Aircraft-Induced Psychoacoustic Annoyance Quantification Using Artificial Intelligence

by

Vlad Ştefan Buzeţelu
Delft University of Technology, Faculty of Aerospace Engineering, Delft, Zuid-Holland, 2629 HS, the Netherlands

to obtain the degree of Master's of Science
at Delft University of Technology,
to be defended publicly on Tuesday March 18th, 2025 at 14:00 PM.

Student number:	5040604	
Project duration:	April 15, 2024 – February 1, 2025	
Thesis committee:	Dr. R. Merino-Martínez,	TU Delft, daily supervisor
	Dr. M. Hoogreef,	TU Delft, chairman of thesis evaluation committee
	Dr. M. Ribeiro,	TU Delft, external member of thesis evaluation committee

TU Delft

Acknowledgements

My journey at the Faculty of Aerospace Engineering of TU Delft goes way back, since the fall of 2019, when I started my BSc degree. That was a difficult, yet extremely fulfilling endeavor. It equipped me with the necessary skills, discipline and passion for the second part of my Aerospace Engineering journey, which started in 2022: the MSc degree program in the Flight Performance & Propulsion track. This second half certainly had its own ups and downs, but the challenges made the rewards even more fruitful. The various courses I took during this time demanded constant effort, but once the work was put in, I quickly realised that I came out of each quarter one step closer to being the well-rounded engineer I always dreamed to become. My curiosity was always driven by the desire to know more about aviation and the multitude of fields it encompasses. After all, the Master's program is a great opportunity to explore many topics of interest - and the Faculty of Aerospace Engineering offers you just that.

This brings me to the final part of my journey: the Master's thesis. During the past 10 months, I have had the pleasure of working within the ANCE department, applying the skills gained throughout my years of studying, and using them to gain new knowledge in fields in which I had little to no experience before: aircraft noise and psychoacoustics. Thus, I would like to thank my supervisor, Dr. Roberto Merino-Martínez, for the constant support and guidance offered throughout my research. I would like to put emphasis on the freedom he has offered me in making my own decisions in terms of the direction in which to take the research. This has been an invaluable experience for me, which I am sure will prove extremely useful in my career from now on. During our meetings, the sharing of thoughts and ideas was always inspiring and it motivated me to push myself harder in order to innovate. The collaboration was made even more enjoyable by Roberto's positive and friendly personality, further motivating me to work with great pleasure.

Moreover, I would like to thank all of my friends and peers with whom I have shared all the wonderful experiences during these 5 and a half years spent in Delft. Special thanks also ought to go to my good friend, Sergiu Priboi, with whom I have shared most of these moments. Having worked on closely related topics for our theses (both within ANCE, under Dr. Roberto Merino-Martínez's supervision), it was a huge help to have yet another person to critically exchange ideas with. I would like to express my gratitude and love towards my family, who have always supported me through the rougher patches, but have also celebrated my achievements with the sincerest joy.

I am hopeful that the future will bring me on many other interesting and rewarding journeys, just as this one has been.

Vlad Ştefan Buzeţelu
Delft, January 2025

Contents

Aircraft-Induced Psychoacoustic Annoyance Quantification Using Artificial Intelligence

Vlad Ştefan Buzeţelu
Delft University of Technology, Faculty of Aerospace Engineering, Delft, Zuid-Holland, 2629 HS, the Netherlands

With the continuous growth of the aviation sector, concerns regarding the effects of aircraft noise on the health and well-being of communities living in the vicinity of airports have been increasing. Aircraft noise annoyance is inherently subjective and its accurate prediction and quantification represent challenging tasks. There is a lack of consensus in the scientific community regarding which metrics are the best predictors for this type of annoyance. Additionally, many of the metrics employed in the field of (psycho)acoustics are typically computationally expensive. This study aims at developing a methodology which leverages machine learning techniques for instant predictions of various sound metrics and for annoyance rating predictions from input aircraft flyover recordings. The two-step framework involves a Convolutional Neural Network (CNN) for the former, followed by artificial intelligence (AI) models which use the CNN predictions as input for the latter, such as the Support Vector Machine and Random Forest. A listening experiment was conducted in order to gather labeled annoyance data from 60 aircraft flyover recordings of both landings and take-offs and a correlation analysis was subsequently made considering a large pool of sound metrics. The results show that, in general, metrics derived from Psychoacoustic Annoyance models, especially those of Zwicker and Di et al., present better performance as predictors compared to conventional metrics and most Sound Quality Metrics taken individually. Moreover, the AI framework achieves very promising results for both the annoyance ratings and metric predictions (overall, mean absolute errors of the annoyance ratings of approximately 0.4 and below, and R^2 values above 0.85), highlighting the potential for bypassing the typically long overhead associated with computing SQMs (which involve expensive and complex algorithms). The drawn conclusions are dependent on the limited amount of data that was available. Hence, the results require further validation using more recordings of various types of aircraft.

I. Introduction

Aircraft noise is the main source of annoyance for communities living in the vicinity of airports and it has seen an increasing trend in the past decades, which is in line with the prospects for the continuous growth of aviation [1, 2]. Hence, people are more annoyed by aircraft noise nowadays than they were 30 years ago [3], despite the considerable advances in aircraft noise reduction [4]. In spite of the various engine technologies that have been implemented in aircraft since the 1970s, which had as a side effect significant reductions in noise [5, 6], the airframes of aircraft are also responsible for a large portion of the produced noise during landing. For example, in the work of Merino-Martínez et al. [7] the strong tonal components arising from the A320 aircraft family's nose landing gear system were investigated and it was found that they are strongly correlated with the airspeed of the aircraft. Individual annoyance reports from people living around the area of Schiphol airport, Amsterdam, also confirm these trends [1]. Previous studies have shown that environmental noise exposure is correlated with severe health risks of strokes, coronary heart disease and cardiovascular disease, but also with psychosocial health concerns [8–10]. Additionally, it seems that aircraft-induced noise elicits higher degrees of annoyance

[1] https://bezoekbas.nl/nieuws/jaarrapportage-2023-beschikbaar/

compared to road and rail noise [11, 12], which makes the necessity to address this issue particularly important.

Since annoyance is inherently subjective, there is still a lack of consensus in the scientific community regarding which sound metrics best capture the variance in annoyance responses. In some cases, some metrics are preferred over others with better performance due to their simplicity in implementation [13]. Traditional energy-based metrics such as the Sound Pressure Level (SPL) have been augmented via more complex metrics, such as the Effective Perceived Noise Level (EPNL) and the A-weighted sound level [14], which take into account, at least to some extent, the human perception of noise through spectral irregularities, presence of tones in one-third-octave band spectra, etc. However, it seems that even these enhanced metrics fail to properly capture the large variance in annoyance responses [15, 16]. On top of all this, there is a correlation between demographic factors and the reported noise-induced annoyance, such as age, gender, background, individual noise sensitivity [17, 18], as well as between visual factors and annoyance [19], which makes it even more difficult to isolate the main contributors towards the perceived annoyance.

In order to tackle the limitations imposed by conventional sound metrics, researchers have turned their attention towards psychoacoustics, the field of study "concerned with the relationships between the physical characteristics of sounds and their perceptual attributes" [20]. Research in this field has bridged a significant portion of the knowledge gap concerning the subjective perception of sound through the emergence of so-called sound quality metrics (SQMs) [21], which are computed based on the human's auditory system's characteristics, such as its varying sensitivity to different frequency bands. Starting from the five individual sound quality metrics, described in Appendix B, several psychoacoustic annoyance models have been developed throughout the years. Of these, the most widely used in aircraft-induced annoyance quantification were Zwicker and Fastl's [21], Di's (which is an improvement to Zwicker's model through the inclusion of tonality and was based on audios from electrical transformers instead) [22], and More's model, also including tonality, which was developed in close relation to conventional aircraft noise characteristics [15].

Several promising attempts to generating predictive models for noise-induced annoyance have been identified in literature, combining some of the metrics currently used in certification with psychoacoustic metrics, within Machine Learning frameworks. Some studies are concerned with the more general traffic or urban noise sources, such as the work of Song et al. [23], while others were targeted specifically at aircraft noise, like that of Gille et al. [24]. Furthermore, Sottek et al. investigated the use of AI on actuator sounds [25]. The main limitation concerning the predictive models identified in literature stems from the lack of large and diverse enough datasets available for training and testing, which is crucial for fitting models capable of generalizing well on new, real-life data.

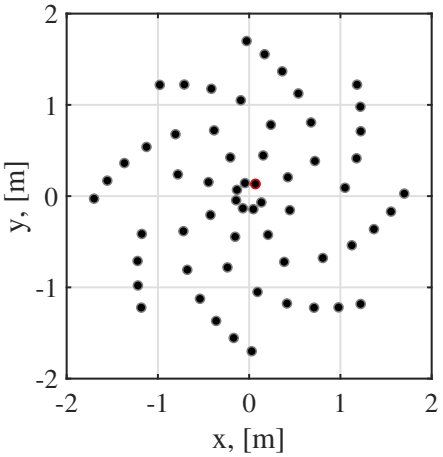
Considering the aspects mentioned above, the aim of this research is to answer the following research questions: "What are the main sound characteristics responsible for annoyance in conventional turbofan aircraft?" and "To what extent can Artificial Intelligence be employed for predicting conventional turbofan aircraft noise annoyance?". By answering these questions simultaneously, the presented research could result in a methodology which can be extrapolated to various types of aircraft, such as UAV's, on which many studies have focused extensively in recent years, such as the works of Torija et al. [26, 27]. Furthermore, such a methodology could be coupled with aircraft noise prediction tools (such as PANAM [28], ANOPP2 [29] and the model proposed by Filippone [30]), in order to have a measure of the expected annoyance of future aircraft already from the design phase, such that mitigation actions can be taken well in advance. In order to integrate it in design loops, it is critical to ensure large computational speeds of both sound metrics and annoyance rating predictions.

The paper is structured as follows. In section II an overview of the methodology used in the steps performed during this research, from data preprocessing, the experimental set-up and correlation analysis, to the development of the AI framework is given. The results are then presented and discussed in section III, followed by the final conclusions and recommendations in section IV. Additional results and details about the work can be found in Appendix A and a comprehensive literature study is given in Appendix B.

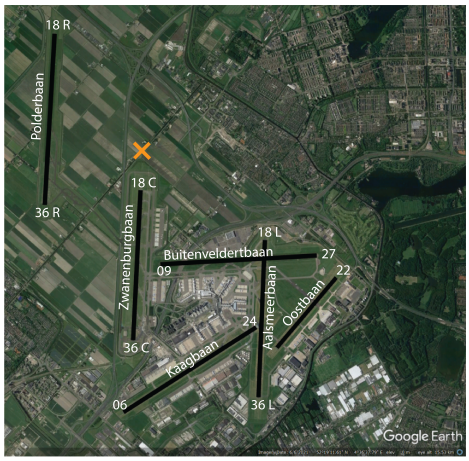
II. Methodology

A. Aircraft Flyover Recordings Preprocessing

The data which was used throughout the research consists of 309 aircraft flyover recordings taken at Schiphol Amsterdam Airport, as part of an extensive campaign within the department of Aircraft Noise and Climate Effects at the Faculty of Aerospace Engineering of TU Delft. Of these, 173 correspond to turbofan aircraft during take-off, while the remaining 136 are during landing. Fig. 1a and Fig. 1b show the 64-microphone array used for the data acquisition (only the data recorded by the highlighted microphone in the figure was used) and the location at which the data was obtained, respectively.



(a) Schematic of the microphone array used for data acquisition [31]



(b) Location of data acquisition [7]

Figure 1. (a) Schematic of the microphone array used for data acquisition, and (b) location of microphone array, denoted with an orange cross.

The resulting data is in the form of time-pressure signals. In order to avoid clipping issues and to limit the exposure of the participants in the listening experiments, the range of amplitudes of all signals was brought to an acceptable interval, which translates to values of the A-weighted equivalent sound pressure level ($L_{p,A,eq}$) ranging from 49.42 dBA to 70 dBA. This was achieved by scaling the recordings to an overhead altitude of 1500 m via two corrections: a spherical spreading correction as in Eq. 1 (where P_{raw} is the raw pressure signal, h_{init} is the real overhead altitude corresponding to the flyover and h_{new} is the altitude for which the scaling is made), and an atmospheric absorption correction. The latter is more complex and follows the ISO-9613-1-1993 standard outlined in [32]. The initial overhead altitudes ranged from a minimum of around 100 m up to more than 400 m for take-off recordings, and from approximately 40 m to just over 100 m in the case of landings. A sampling frequency of 48 kHz was used.

$$P_{corrected} = P_{raw} \cdot \frac{h_{initial}}{h_{new}} \tag{1}$$

Furthermore, as preparation for the subsequent use of the data within the listening experiment campaign and the Machine Learning framework, the variable lengths of the recordings was addressed for consistency reasons. As such, the recordings corresponding to take-offs were cropped to a length of 16 s, while the landings were cropped to 10 s. These values were considered such that recordings of appropriate durations were obtained, while ensuring that not too many of them were eliminated. As a result, a total of 20 recordings had to be discarded (15 take-offs and 5 landings) because they were shorter than the mentioned thresholds. Consequently, those were deemed as being not suitable for further analysis within the scope of this research.

B. Listening Experiment Setup

A listening experiment campaign was conducted in order to obtain short-term annoyance responses for 60 of the available recordings (30 take-offs and 30 landings). These annoyance responses are subsequently used as labeled data for training the Machine Learning models in the second step of the AI framework. To this end, the Psychoacoustic Listening Laboratory (PALILA) at the Faculty of Aerospace Engineering of TU Delft was used. It is an extremely quiet, highly insulated facility made from recycled plastic materials. Its large transmission loss together with its extremely low reverberation times makes it essentially an anechoic chamber, which is very well suited for such experiments, where the participants need to focus on recordings without external interference [33].

A graphical user interface [2] installed on the laptop in the room ensured a smooth way of presenting the recordings while collecting the annoyance responses. The ICBEN 11-point scale was used to answer the following question: "What grade from 0 to 10 best shows how much you would be bothered, disturbed or annoyed by the sound of the aircraft in this recording?". Before starting the experiment, the participants were verbally asked to imagine they are hearing the flyovers while at home or somewhere in their (hypothetical) residential area in the vicinity of an airport.

In order to create an acceptably large dataset of labeled annoyance data while ensuring the responses are statistically relevant, a total of 30 people participated in the experiment. Of these, 21 were men and 9 female, with an average age of 23 years old (and standard deviation of 3). Moreover, 21 were students and the other 9 were employed at the time of the experiment. All participants were in good health condition and had good hearing. The mean duration of the individual experiments was just above 20 min and 30 s (with a standard deviation of 2 min and 49 s). Breaks were also given to the participants every 5 recordings to reduce fatigue.

The 60 flyover recordings were split in three subsets of 40 recordings with a 50% overlap, and these were rotated equally among all test subjects. In other words, each person listened to 40 (20 take-offs and 20 landings) of the 60 recordings and each individual recording was evaluated by 20 people. In the absence of an anchor recording (it was believed that having one would have given the participants a bias), some participants started off with the 20 take-offs and ended with the 20 landings, and some vice versa. Additionally, within each of these two sets (take-offs and landings) the order of the recordings was completely randomised for each subject. The latter two measures were taken in order to average out as much as possible any potential learning effects of the listening order on the received annoyance responses. Finally, all participants were compensated for their time with a 10 € universal voucher upon completing the experiment.

C. Annoyance Responses Analysis

The large pool of conventional noise metrics and SQMs was calculated using the Sound Quality Analysis Toolbox (SQAT) [3] developed within ANCE. An exhaustive overview of the software, encompassing all the standards used in the calculations and performed validations can be found in the work of Greco et al. [34]. This data, combined with that obtained through the experiment campaign, was subsequently used to assess the level of correlation of each metric with the annoyance responses.

To this aim, both Pearson's and Spearman's correlation coefficients are computed in order to reduce an extremely large pool of (statistical variations of the same) features to one which is more likely to achieve good prediction performance in the Machine Learning framework. The former shows the linear correlation, while the latter coefficient captures more complex, non-linear dependencies. Their formulations are given in Eq. 2 and Eq. 3, respectively. x_i and y_i are, respectively, the i-th data points of variables x and y; $\bar{x}$ and $\bar{y}$ are the mean of the variables in consideration; d_i denotes the difference between the two ranks of each observation and n gives the number of observations. Furthermore, an overview of the extracted conventional metrics and

[2] https://zenodo.org/records/11546254

[3] https://github.com/ggrecow/SQAT/tree/main

SQMs using SQAT is provided in Table 1 [4] and Table 2, respectively. In total, 173 metrics and variations of them were computed.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2}$$

$$r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{3}$$

Table 1. Psychoacoustic metrics extracted from SQAT.

Metric	Explanation
N	Loudness, as per the ISO532-1 standard [35]
S	Sharpness, as per DIN45692 [36]
F	Fluctuation Strength, as per Osses et al. [37]
R	Roughness, as per Daniel and Webber [38]
K	Tonality, as per Aures [39]
$PA_{Zwicker}$	Psychoacoustic Annoyance, as per Zwicker's model [21]
PA_{Di}	Psychoacoustic Annoyance, as per the model of Di et al. [22]
PA_{More}	Psychoacoustic Annoyance, as per More's model [15]

Table 2. Conventional noise metrics extracted from SQAT.

Metric	Explanation
EPNL	Effective Perceived Noise Level [EPNdB]
PNLM	Maximum Perceived Noise Level [PNdB]
PNLTM	Maximum Tone-Corrected Perceived Noise Level [PNTdB]
$L_{A_{eq}}$, $L_{B_{eq}}$, $L_{C_{eq}}$, $L_{D_{eq}}$, $L_{Z_{eq}}$	A, B, C, D and Z weighted Sound Pressure Level [dB], respectively
LAF_{max}, LBF_{max}, LCF_{max}, LDF_{max}, LZF_{max}	Maximum values of A, B, C, D and Z weighted Fast Response Sound Pressure Level [dB], respectively
SEL_A, SEL_B, SEL_C, SEL_D, SEL_Z	A, B, C, D and Z weighted Sound Exposure Level [dB], respectively

As a further proxy for the predictive power of the considered metrics, a few rather simple functions were fitted between them and the annoyance responses. The functions are given in Eq. 4 through Eq. 6 [5]. The first is the 10-base logarithmic function (Eq. 4). This function is considered as a consequence of the logarithmic nature of many of the noise certification metrics. Furthermore, the logistic (Eq. 5) and hyperbolic tangent power (Eq. 6) functions were fitted. All three functions are quite simple since the fits only involve two parameters for tuning. The logistic function was observed to be particularly well-suited for relating psychoacoustic annoyance models to the percentage of highly annoyed people by substation noise [40], while the hyperbolic tangent has a similar, S-like shape. Thus, it was considered worth investigating whether this finding could be extrapolated successfully to relating annoyance ratings to various sound metrics.

[4] It should be noted that, for the SQMs and psychoacoustic metrics, several variations from the standard metrics were computed, which include the 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95^{th} percentiles, as well as the maximum, minimum, standard deviation and mean values. In the case of the psychoacoustic models, the scalar values (values calculated based on the 5th percentile values of the SQMs used in their respective calculations) were also added. This was done to ensure that, instead of continuous values, single values are used for further analysis.

[5] Note that, in this case, PA denotes an annoyance rating (further emphasized through the subscript "exp", from experimental), while in the context of Table 1 it denotes metrics obtained from psychoacoustic annoyance models. Depending on the context at hand, the abbreviation can take either of the two meanings throughout this paper.

$$\mathrm{PA_{exp}} = b \cdot log_{10}(\mathrm{x}) + a \tag{4}$$

$$\mathrm{PA_{exp}} = \frac{10}{1 + e^{-k \cdot (x - x_0)}} \tag{5}$$

$$\mathrm{PA_{exp}} = 10 \cdot |\tanh(k \cdot x)|^b \tag{6}$$

D. Machine Learning Framework

1. Convolutional Neural Network for Metric Predictions

The Sound Quality Analysis Toolbox used for calculating a wide variety of conventional and psychoacoustic metrics is a very powerful tool. However, it was observed that generating all the above-mentioned data for a single aircraft flyover recording of 10-20 s takes around 5-10 min. This means that processing large amounts of data can take several days or even longer, depending on the size of the dataset, on a normal computer, defeating the purpose of implementing this approach in design loops. This large overhead was the reason for turning the attention towards leveraging deep learning techniques for bypassing these computational times. Although some attempts have already been made in the existing literature to start looking into this aspect, for example in the work of López-Ballester et al. [41], this is still considered quite a novel approach due to the small number of studies available on a larger scale and with good applicability on a wide variety of metrics.

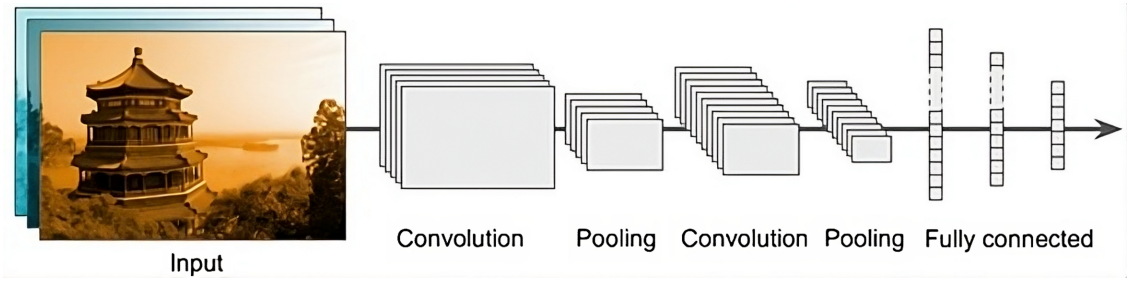
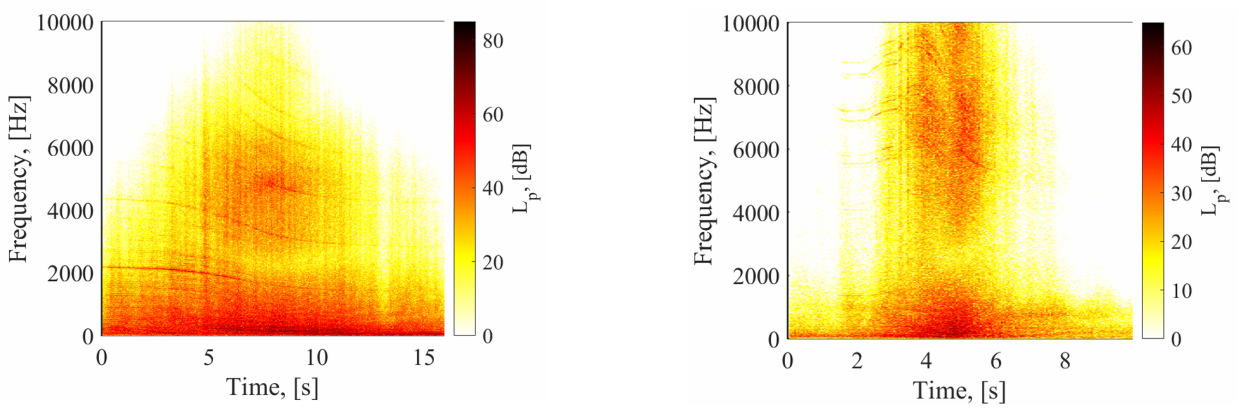


Figure 2. Workflow of generic CNN - figure extracted from [42].



(a) Take-off. **(b) Landing.**

Figure 3. Example of spectrograms from two recordings used in the listening experiment.

Two examples of spectrograms from the recordings used in the listening experiment can be found in Fig. 3. In general, most of the acoustic energy is concentrated around a shorter time span for the landings

compared to the take-offs, due to the relatively shorter distance to the observer in practice, before the distance scaling mentioned in Eq. 1. For the same reason, the peak frequency tends to be higher for landings (although these are general trends, they might differ from recording to recording). The two-dimensional array-like nature of spectrograms - which capture the spectral-temporal patterns of a sound sample - makes the use of Convolutional Neural Networks particularly appealing for obtaining near-instantaneous predictions. This type of neural network architecture leverages the power of convolution for narrowing large amounts of data down to the relevant patterns. The convolutional layers are followed by pooling layers which break down the large arrays into smaller parts, and these cycles repeat, as can be seen in Fig. 2. This ensures a more efficient and easier way of processing large amounts of data and finding relevant patterns. Input data was generated in the form of two-dimensional matrices containing the Sound Pressure Levels (SPL) corresponding to the frequency range 10 Hz to 20 kHz. To this end, a Hanning windowing algorithm with dt = 0.1 s and 50% overlap between the data blocks was used. It should be noted that, in order to increase the size of the available datasets, from each recording (apart from those used in the listening experiments) a "sister" recording was created by multiplying the time-pressure signals by a random factor between 0 and the minimum value that would have resulted in clipping. Hence, the available number of recordings was doubled. Because of the different lengths of the recordings within the take-offs and landings datasets and the requirement that a CNN should receive inputs of constant shape, the two datasets had to be trained and tested separately. An overview of the data used in the training, validation, and testing of the CNNs can be found in Table 3.

Table 3. Overview of CNN input data - values inside parantheses on the first row are the array dimensions resulting from the Hanning windowing algorithm described above.

	Take-offs	**Landings**
Input array dimensions	(319, 2401)	(199, 2401)
Total number of recordings available	256	204
Recordings used for training	171 (66%)	136 (66%)
Recordings used for validation	42 (17%)	34 (17%)
Recordings used for testing	43 (17%)	34 (17%)

Lastly, several principles were applied within the architecture in order to prevent overfitting, to improve convergence, and to reduce the training times. For avoiding overfitting and large training times, early stopping after a number of epochs with no improvement in the validation loss, and a so-called custom callback mechanism, where the weights corresponding to the epoch with the best trade-off between the training and validation losses are used, were implemented. Additionally, dropout layers were included and regularization was applied to other layers. As far as improving convergence is concerned, the learning rate was reduced dynamically when no improvement was seen for a couple epochs in the validation mean absolute error (MAE) by a factor of 0.25 (which was considered to be a suitable threshold for the improvement margin). Prior to training the neural networks, the SPL input data was scaled globally to the range from 0 to 1.

2. Models for Annoyance Rating Predictions

For the second step in the framework, several models were investigated, including a regression Support Vector Machine (SVM) and a regression Random Forest (RF). These were believed to be particularly interesting to analyse considering the nature of the regression task, since they might be able to capture relevant, complex patterns in the available data, which could be well-suited for future generalization on new data. A total of 48 of the 60 recordings used in the listening experiments were used for training (80%) and the remaining 12 (20%) were kept for testing.

For the sake of an exhaustive analysis, several functions were also implemented besides those mentioned in subsection II.C, all of which meet, to an extent, the aforementioned S-shaped property. Hence, the logistic, hyperbolic tangent power, hyperbolic tangent, algebraic sigmoid, and the softsign functions (which can be

visualized in Table 4 and Fig. 4) were fitted on 48 of the 60 recordings as well - using the $curve\_fit$ method of the $scipy.optimize$ module in Python. The generic shapes of these functions can be seen in Fig. 4 (including the aforementioned base 10 logarithmic function). The difference compared to the approach in subsection II.C is that, instead of using just one feature as the argument of the function, linear combinations of several features were used (i.e. $w_1 \cdot x_1 + w_2 \cdot x_2 + ...w_n \cdot x_n$), in order to maximize the predictive power of these rather simple models. A comparison between the obtained results using the models mentioned in this section is provided in section III. Lastly, Fig. 5 shows the workflow of the entire framework involving the steps mentioned above.

Table 4. Functions fitted to annoyance responses.

Equation	Formula
Logistic Equation	$\mathrm{PA}_{\mathrm{exp}} = \frac{10}{1+e^{-k\cdot(x_{\text{linear combination}}-x_0)}}$
Tanh Power	$\mathrm{PA}_{\mathrm{exp}} = 10 \cdot \lvert\tanh(k \cdot x_{\text{linear combination}})\rvert^{b}$
Tanh Linear Combination	$\mathrm{PA}_{\mathrm{exp}} = b \cdot \tanh(x_{\text{linear combination}}) + a$
Algebraic Sigmoid	$\mathrm{PA}_{\mathrm{exp}} = a \cdot \frac{x_{\text{linear combination}}}{\sqrt{1+x^2_{\text{linear combination}}}} + b$
Softsign	$\mathrm{PA}_{\mathrm{exp}} = a \cdot \left(\frac{x_{\text{linear combination}}}{1+\lvert x_{\text{linear combination}}\rvert}\right) + b$

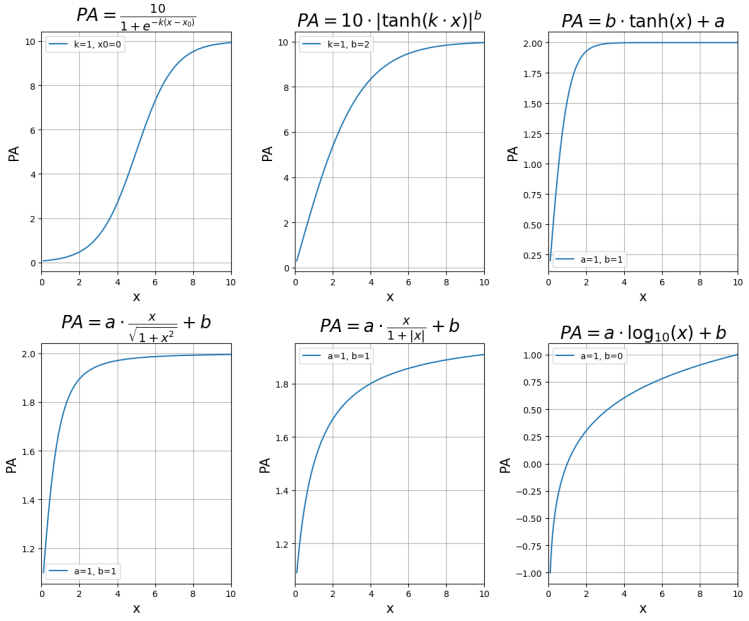


Figure 4. Exemplification of investigated functions (parameters do not correspond to those after fitting the listening experiment data).

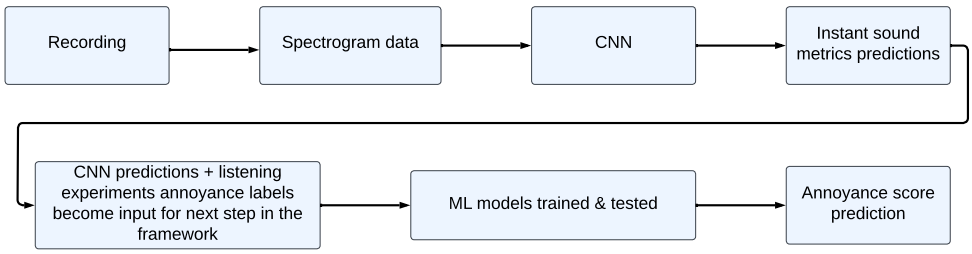


Figure 5. Flowchart of the Machine Learning framework.

III. Results & Discussion

A. Listening Experiment & Correlation Analysis

The annoyance ratings (averaged within all participants of the listening experiment) are summarized in the form of a violin plot in Fig. 6. Upon a quick inspection, a few things become visible: the take-offs were perceived, on average, as being slightly more annoying than the landings - the average annoyance ratings were 6.11 and 5.75, respectively. Also, lower annoyance ratings are less frequent for take-off recordings, as can be seen through the sudden squeeze starting below 6. Although the order of the datasets was randomized for each participant and the recordings were shuffled within each dataset, the longer duration of the take-offs (16 vs 10 s) might have influenced the obtained results to a limited extent. Nevertheless, the results are mostly in line with the expectations, since the take-offs were also generally louder than the landings, as can be visualized in Fig. 7 for the median value of loudness. Previous studies have shown that metrics related to the magnitude of noise - as is the case for loudness - tend to be crucial predictors for annoyance as far as environmental noise is concerned [21, 43]. It was also noticed that the spread in the obtained annoyance ratings is larger in the case of take-offs, while most of the responses are generally concentrated between 5 and 7, indicating that aircraft noise elicits medium to large annoyance in the conditions tested within this research.

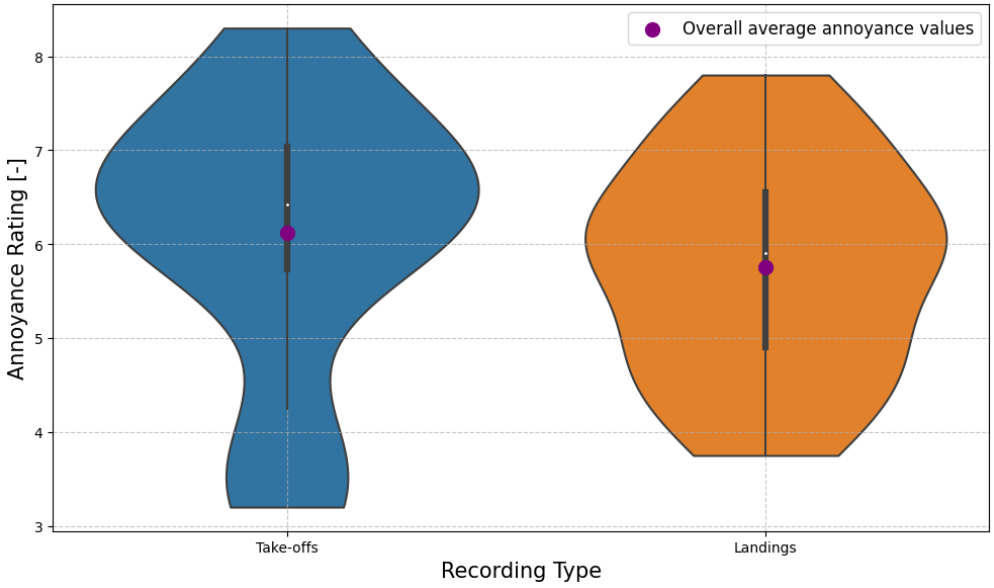


Figure 6. Listening experiment results - width of the plot reflects density of the averaged (per recording) annoyance data at each point; the box plot within the violin plots consist of: white dot (median value), the box (which is the interquartile range), and the whiskers (lines extending from the box to show data points within ± 1.5 times the interquartile range); the purple dots denote the average annoyance ratings.

As a validation of the results prior to the correlation analysis, the mean annoyance corresponding to the major aircraft families as a function of their average MTOW (for the recordings used in the listening experiment) is given in Fig. 8. Although the MTOW of each aircraft family is an averaged value - as it slightly differs depending on the configurations used by each operator - the overall increasing trend also suggests that the results are in line with what one would expect: larger aircraft are typically noisier, and hence, more annoying. Fitting the aforementioned functions was decided upon after a visual analysis of the average annoyance distribution, using Fig. 4 as a starting point. The logarithmic model fitted the data the best and it can explain almost 80% (R^2 = 0.79) of the variance only in terms of MTOW, making the subsequent analyses, using more complex metrics, even more promising.

As for the correlation analysis results, these are summarized in Table 5, using a correlation coefficient

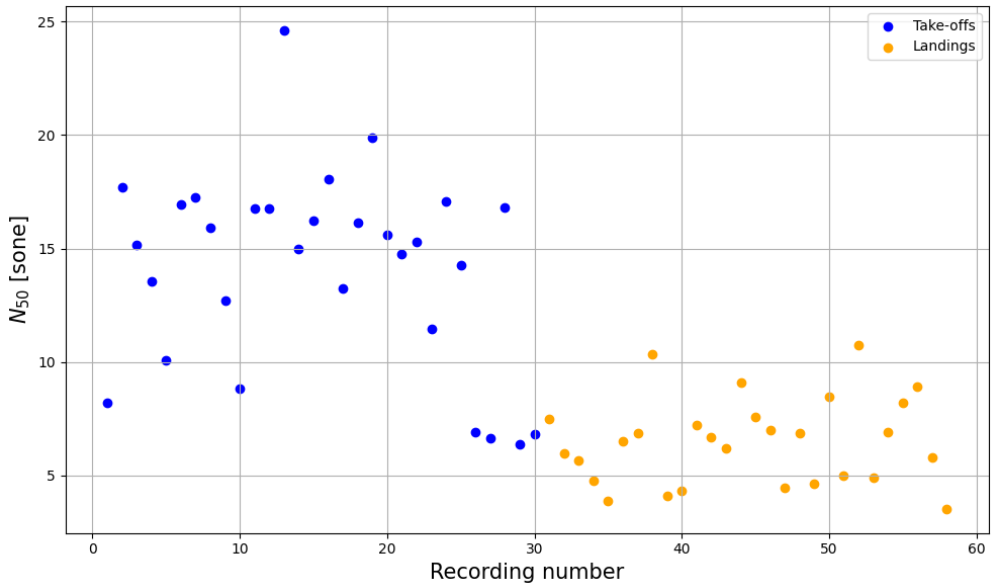


Figure 7. Median values of loudness for the take-offs and landings recordings used in the listening experiment.

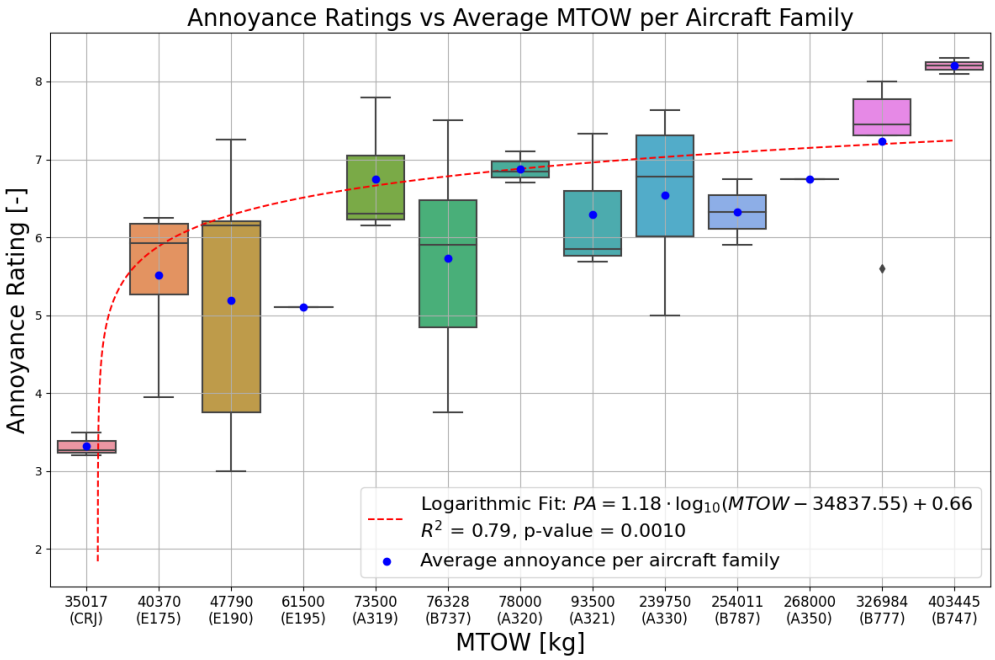


Figure 8. Annoyance ratings vs. average MTOW per aircraft family for recordings used in the listening experiment with logistic function fit - the box plots show the interquartile range (denoted by the boxes), the whiskers extending from the boxes indicate the ± 1.5 times the interquartile range, the horizontal lines within the boxes represent the median values, the diamond shapes denote the outliers, and the blue dots show the average annoyance rating calculated per aircraft family.

(i.e. R-value) threshold of 0.80, so as to narrow down the large initial pool of metrics to those most likely to explain a large portion of the annoyance rating variance[6]. Generally, metrics derived from the field of psychoacoustics, along with the more complex conventional metrics, such as EPNL, PNLM, PNLTM, and the various frequency weightings of the SPL show the greatest predictive potential for the annoyance ratings. Interestingly, apart from loudness, N, none of the other four SQMs made it past the threshold. These metrics are, however, taken into account within the three psychoacoustic models, so their contribution is indirectly considered. This finding perhaps explains that, on their own, these other SQMs do not have a great predictive potential, but, when combined, they are able to explain a large portion of the variance. It is also a further confirmation of the fact that environmental noise annoyance is difficult to quantify and requires rather complex models, supporting the possibility of bridging this gap with the use of AI.

Table 5. Correlation Analysis Results (R-value) - shown in decreasing order of the average value between Pearson's and Spearman's coefficients (all p-values corresponding to the correlation coefficients are lower than 0.05).

Metric	Pearson Correlation Coefficient	Spearman Correlation Coefficient
PNLM	0.94	0.91
$PA_{5_{Zwicker}}$	0.93	0.92
$PA_{5_{Di}}$	0.93	0.92
N_5	0.93	0.91
$PA_{Zwicker}$	0.93	0.91
$PA_{max_{Di}}$	0.93	0.91
$PA_{5_{More}}$	0.92	0.91
PA_{Di}	0.92	0.91
N_{max}	0.92	0.90
$PA_{max_{Zwicker}}$	0.93	0.90
PNLTM	0.93	0.89
LDF_{max}	0.92	0.86
LAF_{max}	0.91	0.87
$L_{D_{eq}}$	0.86	0.84
$PA_{max_{More}}$	0.86	0.83
EPNL	0.84	0.83
$L_{A_{eq}}$	0.84	0.84
PA_{More}	0.84	0.84
N_{std}	0.82	0.82
LZF_{max}	0.82	0.81
LBF_{max}	0.83	0.81
LCF_{max}	0.81	0.80
SEL_D	0.80	0.81

[6]In the case of N and the metrics resulting from the psychoacoustic models of Zwicker, Di and More there were multiple percentile values with correlation factors above 0.80 (from the 1st up to the 30th percentile). However, these metrics were all considered very similar and their correlation factors were basically equal up to the 10th percentile (and rapidly decreasing beyond), hence only the 5th percentile values (which are also the ones normally used in literature), here denoted with a 5 subindex, were kept for further analysis and are presented in the table. $PA_{Zwicker}$, PA_{Di}, PA_{More} are the scalar values of Zwicker's, Di's and More's models respectively and their maximum values are denoted by the subscript "max".

The superiority of psychoacoustic metrics is further confirmed by the results shown in the figures below, which resulted from fitting the linear (Fig. 9, included for a baseline comparison, as linear fits are the most common in the literature), logarithmic (Fig. 10), logistic (Fig. 11), and hyperbolic tangent power function (Fig. 12), respectively. Overall, the (variations of) metrics derived from the PA models of Zwicker, Di et al., and More show the largest R^2 and lowest $RMSE$ values, which means that they explain the largest amount of variance in the obtained annoyance responses. More's PA model performs slightly worse on the data compared to Zwicker's and Di's models. This is unexpected, since More's model was created in close relation to aircraft noise characteristics - although limited to quite a few types of older aircraft (mostly Boeing 757, MD-80, and Beechcraft 1900) [15].

In comparison, many of the more conventional noise certification metrics show a weaker predictive potential (although they are still strongly correlated with the annoyance ratings, as per Table 5), such as the Effective Perceived Noise Level (EPNL) and the Sound Exposure Level (SEL). It might be that a much larger dataset of flyover recordings is required for validating these models' performance, as there is a wide variety of aircraft, all of which have their particular noise patterns. These results also convey that even relatively simple functions are quite versatile and show a solid potential for relating perceived annoyance to a multitude of metrics. An overview of the coefficients of the functions corresponding to the respective best individual fits is provided in Table 6. Hence, it seems that when using one single metric to fit the annoyance ratings, the logarithmic and hyperbolic tangent power functions show marginally better performance compared to the logistic and linear functions - although, for such a small dataset, noise could also be fitted to some extent.

The analysis is further expanded to fitting the same functions (including the linear function for a baseline comparison), to relate sound metrics to the percentage of highly annoyed people (%HA, defined as the percentage of individual annoyance ratings larger than or equal to 7), since this metric is more commonly used in legislation regarding environmental noise annoyance. As such, in Table 7 the best fits in terms of the R^2 value are given. The findings confirm the fact that the logistic function is well-suited for this particular purpose (R^2 = 0.8557), as mentioned previously in subsection II.C, but they also emphasize the slightly better performance achievable via the logarithmic and hyperbolic tangent power functions (over 86% and 87% of the variance is explained, respectively) - but these results are also subject to the possibility of some noise being fitted using the limited data. More detailed results, including the top three best and poorest fits per function, are given in Appendix A. Interestingly, the hierarchy among these functions remains the same as for the annoyance rating case in Table 6, although their corresponding metrics slightly differ.

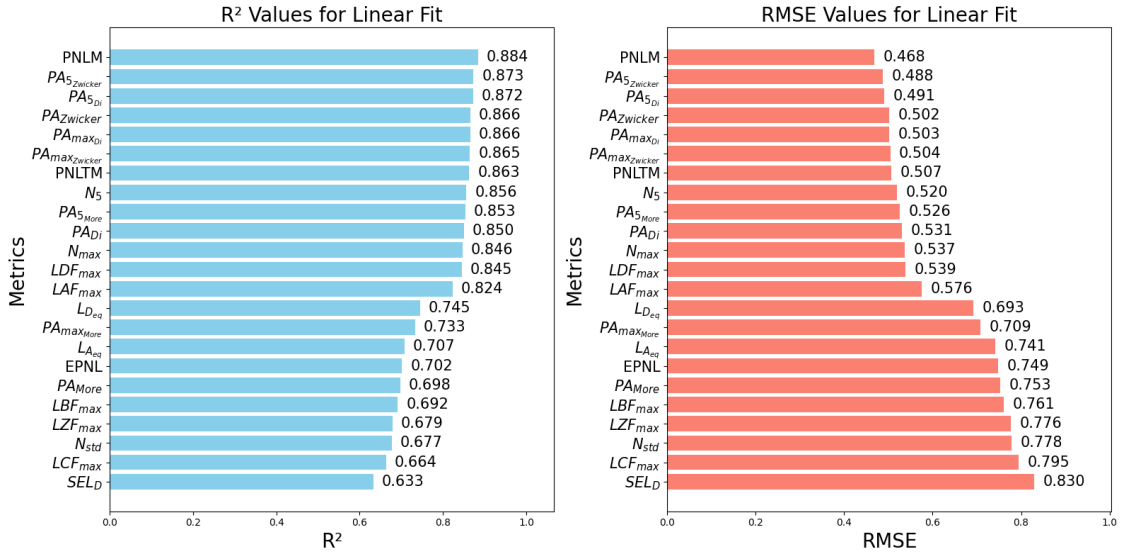


Figure 9. Linear function fit results (included solely as a baseline comparison to the other functions).

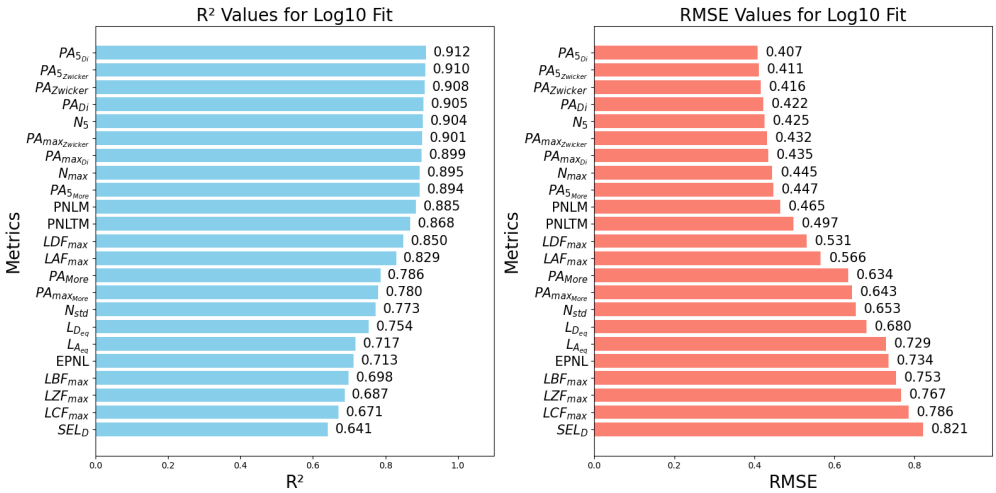


Figure 10. Logarithmic (base 10) function fit results.

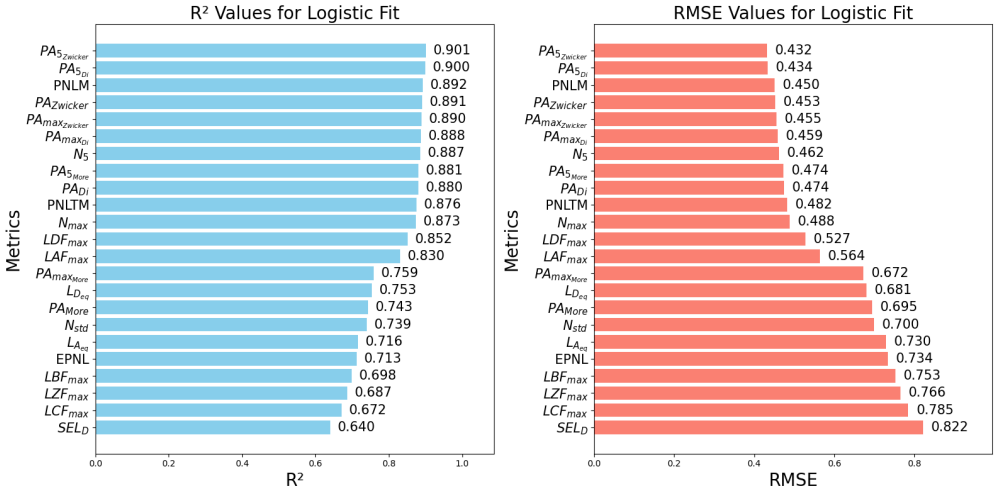


Figure 11. Logistic function fit results.

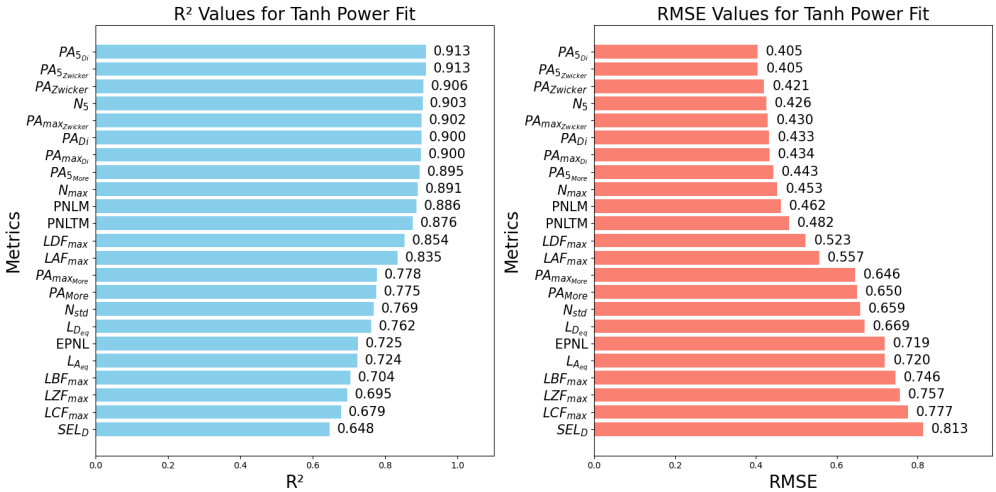


Figure 12. Hyperbolic tangent power function fit results.

Table 6. Parameters of functions corresponding to best fits, for the average annoyance ratings.

Function	R^2 value	Metric	Coefficients
Linear	0.8835	PNLM	intercept = -15.1430, slope = 0.2648
Logistic	0.9007	$PA_{5_{Zwicker}}$	x_0 = 19.8389, k = 0.0827
Log_{10}	0.9116	$PA_{5_{Di}}$	a = -7.6011, b = 9.8045
Tanh Power	0.9128	$PA_{5_{Di}}$	k = 0.0295, b = 1.0698

Table 7. Parameters of functions corresponding to best fits, for the percentage of highly annoyed people.

Function	R^2 value	Metric	Coefficients
Linear	0.8479	$PA_{max_{Di}}$	intercept = -45.6688, slope = 3.0827
Logistic	0.8557	PNLM	x_0 = 81.1810, k = 0.2754
Log_{10}	0.8654	PA_{Di}	a = -224.5500, b = 191.0600
Tanh Power	0.8727	$PA_{5_{Di}}$	k = 0.0599, b = 8.4297

B. Convolutional Neural Network Predictions

For each of the metrics used in the annoyance rating prediction models (see subsection III.C) two neural networks were trained, because of the different input dimensions of the spectrograms corresponding to the landings and take-offs. The obtained results in terms of MAE on the recordings used in the experiments (in order to simulate the behavior on "unseen", real-life data) can be found in Fig. 13, which includes a wider selection of metrics compared to those used in fitting models in subsection III.C - such that a more general view of the CNN behavior is given. The overall precision of the predictions can vary by a few units, but the values are mostly within the same order of magnitude, which indicates that even with a limited amount of training data the architecture can be made quite robust.

In general, it seems that psychoacoustic metrics derived from Zwicker's, Di's and More's models are slightly more difficult to predict, especially in the case of take-offs (all metrics emerging from PA models are located in the upper halves of Fig. 13). One could intuitively partly explain this behavior through the fact that these metrics are quite complex, as they represent the combined contributions of various SQMs, which are also complex in terms of the computation algorithms they use, so their individual prediction errors are combined - although in a way which would be impossible to quantify, due to the "black box" nature of the CNN.

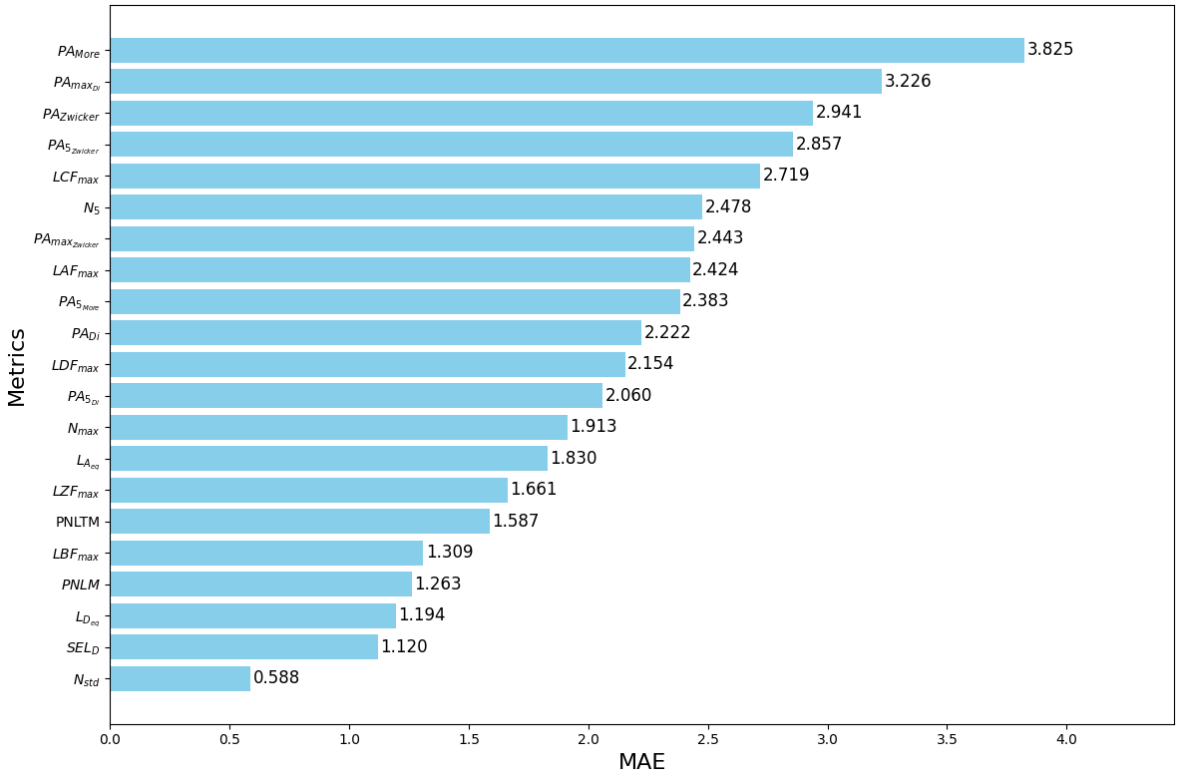
Additionally, the errors corresponding to the landings are, on average, slightly larger than those obtained for take-offs (2.22 dB with a standard deviation of 0.766 dB vs 2.1 dB with a standard deviation of 0.72 dB). This might be explained by the lower amount of available training data for landings, as previously seen in Table 3, which might lead to more limited generalization capabilities. However, the difference is too small to generalize this behavior. Landing recordings are shorter and hence vary more rapidly, which might also add difficulty in making predictions. This is of great importance, since, especially for landings, the results differed depending on the training-validation-testing splits. The take-off CNNs behaved in a much more stable manner for different metrics using the same data splits (and were thus trained on the same split), but the landing CNNs showed a much larger dependency on this aspect in terms of generalization performance and training stability - this resulted in more of a trial and error approach for obtaining satisfying metric predictions for the landings.

Moreover, it seems that the landing CNNs have the tendency of underfitting and they struggle with converging on the training loss value. Although these effects might be eliminated with efforts towards further tailoring the current CNN architecture to create a separate one for landing flyover recordings, it was still preferred to use one CNN architecture for both take-offs and landings, for consistency consider-

ations. The architecture can be visualized in Fig. 14[7] and it was implemented using the *keras* high-level API wrapped in Python's *tensorflow* module[8]. Figure 15 shows the training history of the training and validation loss function (mean squared error, or MSE) for the 5th percentile value of Zwicker's model. The CNN behaves as desired, since it narrows down the gap between the validation and training loss in a stable manner, ensuring that the model not only learns the training data well, but that it also prevents overfitting. It can also be observed that the training history corresponding to the take-offs is more stable, while the landing CNN exhibits a somewhat more erratic behavior. The figures corresponding to the other metrics' CNN training histories can be found in Appendix A and they further exemplify the observations made here.

The custom callback (see red dashed line in the plots) further prevents both overfitting and underfitting by restoring the weights corresponding to the epoch with the best trade-off between the two losses. The trade-off was set through trial and error and it was found that robust enough results are always obtained on the available data using the minimum value of Eq. 7 - the 15 factor was included in order to limit the difference between the training and validation sets performance more drastically. The one major difference between the take-off and landing CNNs is the batch sizes which were used: 7 for the former and 5 for the latter. This was implemented by trial and error as well, in order to account for the different sizes of the datasets used during training, such that proper convergence was obtainable, and for staying within a reasonable margin from both overfitting and underfitting. Finally, the performance on the experiment recordings confirms the achievable robustness of the trained models, with the MAE remaining close to the MAE observed on the datasets used for testing.

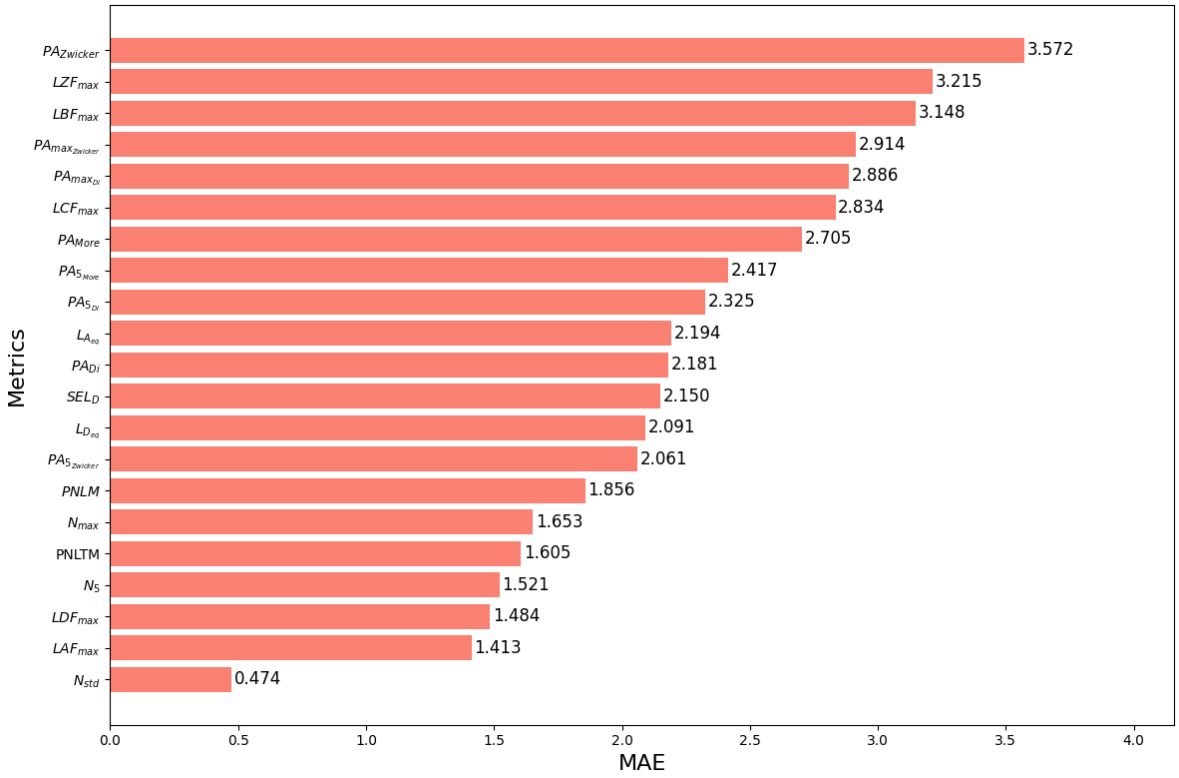
$$\Delta_{loss} = MSE_{val} + 15 \cdot |MSE_{train} - MSE_{val}| \tag{7}$$



(a) Take-offs (descending order of MAE).

[7]The general architecture given in the figure does not include the various weights and biases - the actual weights and biases on each layer differ for each metric and dataset combination and more details about the tunable parameters can be found in Appendix A

[8]`https://www.tensorflow.org/guide/keras`



(b) Landings (descending order of MAE).

Figure 13. MAE comparison of metrics predicted by CNN models on all listening experiment recordings - MAE is expressed in the respective units of the corresponding metrics.

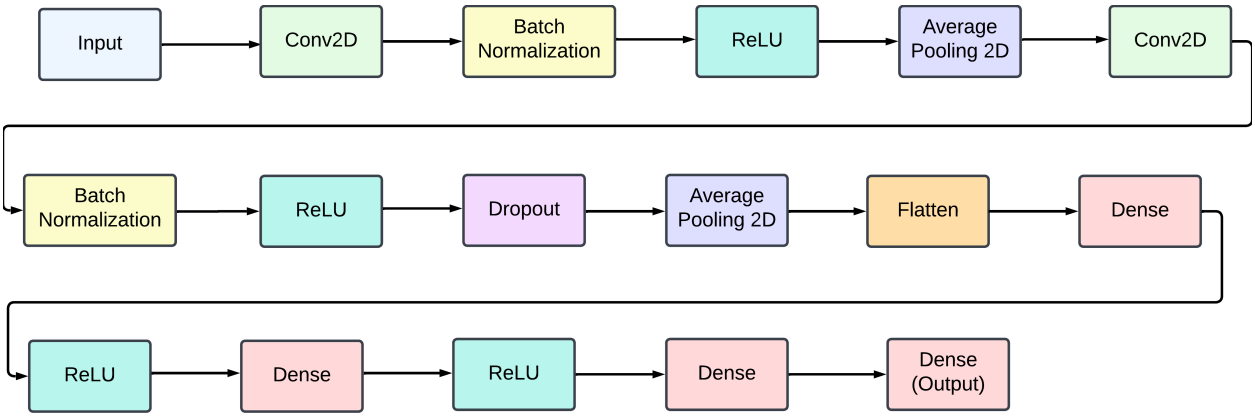
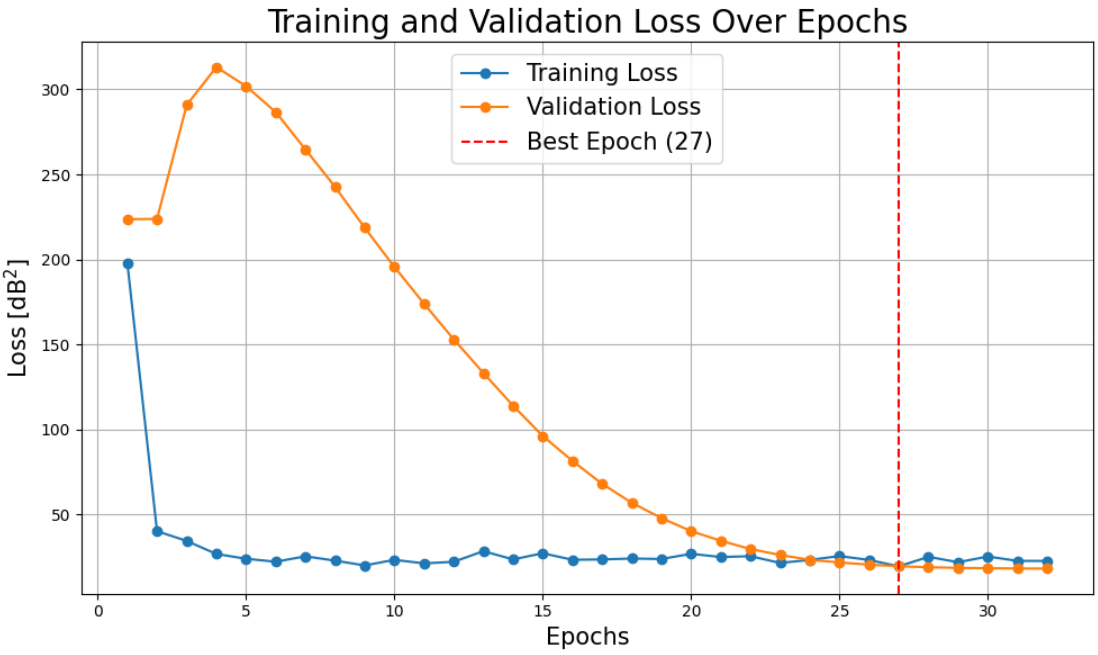


Figure 14. CNN architecture - more details about the parameters and their values can be found in Appendix A.

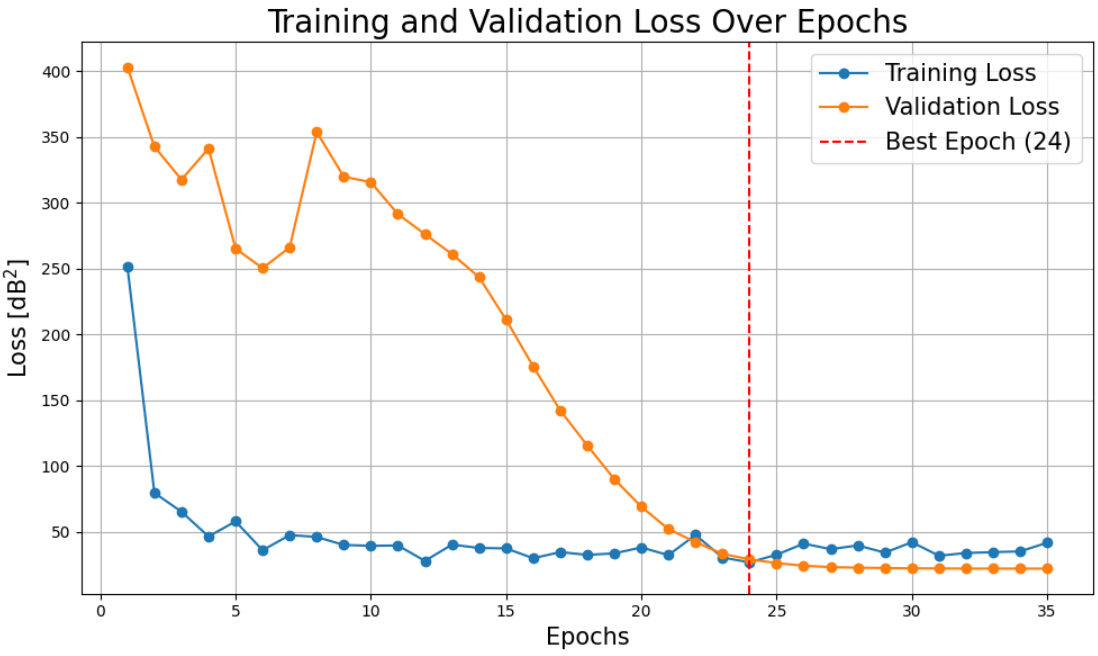
C. Annoyance Rating Prediction Models

Following the logic presented in subsubsection II.D.2, the functions and ML models were fitted and tested first on the data directly obtained from SQAT. The results on the test set (which was the same for all models and comprised 12 recordings, 6 landings and 6 take-offs - making up 20% of the data used in the experiments) can be inspected in Table 8 [9]. Since the available training data consisted of 48 data points, it was decided to

[9]The logarithmic model represented by Eq. 4 was not included in the analysis as it was found that it did not perform as well on the data as the other models.



(a) Take-offs



(b) Landings

Figure 15. Training history of CNN for $PA_{5_{Di}}$.

reduce the search grid of feature combinations to combinations of 4, such that the number of features does not exceed 10% of the number of data points. This measure was taken in order to make sure that overfitting is limited as much as possible from the start.

The chosen feature combinations used throughout this analysis (see Table 8) were selected mainly by considering the correlation factors in Table 5, by making sure that the used metrics are as diverse as possible, and by checking whether an acceptable training/testing performance trade-off could be achieved on all tested models. This was done in order to make the comparisons fair, while acknowledging that the size of the dataset used for training and testing leads to different combinations showing the best behavior on different models. By comparing the models on the same feature combinations, the effects on the overall performance of the framework of the CNN predictions will not affect the interpretability of the results too much, as will be seen in subsection III.D. Had the models been tested on different feature combinations, combining this step with the CNN prediction step would have led to difficulties in making a fair comparison and in assessing which models are the most suitable for annoyance predictions. Additionally, two different combinations were employed such that a more robust analysis could be made.

A stratified learning approach was used during training: the data was split such that the training and testing data covered the whole range of annoyance responses from the listening experiments as uniformly as possible. The exact label distribution can be observed in Table 9, which is in line with the violin plots in Fig. 6. Moreover, apart from the Random Forest model, all other models were trained and tested on scaled data, using a mapping from 0 to 1, thus improving convergence and performance. The RF (Random Forest) does not require scaling because of the inherent nature of its prediction mechanism. Furthermore, the SVM and RF models' hyperparameters were tuned by means of a grid-search (combined with cross-validation), which can be found in Table 10 and Table 11, respectively. More details about this process can be found in Appendix A.

Generally, it seems that all of the explored models can achieve satisfactory performance when combined carefully with hyperparameter tuning and feature selection. The hyperbolic tangent, algebraic sigmoid, and softsign functions exhibit, on average, considerably smaller errors compared to the SVM, RF, logistic, and hyperbolic tangent power functions. It remains to be confirmed how these results are influenced when using the exact same models from Table 8 on the metrics predicted by the CNN, especially since some of these models have a slight underfitting tendency on the used dataset and features. One interesting aspect is that the SVM results are quite similar to those obtained in the work of Sottek et al. [25], in terms of MAE and R^2 values, even though the cited work used actuator sounds, which are fundamentally different from aircraft noise. Of course, the results would have to be validated in the future with larger datasets. It might be that different feature combinations would work better - and also that simpler models would be outperformed by the SVM and RF on larger datasets. Hence, the qualitative aspect of this analysis is considered of greater value in this research, as it stands to show that there is no perfect solution to predicting such a complex aspect as aircraft noise-induced annoyance. These results are also a proof for the possibility to combine SQMs, PA models, and more conventional noise certification metrics within the frameworks of Machine Learning to leverage their individual strengths. Perhaps, under the condition that a lot of data can be gathered, one extremely robust model could be created for e.g. predicting turbofan aircraft noise annoyance (but also for other types of aircraft, like UAVs). In subsection III.D the behavior of these models on the test set is shown, but this time using the CNN metrics' predictions (see subsection III.B) instead of the values obtained through direct calculation employed in the current section.

D. Overall Framework Performance

Lastly, the entire framework's performance is assessed. An overview of the obtained results when combining the CNN metric predictions with the other models for annoyance prediction on the 12 recordings used for testing in subsection III.C is given in Table 12.

Table 8. Parameters and performance of the models and functions employed in the second step of the framework on the experiment recordings - predictions made on metrics' values as obtained from SQAT (the best performing configuration is highlighted in bold text) - numbering of the weight parameters (subscripts 1 to 4) follows the order in which the used features are mentioned in the table.

Model	Parameters	Train MSE	Train MAE	Test MSE	Test MAE	Test R^2
Feature Combination 1: PNLM, $PA_{5_{Di}}$, LDF_{max}, N_{max}						
SVM	C = 10, gamma = 'auto' kernel = 'rbf', epsilon = 0.5	0.1745	0.3540	0.2098	0.3889	0.8863
Random Forest	max_depth = None n_estimators = 1000 min_samples_split = 5 min_samples_leaf = 5	0.1844	0.3481	0.2235	0.3786	0.8789
Logistic function	x_0 = 2.8766, k = 0.2685 w_1 = 4.2827, w_2 = 3.6890 w_3 = -1.4451, w_4 = 2.1583	0.1687	0.3155	0.2420	0.3936	0.8689
Hyperbolic tangent power function	k = 0.1247, b = 0.2981 w_1 = 1.0630, w_2 = -1.4327 w_3 = -1.1604, w_4 = 5.0261	0.4180	0.4427	0.2678	0.4001	0.8550
Hyperbolic tangent function	**b = 5.0057, a = 3.2006** **w_1 = -0.2357, w_2 = 1.5786** **w_3 = -0.5577, w_4 = 0.8620**	**0.1426**	**0.3105**	**0.1310**	**0.2908**	**0.9291**
Algebraic sigmoid function	**a = 5.5822, b = 3.1986** **$w_1 = -0.2522$, w_2 = 1.4519** **$w_3 = -0.5295$, w_4 = 0.8645**	**0.1427**	**0.3113**	**0.1345**	**0.2942**	**0.9271**
Softsign function	**a = 7.7517, b = 3.1965** **$w_1 = -0.4951$, w_2 = 1.4731** **$w_3 = -0.5169$, w_4 = 1.1204**	**0.1452**	**0.3137**	**0.1519**	**0.3107**	**0.9177**
Feature Combination 2: PNLTM, $PA_{Zwicker}$, LAF_{max}, N_5						
SVM	C = 50, gamma = 0.1 kernel = 'rbf', epsilon = 0.5	0.1793	0.3425	0.2087	0.3822	0.8869
Random Forest	max_depth = None n_estimators = 1000 min_samples_split = 5 min_samples_leaf = 5	0.1949	0.3618	0.2646	0.4124	0.8567
Logistic function	x_0 = 0.5927, k = 1.2368 w_1 = 0.5878, w_2 = 0.8939 w_3 = 0.2006, w_4 = 1.1361	0.1880	0.3391	0.2551	0.4174	0.8618
Hyperbolic tangent power function	k = 0.0636, b = 0.2889 w_1 = -3.1886, w_2 = 0.4200 w_3 = 2.7907, w_4 = 6.7616	0.2210	0.3801	0.1650	0.3081	0.9107
Hyperbolic tangent function	**b = 5.2802, a = 3.1070** **w_1 = -0.1749, w_2 = 0.8131** **w_3 = -0.1028, w_4 = 0.8949**	**0.1656**	**0.3286**	**0.1276**	**0.2841**	**0.9309**
Algebraic sigmoid function	**a = 5.9733, b = 3.1040** **$w_1 = -0.1898$, w_2 = 0.7783** **$w_3 = -0.0761$, w_4 = 0.7843**	**0.1655**	**0.3279**	**0.1305**	**0.2862**	**0.9293**
Softsign function	**a = 9.1328, b = 3.0845** **$w_1 = -0.2946$, w_2 = 0.8962** **w_3 = 0.0142, w_4 = 0.4995**	**0.1684**	**0.3269**	**0.1392**	**0.2948**	**0.9246**

Table 9. Stratified learning - bins and number of samples per bin used in training.

Annoyance rating bins range	3.0-5.0	5.0-6.0	6.0-7.0	7.0-9.0
Number of samples	15	10	19	15

Table 10. Grid search parameters for SVM.

Parameter	Values
`C`	0.1, 1, 10, 50
γ `(gamma)`	scale, auto, 0.1, 0.01, 0.001
`kernel`	rbf, linear
ε `(epsilon)`	0.01, 0.1, 0.2, 0.5, 1.0

Table 11. Grid search parameters for Random Forest.

Parameter	Values
`n_estimators`	100, 200, 1000
`max_depth`	None, 5, 10
`min_samples_split`	None, 5, 10
`min_samples_leaf`	None, 5, 10

Table 12. Framework annoyance rating prediction performance on 20% of the data used in the experiment recordings (best models highlighted in bold) - results obtained using the CNN metric predictions.

Model	MSE	MAE	Maximum MAE	R^2
Feature Combination 1: PNLM, $PA_{5_{Di}}$, LDF_{max}, N_{max}				
SVM	**0.2362**	**0.3686**	**1.12**	**0.8721**
Random Forest	**0.2706**	**0.3996**	**1.21**	**0.8534**
Logistic function	**0.2585**	**0.3644**	**1.16**	**0.8600**
Hyperbolic tangent power function	0.1995	0.3598	0.94	0.8919
Hyperbolic tangent function	0.3660	0.4542	1.34	0.8018
Algebraic sigmoid function	0.3690	0.4605	1.33	0.8001
Softsign function	0.4027	0.4841	1.33	0.7819
Feature Combination 2: PNLTM, $PA_{Zwicker}$, LAF_{max}, N_5				
SVM	**0.2643**	**0.3707**	**1.05**	**0.8569**
Random Forest	**0.3123**	**0.4342**	**1.25**	**0.8309**
Logistic function	**0.2651**	**0.3770**	**0.95**	**0.8564**
Hyperbolic tangent power function	0.4474	0.4395	1.72	0.7577
Hyperbolic tangent function	0.2075	0.3167	1.04	0.8876
Algebraic sigmoid function	0.2128	0.3195	1.04	0.8847
Softsign function	0.2642	0.3425	1.08	0.8569

In order to analyse the framework as a whole, one needs to keep two main aspects in mind: the prediction accuracy of the models in the second step (that of annoyance predictions), as well as the influence of the first step's (CNN) metric predictions on the performance of the second step. In other words, it is crucial that the annoyance prediction model employed in such a ML framework is somewhat robust to the variability of the CNN's accuracy. Because of this variability, one should not be immediately tempted to consider the models with very small errors on one particular feature combination (which might be, after all, a matter of chance) as being the best. Consequently, it would seem that, in terms of MSE, MAE, and R^2 values, the SVM, the RF, and the logistic function are the most robust of the models, when comparing Table 8 with Table 12, for both considered four-feature combinations. The portions of explained variance differ by no more than 3% for the

corresponding pairs of model-features, and the MAEs and MSEs also remain fairly similar. This behavior is expected for the SVM and the RF, as they are more complex and can effectively learn patterns better, which can be translated to new, unseen data.

As the CNN prediction accuracy tends to be quite variable among different metrics (see Fig. 13), the random forest is inherently less prone to being influenced by inaccurate CNN predictions, as long as the relative relationships between the metrics/features it employs remain relatively constant. Also, its predictions are made by averaging the results of multiple decision trees which work independently, further enhancing its robustness. The SVM's robustness to the CNN's inaccuracies also depends on the hyperparameters' values which directly influence its tolerance to prediction errors (such as ε, see Appendix A). The larger this tolerance is, the more "flexible" is the regression function on which the predictions are based. The logistic function's robustness to the effects of the CNN predictions further confirms the initial hypothesis in subsection II.C - that the extrapolation of the results in the work of Di et al. [40], from relating PA models to the percentage of highly annoyed people, to relating various metrics to psychoacoustic annoyance ratings from listening experiments, is indeed possible. Lastly, the maximum MAE was added to this analysis for further context regarding the prediction power which can be achieved with this kind of ML framework. As can be observed, the best performing models show a maximum value ranging from around 1 to 1.25. Considering that the annoyance responses range from 0 to 10, such an error, while not entirely negligible, does not represent an extreme deviation from the actual observed annoyance - after all, the used annoyance ratings were averaged values obtained from 20 subjective responses for each recording.

As far as the remaining simpler models are concerned, although their initial performance on the test set using the metrics' values calculated directly (using SQAT) was, generally, considerably better than that of the SVM and RF (according to Table 12), they exhibit quite an unpredictable behavior when used together with the CNN predictions. For the first feature combination, they show a drastic decrease in prediction accuracy and in the amount of variance they are able to capture - which might also be due to the CNN accuracy on its own, so further analysis would be required to validate these results. For the second combination, the MAEs of the hyperbolic tangent, algebraic sigmoid and softsign functions are quite similar between Table 8 and Table 12. However, the MSEs and R^2 values show larger variations. The tendency to slightly underfit (i.e. the training errors are slightly smaller than those on the test set) of these models might also be at play here, but the extent to which this happens when fitting the models is considered too small to draw such a strong conclusion. The rather small dataset of 12 recordings on which the final predictions are evaluated must also influence these results, but the effects are assumed to have been reduced to a large extent through the stratified learning approach.

Thus, the overall results for the entire framework are, from a quantitative point of view, somewhat inconclusive on such a small available dataset. On the other hand, qualitatively speaking, these results are far more useful. One conclusion can be drawn for sure: more complex models, and perhaps especially models which base their predictions on relative relationships between features, as is the case of the random forest, are better suited for the specific framework proposed in this research, in terms of their sensitivity to the inaccuracies arising from predicting the metrics they use as features, instead of calculating them directly. Furthermore, the logistic function seems to be particularly well-suited for capturing the way annoyance scales with the combined contributions of multiple noise-related metrics.

IV. Conclusions, Limitations & Recommendations

A. Limitations & Recommendations

The main limitation of the study is the lack of data to further validate the results on a larger dataset. The CNN predictions vary for each metric, and the reduced dataset size that was available means that different splits of data can lead to different results and generalization capabilities (the dataset for the training of the CNNs were artificially enlarged by multiplying each recording's time-pressure signal by a factor, which can lead to overly

optimistic results on some data splits, while performing poorly on completely new, unseen data - however, this measure generally helped in obtaining more robust results on the available recordings). The same holds for the models used for annoyance predictions.

Consequently, it is highly recommended for future work to enlarge the available aircraft recordings dataset and to apply the methodology presented in this paper (re-tuning the models would definitely be required on a new dataset), to further quantify the achievable robustness of such a Machine Learning framework. In addition, the psychoacoustic metrics are extremely correlated to one another since they combine many of the same sub-metrics in slightly different ways. Hence, if more data were available, techniques such as Principal Component Analysis could also be applied within the second step of the framework in order to reduce dimensionality and isolate the most relevant directions for annoyance prediction.

B. Conclusion

Predicting aircraft-induced noise annoyance is a crucial challenge that needs to be addressed in order to mitigate the consequences of environmental noise pollution on the affected communities. The task is anything but trivial, since the mechanisms associated with psychoacoustic annoyance are extremely complex and also computationally expensive. Hence, the aim of this research was to investigate the possibility to combine various Sound Quality Metrics, Psychoacoustic Annoyance models, and conventional noise certification metrics with Artificial Intelligence in order to obtain both instant metric predictions and annoyance score predictions from an input flyover recording. To this end, a listening experiment campaign was conducted to gather labeled annoyance data (on the 11-point ICBEN scale), and a two-step AI framework was created. The latter consists of a Convolutional Neural Network which predicts complex metrics from input spectrogram data as a first step, which is followed by a second step concerned with making the annoyance predictions using as input the metrics predicted by the CNN. This methodology could be further developed and combined with, for example, auralization into a powerful tool for assessing the expected annoyance caused by future aircraft configurations from the earlier development phases. Thus, noise mitigation actions could be taken in advance, bypassing the need of retrofitting existing designs.

The results showed that psychoacoustic metrics and their statistical variations are generally much better correlated to the received annoyance ratings compared to more conventional metrics. This shows that metrics which take into account the characteristics of the human auditory system to a deeper level, such as the increased sensitivity to certain frequency ranges, amplitude modulations, or the presence of tones are of paramount importance for the task of annoyance prediction. Additionally, it was seen that, even with a dataset limited in size, satisfactory predictions of these complex metrics can be made with properly tuned CNNs. Moreover, the strengths of random forests, SVMs, and logistic functions in particular can be leveraged in combination with the CNN predictions for further predicting annoyance ratings based on various combinations of metrics. Overall, the framework can achieve Mean Absolute Errors of the annoyance ratings of around 0.4 and below, which represents roughly 4% of the range of the scale used for the ratings, and R^2 values above 0.85.

Acknowledgments

The author would like to thank Dr. Roberto Merino-Martínez for the supervision and constant support offered during this research. Moreover, special thanks are extended to Ir. Irina Besnea for kindly offering the flyover recordings used in this research and to Ir. Josephine Pockelé for the help given during the listening experiment campaign on using the GUI installed on the laptop from the listening lab. Lastly, special thanks are extended to the 30 participants in the listening experiment campaign.

A. Supporting Work

A. Details on Aircraft Flyover Recordings and Listening Experiment Results

The distribution of aircraft types for all the initial recordings (309 in total) and for those used in the listening experiment (60 in total) can be found in Fig. A1. The majority of the aircraft are Boeing 737 models (of various series). For the listening experiment, it was aimed to follow the overall distribution of the 309 recordings as closely as possible, such that the datasets used for fitting the ML models and for training the CNNs are as similar as possible (each aircraft family could possibly exhibit its own particular noise properties). Not all types of aircraft were included in the experiment, but most of them were. It was also a priority to include in the experiment recordings corresponding to aircraft of various sizes, from small to very large - so the distributions are not identical, but fairly similar nevertheless. In addition, the overhead velocity distributions for both datasets are presented in Fig. A2 - note that, since the plots were generated using ADS-B data, which was not always extremely accurate, some data points were filtered out, thus the number of data points does not match the sizes of the datasets. The recordings corresponding to take-offs have significantly larger velocities, as expected. This might be yet another influencing factor in the perceived annoyance, but it was assumed that this effect is partly offset by the lower altitude of the aircraft during landing, as mentioned in the main chapters.

A more in-depth look into the obtained results from the listening experiments can be taken through Fig. A3. Unlike the violin plots, the box plots also show the outliers for each individual recording. These were eliminated from all subsequent analyses - in order to use the most accurate and meaningful data available - through the following process:

- The first and third quartiles (25^{th} and 75^{th} percentiles, respectively) were computed
- The interquartile range (IQR, i.e. the difference between third and first quartiles) was calculated
- The upper and lower bounds were calculated per recording, defined as $b_u = Q_1 + 1.5 \cdot IQR$ and $b_l = Q_1 - 1.5 \cdot IQR$, respectively
- Responses falling outside of these bounds were eliminated for each recording and the average annoyance rating was calculated from the remaining values

In Fig. A4 the cross-correlation matrices for the pairs of individual responses is given (one for each of the three batches of recordings used in the listening experiment - hence 10 participants per batch out of the total of 30 participants in the experiment). With the exception of one slightly negatively correlated pair in the first batch, all of the pair-wise correlations are positive, indicating a good alignment between all participants.

Additionally, using the average responses, the three best and worst combinations of metrics (calculated via SQAT) and function (linear, 10-base logarithmic, logistic and hyperbolic tangent power) are given in Fig. A5 - Fig. A8 and Fig. A9 - Fig. A12, respectively. The standard errors are also plotted on top of each recording's average annoyance rating, defined as the standard deviation of the responses corresponding to each recording divided by the square root of the total number of responses per recording. The plain linear function is also included here as a means to showcase the slight improvement achieved via more complex functions and the data's tendency of having a slightly S-shaped distribution.

Interestingly, some of the metrics not able to explain a very large portion of the variance on an individual level, such as SEL_A and SEL_D, are commonly used in current certifications. This showcases the importance of the current research for bridging the existing gaps between the observed annoyance responses and the ability of mainstream metrics to capture their variance to a large extent. SEL_A was included in this analysis, although it did not pass the 0.8 correlation thresholds mentioned in the main chapters. This was done purely for comparison purposes, since it is currently a widely popular metric in noise certification. Additionally, its correlation factors were not below the threshold by much (above 0.7, so it still exhibited quite a strong correlation to the annoyance ratings). It is also crucial to note that psychoacoustic annoyance models (and

their statistical derivatives) generally exhibit the most promising predictive behaviors on these functions. These observations reinforce the need for finding more suitable models for certification purposes, as well as they offer promising directions in which this could be explored in future research.

Furthermore, the same analysis (using the same four functions) is made for the metrics calculated with SQAT, but this time they are related to the percentage of highly annoyed (denoted by %HA) people across all recordings (defined as the percentage of annoyance ratings of at least 7, after having eliminated the outliers). Once again, the three best and worst fits per function are given in Fig. A13 - Fig. A16, and Fig. A17 - Fig. A20, respectively. Similarly to the annoyance ratings case, metrics derived from PA models obtain the best R^2 - although the maximum values are slightly lower this time (no more than 87% of the variance is covered, compared to 91% in the previous analysis). The R^2 values obtained in the present study are slightly lower than those obtained by Di et al. in their similar work on relating %HA to the PA model of Di et al. [40] - the cited study was performed on substation noise, however. One key difference compared to Fig. A5 - Fig. A8 comes from the fact that PNLM and PNLTM occurences have been mostly replaced by the maximum loudness among the top three best fits. Lastly, the same three metrics belong to all the top three weakest fits: SEL_A, SEL_D (also found among the weakest fits using annoyance ratings and with similar results), and LCF_{max}.

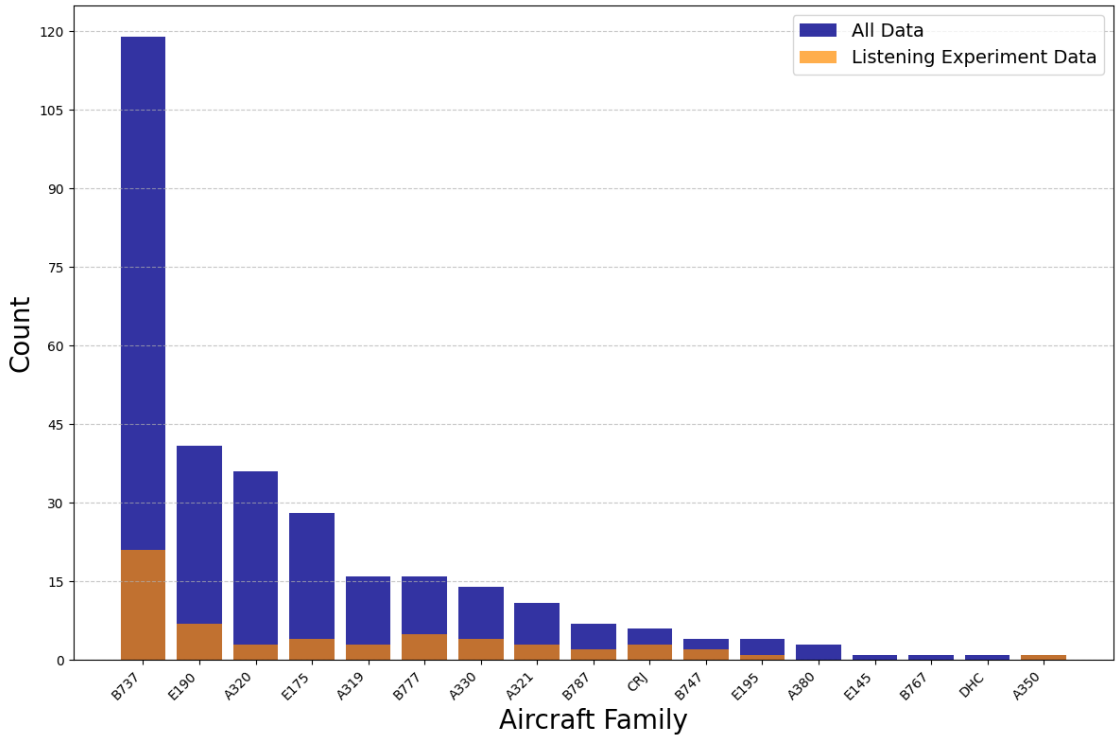
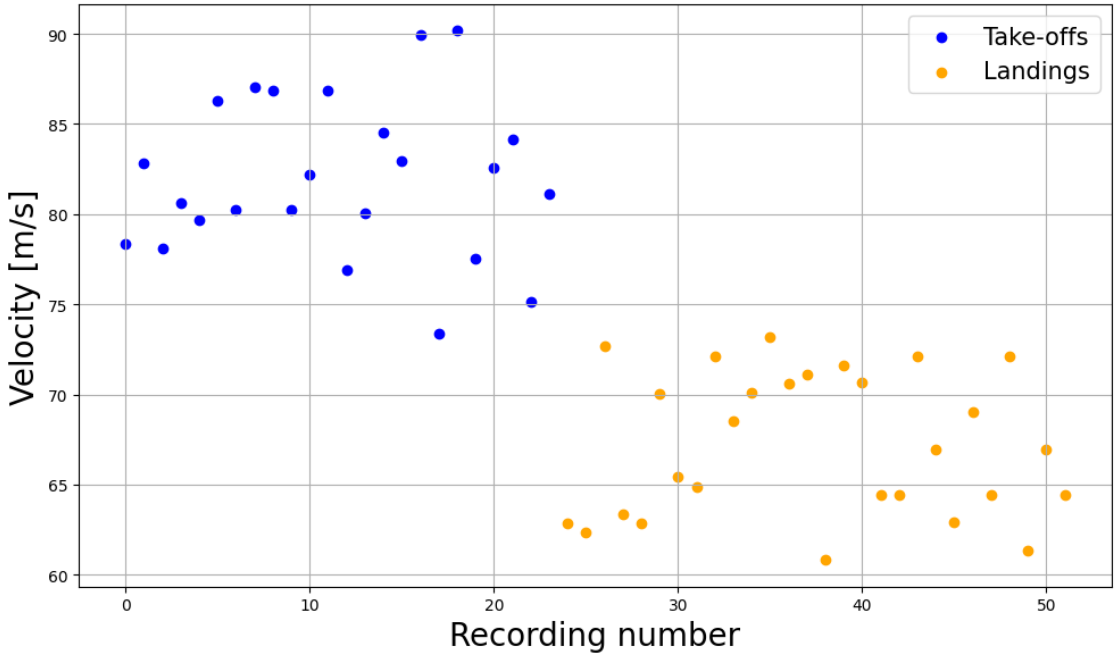
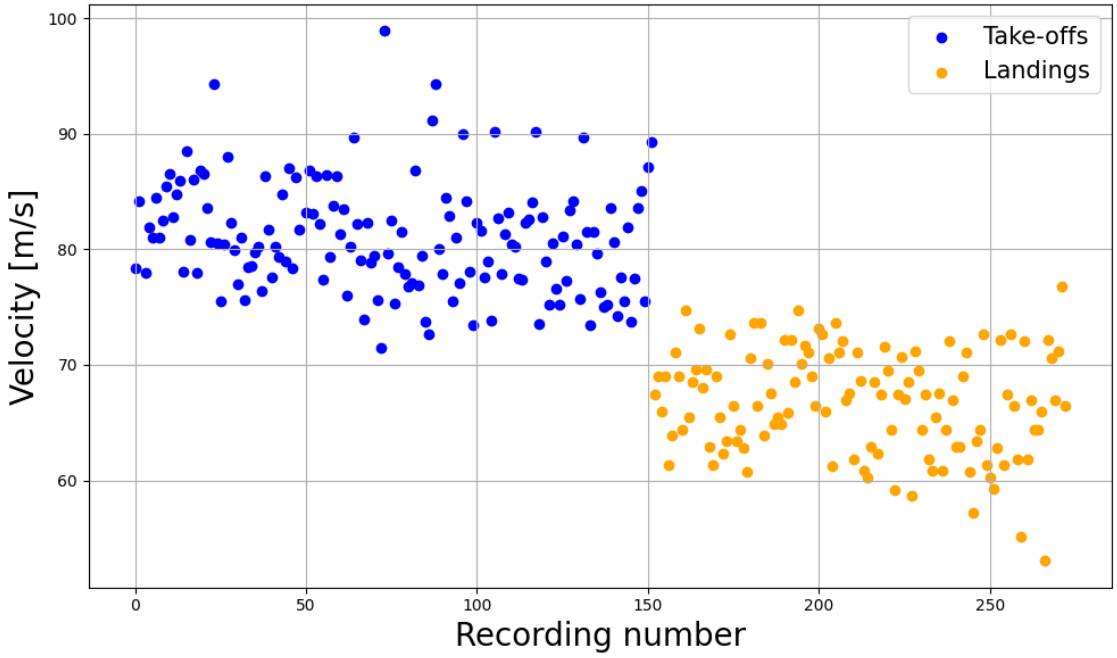


Figure A1. Comparison of aircraft type distribution on the entire recordings dataset and on the dataset used for the listening experiment.



(a) Aircraft overhead velocity distribution of the dataset used in the listening experiment.



(b) Aircraft overhead velocity distribution of the entire recording dataset.

Figure A2. Overhead velocity distributions (inconclusive ADSB data points are filtered out).

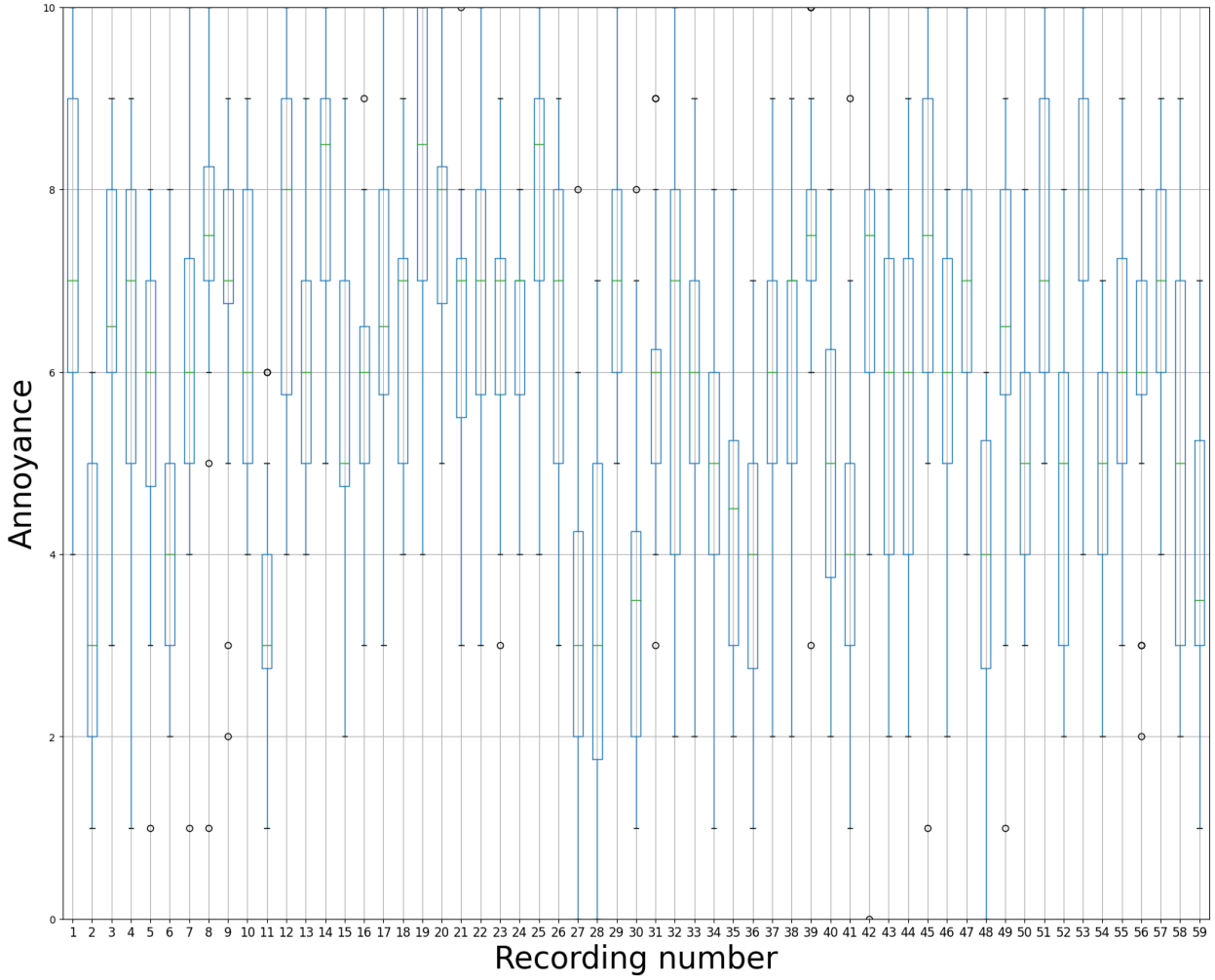
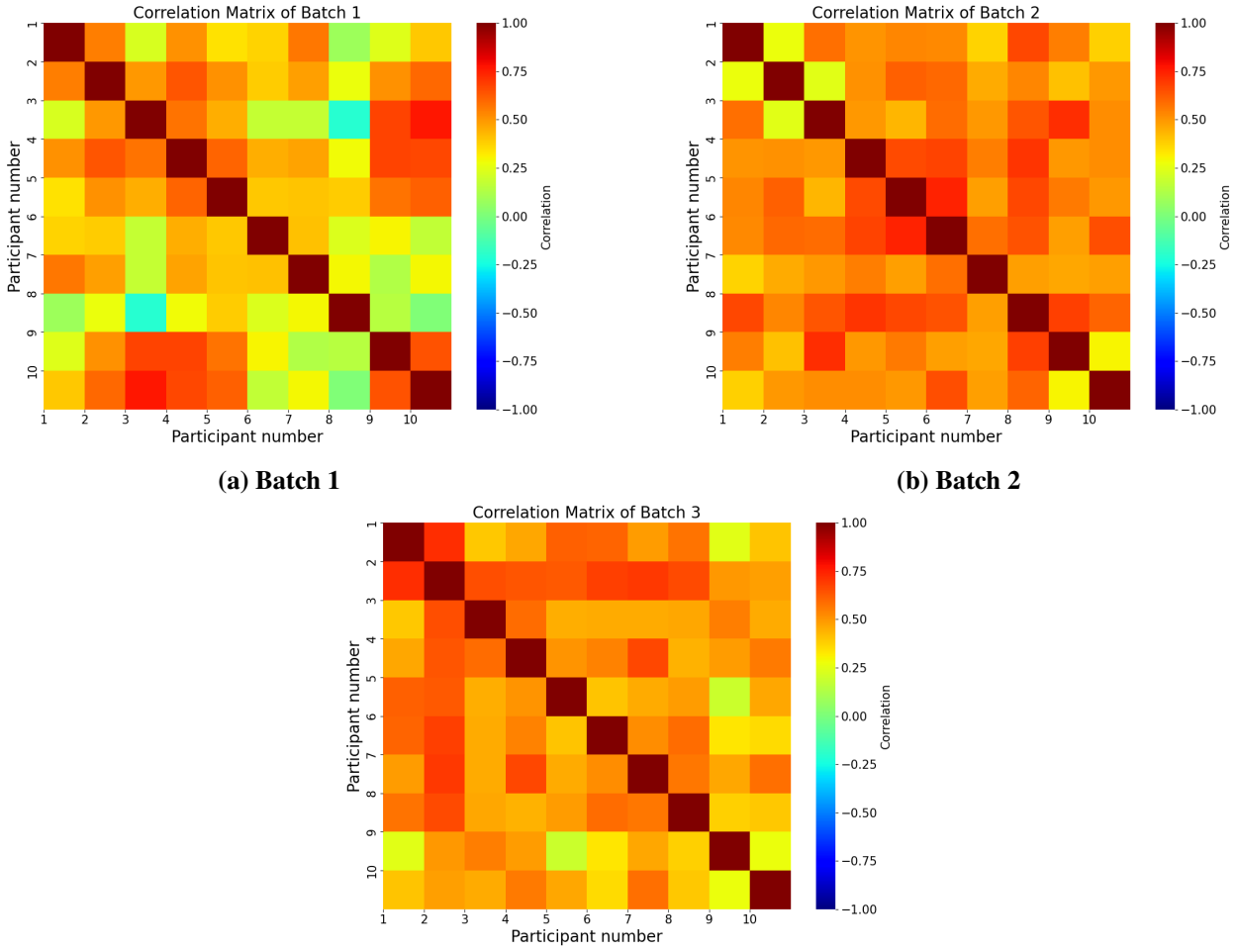


Figure A3. Box plots of obtained results during the experiment campaign.

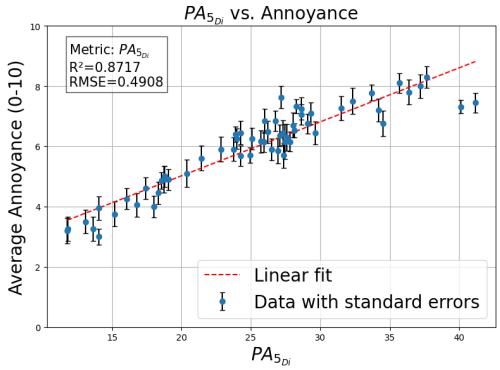


(a) Batch 1

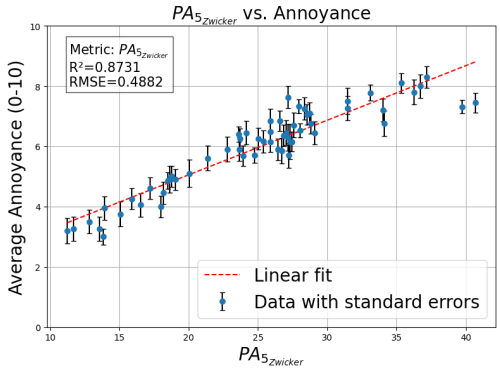
(b) Batch 2

(c) Batch 3

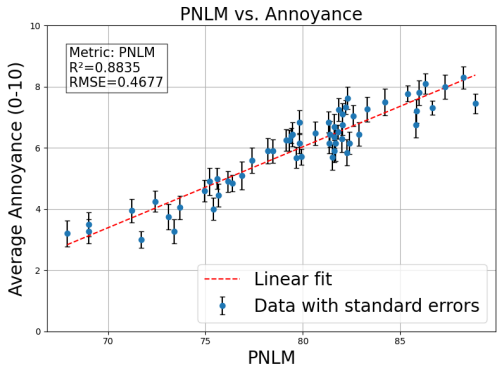
Figure A4. Cross-correlation matrices for the responses of each pair of individual participants for the three batches used in the listening experiment campaign - showing Pearson's correlation coefficient (ρ).



(a) $PA_{5_{Di}}$ linear fit

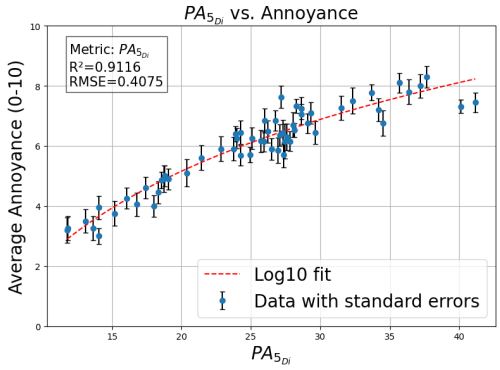


(b) $PA_{5_{Zwicker}}$ linear fit

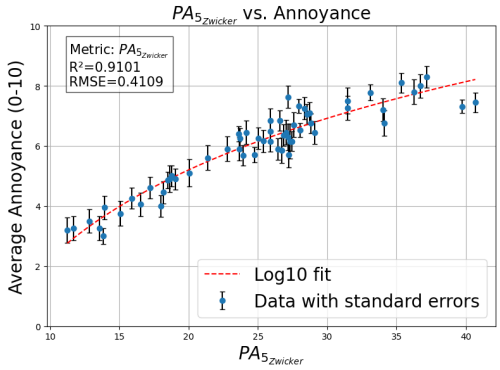


(c) PNLM linear fit

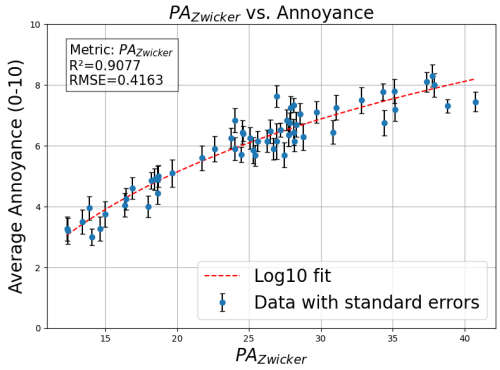
Figure A5. Best linear function fits using annoyance ratings.



(a) $PA_{5_{Di}}$ logarithmic fit

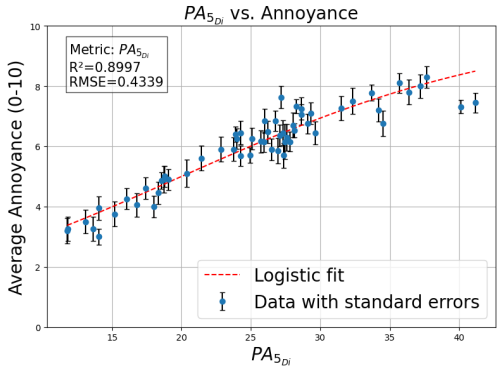


(b) $PA_{5_{Zwicker}}$ logarithmic fit

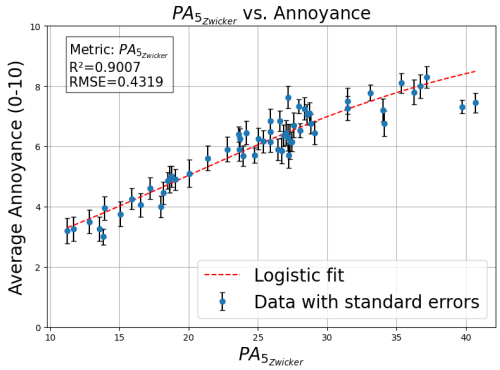


(c) $PA_{Zwicker}$ logarithmic fit

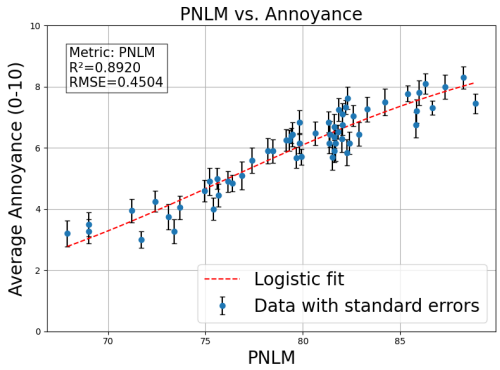
Figure A6. Best logarithmic function fits using annoyance ratings.



(a) $PA_{5_{Di}}$ logistic fit

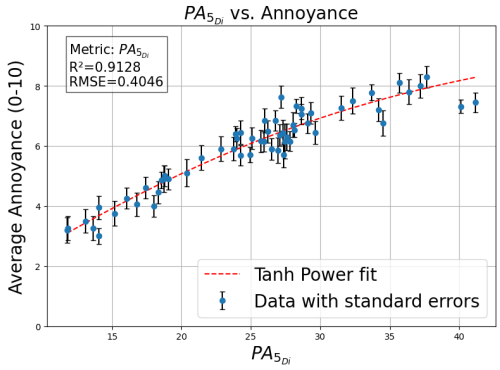


(b) $PA_{5_{Zwicker}}$ logistic fit

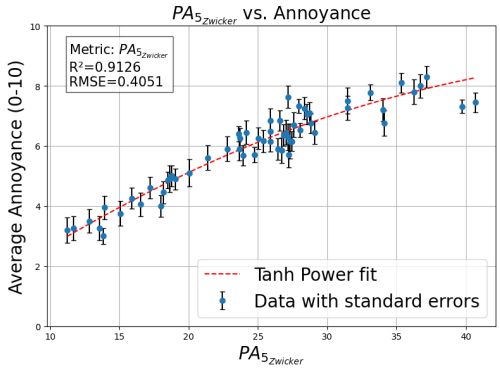


(c) PNLM logistic fit

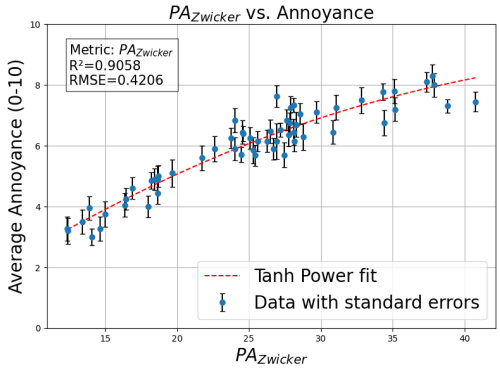
Figure A7. Best logistic function fits using annoyance ratings.



(a) $PA_{5_{Di}}$ tanh power fit

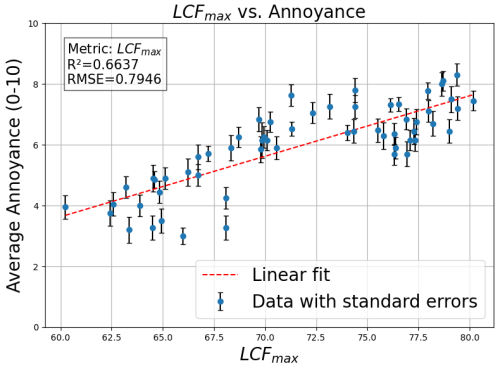


(b) $PA_{5_{Zwicker}}$ tanh power fit

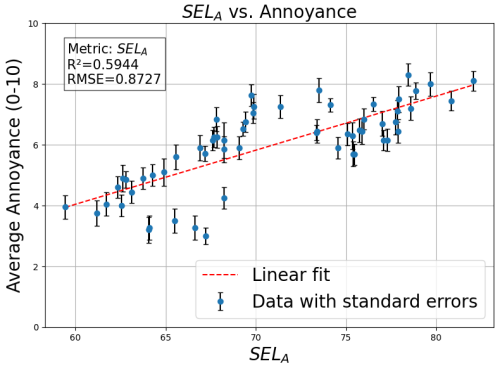


(c) $PA_{Zwicker}$ tanh power fit

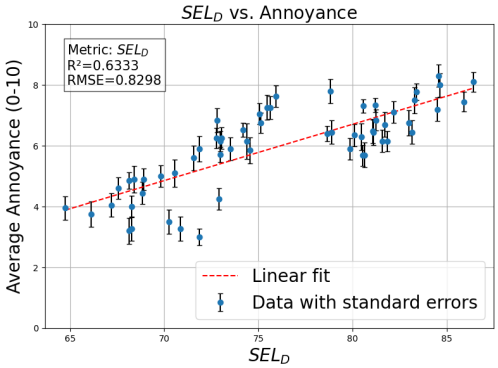
Figure A8. Best hyperbolic tangent power function fits using annoyance ratings.



(a) LCF_{max} linear fit

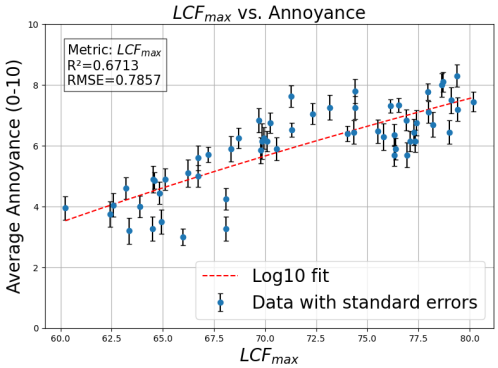


(b) SEL_A linear fit

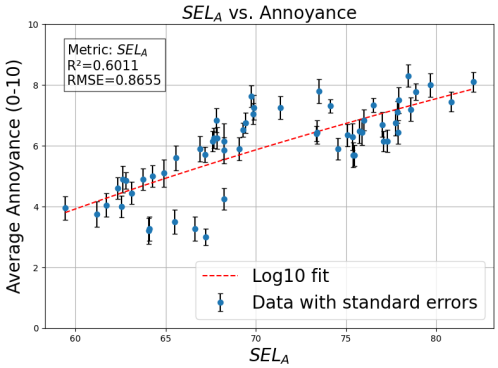


(c) SEL_D linear fit

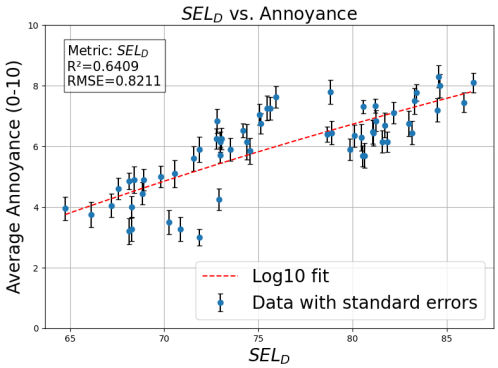
Figure A9. Worst linear function fits using annoyance ratings.



(a) LCF_{max} logarithmic fit

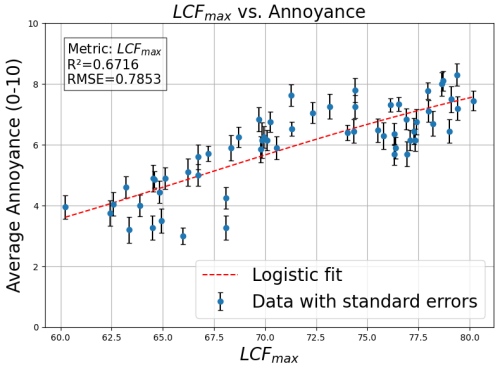


(b) SEL_A logarithmic fit

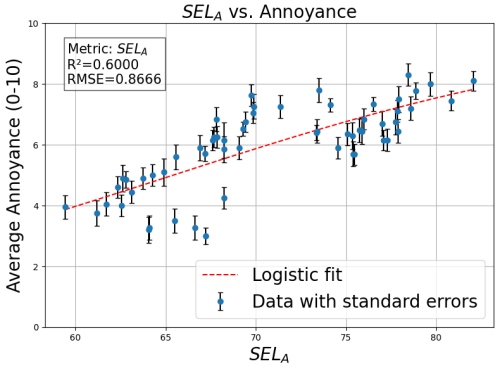


(c) SEL_D logarithmic fit

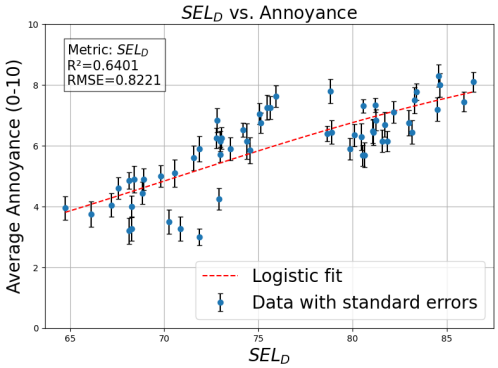
Figure A10. Worst logarithmic function fits using annoyance ratings.



(a) LCF_{max} logistic fit

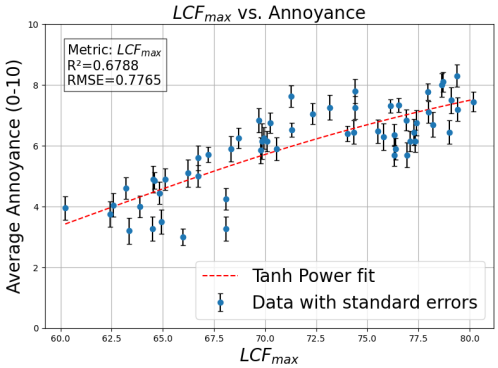


(b) SEL_A logistic fit

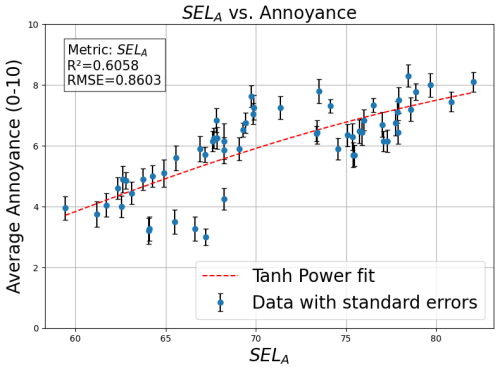


(c) SEL_D logistic fit

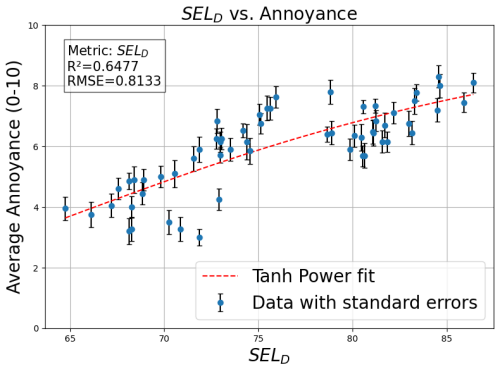
Figure A11. Worst logistic function fits using annoyance ratings.



(a) LCF_{max} tanh power fit

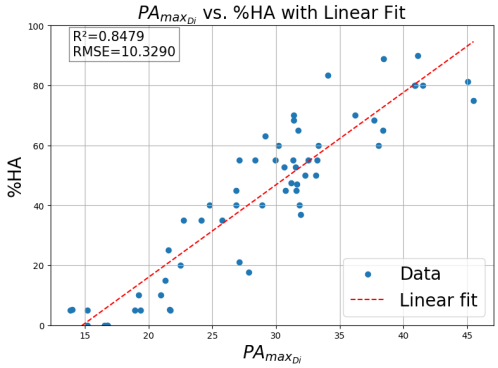


(b) SEL_A tanh power fit

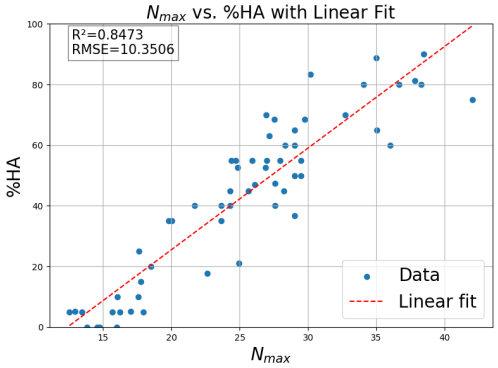


(c) SEL_D tanh power fit

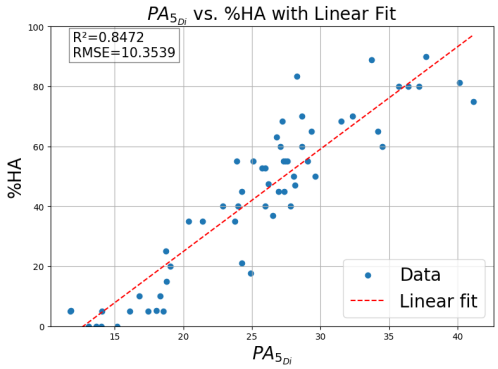
Figure A12. Worst hyperbolic tangent power function fits using annoyance ratings.



(a) $PA_{max_{Di}}$ linear fit

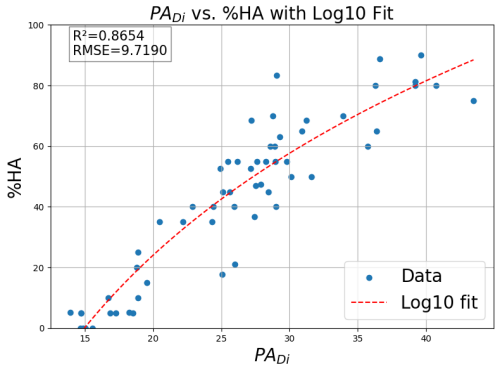


(b) N_{max} linear fit

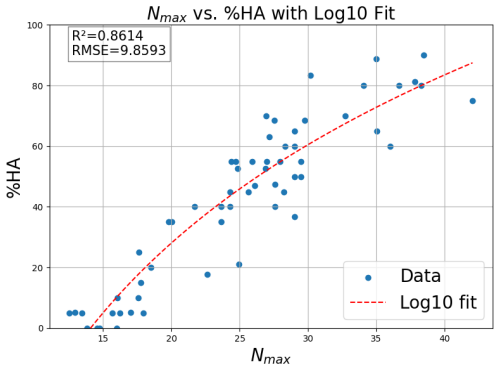


(c) $PA_{5_{Di}}$ linear fit

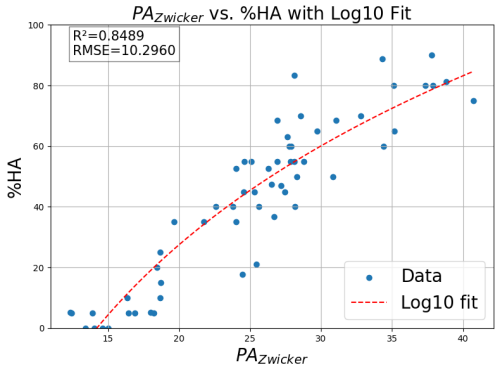
Figure A13. Best linear function fits using %HA.



(a) PA_{Di} logarithmic fit

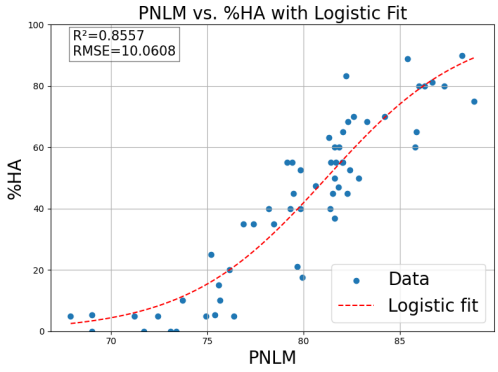


(b) N_{max} logarithmic fit

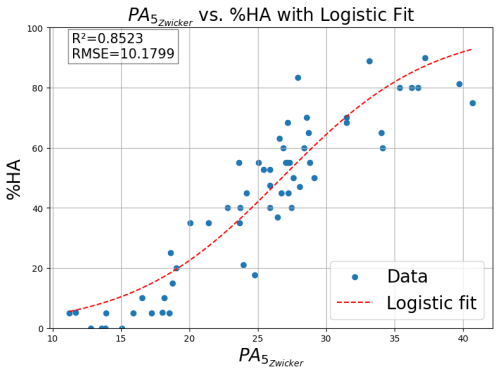


(c) $PA_{Zwicker}$ logarithmic fit

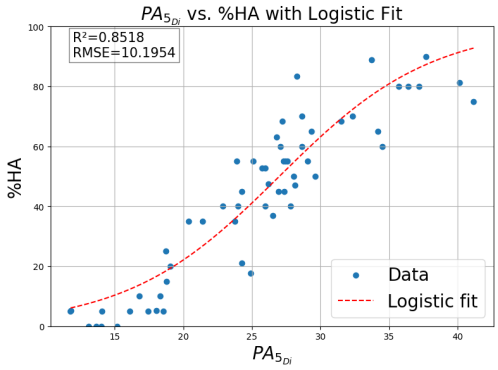
Figure A14. Best logarithmic function fits using %HA.



(a) PNLM logistic fit

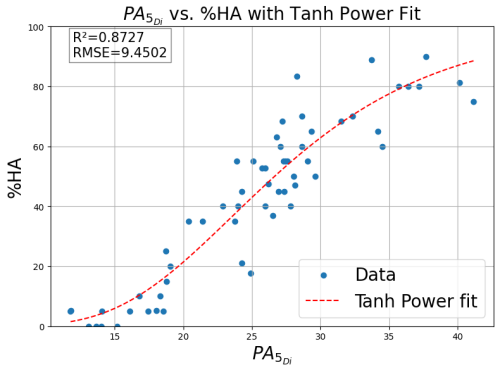


(b) $PA5_{Zwicker}$ logistic fit

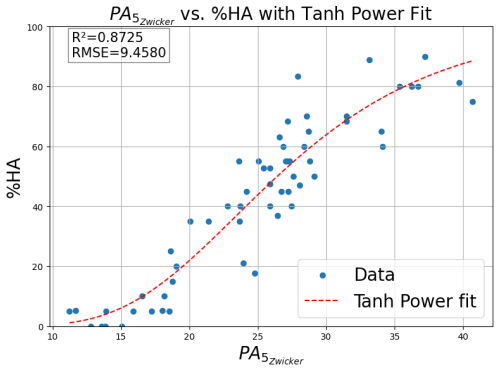


(c) $PA5_{Di}$ logistic fit

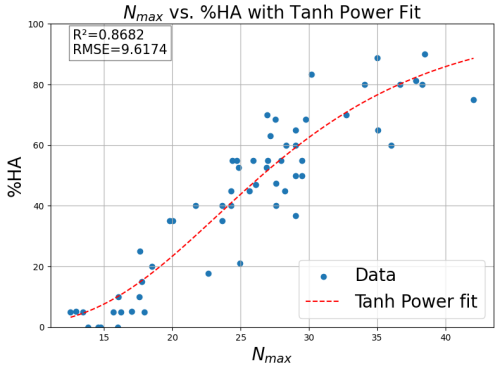
Figure A15. Best logistic function fits using %HA.



(a) $PA5_{Di}$ tanh power fit

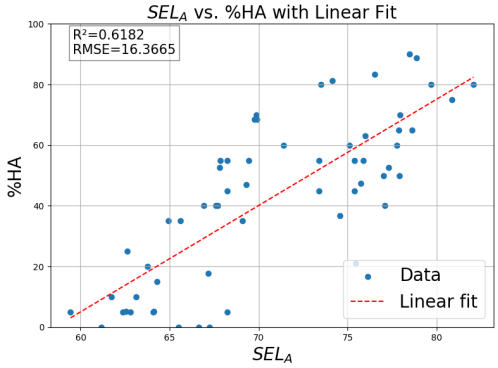


(b) $PA5_{Zwicker}$ tanh power fit

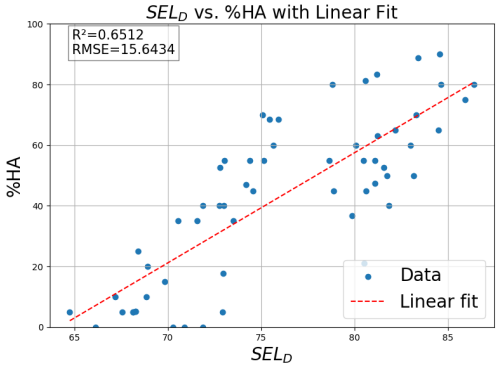


(c) N_{max} tanh power fit

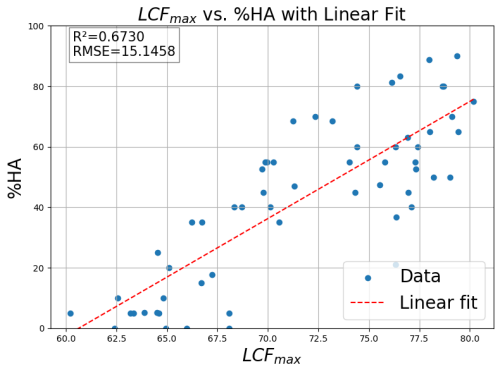
Figure A16. Best hyperbolic tangent power function fits using %HA.



(a) SEL_A linear fit

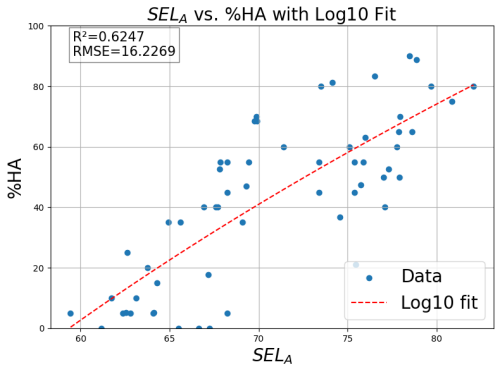


(b) SEL_D linear fit

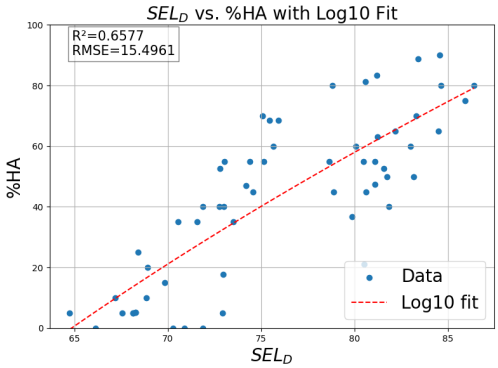


(c) LCF_{max} linear fit

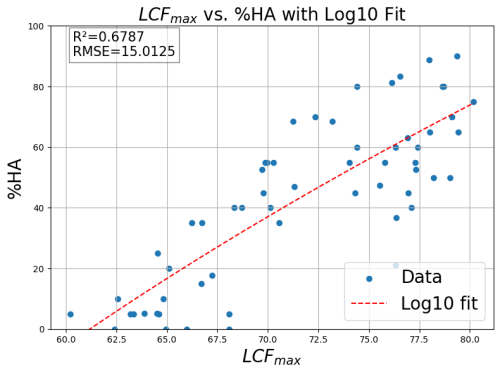
Figure A17. Worst linear function fits using %HA.



(a) LCF_{max} logarithmic fit

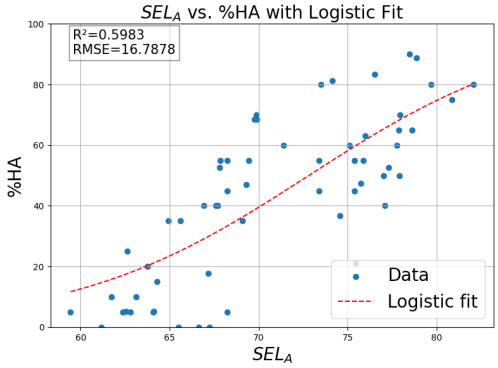


(b) SEL_D logarithmic fit

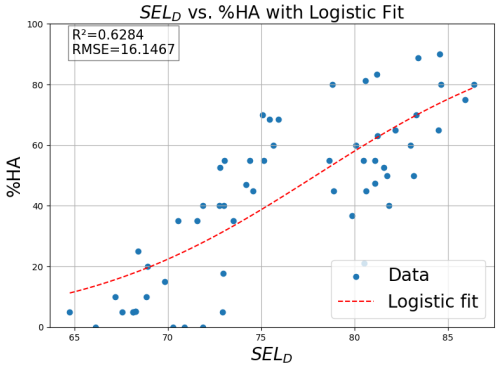


(c) LCF_{max} logarithmic fit

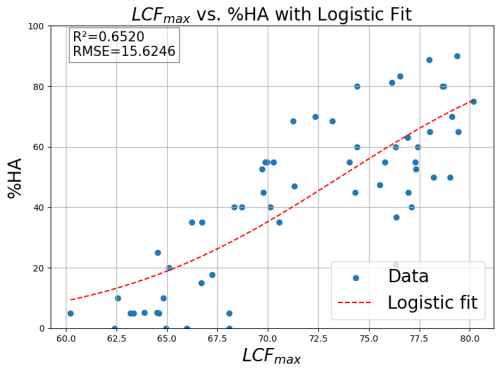
Figure A18. Worst logarithmic function fits using %HA.



(a) SEL_A logistic fit

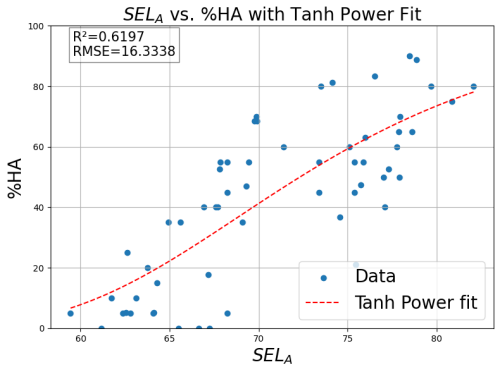


(b) SEL_D logistic fit

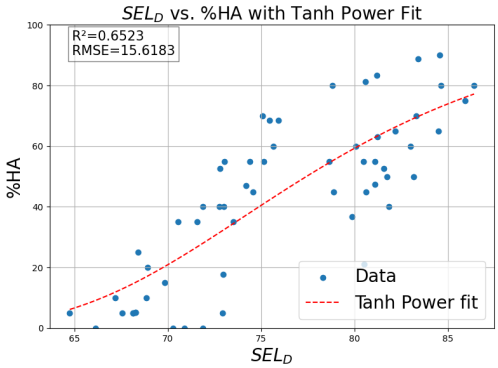


(c) LCF_{max} logistic fit

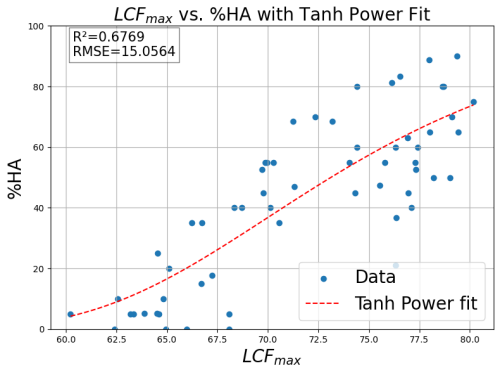
Figure A19. Worst logistic function fits using %HA.



(a) SEL_A tanh power fit



(b) SEL_D tanh power fit



(c) LCF_{max} tanh power fit

Figure A20. Worst hyperbolic tangent power function fits using %HA.

B. Details on CNN for Metric Predictions

The training histories of all the trained Convolutional Neural Networks can be found in Fig. A21 to Fig. A41. Most of them behave in quite a stable manner, especially from a certain epoch onward. The gap between the training and validation losses (represented by the MSE) is gradually decreased and the early stopping mechanism kicks in when convergence is acceptable and no improvement is observed - the suitable number of epochs with no improvement (or with improvements which are too small) after which training is stopped differs from metric to metric, perhaps because of a combination between the variability of each metric and the limited training data. Training can either stop before the maximum number of epochs is reached either if the validation set's MAE does not see enough improvement (at least of 0.25 dB), or if the validation MAE is smaller than the training MAE for multiple epochs in a row. Both are generally applied using a so-called "patience" parameter of at least 6 epochs. The custom callback mechanism ensures that the over/underfitting trade-off is such that the model learns the training data as much as possible while being able to generalize to a similar degree.

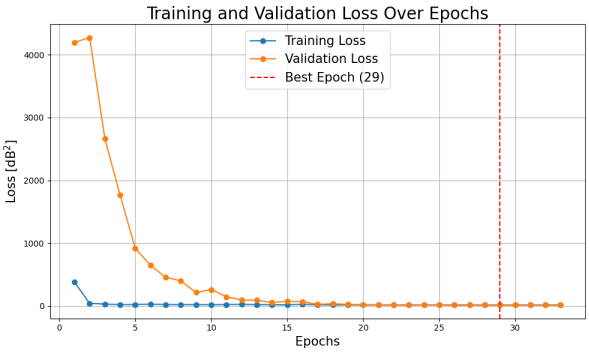
Moreover, during training, the learning rate is gradually reduced whenever the validation MAE does not decrease by at least 0.2 for two consecutive epochs. This measure helps with the model's convergence, since a smaller learning rate ensures that, during the model's weight updates, the gradient descent does not overshoot the direction which results in reducing the error. Every time the learning rate is reduced, this is done by a factor of 0.25. Additionally, the starting learning rate is 5e-5 and the minimum learning rate is constrained to 1e-9 - if it becomes too small, the weights can get stuck around the same values at each epoch and the model fails to converge (also known "as vanishing gradients").

Finally, an exhaustive overview of the used CNN architecture can be found in Table A1, which gives the layer types (in the order in which they appear in the model), their parameters and values, followed by the explanations of the working principle of the layer and of the parameters' influence. The presented architecture has a satisfactory robustness, although there is a certain variability observed among the obtained errors across all the metrics used in training the CNNs. The different complexities of the explored metrics, together with the limited available training data, could explain this observed variability. The variability also influences the final results obtained when predicting annoyance using the models observed in subsection A.3 using the CNN predictions.

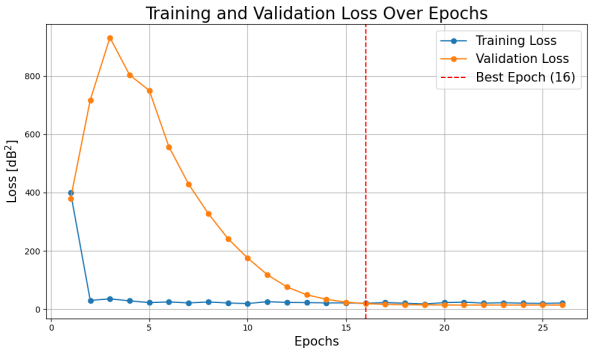
Table A1. Overview of the layers in the neural network model, including parameters and their explanations.

Layer Type	Parameters	Explanation
InputLayer	`input_shape=(319, 2401, 1)` for take-offs or `(199, 2401, 1)` for landings	Specifies the shape of the input data. Ensures the model knows what input dimensions to expect.
Conv2D	`filters=10, kernel_size=(3, 3), kernel_initializer=HeNormal, padding='same'`	Performs 2D convolution with 10 filters, each of size 3×3. `HeNormal` initializes weights for improved convergence. Padding='same' ensures the output size matches the input size.
BatchNormalization	No trainable parameters; normalizes the output of the previous layer.	Normalizes inputs across the batch to improve stability and accelerate training.
ReLU	No parameters.	Applies the Rectified Linear Unit activation, setting negative values to zero.

AveragePooling2D	`pool_size=(2, 2)`	Reduces the spatial dimensions of the input by averaging values over 2×2 regions.
Conv2D	`filters=20, kernel_size=(3, 3), kernel_initializer=HeNormal, padding='same'`	Similar to the previous `Conv2D` layer but with 20 filters for greater feature extraction.
BatchNormalization	No trainable parameters; normalizes the output of the previous layer.	Maintains consistent distributions in the activations of the previous layer.
ReLU	No parameters.	Applies the ReLU activation, adding nonlinearity to the model.
Dropout	`rate=0.3`	Randomly sets 30% of the neurons to zero during training to prevent overfitting.
AveragePooling2D	`pool_size=(2, 2)`	Further reduces spatial dimensions through 2×2 averaging.
Flatten	No parameters.	Converts the multidimensional tensor into a 1D vector for input into dense layers.
Dense	`units=24, kernel_regularizer=L2 (0.2)`	Fully connected layer with 24 neurons. L2 regularization penalizes large weights to prevent overfitting.
ReLU	No parameters.	Introduces nonlinearity using ReLU activation.
Dense	`units=12, kernel_regularizer=L2 (0.2)`	Fully connected layer with 12 neurons, applying L2 regularization.
ReLU	No parameters.	Adds nonlinearity using ReLU activation.
Dense	`units=6, kernel_regularizer=L2 (0.2)`	Fully connected layer with 6 neurons, regularized with L2.
Dense	`units=1`	Final output layer with 1 neuron for regression. Produces the predicted value.

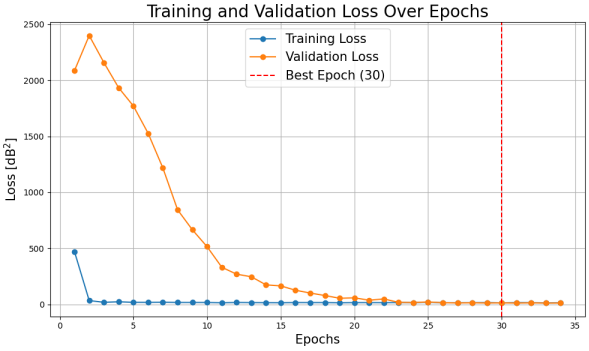


(a) Take-offs

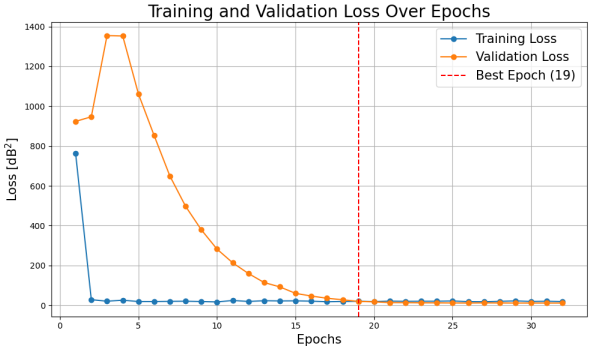


(b) Landings

Figure A21. $L_{A_{eq}}$ CNN training history.

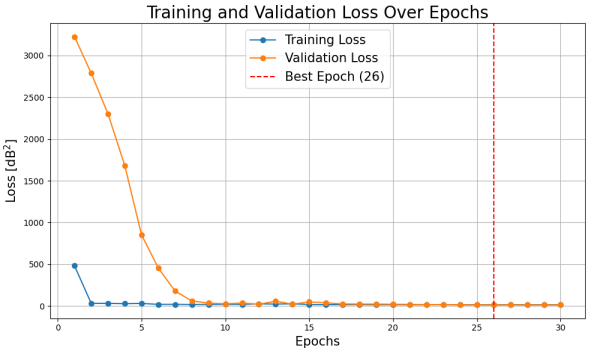


(a) Take-offs

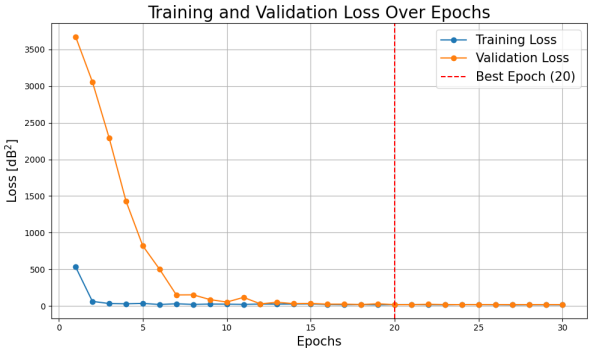


(b) Landings

Figure A22. $L_{D_{eq}}$ CNN training history.

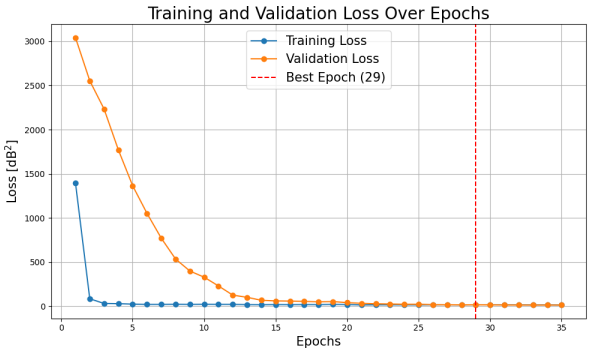


(a) Take-offs

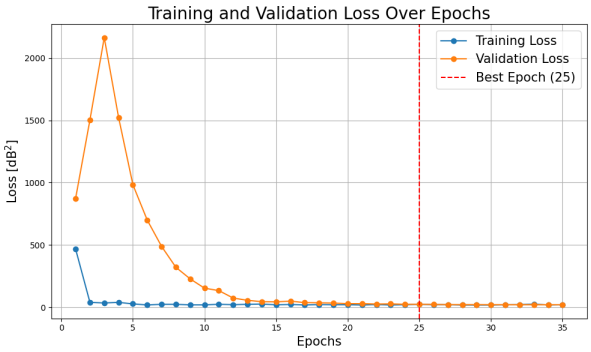


(b) Landings

Figure A23. LAF_{max} CNN training history.

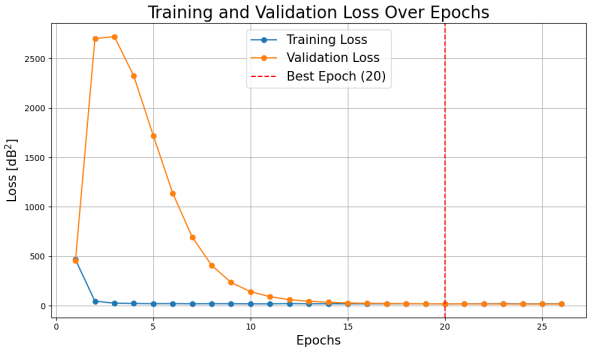


(a) Take-offs

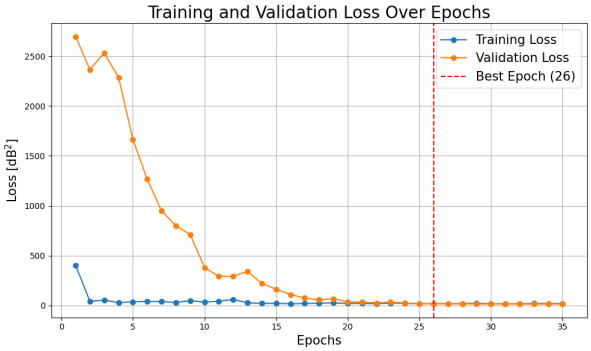


(b) Landings

Figure A24. LBF_{max} CNN training history.

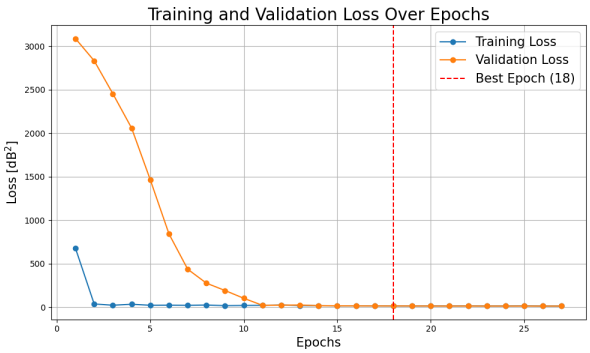


(a) Take-offs

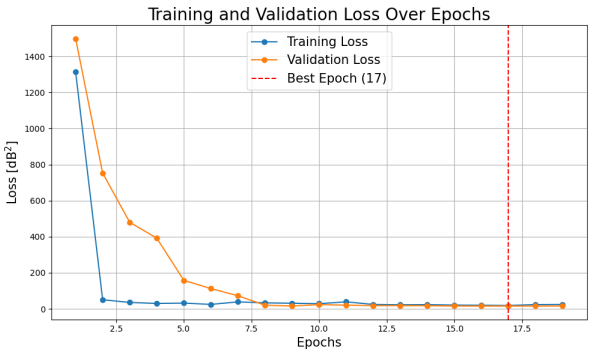


(b) Landings

Figure A25. LCF_{max} CNN training history.

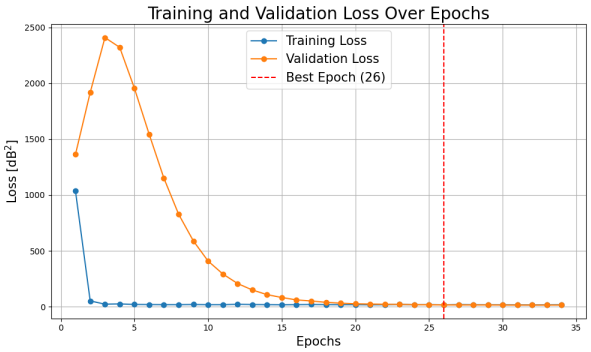


(a) Take-offs

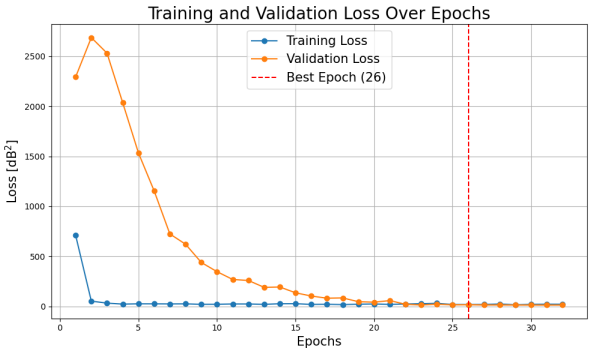


(b) Landings

Figure A26. LDF_{max} CNN training history.

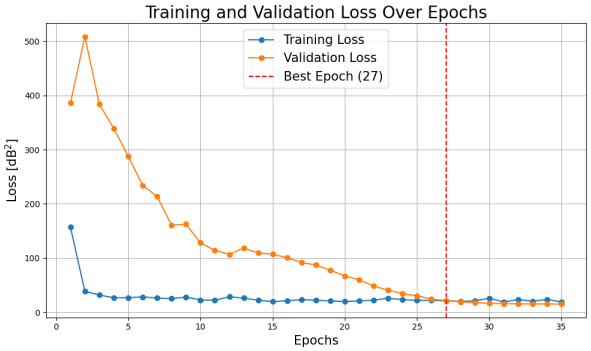


(a) Take-offs

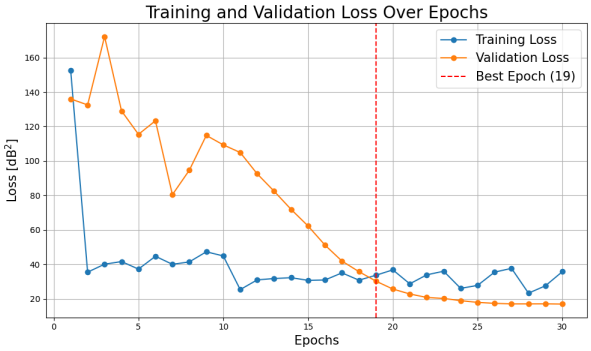


(b) Landings

Figure A27. LZF_{max} CNN training history.

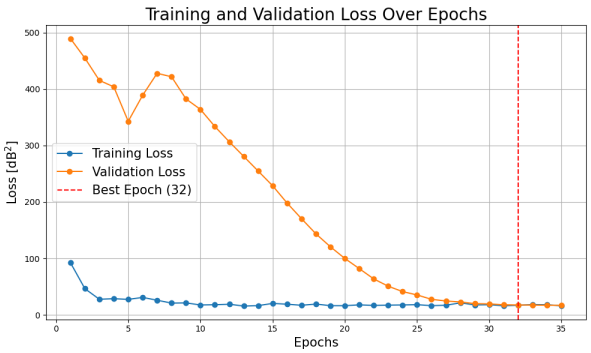


(a) Take-offs

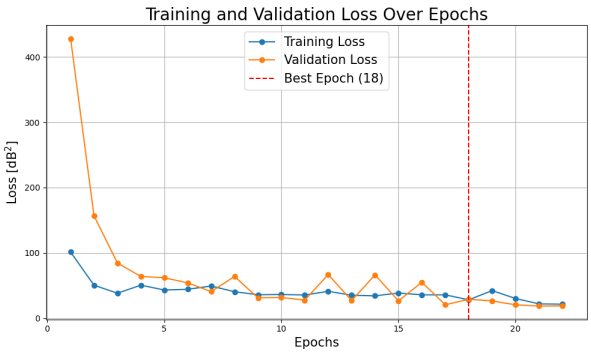


(b) Landings

Figure A28. L_{max} CNN training history.



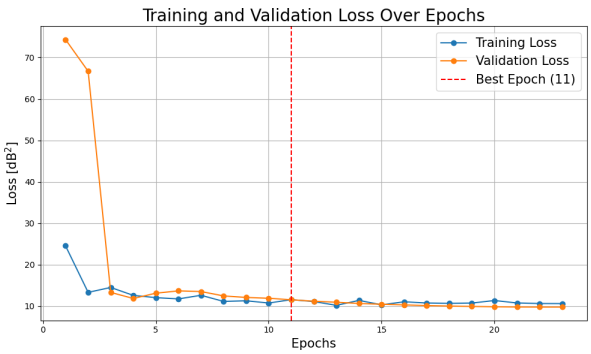
(a) Take-offs



(b) Landings

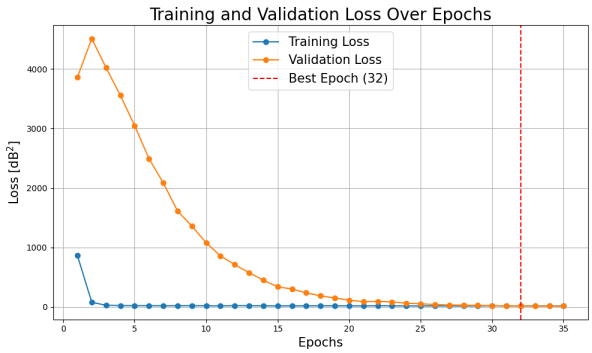
Figure A29. L_5 CNN training history.

(a) Take-offs

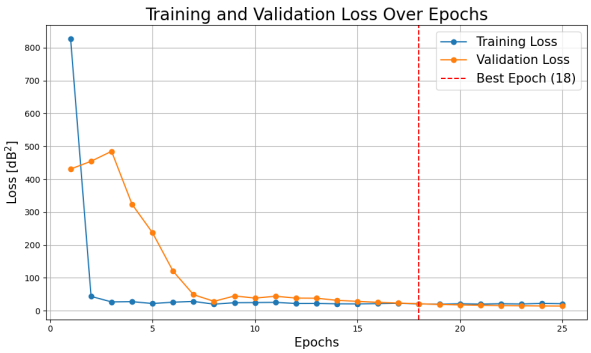


(b) Landings

Figure A30. L_{std} CNN training history.

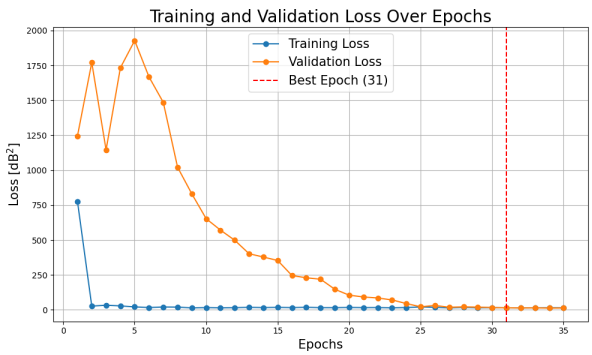


(a) Take-offs

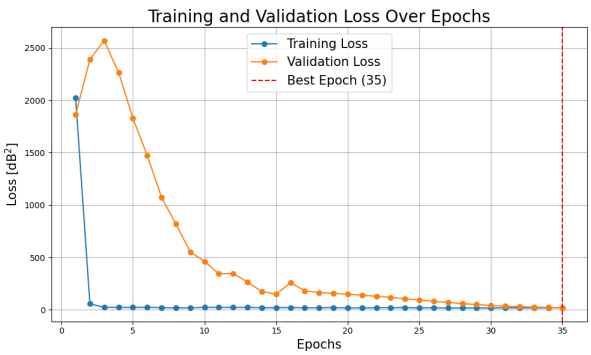


(b) Landings

Figure A31. PNLM CNN training history.

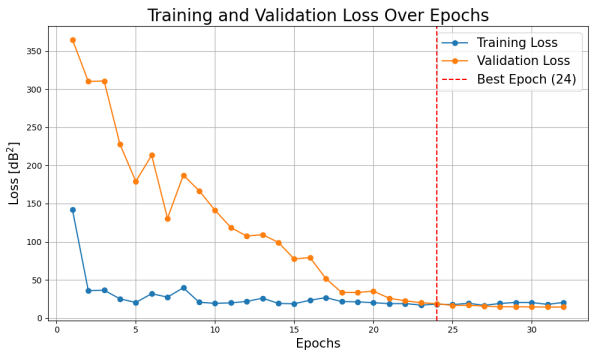


(a) Take-offs

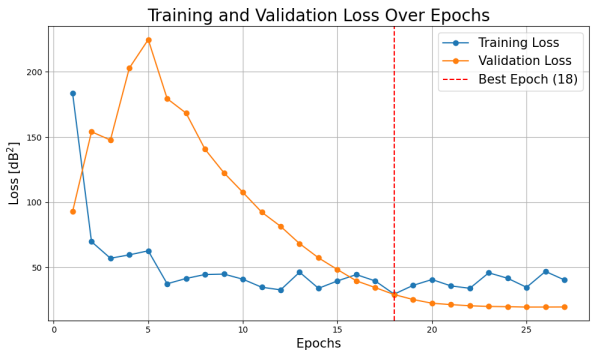


(b) Landings

Figure A32. PNLTM CNN training history.

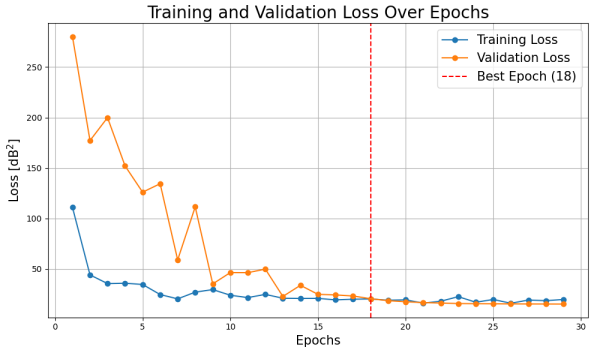


(a) Take-offs

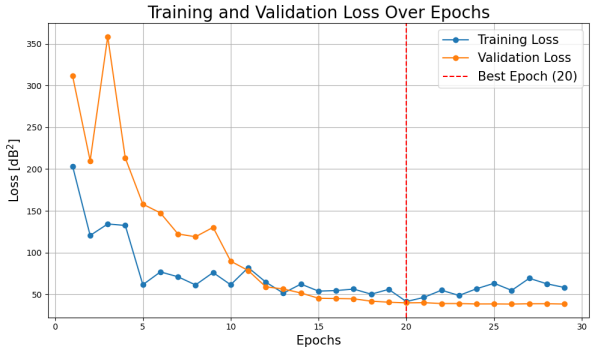


(b) Landings

Figure A33. $PA_{5_{Di}}$ CNN training history.

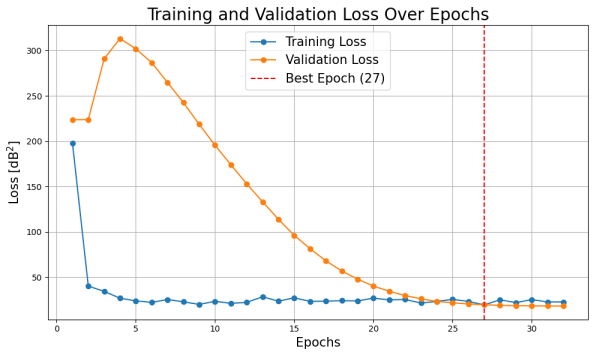


(a) Take-offs

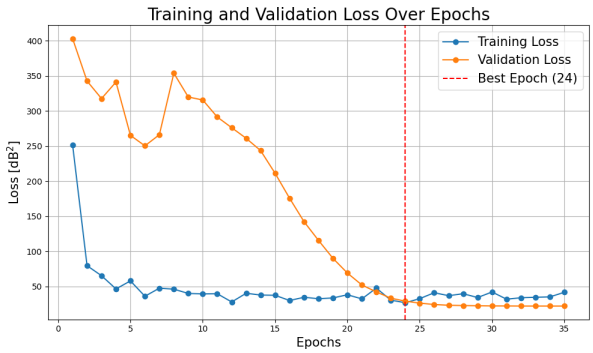


(b) Landings

Figure A34. $PA_{5_{More}}$ CNN training history.

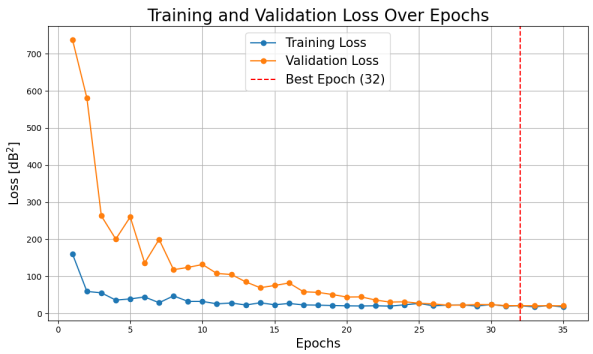


(a) Take-offs

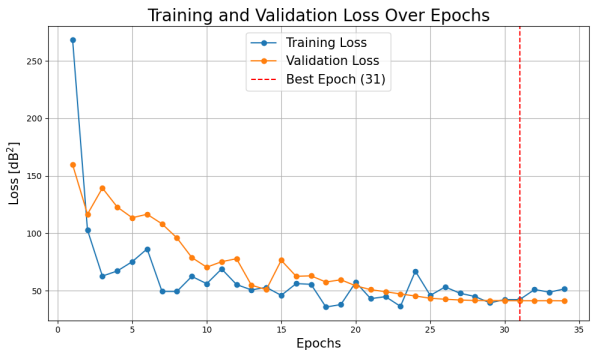


(b) Landings

Figure A35. $PA_{5_{Zwicker}}$ CNN training history.

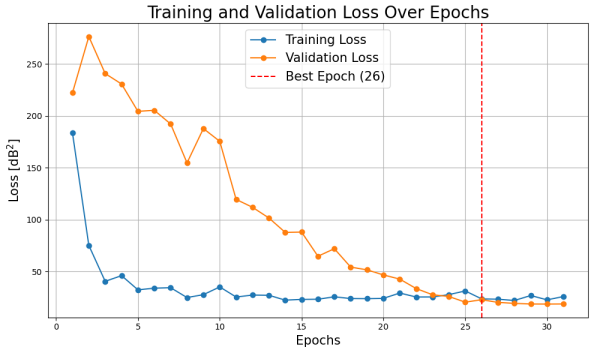


(a) Take-offs

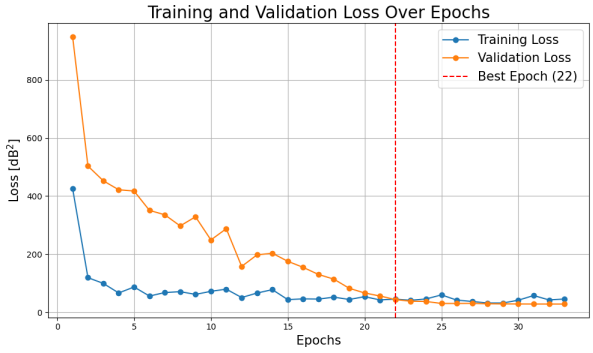


(b) Landings

Figure A36. $PA_{max_{Di}}$ CNN training history.

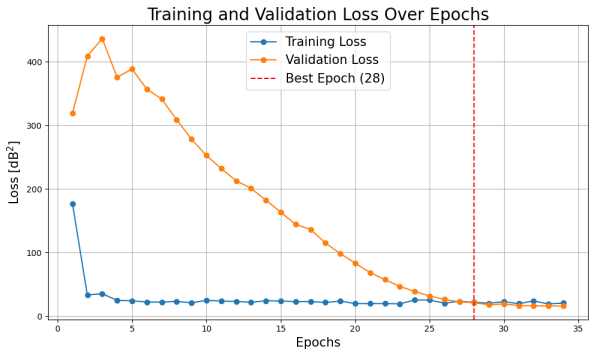


(a) Take-offs

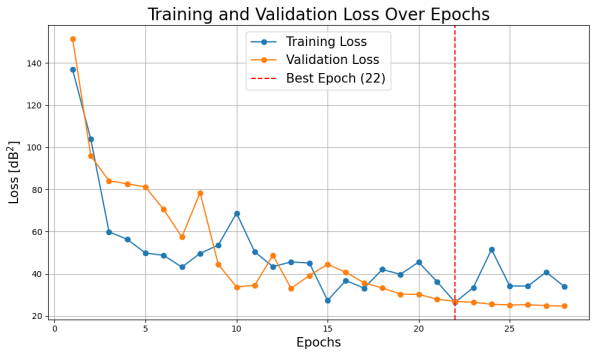


(b) Landings

Figure A37. $PA_{max_{Zwicker}}$ CNN training history.

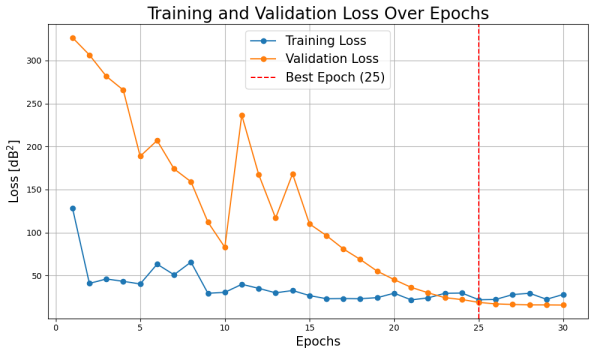


(a) Take-offs

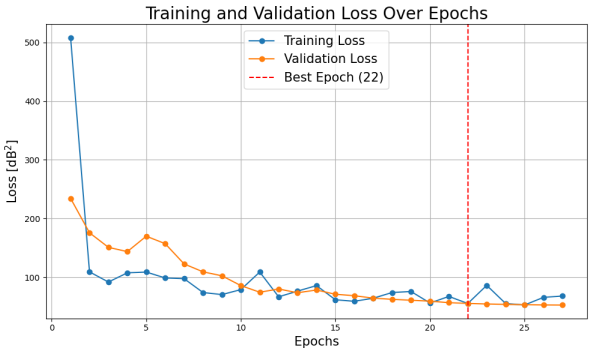


(b) Landings

Figure A38. PA_{Di} CNN training history.

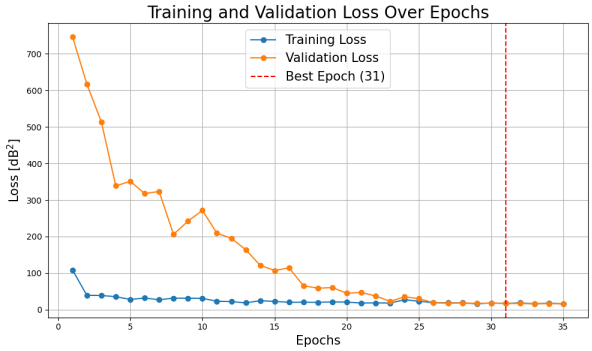


(a) Take-offs

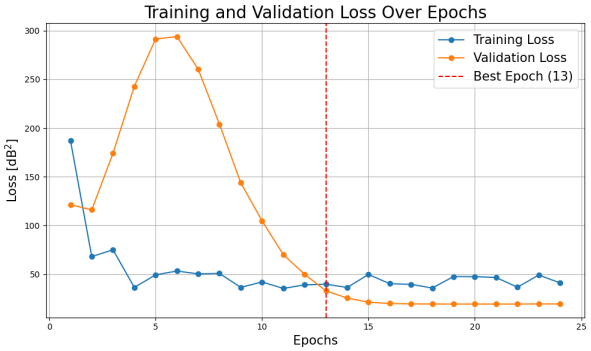


(b) Landings

Figure A39. PA_{More} CNN training history.

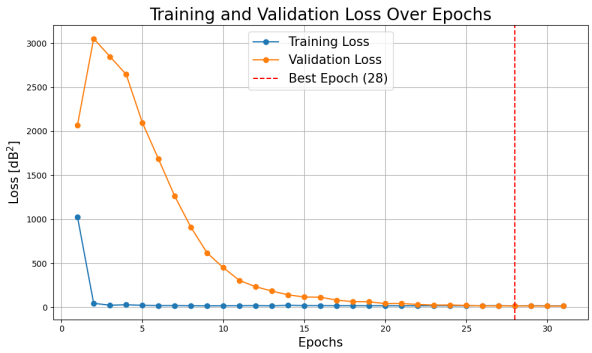


(a) Take-offs

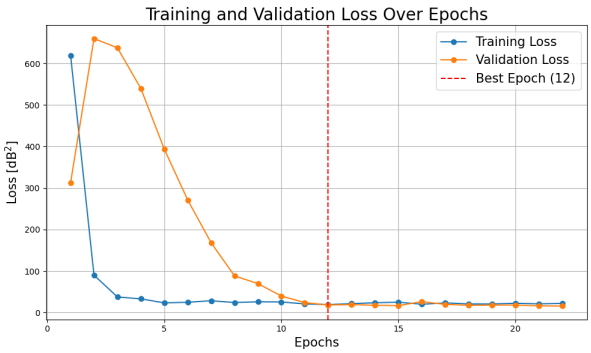


(b) Landings

Figure A40. $PA_{Zwicker}$ CNN training history.



(a) Take-offs



(b) Landings

Figure A41. SEL_D CNN training history.

C. Details on Annoyance Prediction Models

The annoyance prediction models were initially fitted on two combinations of four metrics from the pool of metrics which were strongly correlated to the annoyance score ratings obtained from the listening experiment. For the logistic, hyperbolic tangent power, hyperbolic tangent, sigmoid and softsign functions, linear combinations of the four metrics were used as arguments. For these functions, the main considerations were the trade-off between the training and testing performance metrics (MAE and MSE) observed on the experiment aircraft flyover recordings. The weight tuning for these models was framed as an optimization problem using Python's "scipy.optimize" module. It should once more be noted that, because of the small dataset of labeled flyovers available, the obtained results might change quite drastically when evaluating a different, possibly larger dataset. The stratified learning approach ensured that the small amount of data was distributed such that each of the annoyance rating intervals was represented by the training and testing datasets as similarly as possible. The main purpose of this methodology was to investigate whether applying a Machine Learning framework by using multiple metrics could lead to better annoyance quantification than employing the currently used aircraft noise certification metrics.

As far as the random forest and SVM models are concerned, the process of selecting the final model parameters was more elaborate, due to the more complex nature of these models. This involved a grid search on multiple values of the most important parameters of each model, respectively, as emphasized in the main chapters of this work. A five-fold cross validation was also performed on each parameter combination. The best model was identified by considering the average MSE on the 5 test sets resulting from each cross-validation step in the parameter grid. In Table A2, the overview of the values investigated during this tuning process is provided. Moreover, brief explanations of the models' parameters and their meaning, for a better understanding of the mechanisms governing these models, are included. The same stratified learning approach was used for the SVM and RF (matching the random seeds used for training the previous functions), such that the evaluated performance metrics in the main chapters correspond to precisely the same data points - and, consequently, implying that all models were trained on the same dataset.

Table A2. Hyperparameters used in grid search, along with descriptions and their effects (values used in the final models are highlighted in bold).

Algorithm	Hyperparameter (Values Explored)	Description and Effects
Random Forest A Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions (by averaging for regression or majority voting for classification) to improve accuracy and reduce overfitting.	`n_estimators` (100, **200**, **1000**)	Number of trees in the forest. Increasing this typically improves performance but increases computation time.
	`max_depth` (**None**, 5, 10)	Maximum depth of each tree. Limits the growth of trees. A smaller value reduces overfitting but may underfit.
	`min_samples_split` (None, **5**, 10)	Minimum number of samples required to split a node. Higher values prevent overfitting by requiring more samples at each split.
	`min_samples_leaf` (None, **5**, 10)	Minimum number of samples required to form a leaf node. Higher values create simpler trees and reduce overfitting.

Algorithm	Hyperparameter (Values Explored)	Description and Effects
SVM Support Vector Machines (SVMs) aim to find the hyperplane that best separates data into classes (or approximates continuous values in regression). In the case of regression, the algorithm fits a regression function within a margin of tolerance defined by ε.	`C` (0.1, 1, **10**, **50**)	Regularization parameter. Controls the trade-off between achieving a low error on training data and minimizing model complexity. Higher values focus on reducing training error but risk overfitting.
	γ (**'auto'**, 'scale', **0.1**, 0.01, 0.001)	Kernel coefficient for 'rbf' and 'poly' kernels. 'scale' adjusts gamma based on feature variance, while 'auto' uses the number of features directly. Lower values create smoother decision boundaries; higher values focus on individual points, potentially overfitting.
	`kernel` (**'rbf'**, 'linear')	Specifies the kernel type for the model. 'rbf' captures non-linear relationships, while 'linear' assumes linear separability.
	ε (0.01, 0.1, 0.2, **0.5**, 1.0)	Defines a margin of tolerance for error in the model's predictions. Smaller values aim for higher precision but may lead to overfitting.

B. Literature Review

A. Introduction & Motivation

Aircraft noise is the main source of annoyance for communities living in the vicinity of airports [44] and it elicits a higher degree of annoyance than road and rail noise [45] (as can be seen in Fig. B1). Generally speaking, noise annoyance poses threats to people's physical and psychosocial/mental health [8–10]. With the air traffic in a continuous growth over the last decades [2], people are more annoyed by aircraft noise nowadays than they were 30 years ago [3]. Moreover, the rise in popularity of Unmanned Aerial Vehicles (UAV) has led to an increased interest towards researching their perceived annoyance. It has been shown that their noise exhibits particular characteristics which contribute to their perceived annoyance and which set UAV's apart from conventional aircraft in terms of the mechanism responsible for causing most of the annoyance [26, 46]. Still, the topic of UAV noise-induced annoyance requires further research in the future [27, 47].

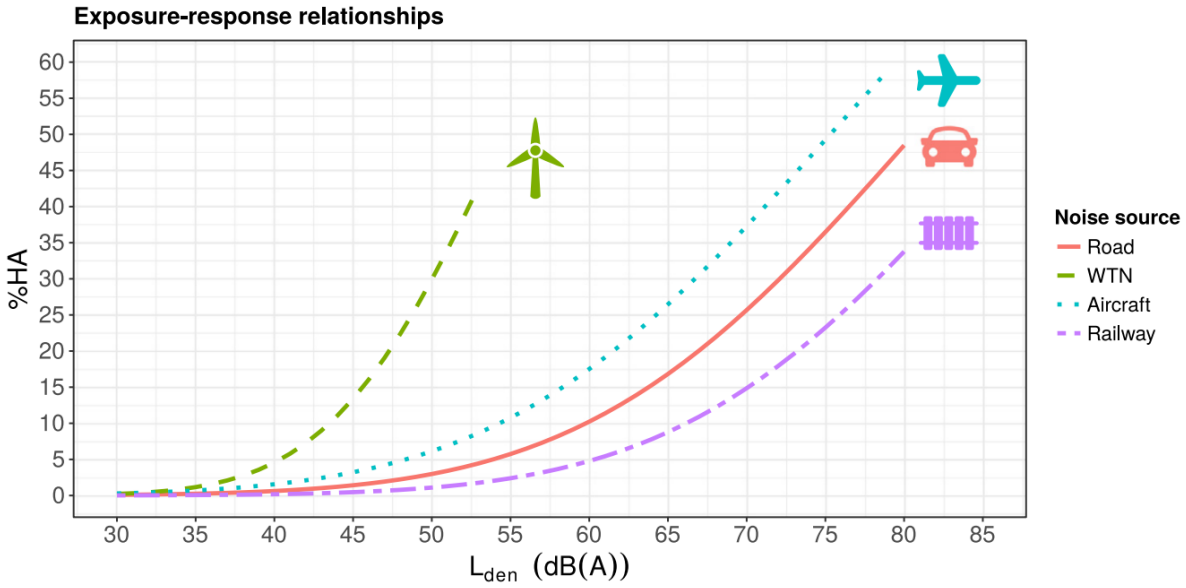


Figure B1. Annoyance curves for aircraft, road and rail noise, respectively [12].

In spite of the noise-reducing engine technologies which have been implemented in aircraft since the 1970s [5, 6], there still remains the airframe, which can still cause significant noise levels [48]. In addition, it seems that currently employed metrics for aircraft noise certification fail to fully capture the annoyance response [15]. Traditional energy-based metrics such as the Sound Pressure Level (SPL) have been augmented via more complex metrics, such as EPNL and the A-weighted sound level [14], which take into account (to some extent) the human perception of noise through spectral irregularities, tonality etc. However, studies have shown that this type of metrics cannot explain significant portions of the variance in annoyance caused by aircraft noise, as reported by test subjects [15, 16]. These findings lead to the need to incorporate aspects related to the sensitivity of the human auditory system into the environmental noise metrics in order to properly assess aircraft noise-induced annoyance [49]. On top of that, there is a correlation between non-acoustical factors and the reported noise-induced annoyance, such as age, gender, background, individual noise sensitivity [17, 18], visual effects [19] etc.

In order to tackle the limitations imposed by conventional sound metrics, one needs to turn their attention towards psychoacoustics, the field of study "concerned with the relationships between the physical characteristics of sounds and their perceptual attributes" [20]. Research in this field has bridged a significant portion of the knowledge gap concerning the subjective perception of sound through the emergence of so-called sound quality metrics [21], which are computed based on the human's auditory system's characteristics, such as its varying

sensitivity to different frequency bands. A further breakdown of these metrics and of the psychoacoustic annoyance models that have been created based on them is made in subsequent sections of this literature review.

Several promising attempts to generating predictive models for noise-induced annoyance have been identified in literature, combining some of the currently used metrics in certification with psychoacoustic metrics, within Machine Learning frameworks. Some studies are concerned with the more general traffic or urban noise sources, such as [23], while others were targeted specifically at aircraft noise, like [24]. The main limitation concerning predictive models identified in literature stems from the lack of large and diverse enough datasets available for training and testing.

Such a predictive model focused on aircraft noise with extremely good accuracy and solid generalization performance would prove to be a crucial tool during the design and certification of future aircraft configurations. With a better understanding of the aircraft sound characteristics which determine annoyance, together with a robust model incorporating these characteristics, the scientific community could make a great leap towards the mitigation of aircraft noise-induced annoyance on the communities most affected by it. Still, there is a lot of ongoing debate on the selection of the most suitable metrics for predicting noise annoyance, combined with the fact that there is such a strong variability in the annoyance responses [11], as seen in Fig. B2.

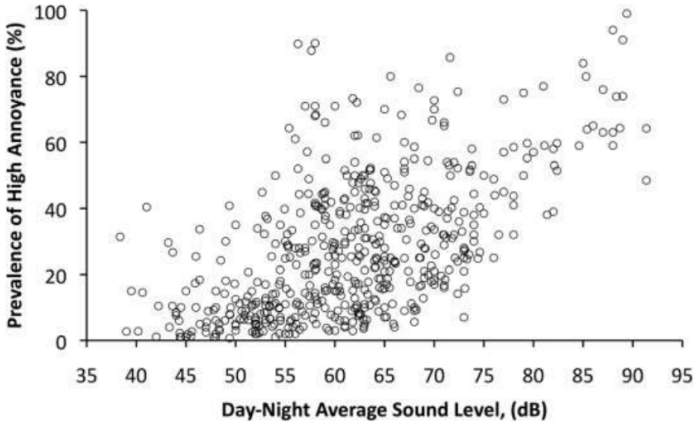


Figure B2. "Illustration of variability in annoyance prevalence rates as a function of cumulative noise exposure. Each point represents an estimate of the prevalence of high annoyance at a single interviewing site" [3].

B. Noise Quantification Metrics

This chapter is concerned with the investigation of the most important noise metrics used in aircraft certification and in the assessment of environmental noise, as well as of the metrics derived from the field of psychoacoustics. It should be noted that emphasis was put mostly on single-event metrics. Multi-event metrics require more complicated procedures for experiments in order to accurately simulate complex psychological processes occurring in real life. Moreover, as the scope of this research is to model psychoacoustic annoyance based on purely acoustical data combined with subjective perception metrics, it was considered superfluous to put a lot of effort into multi-event metrics. Doing that would have introduced a lot of variability which cannot be captured with the available resources for this research, and possibly affecting the reliability of future results.

1. Traditional Metrics Used in Certification and Assessment of Environmental Noise

According to International Standards [14], different categories of aircraft can be certified in terms of noise via different metrics. There are lots of noise metrics employed in various contexts, many of which were created such that they account for subjective effects of aircraft noise on human beings. Nevertheless, the

extent to which these metrics cover the subjective spectrum of the human response to noise is very limited. An overview of the main metrics used in certification and environmental evaluation is given below, using the categorization employed by More in his doctoral thesis [15].

Weighted Sound Pressure Level based ratings

- A and C-weighted Sound Pressure Level (L_A and L_C):
  These weighted metrics are based on equal loudness contours at 40 phon (the concept of loudness is further explained in the next section) and they take into account the human ear's sensitivity in different octave bands; their general formula is given by $10log_{10}\Sigma_i \left(\frac{w_i \cdot p_i}{p_0}\right)^2$, where w_i are the weighting coefficients (different for both metrics), p_i is the average pressure in each octave band and p_0 is a reference pressure of 20 μPa. In spite of L_A being widely used in community noise measurement, it is very limited when it comes to quantifying the impact of the noise on said communities due to the lack of importance assigned to frequencies below 400 Hz and above 4000 Hz [50]. This is of particular importance because it is believed that most of the energy content of aircraft noise is located within the lower frequency bands [51].

Average energy level based ratings

- Average A-weighted Sound Level ($L_{A_{eq}T}$):
  It is the average A-weighted sound level measured in dB(A) over a fixed period of observation and it is calculated via Eq. 8, where L_i is the sound level in dBA, τ_i is a penalty factor dependent on day or night time and T is the averaging time (typically 15h for day-time and 9h for night-time). Even though this metric has been widely used as a measure of aircraft noise impact, it was shown that this metric is not enough on its own to capture all of the variance in annoyance to railway noise responses [52].
- Maximum Noise Level (L_{max}):
  It is the maximum noise level in dB over the observation time, as can be seen in Fig. B3. Later, $L_{A_{max}}$ was proposed, or its A-weighted counterpart, showing better correlation to annoyance responses [53].
- Sound Exposure Level (SEL or L_{AX}):
  This metric (measured in dB) is usually used for assessing environmental noise from different sources including aircraft. It is calculated via Eq. 9, where p_{ref}= 20 μPa and p is the sound pressure. The A and C-weighted counterparts can be obtained by substituting p with p_A and p_C, respectively.

$$L_{A_{eq}T} = 10log_{10}\left(\frac{1}{T}\right)\left[\Sigma_i\left(\tau_i \cdot 10^{(0.1\cdot L_i)}\right)\right] \tag{8}$$

$$SEL = 10\log_{10}\int_{t_1}^{t_2}\frac{p^2(t)}{p_{\text{ref}}^2}dt \tag{9}$$

Average level and time of day based metrics

- Day-Night Average Sound Level (DNL or L_{dn}):
  DNL (or L_{dn}) has been long used in the US as a health and welfare criterion, while L_{den} (or DENL, where E stands for evening) has been introduced more recently in the European Union for assessing community noise impact. The metric is very similar to A-weighted sound pressure level, with added penalties for evening and night-time. L_{den} is calculated as shown in Eq. 10.
  A normalized version of DNL (NDNL) was proposed [54] in order to bypass some of the limitations of DNL associated with neglecting the effect of pure tones and isolated loud events [15].

$$L_{\text{den}} = 10\log_{10}\left[(1/24)\left[12\left(10^{L_d/10}\right)+3\left(10^{(L_e+5)/10}\right)+9\left(10^{(L_n+10)/10}\right)\right]\right] \tag{10}$$

Loudness Based Metrics

- Stevens' Loudness:
  Loudness can be regarded as the sensation produced by the human auditory system as a response to the sound level, accounting for both the spectral content and the SPL. Hence, it falls under the category of subjective metrics. Multiple other metrics are derived using Stevens' loudness as a starting point. As

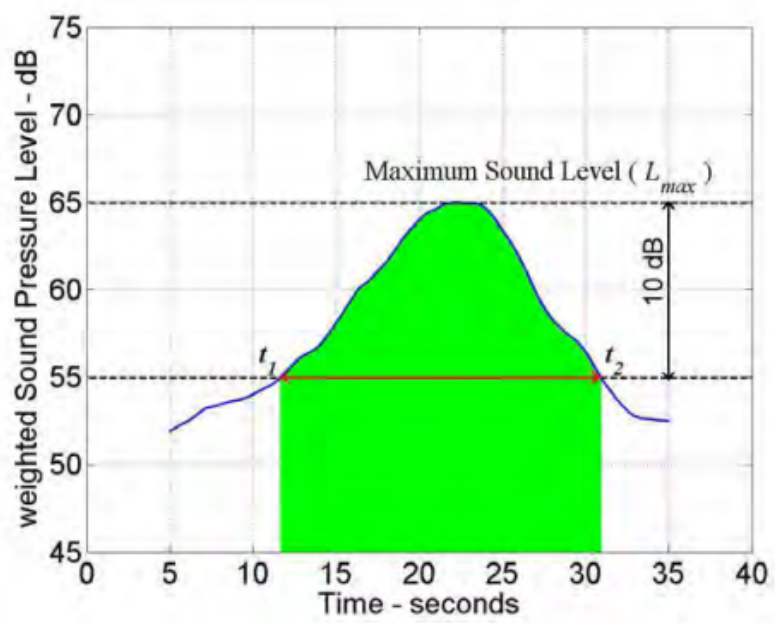


Figure B3. A-weighted sound pressure level time history of an aircraft noise event [15].

will be seen in the next section, there is a difference between the loudness scale (measured in "sone"), presented here, and Zwicker's loudness level (measured in "phon"), but they are very strongly related to each other. Stevens' loudness is given by Eq. 11, where I is the sound intensity and k and p are constants which depend on units and the type of sound stimulus, respectively.

- Loudness Level Weighted Sound Exposure Level (LLSEL):
  Unlike SEL, this metric puts emphasis on the low-frequency content of a stimulus and on its impulsiveness [55] and it is based on equal-loudness-level-contours, as depicted in Fig. B4. It can be computed via Eq. 12, where $L_{L_{ij}}$ is the phon level corresponding to the i-th one-third octave band and j-th time sample. LLSEL is thus also a subjective metric.
- Perceived Noise Level (PNL):
  This metric (measured in PNdB) is derived from Federal Aviation Regulations [56], and equal noisiness curves are used to convert the third-octave SPL values to noise levels through so-called 'noy' values ("the perceived noisiness of a one-third octave band sound pressure level in a given spectrum" [14]), as shown in Eq. 13 and Eq. 14, where n is the noy value corresponding to each frequency band from 50 Hz to 10 kHz and sound pressure level, n_{max} is the maximum of all the "noy" values, and k is the index of third-octave bands from 50 Hz to 10 kHz. Hence, it is obvious that this is also a subjective metric, as it is based on the human perception of sound.
- Tone-corrected Perceived Noise Level (PNLT):
  This metric (measured in TPNdB) is simply obtained by applying some corrections for the spectral irregularities to the PNL (or tonal correction factors, where sound pressure levels in third-octave frequency bands from 80 Hz to 10 kHz are considered), as described by the FAA [56]. The calculation procedure involves quite a lengthy iterative process, which can be seen in full in [15].
- Effective Perceived Noise Level (EPNL):
  Once again, according to the FAA procedure [56], the EPNL is obtained starting from PNLT, to which a correction for the duration of the flyover is applied. The correction factor is obtained via Eq. 15, where PNLTM is the maximum value of the PNLT time history. Finally, EPNL is obtained through Eq. 16.

$$L = k \cdot I^p \tag{11}$$

$$LLSEL = 10\log_{10}\left(\sum_{j}\sum_{i}\left(10^{\frac{L_{Lij}}{10}}\right)\right) \tag{12}$$

$$N_t = n_{\max} + 0.15\sum_{i=1}^{k}(n_i - n_{\max}) \tag{13}$$

$$PNL = 40 + \frac{10\log_{10} N_t}{\log_{10} 2} \tag{14}$$

$$D = 10\log_{10}\left[\sum_{k=0}^{2d}\left(10^{\frac{PNLT(k)}{10}}\right)\right] - PNLTM - 13 \tag{15}$$

$$EPNL = PNLTM + D \tag{16}$$

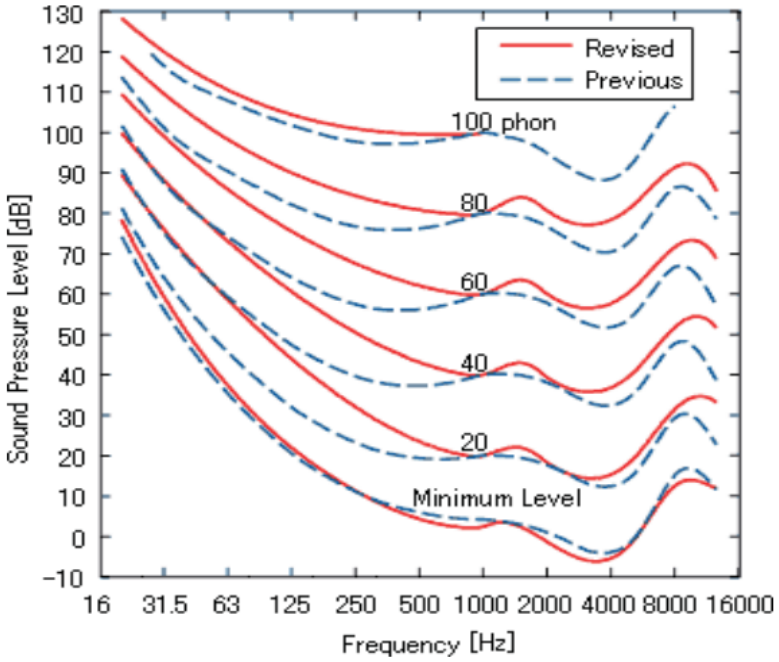


Figure B4. Equal loudness level contours presented in standards ISO 226 (1987) and ISO 226 (2003) [15][10].

Multiple studies have shown that most of the energy content of aircraft noise is located within the lower frequency regions, hence contributing relatively more towards the perceived annoyance [51, 58]. On the other hand, a few decades have passed since those studies were made, time in which aircraft noise properties might have changed, as well as the human perception. Other studies have shown that some aircraft have very specific characteristics which lead to noise-induced annoyance for other reasons than for the low-frequency content, such as highly pitched tonal components [44]. However, it is still important to investigate some noise metrics specific to the low frequency spectrum, so as to cover the sources for annoyance in an exhaustive manner, especially since low frequency noise energy can pass through the walls of buildings more easily, which can cause increased annoyance, especially at night time [58]. Below, a list of the most important such metrics is given, as identified in [15].

[10]It should be noted that ISO 226 was revised again in 2023, but the differences are negligible for most practical purposes [57].

Low Frequency Noise Metrics

- Low-Frequency Sound Level (LFSL):
  This metric represents the summation of the maximum noise levels in each of the one-third octave bands centered between 25 - 80 Hz. It has been proposed as a predictor for the rattle effect (low frequency noise causes vibration in households in the vicinity of the events) [58, 59]. However, researchers have argued that the 25-80 Hz interval is too small to fully account for all levels of noise-induced vibrations [60].
- Low-Frequency Sound Pressure Level (L_{LF}):
  This metric is the summation of the mean-square sound pressure levels in the 16, 31.5 and 63 Hz octave bands.
- Adjusted Sound Exposure level (L_{NE}):
  The L_{NE} is calculated via Eq. 17 (where T is the duration of the signal) and is based on the L_{LF} described above. The main point of interest here is that, compared to L_{LF}, L_{NE} accounts for the rapid increase of annoyance for low-frequency sound pressure level larger than 65 dB, through the 2 multiplication factor.

$$L_{NE} = 2(L_{LF} - 65) + 55 + 10log_{10}(T) \tag{17}$$

2. Sound Quality Metrics (SQMs)

Apart from the more conventional metrics described in the previous section, there are other, highly subjective metrics, which were defined with a strong emphasis on the human perception of sound. As will be seen, several models have been created by taking into account these SQMs. The main starting point is Zwicker and Fastl's psychoacoustic annoyance (PA) model [21], based on which other variants have been developed [15, 22]. A list of the most important SQMs is given, followed by a characterization of the mentioned PA models.

Sound Quality Metrics

- Loudness (N):
  As mentioned in the previous section, loudness can be understood as the subjective human perception of sound intensity. It is based on intensity, frequency and duration [61]. Because of its inherent subjective nature, it was experimentally determined from a 1 kHz pure tone [21]. The relation between Stevens' loudness scale, measured in sone, and Barkhausen's loudness level, measured in phon, is given in Fig. B5. In short, the loudness is computed as the integral of the specific loudness over all critical bands. According to Zwicker's model - multiple models were created throughout time, such as the one by Moore and Glasberg [62] - loudness is calculated through Eq. 18-Eq. 21, where N' is the specific loudness for one critical band, z is the number of contiguous critical bands for a frequency f, and CBW is the critical bandwidth. These calculations employ the Bark scale, which is a nonlinear scale used in psychoacoustics, as it was created to reflect the nonlinear response of the human ear to different frequencies [63].

$$\text{CBW} = 25 + 75\left(1 + 1.4\left(\frac{f_c}{1000}\right)^2\right)^{0.69} \text{ Hz} \tag{18}$$

$$z = 13\arctan\left(0.76\frac{f}{1000}\right) + 3.5\arctan\left(\frac{f}{7500}\right)^2 \text{ Bark} \tag{19}$$

$$N' = 0.08\left(\frac{E_{TQ}}{E_0}\right)^{0.23}\left[\left(0.5 + 0.5\frac{E}{E_{TQ}}\right)^{0.23} - 1\right] \text{ sone/Bark} \tag{20}$$

$$N = \int_0^{24Bark} N'(z)dz \tag{21}$$

- Sharpness (S):
  Sharpness is a metric which reflects (parts of) the spectral characteristics of a sound and it is measured in acum [21] - "a narrow band noise with 1 kHz center frequency and 160 Hz bandwidth, with a sound pressure level of 60 dB, would produce a Sharpness of 1 acum" [15]. Larger emphasis is put on the higher frequency content. Sharpness is calculated via Eq. 22 (based on von Bismarck's model [64]), where $g(z)$ (see Eq. 23) is the weighting factor emphasizing these higher frequencies, and c is a constant dependent on the normalization of the reference sound (a narrow-band noise one critical-band wide at a centre frequency of 1 kHz having a level of 60 dB [21]).

$$S = \frac{c \cdot \int_0^{24} N'(z) \cdot g(z) \cdot z dz}{N} \tag{22}$$

$$g(z) = \left\{ \begin{array}{ccc} 1 & \text{for} & z \leq 16 \\ 0.066e^{0.171z} & \text{for} & z > 16 \end{array} \right\} \tag{23}$$

- Roughness (R):
  This metric (measured in asper) quantifies the fast time fluctuations of loudness of an audio signal (a tone with a center frequency 1 kHz, SPL 60 dB and a 100%, 70 Hz amplitude-modulation, produces a Roughness of 1 asper). According to Zwicker and Fastl's model [21], it is calculated via Eq. 24, where $\Delta L(z)$ (see Eq. 25) is the modulation depth of the specific loudness at critical band rate z after applying a temporal filter. At the time the doctoral thesis of More was published, there was still not a consensus regarding complex and modulated signals and "how to combine different modulations in a way that reflects roughness perception" [15].

$$R = 0.3 \int_0^{24} \Delta L(z) dz \tag{24}$$

$$\Delta L(z) = 20 log_{10} \left(\frac{F_{N'_{max}}(z)}{F_{N'_{min}}(z)} \right) \tag{25}$$

- Fluctuation Strength (F):
  Similarly to roughness, fluctuation strength also deals with time variations of loudness, but this time the slow fluctuations (1 to 16 cycles per second) are considered, and it is measured in vacil. This sensation appears to reach its maximum around a fluctuation of 4 cycles per second [21]. "A tone with sound pressure level 60 dB, 1 kHz center frequency and with a 100% amplitude modulation at 4 Hz, produces a Fluctuation Strength of 1 vacil" [15]. Two models, one for broad-band noise and another for tones, were proposed in the same work that has been extensively cited in this chapter, by Zwicker and Fastl, both of which can be seen in Eq. 26 and Eq. 27, respectively. Here, m is the modulation factor, f_{mod} is the modulation frequency and $\Delta L(z)$ is the modulation depth, as mentioned previously. The same limitations as those mentioned for Roughness calculation hold for Fluctuation Strength, since they are very similar metrics.

$$FS_{BBN} = \frac{5.8(1.25m - 0.25)(0.05L_{BBN} - 1)}{(f_{mod}/5)^2 + (4/f_{mod}) + 1.5} \tag{26}$$

$$FS_{tone} = \frac{0.008 \int_0^{24} \Delta L(z) dz}{(4/f_{mod}) + (f_{mod}/4)} \tag{27}$$

- Tonality (K):
  Several metrics taking into account the highest tonality component have been created, such as the Tone-to-Noise Ratio and Prominence Ratio. However, Aures' model for tonality sums all identified tonal components, making it a more general metric for the purpose of the intended research. More specifically, it is a function of "the bandwidth, frequency, the prominence of the tonal component, and

the level of the tonal content relative to the level of the entire signal" [15]. The calculation procedure is quite lengthy, so the equations are not reproduced here. Instead, an interested reader can consult the paper where Aures first introduced his tonality model for an inspection of the involved equations [65]. Another potentially tonality-related metric is the Tonality Audibility (L_{ta}), which measures the prominence of tones in sound.

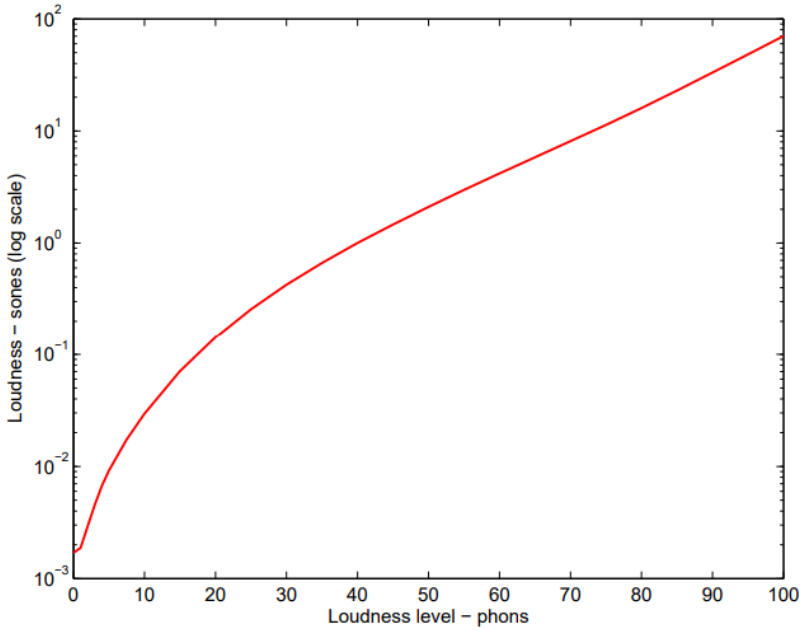


Figure B5. Relationship between loudness in sones and loudness level in phons [15].

3. Psychoacoustic Annoyance Models

Combining the sound quality metrics discussed in subsubsection B.2.2 leads to more complex models which reflect the human annoyance response to noise stimuli. One of the first, and perhaps the most well-known PA models, was first introduced by Zwicker and Fastl [21] and it combines the contributions of the following hearing sensations: loudness, sharpness, fluctuation strength and roughness. Its mathematical formulation is shown in Eq. 28, where N_5 is the 5$^{\text{th}}$ percentile loudness in sone (or the value of loudness exceeded during 5% of the signal), w_S describes the effect of sharpness, and w_{FR} describes the combined effect of fluctuation strength and roughness. The latter are calculated via Eq. 29 and Eq. 30, respectively.

Zwicker's PA model

$$PA = N_5 + \left(1 + \sqrt{w_S^2 + w_{FR}^2}\right) \tag{28}$$

$$w_S = \left\{ \begin{array}{cc} 0.25(S - 1.75) \cdot log_{10}(N + 10) & \text{for} \quad S > 1.75 \\ 0 & \text{for} \quad S \leq 1.75 \end{array} \right\} \tag{29}$$

$$w_{FR} = \frac{2.18}{N_5} \cdot (0.4F + 0.6R) \tag{30}$$

As can be observed, Zicker's PA model does not account for the tonality sensation of the noise. It was observed, however, that tonalness is associated with increased annoyance [44], and that none of the metrics currently used for quantifying aircraft noise-induced annoyance incorporate the contributions of loudness,

tonality, and roughness all together (for example, the FAA's EPNL takes into account the level, tonalness, and duration of aircraft noise, without considering any tonality-related aspect) [15]. For this reason, several attempts at improving Zwicker's original PA model, such that tonality is also considered, were identified. The first improvement comes from the doctoral thesis of More [15] and its formulation is described below.

More's improved version of Zwicker's PA model

$$PA_{More} = N_5 \left(1 + \sqrt{\gamma_0 + \gamma_1 w_s^2 + \gamma_2 w_{FR}^2 + \gamma_3 w_T^2}\right) \tag{31}$$

Here, w_S and w_{FR} represent the same contributions as in Zwicker's model and are calculated identically, while w_T is given via Eq. 32, and it represents the combined contribution of Aures' tonality (K) and loudness (N). The terms γ_i are the model weight parameters and they have been determined in the work of More by fitting the model to experimental data ($\gamma_0 = -0.16, \gamma_1 = 11.48, \gamma_2 = 0.84, \gamma_3 = 1.25$). It was also seen in the same work that this revised model is able to fit data from multiple dedicated aircraft noise listening experiments better than Zwicker's original model [15].

$$w_T = (1 - e^{-0.29N})(1 - e^{-5.49K}) \tag{32}$$

Additionally, Di et al. [22] have also proposed a modified version of Zwicker and Fastl's model, also by taking into account the tonality of noise stimuli. This was achieved by analysing tonal sound samples corresponding to low, mid and high frequencies, of varying loudness levels and A-weighted SPL. The resulting model, summarized below, was able to perform better when related to tonal noise annoyance compared to the original model.

Di et al. improved version of Zwicker's PA model

$$PA_{Di} = N_5 \left(1 + \sqrt{w_S^2 + w_{FR}^2 + w_T^2}\right) \tag{33}$$

Here, N_5, w_S and w_{FR} are the same as before, and w_T is given by Eq. 34, where K is Aures' tonality described previously.

$$w_T = \frac{6.41}{N_5^{0.52}} K \tag{34}$$

C. Aircraft Noise Characteristics and Predictive Models for Noise-Induced Annoyance

This chapter summarizes the findings made in literature concerning the observed relations between the characteristics of noise generated by various types of aircraft, with a strong focus on conventional commercial aircraft and on drones, and the human annoyance responses. Moreover, the state-of-the-art research in terms of using machine learning techniques for predicting noise-induced annoyance is outlined.

1. Aircraft Noise Characteristics

It has been widely agreed that noise-induced annoyance is generally attributed to several factors: loudness, or the sound intensity as perceived by the human ear, and the temporal and the spectral distribution of the sound stimuli [21, 43]. This means that sound intensity-related metrics only partly explain the spread in annoyance responses from psychoacoustic experiments. In some cases, the variation of the sound level related metrics and the distribution of the sound energy in frequency bands might contribute more towards the perceived annoyance than just the simpler measure of maximum sound level or other such metrics, as enumerated in subsection B.2. These findings are in line with the fact that the PA models outlined in the previous chapter use metrics which describe the aspects presented here.

The same holds for aircraft-generated noise [24, 66]. Numerous studies have been made, which aim at identifying the main sources of annoyance, for both conventional commercial aircraft and for drones. The

latter has been the more recent subject of research, as the emergence of more such aircraft is expected to grow in the future. Hence, it is crucial to evaluate their dominant characteristics in terms of annoyance, such that preventive measures can be taken in the design process. An overview of the most relevant studies is given below.

Conventional Aircraft Noise
Institutions and companies such as NASA and Boeing have long been concerned with studying commercial aircraft-induced noise annoyance. However, the large variability of aircraft sound stimuli, combined with the limited amount of data used throughout most of the studies, lead to a degree of variability of the results. Nevertheless, meaningful conclusions can be drawn by combining the results of the available research on conventional aircraft noise-induced annoyance.

Ever since the seventies it was found at NASA Langley that aspects like the tonal content and duration of an aircraft flyover are strong predictors for annoyance. Additionally, the interaction between the tonal components and the sound level was also identified as an important factor. Surprisingly, the rate and magnitude of the level fluctuation did not seem to affect the annoyance response in McCurdy's study [67]. The influence of the tonal components is also backed up by the findings of John W. Little at Boeing [68].

More recent studies have also found that tonality, in the form of high-pitched noise, is not desirable in terms of perceived annoyance. The psychoacoustic metrics sharpness and tonality were observed to be capable of explaining most of the variance in the annoyance responses in the experiment conducted by Torija et al. [69]. The work of More and Davies [49] also assessed the influence of tonality, by artificially varying the tonalness first alone, and then the tonalness and loudness together. The conclusion was that, alongside loudness, variations in tonalness also lead to changes in the annoyance ratings, and that sound metrics accounting for both are necessary in assessing aircraft environmental noise annoyance. A good example of tonality possibly influencing the annoyance response is found in the work of Merino-Martínez et al. [70], which showed that the effect of a strong tonal component associated with the landing gear of an Airbus A320 during approach can lead to a steep increase in annoyance (even surpassing the annoyance caused by the engines).

It seems that conventional aircraft noise annoyance is heavily influenced by the stimuli's spectral distribution, on top of metrics highly correlated with the sound intensity and its temporal variations. However, currently employed metrics for aircraft noise assessment, such as PNLT and EPNL, do not reflect the annoyance responses as well as desired. It is imperative to note that low noise levels do not automatically translate to low perceived annoyance [16]. It is therefore critical to consider metrics covering a wide range of sound perception dimensions in order to capture and understand the spread in human annoyance responses. Finally, one should be aware of the fact that extremely subjective aspects also play a role in the individual noise-induced annoyance [24], such as the individual's own sensitivity to noise or to certain characteristics of noise. However, these are very difficult to capture during limited listening experiments - to the extent at which a variety of participants representative of the global population is achieved.

Drone Noise
As far as drone noise is concerned, this category has been emerging as an important topic for research in recent years, because of the increase in popularity of Unmanned Aerial Vehicles (UAVs) for various purposes within the urban environments, at altitudes much closer to the ground than conventional aircraft. While research might still be necessary to fully quantify the effect of drone noise on annoyance, it has already been concluded that the metrics used for conventional aircraft noise certifications may not adequately capture the noise effects of UAVs [71], pointing towards the idea that drone noise is significantly different from that produced by civil aircraft.

The work of Torija, focused on drone noise-induced annoyance, has brought to light several characteristics of UAV noise. Firstly, experiments showed that masking of drone noise in urban environments highly affected by road noise is most likely to occur, leading to an overall lower annoyance caused by drone noise itself [26].

Torija observed in his experiments that the annoyance ratings were 1.3 times higher than those where no drone noise is present for soundscapes with heavy road traffic noise, while the ratings increased to 6.4 times the value corresponding to no drone noise for soundscapes with low levels of road traffic noise. This insight shows that communities living in quieter areas could potentially be affected most by the emergence of drones in a wide variety of uses. However, the contribution of drone noise to the annoyance of communities living in louder areas is also not to be neglected.

The fundamental difference between drone and conventional aircraft noise is the larger high frequency energy content of the former [72] (see Fig. B6). This is partly because of the larger atmospheric attenuation of higher frequencies in the case of the latter, since drones operate much closer to the ground. Additionally, drone noise is highly tonal and sharp [73], characteristics which were strongly correlated to annoyance in the case of conventional aircraft. The variation of the drones' rotor RPM to compensate wind gusts leads to an unsteady acoustic signature "with rapid temporal fluctuations of the tonal components" [27].

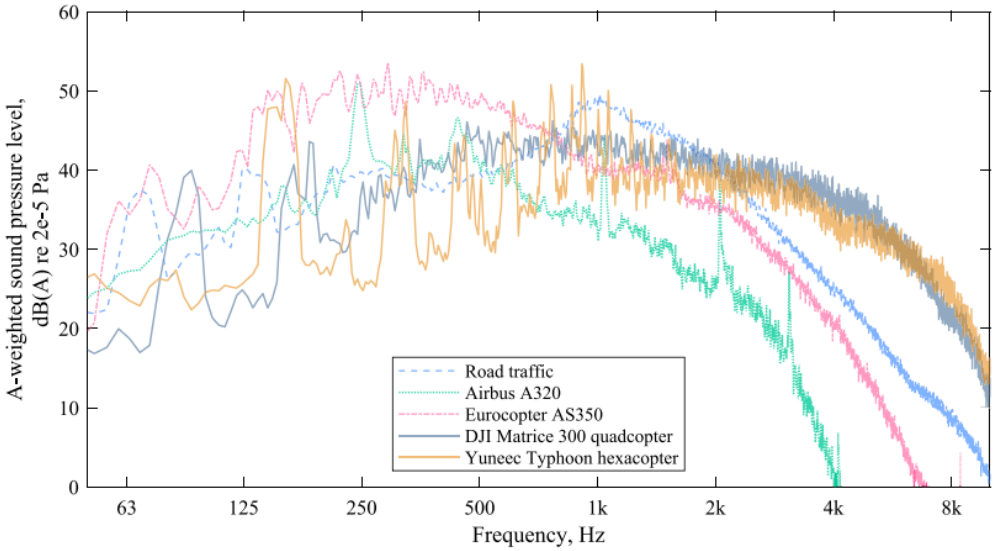


Figure B6. Two drone noise vs. other sources' spectral energy distributions, showing a higher energy content of drone noise for f>2kHz [46].

An even more recent study of Lotinga et al. [46] has concluded that, just as for conventional aircraft, drone noise-induced annoyance is mainly caused by loudness, followed by contributions from the temporal variation in loudness and spectral distribution of energy. Even so, a comprehensive report made at NASA Langley Research Center [47] concludes that noise certification standards should be revised for drones before they become widely used in more shapes, sizes and configurations. It also emphasizes the need for sharing available data between industries and regulators, in order to understand the relations between vehicle design and noise characteristics. Gathering as much and as diverse data as possible is of utmost importance, as will be seen in the next section on annoyance predictive models.

2. Predictive Models for Noise-Induced Annoyance

The metrics and sound characteristics that have been discussed above have been used to some extent in multiple studies to create predictive models for psychoacoustic annoyance caused by urban noise. These include the assessment of road, railway, commercial aircraft and drone noise alone, as well as multiple of these combined. The main findings and limitations are discussed in this section, based on which one of the research questions will then be formulated.

Before diving into complex Machine Learning models for annoyance prediction, it is important to discuss

some findings made in relation to the correlation of various SQMs and acoustical metrics with reported annoyance in various listening experiments. In terms of urban noise in general, the work of Song et al. [23] showed that it is possible to explain a significant portion of human annoyance responses to sound stimuli with a combination of a few parameters, starting from a large pool of features. Even though the sounds used in this work are fundamentally different from aircraft flyover noise, it stands as proof that properly performing correlation and dimensional analyses can drastically reduce the feature space while retaining acceptable accuracy. This is confirmed by the work of Bonebright [74], where multidimensional scaling (MDS) was performed in order to find the correlation between and among acoustic measurements and verbal attribute ratings used in listening experiments. It was found that even a 3-dimensional predictor space can explain a large portion of the observed annoyance response variance.

Concerning studies focusing on aircraft-generated noise, research has shown how different (combinations of) psychoacoustic parameters can be used to explain annoyance responses from various types of aircraft. For example, Torija et al. [75] showed that, for a limited sample of drone noise recordings, the main contributors towards annoyance prediction among the considered SQMs and acoustic parameters are the PNL and sharpness, which is in line with the findings enumerated in the previous sections of this chapter. Furthermore, Krishnamurthy et al. [76] and Boucher et al. [77] studied the main contributors towards annoyance caused by rotorcraft systems and found that sharpness, fluctuation strength and tonality are key parameters in explaining this effect. Additionally, it was also noted that the main predictors for annoyance can vary with the position (outdoor/indoor) of the recipient of the sound stimulus [78].

Nevertheless, since the relations between parameters affecting annoyance are complex and highly non-linear, the need for more complex regression ML models is required in order to capture the complexity of the soundscape generated by aircraft flyovers. There will always be limitations caused by subjective, individual-level based parameters, such as noise sensitivity [24], which are very difficult to capture during listening experiments in such a way that does not overfit the used samples. Hence, creating accurate and robust predictive models is crucial for eliminating as much of the individual's subjectivity from the picture, so as to capture the more "measurable" and relevant parameters/characteristics for annoyance. Consequently, the state-of-the-art in such models is discussed below.

Application of deep learning methods for predicting psychoacoustic annoyance have been applied especially in the context of urban and/or road traffic noise, showing promising performance. Wang et al. [79] have employed a Convolutional Neural Network (CNN) architecture in which the amplitude spectrum of sound recordings is used as input. The study was performed without incorporating subjects' personal information or additional acoustic features. The main drawback of the study is the low number of data points to be used for training and testing the model. To this end, a transfer learning approach is proposed, through which the model parameters are pre-tweaked using a "computer-generated psychoacoustic annoyance degree dataset" [79]. The pre-trained model seems to outperform both the regression models used for comparison and the initial neural network.

Lopez et al. [80] have investigated the use of CNNs for predicting Zwicker's PA score directly from raw sound recordings. Since a large correlation has been observed between the PA scores obtained using PA models and subjective annoyance scores reported in listening experiments, such a model could offer a possible way of generating more labelled data without the need for an unrealistically long experiment campaign. The performance of the model was also very promising, with a mean quadratic error of around 3%.

In a different approach, Shu et al. [81] use a Recurrent Neural Network (RNN) which considers time series effects to predict annoyance of urban noise recordings. The resulting model outperforms (in terms of MSE, Pearson and Spearman correlation coefficients) several dimension reduction regression models, namely a PCA-based linear regression, a Least Absolute Shrinkage and Selection Operator-based regression (LASSO), and a Partial Least Squares (PLS) regression model. The use of RNNs might be limited by the large GPU

requirements it poses, as shown in an attempt to use it for interior cabin noise annoyance prediction, in which the CNN architecture is deemed superior for this purpose [82].

The main limitations of deep neural network approaches is the lack of interpretability of the obtained results, since neural networks are black-box-type models. It is therefore difficult to evaluate the main predictors on which the predictions are based. While they show great versatility and capability to deal with complex patterns, such as those seen in noise annoyance, they might not immediately help in understanding the precise underlying mechanisms of annoyance. However, such complex ML models have yet to be investigated in detail for aircraft noise annoyance prediction. Several studies have shown that other regression methods, especially decision-tree-based methods, combined with dimensionality reduction and feature selection techniques, can still achieve satisfying performance and offer more interpretable results [83, 84].

D. Research Objectives

The presented literature survey was aimed at getting familiar with the topic of psychoacoustics in the context of aircraft noise annoyance and on the use of ML algorithms for predicting it. Promising directions involving deep learning methods and other classical regression techniques have been identified, as well as limitations. The main consensus observed regarding the limitations of such predictive models is the need for more data for training, testing and validating models. It is therefore concluded that there is a need for organising a listening experiment camapaign during this research, in order to gather as much and as diverse data as possible. A large pool of combined objective and subjective noise metrics has also been investigated in the attempt to understand their respective strengths and limitations. Combining this knowledge, the large pool of features can be used as a starting point in identifying a reduced pool of predictors to be used in one or more ML models for PA prediction.

Hence, the aim of the research is to jointly answer the following research questions: "What are the main sound characteristics responsible for annoyance in conventional turbofan aircraft?" and "To what extent can Artificial Intelligence be employed for predicting conventional turbofan aircraft noise annoyance?"

By answering these questions, the research performed can result in a model which could be used in strong connection with aircraft noise prediction tools (as the one presented in [30]), in order to include noise annoyance prediction in the design phase of future aircraft, such as the Flying-V, currently under development at TU Delft. Moreover, the understanding of the main characteristics influencing annoyance could aid the scientific community in further related research and could contribute to novel and more adequate certification metrics and requirements for future configurations.

References

[1] Boeing Commercial Airplanes, "Current Market Outlook," *URL: http://www. boeing. com*, 2005.

[2] Airbus SAS, "„Global Market Forecast-Flying by Numbers 2015-2034,"," *Airbus SAS, Blagnac Cedex, FRA*, 2015.

[3] Civil Aviation Authority, "Aircraft Noise and Annoyance: Recent Findings," *CAA Publication*, 2018.

[4] Mongeau, L., Huff, D., and Tester, B., "Aircraft noise technology review and medium and long term noise reduction goals," *Proceedings of Meetings on Acoustics*, Vol. 19, AIP Publishing, 2013.

[5] Ruijgrok, G. J., "Elements of aviation acoustics," 1993.

[6] Dobrzynski, W., "Almost 40 years of airframe noise research: what did we achieve?" *Journal of aircraft*, Vol. 47, No. 2, 2010, pp. 353–367.

[7] Merino-Martínez, R., Besnea, I., von den Hoff, B., and Snellen, M., "Psychoacoustic Analysis of the Noise Emissions from the Airbus A320 Aircraft Family and its Nose Landing Gear System," *30th AIAA/CEAS Aeroacoustics Conference (2024)*, 2024, p. 3398.

[8] World Health Organization, *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*, World Health Organization. Regional Office for Europe, 2011.

[9] Hansell, A. L., Blangiardo, M., Fortunato, L., Floud, S., De Hoogh, K., Fecht, D., Ghosh, R. E., Laszlo, H. E., Pearson, C., Beale, L., et al., "Aircraft noise and cardiovascular disease near Heathrow airport in London: small area study," *Bmj*, Vol. 347, 2013.

[10] Faiyetole, A. A., and Sivowaku, J. T., "The effects of aircraft noise on psychosocial health," *Journal of Transport & Health*, Vol. 22, 2021, p. 101230.

[11] Miedema, H. M., "Annoyance caused by environmental noise: Elements for evidence-based noise policies," *Journal of social issues*, Vol. 63, No. 1, 2007, pp. 41–57.

[12] Fredianelli, L., Carpita, S., and Licitra, G., "A procedure for deriving wind turbine noise limits by taking into account annoyance," *Science of the total environment*, Vol. 648, 2019, pp. 728–736.

[13] Malaval, G., "Approach to Noise Regulation of Unmanned Aviation in the European Union," 2024.

[14] Protection, E., "Annex 16 to the Convention on International Civil Aviation–Volume I–," *Aircraft noise*, 2008.

[15] More, S. R., *Aircraft noise characteristics and metrics*, Purdue University, 2010.

[16] Rizzi, S. A., and Christian, A., "A psychoacoustic evaluation of noise signatures from advanced civil transport aircraft," *22nd AIAA/CEAS Aeroacoustics Conference*, 2016, p. 2907.

[17] Guski, R., "Personal and social variables as co-determinants of noise annoyance," *Noise and health*, Vol. 1, No. 3, 1999, pp. 45–56.

[18] Miedema, H. M., and Vos, H., "Demographic and attitudinal factors that modify annoyance from transportation noise," *The Journal of the Acoustical Society of America*, Vol. 105, No. 6, 1999, pp. 3336–3344.

[19] Cox, T. J., "The effect of visual stimuli on the horribleness of awful sounds," *Applied acoustics*, Vol. 69, No. 8, 2008, pp. 691–703.

[20] Moore, B. C., "Psychoacoustics," *Springer handbook of acoustics*, 2014, pp. 475–517.

[21] Zwicker, E., and Fastl, H., *Psychoacoustics: Facts and models*, Vol. 22, Springer Science & Business Media, 2013.

[22] Di, G., Chen, X., Song, K., Zhou, B., and Pei, C.-M., "Improvement of Zwicker's psychoacoustic annoyance model aiming at tonal noises," *Applied Acoustics*, Vol. 105, 2016, pp. 164–170.

[23] Song, Y., Zhou, H., and Shu, H., "An efficient feature matrix for urban noise annoyance measurement," *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, IEEE, 2018, pp. 535–539.

[24] Gille, L.-A., Marquis-Favre, C., and Weber, R., "Aircraft noise annoyance modeling: Consideration of noise sensitivity and of different annoying acoustical characteristics," *Applied Acoustics*, Vol. 115, 2017, pp. 139–149.

[25] Sottek, R., and Lobato, T., "AI-SQ Metrics: Artificial Intelligence in Sound Quality Metrics," *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Vol. 261, Institute of Noise Control Engineering, 2020, pp. 3083–3091.

[26] Torija, A. J., Li, Z., and Self, R. H., "Effects of a hovering unmanned aerial vehicle on urban soundscapes perception," *Transportation Research Part D: Transport and Environment*, Vol. 78, 2020, p. 102195.

[27] Torija, A. J., and Clark, C., "A psychoacoustic approach to building knowledge about human response to noise of unmanned aerial vehicles," *International Journal of Environmental Research and Public Health*, Vol. 18, No. 2, 2021, p. 682.

[28] Bertsch, L., "Noise prediction within conceptual aircraft design," Ph.D. Thesis, Technische Universität Braunschweig, 2013.

[29] Lopes, L., and Burley, C., "Design of the next generation aircraft noise prediction program: ANOPP2," *17th AIAA/CEAS aeroacoustics conference (32nd AIAA aeroacoustics conference)*, 2011, p. 2854.

[30] Filippone, A., "Aircraft noise prediction," *Progress in Aerospace Sciences*, Vol. 68, 2014, pp. 27–63.

[31] Merino-Martínez, R., Yupa-Villanueva, R. M., von den Hoff, B., and Pockelé, J. S., "Human response to the flyover noise of different types of drones recorded in field measurements," 2024.

[32] *Acoustics: Attenuation of Sound During Propagation Outdoors*, International Organization for Standardization, 1993.

[33] Merino-Martínez, R., von den Hoff, B., and Simons, D. G., "Design and acoustic characterization of a psychoacoustic listening facility," *Proceedings of the 29th International Congress on Sound and Vibration, ICSV 2023*, Society of Acoustics, 2023.

[34] Greco, G. F., Merino-Martínez, R., Osses, A., and Langer, S. C., "SQAT: a MATLAB-based toolbox for quantitative sound quality analysis," *Inter-Noise and Noise-Con Congress and Conference Proceedings*, Vol. 268, Institute of Noise Control Engineering, 2023, pp. 7172–7183.

[35] ISO, I., "532-1-2017 Acoustics-Methods for Calculating Loudness-Part1: Zwicker Method," *ISO: Geneva, Switzerland*, 2017.

[36] DIN, D., "45692: Measurement Technique for the Simulation of the Auditory Sensation of Sharpness," *DIN: Berlin, Germany*, 2009.

[37] Osses Vecchi, A., García León, R., and Kohlrausch, A., "Modelling the sensation of fluctuation strength," *Proceedings of Meetings on Acoustics*, Vol. 28, AIP Publishing, 2016.

[38] Daniel, P., and Weber, R., "Psychoacoustical roughness: Implementation of an optimized model," *Acta Acustica*

united with Acustica, Vol. 83, No. 1, 1997, pp. 113–123.

[39] Aures, W., "Der sensorische wohlklang als funktion psychoakustischer empfindungsgrössen," *Acta Acustica United with Acustica*, Vol. 58, No. 5, 1985, pp. 282–290.

[40] Di, G., Cong, C., Yao, Y., Li, D., and Jian, W., "A study on the conversion relationship of noise perceived annoyance and psychoacoustic annoyance—a case of substation noise," *Journal of Low Frequency Noise, Vibration and Active Control*, Vol. 41, No. 2, 2022, pp. 810–818.

[41] Lopez-Ballester, J., Pastor-Aparicio, A., Felici-Castell, S., Segura-Garcia, J., and Cobos, M., "Enabling real-time computation of psycho-acoustic parameters in acoustic sensors using convolutional neural networks," *IEEE Sensors Journal*, Vol. 20, No. 19, 2020, pp. 11429–11438.

[42] Géron, A., *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*, " O'Reilly Media, Inc.", 2017.

[43] Ma, K. W., Wong, H. M., and Mak, C. M., "A systematic review of human perceptual dimensions of sound: Meta-analysis of semantic differential method applications to indoor and outdoor sounds," *Building and Environment*, Vol. 133, 2018, pp. 123–150.

[44] Merino-Martínez, R., Snellen, M., and Simons, D., "Analysis of landing gear noise during approach," *22nd AIAA/CEAS Aeroacoustics Conference, 2016: Lyon, France*, 2016.

[45] Beutel, M. E., Jünger, C., Klein, E. M., Wild, P., Lackner, K., Blettner, M., Binder, H., Michal, M., Wiltink, J., Brähler, E., et al., "Noise annoyance is associated with depression and anxiety in the general population-the contribution of aircraft noise," *Plos one*, Vol. 11, No. 5, 2016, p. e0155357.

[46] Lotinga, M. J., Ramos-Romero, C., Green, N., and Torija, A. J., "Noise from Unconventional Aircraft: A Review of Current Measurement Techniques, Psychoacoustics, Metrics and Regulation," *Current Pollution Reports*, Vol. 9, No. 4, 2023, pp. 724–745.

[47] Rizzi, S. A., Huff, D. L., Boyd, D. D., Bent, P., Henderson, B. S., Pascioni, K. A., Sargent, D. C., Josephson, D. L., Marsan, M., He, H. B., et al., "Urban air mobility noise: Current practice, gaps, and recommendations," , 2020.

[48] Lighthill, M. J., "On sound generated aerodynamically I. General theory," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, Vol. 211, No. 1107, 1952, pp. 564–587.

[49] More, S., and Davies, P., "Human responses to the tonalness of aircraft noise," *Noise Control Engineering Journal*, Vol. 58, No. 4, 2010, pp. 420–440.

[50] Fidell, S., Pearsons, K., Tabachnick, B. G., and Howe, R., "Effects on sleep disturbance of changes in aircraft noise near three airports," *The Journal of the Acoustical Society of America*, Vol. 107, No. 5, 2000, pp. 2535–2547.

[51] Leventhall, G., Pelmear, P., and Benton, S., "A review of published research on low frequency noise and its effects," 2003.

[52] Sörensen, S., and Hammar, N., "Annoyance reactions due to railway noise," *Journal of sound and vibration*, Vol. 87, No. 2, 1983, pp. 315–319.

[53] Rylander, R., Sörensen, S., and Berglund, K., "Re-analysis of aircraft noise annoyance data against the dB (A) peak concept," *Journal of Sound and Vibration*, Vol. 36, No. 3, 1974, pp. 399–406.

[54] Schomer, P. D., "On normalizing DNL to provide better correlation with response," *Sound and Vibration*, Vol. 36, No. 12, 2002, pp. 14–25.

[55] Schomer, P. D., "The importance of proper integration of and emphasis on the low-frequency sound energies for environmental noise assessment," *Noise Control Engineering Journal*, Vol. 52, No. 1, 2004, pp. 26–39.

[56] FAR, F. A. R., "Part 36: Noise Standards: Aircraft Type and Airworthiness Certification," *Washington, DC: US Federal Aviation Administration*, 2002.

[57] Suzuki, Y., Takeshima, H., and Kurakata, K., "Revision of ISO 226" Normal Equal-Loudness-Level Contours" from 2003 to 2023 edition: The background and results," *Acoustical Science and Technology*, Vol. 45, No. 1, 2024, pp. 1–8.

[58] Fidell, S., Silvati, L., Pearsons, K., Lind, S., and Howe, R., "Field study of the annoyance of low-frequency runway sideline noise," *The Journal of the Acoustical Society of America*, Vol. 106, No. 3, 1999, pp. 1408–1415.

[59] Fidell, S., Harris, A., and Sutherland, L., "Findings of the Low-Frequency Noise Expert Panel of Richfield-MAC Noise Mitigation Agreement of 17 December 1998," *Report to the City of Richfield, Minnesota and the Minneapolis Metropolitan Airports Commission*, 2000.

[60] Sharp, B. H., Gurovich, Y. A., and Albee, W. W., "Status of Low-Frequency Aircraft Noise Research and Mitigation," *Wyle Report WR*, 2001, pp. 01–21.

[61] Glasberg, B. R., and Moore, B. C., "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, Vol. 50, No. 5, 2002, pp. 331–342.

[62] Moore, B. C., and Glasberg, B. R., "A revision of Zwicker's loudness model," *Acta Acustica united with Acustica*,

Vol. 82, No. 2, 1996, pp. 335–345.

[63] Zwicker, E., and Feldtkeller, R., "The ear as a communication receiver," *Acustica*, 1999.

[64] von Bismarck, G., "Sharpness as an attribute of the timbre of steady sounds," *Acta Acustica united with Acustica*, Vol. 30, No. 3, 1974, pp. 159–172.

[65] Aures, W., "Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale," *Acta Acustica united with Acustica*, Vol. 59, No. 2, 1985, pp. 130–141.

[66] Barbot, B., Lavandier, C., and Cheminée, P., "Perceptual representation of aircraft sounds," *Applied Acoustics*, Vol. 69, No. 11, 2008, pp. 1003–1016.

[67] McCurdy, D. A., *Effects of sound level fluctuations on annoyance caused by aircraft-flyover noise*, National Aeronautics and Space Administration, 1979.

[68] Little, J. W., "Human response to jet engine noises," *Noise Control*, Vol. 7, No. 3, 1961, pp. 11–13.

[69] Torija, A. J., Roberts, S., Woodward, R., Flindell, I. H., McKenzie, A. R., and Self, R. H., "On the assessment of subjective response to tonal content of contemporary aircraft noise," *Applied Acoustics*, Vol. 146, 2019, pp. 190–203.

[70] Merino-Martínez, R., Vieira, A., Snellen, M., and G. Simons, D., "Sound quality metrics applied to aircraft components under operational conditions using a microphone array," *25th AIAA/CEAS aeroacoustics conference*, 2019, p. 2513.

[71] Senzig, D., and Marsan, M., "UAS noise certification," *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Vol. 258, Institute of Noise Control Engineering, 2018, pp. 3718–3726.

[72] Gwak, D. Y., Han, D., and Lee, S., "Sound quality factors influencing annoyance from hovering UAV," *Journal of Sound and Vibration*, Vol. 489, 2020, p. 115651.

[73] Merino-Martínez, R., Torija, A. J., and Li, Z., "Metrics for assessing the perception of drone noise," *Forum Acusticum*, 2020, pp. 3163–3168.

[74] Bonebright, T. L., "Perceptual structure of everyday sounds: A multidimensional scaling approach," *Proceedings of the 2001 International Conference on Auditory Display*, Vol. 35, Laboratory of Acoustics and Audio Signal Processing and the . . . , 2001.

[75] Torija, A. J., and Nicholls, R. K., "Investigation of metrics for assessing human response to drone noise," *International Journal of Environmental Research and Public Health*, Vol. 19, No. 6, 2022, p. 3152.

[76] Krishnamurthy, S., Christian, A., and Rizzi, S., "Psychoacoustic test to determine sound quality metric indicators of rotorcraft noise annoyance," *Inter-Noise and Noise-Con Congress and Conference Proceedings*, Vol. 258, Institute of Noise Control Engineering, 2018, pp. 317–328.

[77] Boucher, M., Krishnamurthy, S., Christian, A., and Rizzi, S. A., "Sound quality metric indicators of rotorcraft noise annoyance using multilevel regression analysis," *Proceedings of Meetings on Acoustics*, Vol. 36, AIP Publishing, 2019.

[78] Green, N., Torija, A. J., and Ramos-Romero, C., "Perception of noise from unmanned aircraft systems: Efficacy of metrics for indoor and outdoor listener positions," *The Journal of the Acoustical Society of America*, Vol. 155, No. 2, 2024, pp. 915–929.

[79] Wang, J., Wang, X., Yuan, M., Hu, W., Hu, X., and Lu, K., "Deep learning-based road traffic noise annoyance assessment," *International Journal of Environmental Research and Public Health*, Vol. 20, No. 6, 2023, p. 5199.

[80] Lopez-Ballester, J., Pastor-Aparicio, A., Segura-Garcia, J., Felici-Castell, S., and Cobos, M., "Computation of psycho-acoustic annoyance using deep neural networks," *Applied Sciences*, Vol. 9, No. 15, 2019, p. 3136.

[81] Shu, H., Song, Y., and Zhou, H., "RNN based noise annoyance measurement for urban noise evaluation," *TENCON 2017-2017 IEEE Region 10 Conference*, IEEE, 2017, pp. 2353–2356.

[82] Hadzalic, D., "Application of neural networks for prediction of subjectively assessed interior aircraft noise," , 2018.

[83] Rafaelof, M., and Schroeder, A., "Investigation of machine learning algorithms to model perception of sound," *Proceedings of Meetings on Acoustics*, Vol. 33, AIP Publishing, 2018.

[84] Zhou, H., Shu, H., and Song, Y., "Using machine learning to predict noise-induced annoyance," *TENCON 2018-2018 IEEE Region 10 Conference*, IEEE, 2018, pp. 0229–0234.