DOI: 10.61782/fa.2025.0542

MACHINE LEARNING-BASED PREDICTION OF AIR SOURCE HEAT PUMP NOISE ANNOYANCE

Volkan Acun[1*] **Simone Graetzer**[1] **Antonio J. Torija Martinez**[1]
[1] Acoustics Research Centre, University of Salford, United Kingdom

ABSTRACT

Air Source Heat Pumps (ASHPs) are pivotal in decarbonising domestic heating and air conditioning, yet noise emissions remain a significant barrier to their wider adoption. This paper presents a machine learning-based approach to predicting annoyance caused by ASHP noise emissions to address the complexity of human perception of noise by integrating psychoacoustic metrics, emotional responses, and demographic factors. Seven predictive models were evaluated, including tree-based methods (Gradient Boosting and Random Forest) and traditional regression approaches (Support Vector Machine, Lasso, Linear Regression, Ridge, and ElasticNet). Among these, the Gradient Boosting method demonstrated superior performance (test R² = 0.846, RMSE = 0.910) compared to linear methods, highlighting the non-linear nature of annoyance response. Feature importance analysis revealed emotional responses (Arousal and Valence) as the dominant predictors, collectively accounting for 92.3% of the model's predictive capability, while Zwicker's Psychoacoustic Annoyance (PA) and Tonality showed moderate correlations with these emotional factors, suggesting indirect influence and mediation effect. Despite limitations in available training data that did not allow full implementation of neural network approaches, the current model provides a robust foundation for predicting annoyance caused by ASHP noise emissions and highlighting the impact of subjective perception

Keywords: *noise perception, air source heat pumps, noise annoyance, machine learning, regression*

1. INTRODUCTION

Switching from fossil fuels to renewable options is key for reducing carbon emissions in domestic heating and air conditioning around the globe. In Europe, buildings are the largest consumers of energy. Most of these buildings waste energy. They account for 36% of greenhouse gas emissions and 40% of total energy use [1]. One way to tackle this issue is by replacing old fossil fuel heating systems with energy-efficient options like air source heat pumps (ASHPs).

ASHPs are easy to maintain and cost-efficient. The thermal energy produced by an ASHP can be well over 100% of the amount of energy they use[2]. However, noise emissions from these units have raised concerns among the public. The spectrum of the ASHP noise varies depending on the operating condition of the unit. Low-frequency content dominates the ASHP signal, but it also typically includes both narrowband and tonal components [3], with becoming particularly more tonal in the low-frequency region. These components can significantly contribute to the annoyance caused by noise emissions.

The most common policy around ASHP noise emissions is regulating the A-weighted sound pressure levels. In addition, some countries, such as Germany, Finland, and the Netherlands, included tonal penalties in their noise assessment regulations to address the annoyance that could be caused by the tonal character of the signal [4]. These policies, in general, fell short of including the human perception of ASHP noise and solely relied on objective measures. People's subjective responses should

[*] ***Corresponding author***: v.acun@salford.ac.uk

be used alongside the A-weighted sound pressure level-based measurement to ensure best practice in the policy.

This paper presents the preliminary results of a more extensive study to incorporate Machine Learning (ML) based models to predict the noise annoyance caused by ASHP noise emissions. This paper aims to investigate which acoustics and non-acoustic factors and ML methods are most suitable for developing a robust model that accurately predicts noise annoyance caused by ASHP noise emissions. By incorporating additional data from diverse datasets in the future, this research aims to develop a fast and accessible tool for supporting decision-making based on relevant acoustic and non-acoustic metrics. This will facilitate effective planning and placement of ASHP units to minimise noise impact on communities.

2. METHODS

The data used for this study were gathered through a listening test conducted at the Acoustics Research Centre of the University of Salford. This listening test was concerned with human perception of ASHP noise at various source distances, background noise levels, and under different loads. It consisted of two parts, each examining different aspects of the ASHP noise under different background noise levels and source distances. The first part of the experiment considered people's reactions to continuous ASHP noise. In contrast, the second part concerned the response to transient ASHP noise. However, the scope of this study does not include the investigation of the responses to continuous and transient noise, source distances or the effect of different background noise levels. Instead, the focus is on understanding how the subjective response and psychoacoustic character of ASHP noise affects perception and how it can be used to develop an ML-based prediction model. Therefore, only a brief overview of the experimental procedure is provided.

2.1 Audio Stimuli

The audio stimuli used in this experiment were based on recordings conducted as part of the IAE HPT Annex 51[4][5]. These recordings had a sampling rate of 96 kHz and were calibrated to the original levels based on a 94 dB(A) 1 kHz calibration tone.

Two recordings were selected from this original recording dataset and trimmed into shorter segments of two different lengths for two distinct parts of the experiment (20 s for Part 1 and 60 s for Part 2). Responses to both parts of the experiment were used in this ML implementation. The sound levels (L_{Aeq20s} and $\mathrm{L}_{\mathrm{Aeq60s}}$) of these excerpts were calibrated in the Listening Room of the University of Salford's Acoustics Labs to ASHP-receiver distance attenuation target levels (36.5 dB(A) at 15 m, 40 dB(A) at 10 m, and 46 dB(A) at 5 m).

In addition, ambient background noise is added to the ASHP signal. Instead of a soundscape recording, a shaped pink noise is used as the ambient background noise. To achieve this, pink noise is filtered to represent a typical traffic noise spectrum based on the filter curve detailed in the BS EN ISO 717-2020 [7]. The shaped-pink noise was then calibrated to 39.5 dB(A) and 31.5 to represent the ambient daytime and nighttime background noise levels of a typical rural area [8].

Lastly, these stimuli were propagated indoors through a partially open window (0.05 m^2) following a filter curve provided by [9]. A total of 114 audio stimuli were used for parts one (84 stimuli) and two (30 stimuli). The participants were pseudo-randomly presented with half of the stimuli (57 stimuli) from the stimulus sets for each part of the experiment. Additionally, three control stimuli (2 stimuli for Part 1 and 1 stimulus for Part 2), which only included the shaped pink noise as ambient background noise, were presented to each participant.

2.2 Experimental Procedure

Listening tests were conducted in the University of Salford's Listening Room following ethical approval (Ref: 2024-0145-228). Participants listened to 60 audio stimuli across two parts (Part 1: 42 audio stimuli + 2 control stimuli; Part 2: 15 Audio stimuli and one control stimulus), with responses submitted via a custom Python GUI. The playback system included a Motu 4Pre interface, Genelec 8030A loudspeaker, and 7020B subwoofer, with participants seated 2 m from the source.

Each session lasted approximately 70 minutes and included breaks between parts. Before starting, participants completed consent forms, demographic questions, and the NoiSeQ noise sensitivity scale. After each stimulus, they rated Valence and Arousal on a 9-point Self-Assessment Manikin (SAM) scale, reflecting pleasantness and alertness. Annoyance was rated using a 0–10 scale (0 – "Not Annoying at All," 10 – "Extremely Annoying") adapted from ISO/TS 15666:2021 [10].

In Part 2, participants responded to two temporal points per stimulus to capture perceptual changes, resulting in 75 total responses per participant.

2.3 Participants

The dataset used for research consisted of 50 volunteers that participated in the experiment: 35 male (70%) and 15 female (30%), aged between 19 and 57 years (mean = 32.1, SD = 8.95). Nearly half of the participants reported having expertise in audio or music (n= 24, 48%), while just over half stated they had no previous experience (n= 26, 52%). Most participants lacked expertise in environmental noise or urban planning (n=40, 80%). In terms of housing, 18 participants resided in apartments (36%), nine in terraced houses (18%), six in semi-detached houses (12%), five in detached houses (10%), and the rest lived in student accommodation or shared housing.

2.4 Data Analysis

ArtemiS Suite 15.6 is used to calculate the sound quality metrics (SQM) of the audio stimuli, including Loudness, Sharpness, Roughness, Fluctuation Strength, and Tonality. Loudness and Sharpness metrics are calculated according to DIN 45631/A1[11] standard, while ECMA 418-2(1st)/(2nd) [12] is used for calculating Roughness, Tonality and Fluctuation Strength.

In addition to SQM, Zwicker's Psychoacoustic Annoyance (PA)model [13] is calculated using the formula:

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

Where:

- N5 percentile loudness in Sone

$$w_S = \left(\frac{S}{\text{acum}} - 1.75\right) \cdot 0.25 \log\left(\frac{N_5}{\text{sone}} + 10\right)$$

for $S > 1.75$ acum (2)

- describing the effect of Sharpness in S and

$$w_{\text{FR}} = \frac{2.18}{(N_5/\text{sone})^{0.4}}\left(0.4 \cdot \frac{F}{\text{vacil}} + 0.6 \cdot \frac{R}{\text{asper}}\right) \tag{3}$$

- describing Roughness R and Fluctuation F.

To quantify the application of ML methods to predict annoyance from ASHP noise emissions, Python 3.13.1 was used with relevant data analysis and visualisation libraries, including scikit-learn, statsmodels, matplotlib, and seaborn.

As previously mentioned, the dataset used for this experiment consisted of responses from 50 participants, each submitting a total of 75 responses, resulting in 3750 data points. Because of this relatively small sample size, ANN-based models were not considered in the prediction model. Instead, the study employed a range of ML-based linear regression models, including Ridge, Lasso, Elastic Net and Linear; tree-based models, including Random Forest and Gradient Boosting; and Support Vector Machine regression (SVR). These models were selected to compare performance across different regression approaches, balancing interpretability and prediction accuracy.

To optimise the hyperparameters of these models, GridSearchCV from the scikit-learn library was employed. This tool is commonly used for tuning the hyperparameters for ML algorithms by automating the process of finding the optimal combination of hyperparameters by exploring a predefined grid of parameter values and evaluating each combination using cross-validation [14]

To ensure the robustness of the results, 5-fold cross-validation (k=5) was used. This method partitions the dataset into five subsets, using each one for validation while the remaining four subsets are used for training. The use of cross-validation helps mitigate overfitting and provides a more reliable estimate of model performance, particularly given the relatively limited sample size in this study.

3. RESULTS AND DISCUSSION

The main focus of developing this prediction model was to incorporate the subjective response and psychoacoustic characteristics of ASHP noise using SQM. Performances of different feature combinations are compared across the previously listed regression and tree-based ML methods to determine which features to include in addition to the subjective response variables of Valence and Arousal and the SQM.

Table 1: Comparison of VIF values for training features that include Psychoacoustic Annoyance (PA) and Sound Quality Metrics (SQM)

Variance Inflation Factor (VIF) values:		
Feature	SQM as Predictors	PA as a Predictor
Tonality	8.673	8.456
Valence	8.08	6.161
Arousal	13.668	5.751
Age	12.258	10.231
PA	NA	10.539
Loudness	27.238	NA
Roughness	23.52	NA
Sharpness	27.557	NA
Fluctuation	4.618	NA

Multicollinearity analysis was conducted as part of the model comparisons, which revealed substantial interdependence among the original SQM. VIF values for Loudness (27.238), Sharpness (27.557), and Roughness (23.52) in the initial model exceeded recommended thresholds, indicating problematic collinearity (Table 1). When these metrics were replaced with the composite Psychoacoustic Annoyance (PA) parameter, the model's multicollinearity was substantially reduced. The PA-inclusive model demonstrated improved VIF values across predictor variables, with Arousal showing the most dramatic reduction (from 13.668 to 5.751). Although PA exhibited a VIF (10.539) value right at the cutoff value, it remains considerably lower than the metrics it replaced. The replacement of individual sound quality metrics with PA not only addresses multicollinearity concerns but also provides a more compact model while preserving essential psychoacoustic information. A-weighted sound levels are not included in this model, as they provided extreme VIF values and did not significantly contribute to the predictive power in any of the tested models.

The only demographic variable included in the models was Age. While other demographic variables were initially considered, such as gender and noise sensitivity data from NoiSeQ responses, they either produced extremely high VIF values (e.g. NoiSeQ VIF = 32.780 when used alongside the final features) or were found to not meaningfully contribute to the model based on correlation and feature importance analysis.

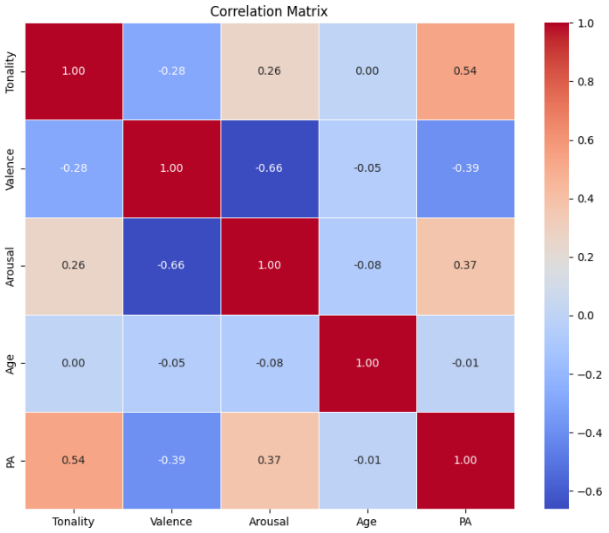

Figure 1: Heatmap representing the correlation matrix between the final set of predictor variables.

Figure 1 shows the correlation matrix between the final set of predictor variables used in the model comparison, providing further justification for the variable selection strategy explained in the VIF analysis. PA and Tonality are the strongest predictors in the matrix. The moderate correlation between PA and subjective response variables (Valence: -0.30, Arousal: 0.37) indicates that while PA effectively consolidates the SQM, it remains sufficiently distinct from other subjective response variables. Age shows the weakest correlation among all the predictors. It is kept as an independent factor in the model since a correlation matrix only captures the linear relationships, while models like Gradient Boosting and Random Forest can capture non-linear patterns and interaction effects that may not be evident in a correlation matrix.

A comprehensive evaluation of predictive models was conducted to identify the optimal approach for estimating acoustic responses. Performance metrics, including R², Mean Absolute Error (MAE), Mean Squared Error (MSE), and Roost Mean Squared Error (RMSE), were calculated for training and test datasets across seven distinct modelling techniques (Table 2). Among all the models compared, the Random Forest achieved the highest training R² (0.917), suggesting excellent data fitting; however, the slightly lower test performance (R² = 0.834) indicated potential overfitting. Linear models (SVR, Lasso, Linear Regression,

Table 2: Comparison of train and test performance of different models.

Model	Train R²	Test R²	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE
Gradient Boost	0.902	0.846	0.543	0.677	0.525	0.829	0.725	0.910
Random Forest	0.917	0.834	0.500	0.703	0.447	0.889	0.668	0.943
SVR	0.758	0.757	0.875	0.877	1.298	1.301	1.139	1.140
Lasso	0.751	0.750	0.892	0.892	1.339	1.341	1.157	1.158
Linear Regression	0.754	0.753	0.888	0.889	1.321	1.326	1.149	1.151
Ridge	0.754	0.753	0.888	0.889	1.321	1.325	1.150	1.151
ElasticNet	0.753	0.752	0.890	0.891	1.330	1.331	1.153	1.154

Ridge, and ElasticNet) exhibited comparable performance patterns, with test R² values ranging from 0.750 to 0.757, substantially lower than the ensemble methods. The consistency between training and test metrics for these linear models, particularly SVR (training R² = 0.758, test R² = 0.757), suggests stable performance despite lower overall accuracy. A significant performance gap between tree-based ensemble methods and linear models was observed, with approximately 10% improvement in explained variance and almost 20% reduction in prediction error, highlighting important non-linear relationships in the data.

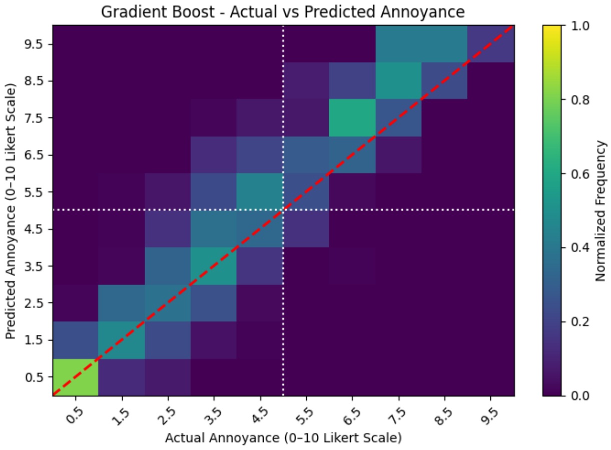

Figure 2: Heatmap showing the normalised frequency of predicted vs annoyance ratings for the Gradient Boost model. Each cell's colour represents the proportion of test samples falling into the corresponding (actual, predicted) bin. The red dashed line diagonal represents the perfect prediction.

The consistency between training and test metrics for these linear models, particularly SVR (training R² = 0.758, test R² = 0.757), suggests stable performance despite lower overall accuracy.

Among the models compared, the Gradient Boosting model demonstrated the best balance between predictive accuracy and generalisation. It achieved the highest test R² (0.846) while maintaining the lowest test error metrics (MAE = 0.677, RMSE = 0.910), indicating superior performance on unseen data and capturing potentially non-linear relationships. While Random Forest showed a slightly better prediction power over the training data, as it showed evidence of overfitting, the Gradient Boosting model was chosen to be used. In addition, the Gradient Boost method effectively handles multicollinearity. Based on the VIF values shown in Table 01, this is a requirement of the model, as both PA and Age are at the critical thresholds for multicollinearity.

Figure 2 illustrates the predictive performance of the model by comparing the actual and predicted values. The heatmap demonstrates a strong linear relationship between predicted and actual Annoyance values. Overall, the model aligns strongly with the prediction diagonal (red dashed line) but shows some deviation and mild underprediction for higher annoyance values.

Finally, feature importance scores were computed for the Gradient Boost model to understand the relative contribution of each predictor. The highest feature importance scores were observed for Arousal (0.550) and Valence (0.373), which accounted for 92.3% of the predictive power, which are also affected by the ASHP noise. Feature importance scores show that Age contributed

0.0659, while PA (0.0078) and Tonality (0.0026) had minimal impact. While these results indicate that psychoacoustic factors and tonality have a minimal direct impact on this model, their role in shaping the primary predictors should not be overlooked. When these results are considered with the insights from the correlation matrix (Figure 1), it shows that PA and Tonality are moderately correlated with Arousal and Valence, which are the most influential features. This indicates a potential mediating effect, where PA and Tonality influence the prediction indirectly via Arousal and Valence. This emphasises the importance of gaining a deeper understanding of the subjective responses to ASHP noise emissions and considering the soundscape approach.

4. LIMITATIONS AND FUTURE WORK

The main aim of this research was to create a robust ML-based prediction model and leverage ANN models. At the time of writing this paper, this is only partially achieved due to the lack of access to relevant datasets. The first step in future work will be supplementing this research with further data from similar acoustic studies. This will allow the development of more sophisticated ANN-based models capable of capturing the complex non-linear relationships observed in the Gradient Boost implementation. The significant performance gap between ensemble and linear models suggests that deep learning approaches offer additional predictive power with sufficient training data.

Future iterations of the model can also include temporal dynamics ASHP operation cycles. While the dataset from this study included different operating conditions and transient events, possibly due to sample size limitations, they were not found to be significant in any of the top-performing models. The temporal dynamics of SMQ and PA should also be further investigated. Incorporating time-varying acoustic features could better represent real-world ASHP operational conditions.

5. CONCLUSION

This study investigated the prediction of Annoyance responses to Air Source Heat Pump (ASHP) noise using machine learning (ML) approaches. The comparison of various regression models demonstrated that tree-based models, particularly Gradient Boosting, outperformed linear regression techniques. This superior performance highlights complex non-linear relationships in the perception of annoyance that simpler models fail to capture adequately. The PA is implemented as a composite metric to address multicollinearity issues present in the original SQM, resulting in a more compact and robust model.

Feature importance analysis revealed that emotional responses, specifically Arousal (0.550) and Valence (0.373), were the dominant predictors of annoyance, collectively accounting for over 92% of the model's predictive power. While PA and Tonality showed minimal direct contributions, their moderate correlations with the primary predictors suggest an indirect influence through emotional responses. This finding underscores the complex interplay between objective acoustic parameters and subjective emotional responses, supporting the adoption of soundscape approaches in noise annoyance research. The observed relationships emphasise the necessity of considering both acoustic and non-acoustic factors when evaluating and mitigating ASHP noise impacts.

The successful development of this predictive model represents a significant step toward practical applications in sustainable urban development. Despite limitations in available training data, the model demonstrates considerable potential for real-world implementation. Future work will focus on expanding the dataset to enable more sophisticated neural network implementations and incorporating temporal variations in acoustic features. The ultimate goal remains the integration of these predictive capabilities into accessible tools for optimal ASHP placement, thereby supporting the widespread adoption of this sustainable technology while minimising community noise impacts. This balanced approach could substantially contribute to meeting climate goals without compromising acoustic comfort in residential environments.

6. ACKNOWLEDGEMENTS

This study was supported by the University of Salford's Pump Prime Funding and the Future Homes Project, which is funded through the Innovation Accelerator programme by Innovate UK (Project No. 10054845).

7. REFERENCES

[1] European Commission, “European Green Deal: Commission proposes to boost renovation and decarbonisation of buildings,” European Commission. Accessed: Dec. 04, 2024. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6683

[2] Natural Resources Canada, “Heating and Cooling With a Heat Pump,” Natural Resources Canada. Accessed: Oct. 04, 2024. [Online]. Available: https://natural-resources.canada.ca/energy-efficiency/energy-star-canada/about/energy-star-announcements/publications/heating-and-cooling-heat-pump/6817

[3] P. Brandstätt, M. Krämer, and B. Kaltbeitzel, “Noise and Vibration Characteristics of Outdoor Heat Pumps,” 2024.

[4] R. Fumagalli *et al.*, “Acoustic Signatures of Heat Pumps Final Report – Part 4 1.2 Regulations - Countries overview,” Heat Pump Centre, 2020.

[5] C. H. Kasess, C. Reichl, H. Waubke, and P. Majdak, “Perception Rating of the Acoustic Emissions of Heat Pumps,” in *Forum Acusticum*, Lyon, France, Dec. 2020, pp. 2453–2458. doi: 10.48465/fa.2020.0363.

[6] C. Reichl, “IEA HPT Annex 51: Acoustic Signatures of Heat Pumps – Final Report,” Heat Pump Centre, HPT-AN51-1, 2020.

[7] *BS EN ISO 717-1:2020 Acoustics. Rating of sound insulation in buildings and of building elements - Airborne sound insulation*, 2020.

[8] C. Skinner and Grimwood, Colin, “The National Noise Incidence Study 2000/2001 (United Kingdom): Volume 1 - Noise Levels,” BRE Environment, United Kingdom, 206344f, 2000.

[9] T. Waters-Fuller and D. Lurcock, “NANR116: ‘Open/Closed Window Research’ Sound Insulation Through Ventilated Domestic Windows,” The Building Performance Centre, School of the Built Environment, Napier University, Apr. 2007.

[10] International Organization for Standardization, *PD ISO-TS 15666-2021- Acoustics — Assessment of noise annoyance by means of social and socio-acoustic surveys*, 2021.

[11] *DIN 45631/A1:2010-03 - Calculation of loudness level and loudness from the sound spectrum - Zwicker method - Amendment 1: Calculation of the loudness of time-variant sound*, 2010. [Online]. Available: https://doi.org/10.31030/1555185

[12] *ECMA-418-2, Psychoacoustic metrics for ITT equipment Part 2 (models based on human perception)*, 2022.

[13] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin: Springer, 2006.

[14] “3.2. Tuning the hyper-parameters of an estimator,” scikit-learn. Accessed: Apr. 07, 2025. [Online]. Available: https://scikit-learn/stable/modules/grid_search.html