking	0.85	0.69	0.76
Throat clearing	1.00	0.94	0.97
Sniffing	0.75	0.94	0.83
Joint cracking	0.80	1.00	0.89
Finger tapping	0.94	0.94	0.94
Humming	1.00	1.00	1.00
Average	0.94	0.91	0.92

Table 2: Performance metrics for the model on **five-second** audio clips, showing precision, recall, and F1-score for each class.

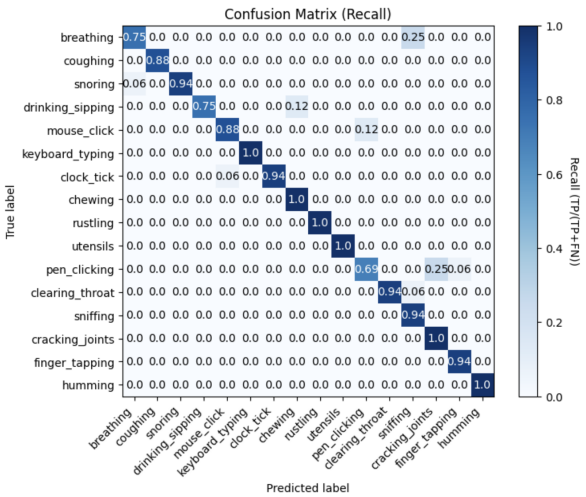


Figure 3: Confusion matrix with results from the **five-second** audio clips, corresponding to the report shown in Table 2.

Categories	Precision	Recall	F1-score
Breathing	0.88	0.94	0.91
Coughing	1.00	0.88	0.93
Snoring	1.00	0.94	0.97
Drink sipping	0.87	0.81	0.84
Mouse clicking	0.87	0.81	0.84
Keyboard typing	1.00	0.81	0.90
Clock ticking	1.00	0.56	0.72
Chewing	0.88	0.94	0.91
Rustling	0.94	1.00	0.97
Utensils	1.00	1.00	1.00
Pen clicking	0.91	0.62	0.74
Throat clearing	1.00	0.94	0.97
Sniffing	0.93	0.88	0.90
Joint cracking	0.89	1.00	0.94
Finger tapping	0.88	0.94	0.91
Humming	1.00	0.94	0.97
Average	0.94	0.88	0.90

Table 3: Performance metrics for the model on **two-second** audio clips, showing precision, recall, and F1-score for each class.

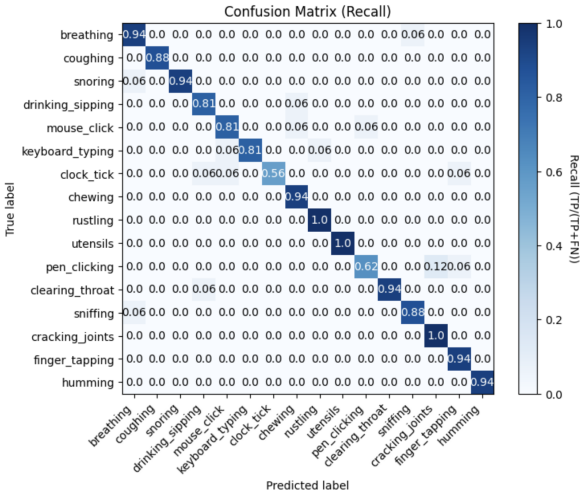


Figure 4: Confusion matrix with results from the **two-second** audio clips, corresponding to the report shown in Table 3.

5.2.2 Shortened audio clip results

Here the average precision, recall, and F1-score results are shown for all audio clips. For audio clips of length 1 s up until 1.5 s, two trimming methods were tested, namely 20% and 50%, to

see if they would yield different outcomes. The results in Table 4 show that the recall remained above 80% for both trimming methods used at a length of 1.5 s and 1.25 s.

Duration	Trim	Precision	Recall	F1-score
5s	20%	0.94	0.91	0.92
2s	20%	0.94	0.88	0.90
1.5s	50%	0.92	0.84	0.87
1.5s	20%	0.92	0.83	0.86
1.25s	50%	0.91	0.83	0.86
1.25s	20%	0.91	0.80	0.84
1s	50%	0.92	0.77	0.83
1s	20%	0.91	0.74	0.80
0.5s	20%	0.87	0.48	0.55

Table 4: Average precision, recall, and F1-score per duration and trimming method.

Neither trimming method achieved this at 1 s. For all durations in which different trimming methods were tested, the performance was higher when the loudest point in the audio was placed in the center of the trimmed segment, so at 50% trimmed, as shown in Table 4. Since the recall scores of audio clips with a duration of 1 s (both trimming methods) are already below 80%, we did not try other trimming methods for the 0.5 s audio clips. Lastly, Table 4 shows that for audio clips with a duration of 0.5 s, the average precision, recall, and F1-scores significantly decrease. More detailed results of the precision, recall, and F1-scores, as well as the corresponding confusion matrices of all audio clips, can be found in the appendix, Section 10.

6 Discussion

In this section, the results of Section 5.2 are interpreted and the limitations of the study are discussed. As stated in the introduction of Section 5.1, we will use the *recall* score as a performance metric to evaluate the shortest audio clips that still perform well. In this study, we consider a recall score of $\geq$ 80% to be sufficient. From Table 2 and Table 3, we can see the average recall of 91% and 88%, respectively. This indicates that our trained model still reaches our desired recall score for the audio clips of 5 s and 2 s.

When the recall score is high (e.g. ≥ 0.8), it suggests that the audio fragments in the category are classified correctly in most cases. In contrast, recall scores below 0.8 may require examination to understand the underlying causes. For example, looking at Figure 4, we can see a 56% recall score for the category *clock ticking* and 62% recall score for the category *pen clicking*. This indicates that only 59% $((0.56 + 0.62)/2)$ of the audio segments labeled as *clock ticking* or *pen clicking* were correctly classified. Although these two categories were misclassified 41% of the time, they were still assigned to another, similar-sounding misophonia trigger category (e.g., *mouse clicking*, *joint cracking*, or *finger tapping*) 19% of the time. The remaining 22% of the time was another non-misophonia trigger sound in the ESC-50 dataset. Furthermore, it is important to note that if we were to calculate recall by grouping all categories into just two classes: "misophonia" and "non-misophonia", the resulting recall would be much higher ($59\% + 19\% = 78\%$ in the example above for *clock ticking/pen clicking*). However, this value would be slightly biased since more than a quarter of the entire dataset of the current categories falls under misophonia triggers.

Table 4 shows that audio samples with a duration of 1.5 s and 1.25 s still reach the desired average recall score of 80%. However, as shown in Figures 7, 8, 9, and 10, some categories start to be misclassified more often. We have already mentioned that *pen clicking* and *clock ticking* have a relatively low recall, which has continued in the tests for these shorter audio segments. Furthermore, the recall scores of *chewing*, *keyboard typing*, *drink sipping*, and *finger tapping* have also fallen below 80%.

Both trimming methods for the one-second audio clips have not reached our required minimum for recall, as the average score is 0.77 when trimmed at 50% and 0.74 when trimmed at 20%, as shown

in Table 4. Lastly, we have also shown at what rate the average recall further drops when testing on 0.5-second audio segments.

Categories	→	**Misclassification**
Snoring Sniffing	→	Breathing
Keyboard typing Joint cracking Pen clicking Finger tapping Clock ticking	→	Mouse clicking
Chewing	→	Drink sipping
Rustling	→	Chewing

Table 5: Categories misclassified as other categories, both misophonia trigger sounds.

Tables 12, 13, and 14 show that not all categories perform weakly, even when given a duration of 0.5 seconds. For example, *breathing*, *snoring*, and *mouse clicking* still have a recall score above 80%. The corresponding confusion matrices shown in Figures 11, 12, and 13 give even more insights; not only the recall scores, but also the categories that perform poorly due to similarities with other sounds, which are summarized in Table 5. For example, Table 5, shows that many short and snappy sounds, such as *joint cracking* or *pen clicking*, are frequently misclassified as *mouse clicking*, which clarifies the relatively low *precision* score for *mouse clicking* in the reports.

Additionally, several limitations of this study should be acknowledged. First, the extended dataset of eight newly recorded categories provides a valuable starting point. However, the number of recordings remains relatively small, which limits the model's exposure to other misophonia trigger sounds and therefore the generalizability of the model as well.

Second, the PaSST model used in this study relied on pretrained weights and was not specifically fine-tuned for the misophonia classification task. While this approach already produced sufficiently strong results, more rigorous fine-tuning could potentially improve performance further, particularly on the shorter audio segments where accuracy declined.

Lastly, the comparison between PaSST and the ViT results should be interpreted with caution. The ViT model reported in earlier work was trained exclusively on misophonia-related sounds, whereas PaSST in this study was trained on both misophonia and non-misophonia classes. This difference in training scope makes the comparison less straightforward, and future studies should aim to evaluate the models under more consistent conditions.

7 Conclusion

This thesis aimed to investigate whether an audio classification model, PaSST (Patchout faSt Spectrogram Transformer), can be effectively applied to misophonia trigger sound recognition. By addressing three key objectives: 1) benchmarking the PaSST model against the ViT model, 2) assessing the PaSST model with the extended ESC-59 dataset, and 3) running experiments on trimmed audio fragments, we aim to answer the research question: *Can an audio transformer model like PaSST be used to produce a state-of-the-art misophonia trigger sound detection system?*

1. We first replicated the setup of previous research using a Vision Transformer (ViT) on a subset of the ESC-50 dataset and found that PaSST reached performance metrics similar to the ViT ((validation accuracy, validation loss) = $(97.1\%, 0.224)$ for PaSSt, the ViT achieved $(92.29\%, 0.1956)$).

2. Next, we extended this dataset by adding eight self-recorded sound categories plus one online-collected category. This yielded a total of 59 classes, 16 of which are related to misophonia. The training results on the

extended dataset showed that the PaSST model still reaches a 91% recall score on five-second audio clips when tested on misophonia trigger sounds, which confirmed the model's robustness.

3. Additionally, we explored the model's performance on shorter audio durations. Although classification performance declined with shorter clips, the model remained sufficient (recall above 80%) down to 1.25-second clips, indicating that the model is suitable for live misophonia trigger detection.

Overall, these results demonstrate that PaSST achieves state-of-the-art performance for misophonia trigger sound recognition, as evaluated on our extended ESC-59 dataset. Even though the model is not specifically fine-tuned for misophonia, its performance still matches ViT-based approaches under the same experimental conditions. In summary, this work demonstrates that PaSST can serve as a strong foundation for building a misophonia sound recognition system, with state-of-the-art performance in audio classification tasks.

8 Future Work

Although this study shows the potential of PaSST for detecting misophonia triggers, several areas still need further exploration. Future work could look at how well the model works in a live detection system. This could be applied in wearable devices, such as noise-cancelling headphones, that would suppress only misophonia trigger sounds, following the work by [Wun24]. To accomplish this, the model must be able to quickly recognize sounds, especially short, high-frequency misophonia trigger sounds. Moreover, the development of targeted noise-cancellation methods is necessary as ANC systems do not yet effectively suppress specific high-frequency sounds. As people differ in their sensitivity to specific triggers, personalization could also improve performance; for example, using an adaptive mobile application that asks for user feedback could improve detection precision.

Furthermore, future studies should test how well the model performs in different recording devices and environmental conditions to improve generalizability. Expanding the dataset with a wider range of both misophonia sounds and non-misophonia sounds would also contribute to model performance and inclusion. In general, future work is essential to apply the experimental models to practical tools that can assist misophonia sufferers in their daily lives.

9 Acknowledgments

I would like to express my gratitude to my supervisors, **Rob Saunders** and **Tessa Verhoef**, for their guidance, encouragement, and constructive feedback throughout this thesis. Your expertise has been invaluable throughout this work. I would also like to thank my **family** for their support, love, and belief in me, and my **friends** for their kindness, good humor, and reality checks during the most challenging moments. Thank you all for standing by me and keeping me motivated. Finally, I would like to thank the **Cleveland Clinic** for providing the misophonia graphics, which I used on the front page of this thesis.[8]

[8] https://my.clevelandclinic.org/health/diseases/24460-misophonia

References

[BBA23] Behnaz Bahmei, Elina Birmingham, and Siamak Arzanpour. Misophonia sound recognition using vision transformer. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[BFP+25] Dinesh Bhatia, Felix V Francis, Pankaj Panging, Banrisuk Khyllait, et al. Revolutionizing noise management: Active noise cancellation headphones in healthcare and beyond. *Journal of Biomedical Science and Engineering*, 17(12):257–272, 2025.

[BL07] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual. Technical report, Technical report B-3. University of Florida, Gainesville, Fl, 2007.

[Clend] Cleveland Clinic. Misophonia: Symptoms & treatment. https://my.clevelandclinic.org/health/diseases/24460-misophonia, n.d. Accessed: 2025-08-09.

[Day12] Dayana Networks Ltd. Voice Record Pro. https://apps.apple.com/us/app/voice-record-pro/id546983235, 2012. iOS app.

[FMH+23] Kazuya Fukazawa, Hiroshi Muto, Haruka Hashimoto, Akihiro Takeuchi, Kentaro Shibuya, Haruki Sato, and Kazuhiko Mori. Smartphone microphones are suitable for use in acoustic analysis of speech: A comparative study with professional microphones. *Journal of Voice*, 37(6):920–927, 2023.

[JJ01] Margaret M Jastreboff and Pawel J Jastreboff. Components of decreased sound tolerance: hyperacusis, misophonia, phonophobia. *ITHS News Lett*, 2(5-7):1–5, 2001.

[KDE+21] Sukhbinder Kumar, Pradeep Dheerendra, Mercede Erfanian, Ester Benzaquén, William Sedley, Phillip E Gander, Meher Lad, Doris E Bamiou, and Timothy D Griffiths. The motor basis for misophonia. *Journal of neuroscience*, 41(26):5762–5770, 2021.

[KEzW21] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. Passt: Efficient training of audio transformers with patchout. https://github.com/kkoutini/PaSST, 2021. [Online; accessed 7-August-2025].

[MDW+23] Seth A Mattson, Johann D'Souza, Katharine D Wojcik, Andrew G Guzick, Wayne K Goodman, and Eric A Storch. A systematic review of treatments for misophonia. *Personalized medicine in psychiatry*, 39:100104, 2023.

[MGMVR20] Jeannette Shijie Ma, Marcello A Gómez Maureira, and Jan N Van Rijn. Eating sound dataset for 20 food types and sound classification using convolutional neural networks. In *Companion publication of the 2020 international conference on multimodal interaction*, pages 348–351, 2020.

[MHI+23] Anne Möllmann, Nina Heinrichs, Lisa Illies, Nadine Potthast, and Hanna Kley. The central role of symptom severity and associated characteristics for functional impairment in misophonia. *Frontiers in Psychiatry*, 14:1112472, 2023.

[Misnd] Misophonia Institute. Misophonia triggers. https://misophoniainstitute.org/misophonia-triggers/, n.d. Accessed: 2025-08-09.

[OBH23] Dean M Orloff, Danielle Benesch, and Heather A Hansen. Curation of foams: a free open-access misophonia stimuli database. *Journal of Open Psychology Data*, 11(1), 2023.

[Pic15] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[SBD+22] SE Swedo, DM Baguley, D Denys, LJ Dixon, M Erfanian, A Fioretti, et al. Consensus definition of misophonia: a delphi study. front neurosci. 2022; 16: 841816, 2022.

[SC25] Marie-Anick Savard and Emily BJ Coffey. Toward cognitive models of misophonia. *Hearing Research*, page 109184, 2025.

[SVD13] Arjan Schröder, Nienke Vulink, and Damiaan Denys. Misophonia: diagnostic criteria for a new psychiatric disorder. *PloS one*, 8(1):e54706, 2013.

[SYA+22] Patrawat Samermit, Michael Young, Allison K Allen, Hannah Trillo, Sandhya Shankar, Abigail Klein, Chris Kay, Ghazaleh Mahzouni, Veda Reddy, Veronica Hamilton, et al. Development and evaluation of a sound-swapped video database for misophonia. *Frontiers in psychology*, 13:890829, 2022.

[TLM+24] Jacqueline Trumbull, Noah Lanier, Katherine McMahon, Rachel Guetta, and M Zachary Rosenthal. Using a standardized sound set to help characterize misophonia: The international affective digitized sounds. *Plos one*, 19(5):e0301105, 2024.

[Vas17] Ashish Vaswani. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017.

[Wun24] Timothy Wunrow. *Selective noise cancelling application for misophonia treatment*. Mississippi State University, 2024.

10 Appendix

10.1 Reports

Categories	Precision	Recall	F1-score
Breathing	0.92	0.75	0.83
Coughing	1.00	0.88	0.93
Snoring	1.00	0.94	0.97
Drink sipping	1.00	0.75	0.86
Mouse clicking	0.93	0.88	0.90
Keyboard typing	1.00	1.00	1.00
Clock ticking	1.00	0.94	0.97
Chewing	0.89	1.00	0.94
Rustling	1.00	1.00	1.00
Utensils	1.00	1.00	1.00
Pen Clicking	0.85	0.69	0.76
Throat clearing	1.00	0.94	0.97
Sniffing	0.75	0.94	0.83
Joint cracking	0.80	1.00	0.89
Finger tapping	0.94	0.94	0.94
Humming	1.00	1.00	1.00
Average	0.94	0.91	0.92

Table 6: Performance metrics for the model on **five-second** audio clips, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.88	0.94	0.91
Coughing	1.00	0.88	0.93
Snoring	1.00	0.94	0.97
Drink sipping	0.87	0.81	0.84
Mouse clicking	0.87	0.81	0.84
Keyboard typing	1.00	0.81	0.90
Clock ticking	1.00	0.56	0.72
Chewing	0.88	0.94	0.91
Rustling	0.94	1.00	0.97
Utensils	1.00	1.00	1.00
Pen clicking	0.91	0.62	0.74
Throat clearing	1.00	0.94	0.97
Sniffing	0.93	0.88	0.90
Joint cracking	0.89	1.00	0.94
Finger tapping	0.88	0.94	0.91
Humming	1.00	0.94	0.97
Average	0.94	0.88	0.90

Table 7: Performance metrics for the model on **two-second** audio clips, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.88	0.94	0.91
Coughing	1.00	0.88	0.93
Snoring	0.93	0.88	0.90
Drink sipping	0.67	0.75	0.71
Mouse clicking	0.81	0.81	0.81
Keyboard typing	1.00	0.81	0.90
Clock ticking	1.00	0.50	0.67
Chewing	0.80	0.75	0.77
Rustling	1.00	0.94	0.97
Utensils	0.94	1.00	0.97
Pen Clicking	0.92	0.75	0.83
Throat clearing	0.93	0.88	0.90
Sniffing	0.93	0.88	0.90
Joint cracking	1.00	1.00	1.00
Finger tapping	0.93	0.81	0.87
Humming	1.00	0.81	0.90
Average	0.92	0.84	0.87

Table 8: Performance metrics for the model on **1.5-second** audio clips **(50% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.88	0.94	0.91
Coughing	1.00	0.88	0.93
Snoring	1.00	0.94	0.97
Drink sipping	0.76	0.81	0.79
Mouse clicking	0.72	0.81	0.76
Keyboard typing	1.00	0.75	0.86
Clock ticking	1.00	0.50	0.67
Chewing	0.80	0.75	0.77
Rustling	1.00	0.94	0.97
Utensils	0.93	0.88	0.90
Pen Clicking	0.91	0.62	0.74
Throat clearing	0.93	0.88	0.90
Sniffing	0.93	0.88	0.90
Joint cracking	0.94	1.00	0.97
Finger tapping	0.87	0.81	0.84
Humming	1.00	0.88	0.93
Average	0.92	0.83	0.86

Table 9: Performance metrics for the model on **1.5-second** audio clips **(20% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.94	0.94	0.94
Coughing	1.00	0.88	0.93
Snoring	0.88	0.88	0.88
Drink sipping	0.75	0.75	0.75
Mouse clicking	0.67	0.88	0.76
Keyboard typing	1.00	0.75	0.86
Clock ticking	1.00	0.56	0.72
Chewing	0.80	0.75	0.77
Rustling	0.94	0.94	0.94
Utensils	1.00	1.00	1.00
Pen Clicking	0.90	0.56	0.69
Throat clearing	1.00	0.94	0.97
Sniffing	0.93	0.88	0.90
Joint cracking	0.89	1.00	0.94
Finger tapping	0.92	0.69	0.79
Humming	1.00	0.88	0.93
Average	0.91	0.83	0.86

Table 10: Performance metrics for the model on **1.25-second** audio clips **(50% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.83	0.94	0.88
Coughing	0.93	0.88	0.90
Snoring	1.00	0.88	0.93
Drink sipping	0.67	0.75	0.71
Mouse clicking	0.70	0.88	0.78
Keyboard typing	1.00	0.69	0.81
Clock ticking	1.00	0.44	0.61
Chewing	0.73	0.69	0.71
Rustling	1.00	0.88	0.93
Utensils	1.00	0.88	0.93
Pen Clicking	0.91	0.62	0.74
Throat clearing	1.00	0.81	0.90
Sniffing	1.00	0.88	0.93
Joint cracking	0.94	1.00	0.97
Finger tapping	0.86	0.75	0.80
Humming	1.00	0.88	0.93
Average	0.91	0.80	0.84

Table 11: Performance metrics for the model on **1.25-second** audio clips **(20% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.88	0.94	0.91
Coughing	1.00	0.81	0.90
Snoring	0.93	0.88	0.90
Drink sipping	0.85	0.69	0.76
Mouse clicking	0.67	0.88	0.76
Keyboard typing	1.00	0.56	0.72
Clock ticking	1.00	0.38	0.55
Chewing	0.67	0.62	0.65
Rustling	1.00	0.75	0.86
Utensils	1.00	0.94	0.97
Pen Clicking	0.92	0.75	0.83
Throat clearing	0.94	0.94	0.94
Sniffing	1.00	0.88	0.93
Joint cracking	1.00	1.00	1.00
Finger tapping	0.89	0.50	0.64
Humming	1.00	0.88	0.93
Average	0.92	0.77	0.83

Table 12: Performance metrics for the model on **one-second** audio clips **(50% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.80	1.00	0.89
Coughing	0.93	0.81	0.87
Snoring	1.00	0.81	0.90
Drink sipping	0.80	0.75	0.77
Mouse clicking	0.68	0.81	0.74
Keyboard typing	0.90	0.56	0.69
Clock ticking	1.00	0.44	0.61
Chewing	0.67	0.62	0.65
Rustling	1.00	0.75	0.86
Utensils	1.00	0.81	0.90
Pen Clicking	0.90	0.56	0.69
Throat clearing	0.81	0.81	0.81
Sniffing	1.00	0.69	0.81
Joint cracking	1.00	1.00	1.00
Finger tapping	1.00	0.62	0.77
Humming	1.00	0.81	0.90
Average	0.91	0.74	0.80

Table 13: Performance metrics for the model on **one-second** audio clips **(20% trim)**, showing precision, recall, and F1-score for each class.

Categories	Precision	Recall	F1-score
Breathing	0.71	0.94	0.81
Coughing	0.92	0.69	0.79
Snoring	0.87	0.81	0.84
Drink sipping	0.50	0.50	0.50
Mouse clicking	0.46	0.81	0.59
Keyboard typing	1.00	0.19	0.32
Clock ticking	1.00	0.19	0.32
Chewing	1.00	0.06	0.12
Rustling	1.00	0.31	0.48
Utensils	1.00	0.69	0.81
Pen Clicking	0.67	0.38	0.48
Throat clearing	0.82	0.56	0.67
Sniffing	1.00	0.50	0.67
Joint cracking	1.00	0.38	0.55
Finger tapping	1.00	0.12	0.22
Humming	1.00	0.56	0.72
Average	0.87	0.48	0.55

Table 14: Performance metrics for the model on **0.5-second** audio clips, showing precision, recall, and F1-score for each class.

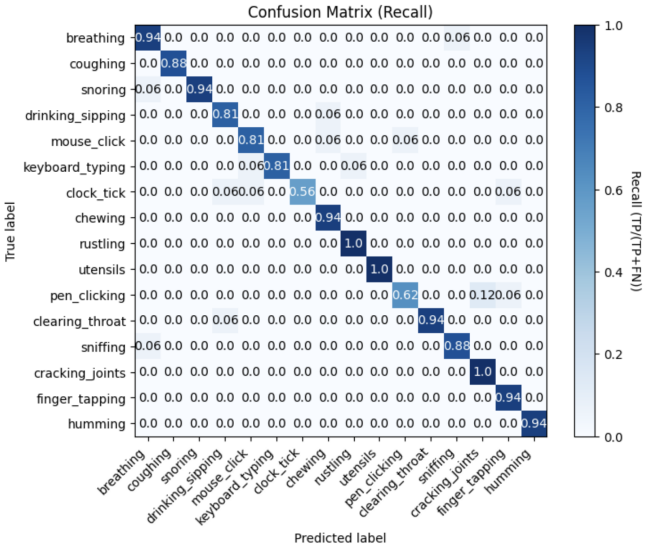


Figure 6: Confusion matrix with results from the **two-second** audio clips, corresponding to the report shown in Table 7.

10.2 Confusion matrices

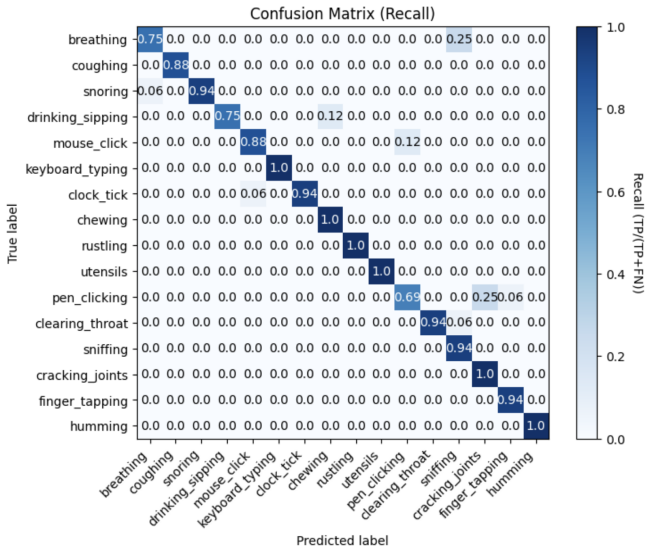


Figure 5: Confusion matrix with results from the **five-second** audio clips, corresponding to the report shown in Table 6.

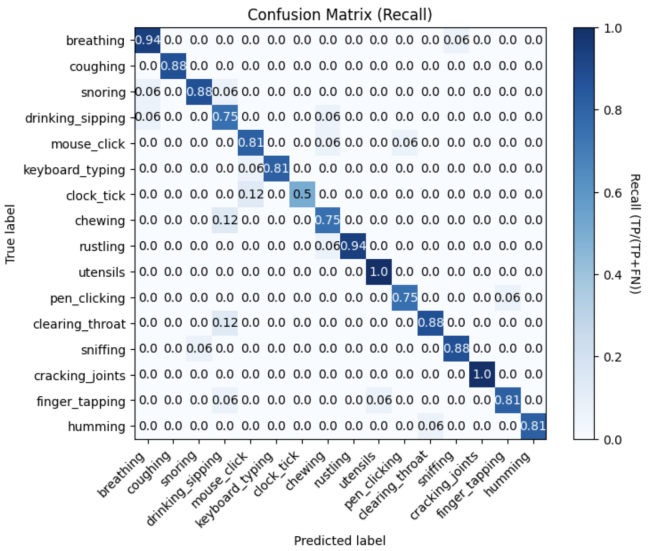


Figure 7: Confusion matrix with results from the **1.5-second** audio clips **(50% trim)**, corresponding to the report shown in Table 8.

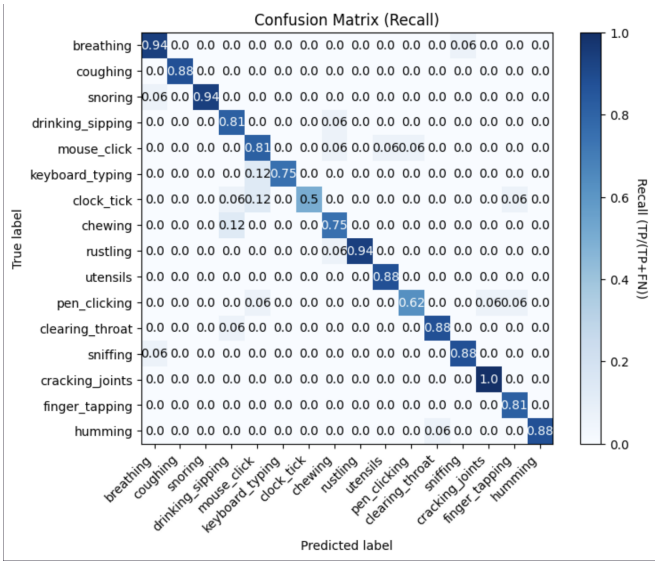


Figure 8: Confusion matrix with results from the **1.5-second** audio clips **(20% trim)**, corresponding to the report shown in Table 9.

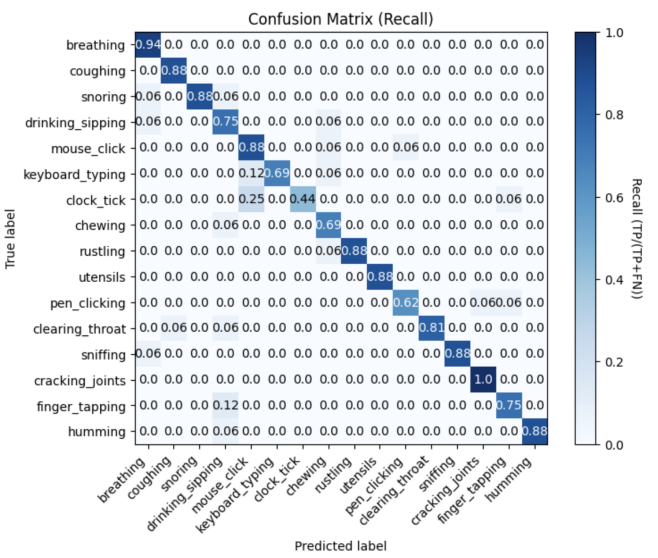


Figure 10: Confusion matrix with results from the **1.25-second** audio clips **(20% trim)**, corresponding to the report shown in Table 11.

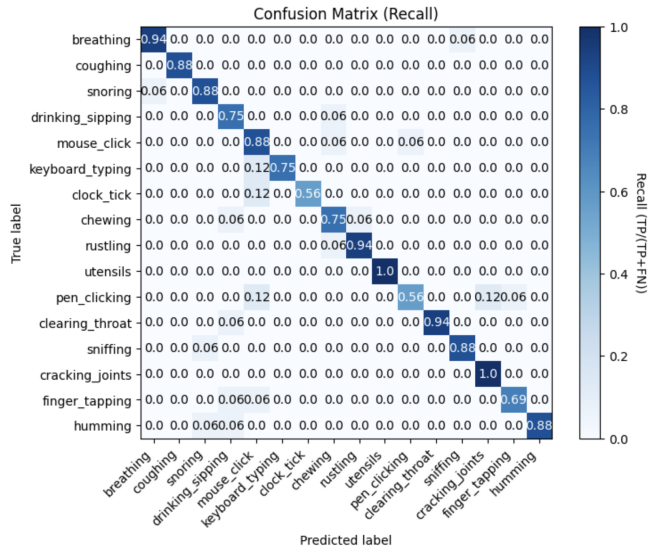


Figure 9: Confusion matrix with results from the **1.25-second** audio clips **(50% trim)**, corresponding to the report shown in Table 10.

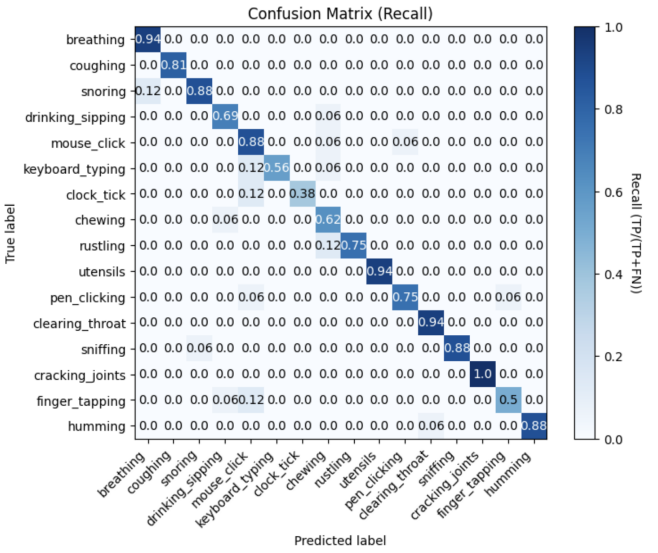


Figure 11: Confusion matrix with results from the **one-second** audio clips **(50% trim)**, corresponding to the report shown in Table 12.

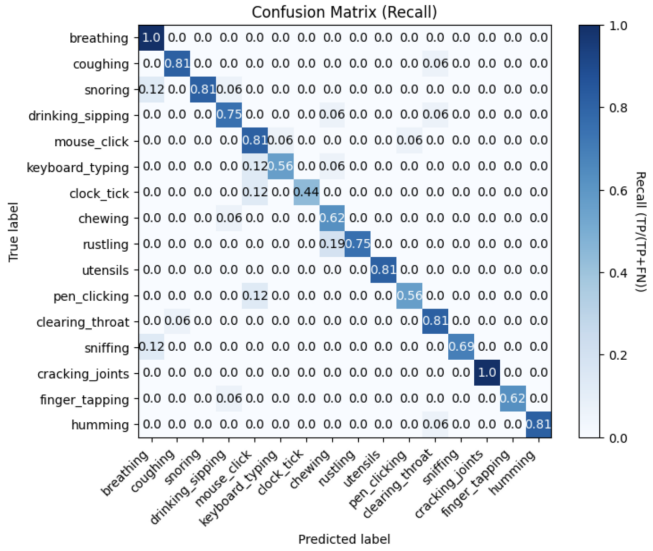


Figure 12: Confusion matrix with results from the **one-second** audio clips **(20% trim)**, corresponding to the report shown in Table 13.

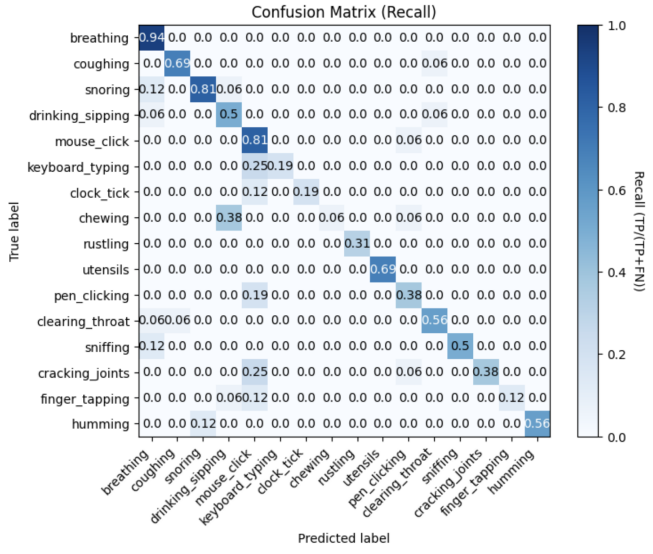


Figure 13: Confusion matrix with results from the **0.5-second** audio clips, corresponding to the report shown in Table 14.

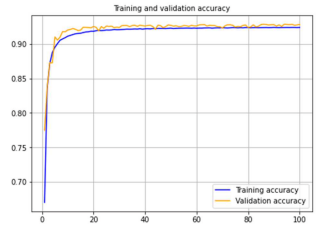


Figure 14: Graph shown in a study by Bahmei, presenting the training and validation accuracy during 100 epochs [BBA23].

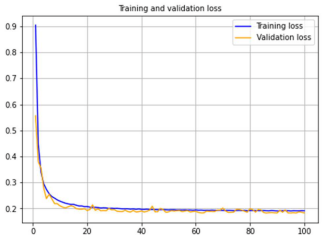


Figure 15: Graph shown in a study by Bahmei, presenting the training and validation loss during 100 epochs [BBA23].

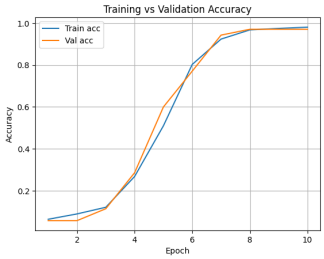


Figure 16: Graph showing the training and validation accuracy during 10 epochs based on our results.

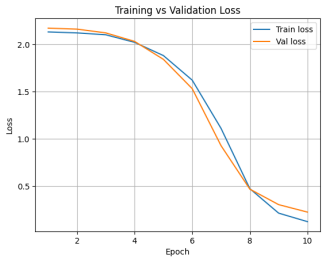


Figure 17: Graph showing the training and validation loss during 10 epochs based on our results.