

EchoFriendly: Analysis On The Classification of Enviromental Soundscape Using Computational Techniques and Machine Learning



Alfred Aubin Fogno Kengne

Department of Computer Science, Bioengineering, Robotics and
System Engineering (DIBRIS)

University of Genova

Supervisor

Gualtiero Volpe,
Nicola Corbellini

In partial fulfillment of the requirements for the degree of

Master of Science in Computer Engineering

October 15, 2026

Acknowledgements

Write here your acknowledgements...

Thesis dedication goes here...

Abstract

Sound is a fundamental component of the environments in which we live. The soundscape of a place can make us feel calm and relaxed, or disrupt our work, celebration or rest. Improving the eco-acoustic quality of these places requires a deep understanding of their soundscapes, and technology offers tools to identify the acoustic features that most affect inhabitants' quality of life. In this work, we combine acoustemology and machine learning, together with a heuristic approach, to automate soundscape analysis. Five classifiers (Random Forest, Radial Radial Basis Function Kernel Support Vector Classifier, Linear kernel Support Vector Classifier, Logistic Regression and LinearSVC with L1 norm) and a Baseline model based on the A-weighted equivalent sound level (LAeq) classify urban soundscapes as noisy or quiet, with the aim of supporting recommendation systems that guide people towards quieter places. Two feature sets are used: Quantitative features extracted from the audio signal, and Hybrid features, which add qualitative information from user questionnaires. The experiments rely on 68 recordings collected with a smartphone and other 68 with a digital recorder, analyzed over windows of 30, 60 and 90 seconds, and evaluated under three cross-validation schemes reproducing realistic usage scenarios: unseen recordings, unseen users, and both unseen users and recordings. Performance is measured by balanced accuracy over 50 repeated splits and compared

through paired significance tests and Friedman-Nemenyi tests. The results show that adding qualitative features significantly improves every machine learning model, and all of them significantly outperform the baseline when Hybrid features are used. The Baseline remains competitive only on smartphone recordings with Quantitative features. No classifier is consistently superior; in the cold-start scenario, i.e, with both new users and new recordings, the five are statistically indistinguishable. Device and window length have little effect on the classifiers, whereas the baseline is sensitive to both.

Contents

1	Introduction	1
1.1	Motivations	1
1.2	Context of the Study	2
1.3	Objectives and Contributions	3
1.4	Overview of the Thesis	4
2	State of the Art	5
3	Methodology	10
3.1	Dataset	10
3.1.1	Qualitative Features	10
3.1.2	Audio Labeling	13
3.1.3	Quantitative Features	14
3.1.4	Hybrid Features	15
3.2	Model Selection	16
3.2.1	Hyperparameter Selection	22
3.3	Model Training And Evaluation	23
3.3.1	Model Training	24
3.3.2	Model Evaluation	25
3.4	Model Comparison	26

3.4.1	Preliminary Considerations for Statistical Testing	26
3.4.2	RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?	29
3.4.3	RQ2 — Does the Hybrid Setting Improve Performance Over Quantitative Alone?	30
3.4.4	RQ3 — Is There a Consistently Best ML Model, or Are They Statistically Equivalent?	30
3.4.5	RQ4 — Do Device and Window Length Affect Performance?	32
4	Experimental Results	34
4.1	Model Evaluation	34
4.1.1	Audios cross-validation	34
4.1.2	Users cross-validation	41
4.1.3	Audio & User	45
4.2	Model Comparison	48
4.2.1	RQ1 - Is the Baseline Sufficient or Do ML Models Improve Performance	51
4.2.2	RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?	57
4.2.3	RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent	59
4.2.4	RQ4 - Do Device and Window Length Affect Performance?	71
4.2.4.1	Device effect (mobile vs zoom)	71
4.2.4.2	Window length effect (30s vs 60s vs 90s)	75
5	Discussion	80
5.1	RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?	81

CONTENTS

5.2	RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?	84
5.3	RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent?	86
5.4	RQ4 - Do Device and Window Length Affect Performance?	89
5.4.1	Device effect (mobile vs zoom)	89
5.4.2	Window length effect (mobile vs zoom)	92
6	Conclusions	94
	References	100

List of Figures

3.1	Number of audio recordings per label. Red indicates "very unpleasant", orange indicates "unpleasant", yellow indicates "pleasant", and Green indicates "very pleasant".	13
3.2	Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 30-second windows. Left: Mobile30 dataset (smartphone recordings). Right: Zoom30 dataset (digital recorder recordings)	19
3.3	Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 60-second windows. Left: Mobile60 dataset (smartphone recordings). Right: Zoom60 dataset (digital recorder recordings)	19
3.4	Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 30-second windows. Left: Mobile90 dataset (smartphone recordings). Right: Zoom90 dataset (digital recorder recordings)	20

LIST OF FIGURES

4.1	RQ3 - Results Nemenyi tests of models built on quantitative datasets in the audios cross-validation scheme	63
4.2	RQ3 - Results Nemenyi tests of models built on hybrid datasets in the audios cross-validation scheme	64
4.3	Results RQ3 tests of models trained using hybrid datasets in the Users cross-validation	67
4.4	Results RQ3 tests of models trained using hybrid datasets in the Audio & User cross-validation	70

List of Tables

3.1	Mapping of the indices appearing in the feature names shown in figs. 3.2 to 3.4	21
4.1	Results for RF for multiclass classification trained & tested on quantitative datasets, obtained from audios cross-validation. . . .	35
4.2	Results for RF trained & tested on quantitative datasets, obtained from audios cross-validation.	36
4.3	Results for RF trained & tested on hybrid datasets, obtained from audios cross-validation.	36
4.4	Results for BASELINE trained & tested on quantitative datasets, obtained from audios cross-validation.	37
4.5	Results for baseline trained & tested on hybrid datasets, obtained from audios cross-validation.	37
4.6	Results for rbf_svc trained & tested on quantitative datasets, ob- tained from audios cross-validation.	38
4.7	Results for rbf_svc trained & tested on hybrid datasets, obtained from audios cross-validation.	38
4.8	Results for LIN_SVC trained & tested on quantitative datasets, obtained from audios cross-validation.	39

LIST OF TABLES

4.9	Results for LIN_SVC trained & tested on hybrid datasets, obtained from audios cross-validation.	39
4.10	Results for LOG_REG trained & tested on quantitative datasets, obtained from audios cross-validation.	40
4.11	Results for LOG_REG trained & tested on hybrid datasets, obtained from audios cross-validation.	40
4.12	Results for LIN_SVC_L1 trained & tested on quantitative datasets, obtained from audios cross-validation.	41
4.13	Results for LIN_SVC_L1 trained & tested on hybrid datasets, obtained from audios cross-validation.	41
4.14	Results for RF trained & tested on hybrid datasets, obtained from Users cross-validation.	42
4.15	Results for baseline trained & tested on hybrid datasets, obtained from Users cross-validation.	42
4.16	Results for rbf_svc trained & tested on hybrid datasets, obtained from Users cross-validation.	43
4.17	Results for LIN_SVC trained & tested on hybrid datasets, obtained from Users cross-validation.	43
4.18	Results for LOG_REG trained & tested on hybrid datasets, obtained from Users cross-validation.	44
4.19	Results for LIN_SVC_L1 trained & tested on hybrid datasets, obtained from Users cross-validation.	44
4.20	Results for RF trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	45
4.21	Results for baseline trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	46

LIST OF TABLES

4.22	Results for rbf_svc trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	46
4.23	Results for LIN_SVC trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	47
4.24	Results for LOG_REG trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	47
4.25	Results for LOG_REG trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	48
4.26	Balanced accuracy of each model trained & tested on quantitative datasets, obtained from audios cross-validation.	49
4.27	Balanced accuracy of each model trained & tested on hybrid datasets, obtained from audios cross-validation.	49
4.28	Balanced accuracy of each model trained & tested on hybrid datasets, obtained from Users cross-validation.	50
4.29	Balanced accuracy of each model trained & tested on hybrid datasets, obtained from Audio & User cross-validation.	51
4.30	Results RQ1 tests of models built on quantitative datasets in the audios cross-validation scheme. * indicates that a significant difference has been found.	52
4.31	Results RQ1 tests of models built on hybrid datasets in the audios cross-validation scheme. * indicates that a significant difference has been found.	53
4.32	Results RQ1 tests of models built on hybrid datasets in the Users cross-validation scheme. * indicates that a significant difference has been found.	55

LIST OF TABLES

4.33 Results RQ1 tests of models built on hybrid datasets in the Audio & User cross-validation scheme. * indicates that a significant difference has been found.	56
4.34 Results RQ2 tests of models built on hybrid and on quantitative datasets (Hyb. Vs. Quant) in the audios cross-validation scheme. * indicates that a significant difference has been found.	59
4.35 device effect on quantitative datasets, obtained from audios cross-validation.	72
4.36 device effect on hybrid datasets, obtained from audios cross-validation. N.S.E.D. stands for No Strong Evidence of Difference	73
4.37 device effect on hybrid datasets, obtained from Users cross-validation. N.S.E.D. stands for No Strong Evidence of Difference	74
4.38 device effect on hybrid datasets, obtained from Audio & User cross-validation. N.S.E.D. stands for No Strong Evidence of Difference	75
4.39 window length effect on quantitative datasets, obtained from audios cross-validation.	76
4.40 window length effect on hybrid datasets, obtained from audios cross-validation.	77
4.41 window length effect on hybrid datasets, obtained from Users cross-validation.	78
4.42 window length effect on hybrid datasets, obtained from Audio & User cross-validation.	79

Chapter 1

Introduction

1.1 Motivations

Daily, we find ourselves immersed in the sounds of all the places in which we go, in which we are, and in which we rest. The soundscape of a place (which can be made up of interactions between natural, human, and artificial activities) therefore plays a fundamental role in the quality of our existence. For example, noise pollution has been associated with increased stress and distraction [1] and can cause sleep disturbances, decreased learning ability, poor reading comprehension, and concentration deficits [2; 3]. In contrast, sounds commonly associated with natural environments have been shown to promote the restoration of perceived attention in healthy subjects [4; 5], promote recovery from stress compared to urban soundscapes [6] and produce an improvement in cognitive functions [7; 8], health and well-being [9; 10; 11]. Technology can help improve the eco-acoustic quality of our lives from different perspectives. On the one hand, it can support research aimed at gaining an in-depth understanding of soundscapes, their effects, and the mechanisms behind those effects. On the other hand, it can provide users with recommendation systems that suggest places and situations characterized by

soundscapes suitable for their personal activities and preferences.

1.2 Context of the Study

Although the continuous evolution of technology can help improve the eco-acoustic quality of our lives, especially with the expansion of machine learning, research is mainly focusing on the detection and classification of music and single sounds (e.g., see [12; 13] for a survey, [14] and [15] for a recent review) or of music performance (e.g., [16]) while leaving underinvestigated the classification of entire soundscapes for which we can find only a few studies (e.g., see [17; 18; 19; 20; 21; 22]). Some of these have focused on human perception and qualitative description (e.g., see [18]), while others have relied on purely quantitative, signal-derived features (e.g., see [17; 19; 20; 21; 22]). However, following these directions, most of them have left three broad axes largely unexplored.

The first concerns the contribution that perceptual information could make if combined with the signal-derived features. This is of particular interest since the two research directions have been largely developed in isolation. Only [19; 22] have partially addressed this question, and even then, they did so separately from the classification model itself.

The second, and most largely unexamined, axis concerns how window length affects classification performance in longer, real-world recordings. Only [20] investigated this, and only within a narrow range (from 4 seconds down to 125 milliseconds).

The third concerns model selection. Aside from [17; 21; 22], who each compare a few algorithms, the reviewed studies generally rely on a single classifier, offering little insight into how model choice interacts with feature type or window length.

1.3 Objectives and Contributions

Thus, from different window lengths (30, 60, 90 seconds) of urban soundscape recordings collected using two different types of devices (smartphones/**Mobile** and digital recorders/**Zoom**), we created several datasets. Some of them were obtained by extracting quantitative features (L_{Aeq} , Zero Crossing Rate, chroma, MFCCs, and many others), while others were obtained by combining these quantitative features with qualitative ones derived from the participants' subjective evaluations, collected through a questionnaire (which included questions such as "What kind of environment are you in right now?" [Park - Monument], "Do you think the sound context is relevant to the appreciation of this place?" [No - I don't know - Yes], among others). These datasets were then used to train a Support Vector Classifier with Radial Basis Function Kernel (hereafter called RBF_SVC), a Support Vector Classifier with Linear Kernel (hereafter called LIN_SVC), a LinearSVC with L1 norm (hereafter called LIN_SVC_L1), a Logistic Regression model (hereafter LOG_REG), a Random Forest (hereafter called RF), and a Baseline Model. The latter was built using, as a discriminator to decide whether a location is noisy or quiet, the average training-set value of the most important quantitative feature of the RF model. With this setup, we can now evaluate: (i) whether the Baseline Model is sufficient, or whether machine learning models genuinely improve performance; (ii) whether hybrid acoustic features (i.e., the combination of quantitative and qualitative acoustic features) improve performance over quantitative features alone; (iii) whether there is a consistently best model, or whether the models are statistically equivalent; (iv) whether device type and window length affect performance.

1.4 Overview of the Thesis

In this direction, Chapter 2 presents the state of the art of what has been done so far regarding the classification of soundscapes using machine learning. Chapter 3 i) describes the methodology of how we made our datasets, ii) presents all the models that we have selected for our study, iii) describes the methodology of training and evaluation of those models, iv) and how we compare them. Finally, the experimental results are presented in Chapter 4 and discussed in Chapter 5.

Chapter 2

State of the Art

In the field of sound, the soundscape should be thought of as a landscape but for your ears, and just as a visual landscape describes what you see, a soundscape describes what you hear in an environment. So, just as a landscape can be described as being composed of trees, mountains, rivers, seas, clouds, sun, stars, etc., the soundscape can also have certain components such as human noise (Anthropophony), wildlife vocalizations (Biophony), meteorological phenomena (Geophony), and calm periods (Quiet) as C.A. Quinn et al. identified [19].

As mentioned previously, only a few studies have addressed soundscapes and Environmental Sound Classification (ESC), and, within this limited body of work, research has evolved along two largely separate tracks: one rooted in human perception and qualitative description, the other in purely quantitative, signal-derived features. **McGregor et al.** [18] established the perceptual foundations of this field by surveying how audio professionals and end users describe soundscapes, finding that both groups rely heavily on source- and action-based dimensions (dynamics, spatial cues, clarity) but share little common terminology, underscoring the difficulty of translating subjective auditory experience into a struc-

tured classification scheme. Building on this perceptual grounding, **Thorogood et al.** [20] operationalized one such dimension — the background/foreground distinction — into a computational task, training an SVM classifier on a small set of handcrafted spectral features derived from a human-labeled soundscape corpus; their results showed that machine classification could approximate average human agreement, but performance degraded sharply for the ambiguous mixed class when the analysis window was shortened, revealing the limits of small, single-source (only quantitative audio features) corpora and fixed short-duration windows.

In 2013, **Torija et al.** [22] extended this window length by using 3-minute recordings of urban locations, from which they extracted 14 descriptors, including an equivalent sound-pressure level and maximum–minimum sound-pressure level, among others. They showed that a Support Vector Machine trained with the Sequential Minimal Optimization algorithm (SMO) outperformed a standard SVM in classifying urban locations into their corresponding soundscape category (e.g., high traffic flow, quiet and fountains, motorway, and others) [23]. In contrast to **Thorogood et al.** [20], they proposed that combining acoustical evaluations — obtained by averaging all acoustical descriptors for each category — with perceptual evaluations — obtained by averaging the 15 semantic-differential (SD) scales for each reference soundscape category (e.g., Quiet (0)–Loud (10), Pleasant (0)–Unpleasant (10); see the full list of 15 SD scales in [23]) — could yield relevant information to properly diagnose a given urban location and, whether deviations between expected and estimated values are found, implement appropriate corrective measures. However, by focusing solely on 3-minute recordings, this study left the question of window length only partially addressed and, like **Thorogood et al.** [20], left open the question of what impact using both quantitative and qualitative features, rather than quantitative features alone, would

have on model performance.

Moving from creative sound design toward ecological monitoring, **Bellisario et al.** [17] applied a multi-label k-nearest neighbor algorithm to tag concurrent soundscape components (e.g., bird, wind, rain, car) in long-term field recordings, combining ecological acoustic indices with audio-profile and Gaussian Mixture Model (GMM) features; while achieving reasonable average precision when generalizing across years, their approach relied on a narrow, fixed feature set computed from only the first minute of each recording and a single classification algorithm, leaving also open how window length and model choice affect predictive accuracy. **Quinn et al.** [19] extended this ecoacoustic direction by applying a transfer-learning convolutional neural network (called ABGQI-CNN) to classify five broad soundscape components (A-Anthropophony, B-Biophony, G-Geophony, Q-Quiet, and I-Interference) from spectrogram images across a much larger citizen-science dataset, achieving strong performance and revealing meaningful temporal and geographic patterns; however their approach relied exclusively on short, fixed 2-second spectrogram windows and a single deep learning architecture, without testing larger or even different window lengths, or incorporating any qualitative input, such as perceptual features, beyond broad presence/absence labels. Moreover, by relying on a CNN applied to spectrogram images, their approach was constrained to leverage only one property of sound, that is frequency, given that a spectrogram is essentially a time-frequency representation, with time displayed along the horizontal axis, frequency along the vertical axis, and color intensity indicating the energy or amplitude associated with each frequency at a given moment; this leaves other important acoustic properties, such as loudness, A-weighted energy-equivalent sound-pressure level, and many others, effectively unexploited by the model. In a parallel but distinct line of work focused on general-purpose environmental sound classification, **Toffa et al.** [21] demonstrated that

texture descriptors borrowed from image processing (local binary patterns, applied to spectrograms and to the 1D signal) could outperform classical audio descriptors such as Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC). By combining these image-derived descriptors with classical audio descriptors — both essentially quantitative in nature — and using short, fixed-length clips, they achieved state-of-the-art results among handcrafted-feature methods on the ESC-10/ESC-50 benchmarks, outperforming classical machine learning models and some CNN-based approaches, though not all deep learning methods; as the authors themselves acknowledge, their method represents “only a good alternative when there is data scarcity or minimal computing power,” rather than a means of surpassing the strongest deep learning approaches. Furthermore, their study did not include perceptual input or systematic investigation of window size.

Taken together, these studies reveal a persistent gap in the literature. Perceptual studies establish which dimensions matter to human listeners (e.g., [18]) but do not test them computationally at scale; ecological and general-purpose ESC studies build increasingly sophisticated quantitative feature sets and classifiers, but rarely incorporate perceptual or qualitative information, nor systematically vary the analysis window length; and no study directly compares whether quantitative acoustic descriptors alone are sufficient, or whether human-perceptual features add meaningful predictive value.

Most of the reviewed studies are more concerned with identifying which sounds are present in a soundscape — pattern or content identification — than with directly characterizing a location’s overall noise level. Two exceptions to this trend are worth noting. **Quinn et al.** [19] represent a partial exception, since their “Quiet” class does touch on this dimension; however, this is not stated explicitly, and the authors themselves acknowledge in Section 4.2.4 that “there may be other

signals above 11 kHz (i.e., insects and bats) that our modeling efforts did not reflect, which would result in lower amounts of Quiet and higher Biophony,” since their spectrograms were restricted to the range 0–11 kHz. In any case, ”Quiet” remains only one label among five rather than the primary focus of the study. The second exception is **Torija et al.** [22], who combined acoustical (quantitative) and perceptual (qualitative) evaluations to obtain relevant information to properly diagnose a given urban location and, where deviations between expected and estimated values are found, to implement appropriate corrective measures. The limitation of their approach, however, is that these evaluations are conducted separately from the classification model rather than integrated within it.

This thesis addresses these gaps by computing a substantially larger set of quantitative descriptors across multiple window lengths (30, 60, and 90 seconds) to assess the effect of window size on classification performance, combining these descriptors with qualitative, perceptually-derived features to directly test their contribution, and benchmarking the resulting models — including newer architectures — to identify the best-performing configuration for classifying whether an environment is noisy or quiet.

Chapter 3

Methodology

3.1 Dataset

The dataset consists of audio recordings of soundscapes and their corresponding subjective evaluations, collected by groups of volunteer participants in the city of Turin. Acquisitions were made in various locations of interest (e.g., parks and monuments). For each site, a pair of simultaneous recordings was collected using two devices: a commercial smartphone (hereafter called **M - Mobile**) and a Zoom H1n digital recorder (hereafter called **Z - Zoom**).

In total, the corpus comprises 138 recordings (69 unique recordings per device). At the end of each recording session, participants individually completed a questionnaire to assess the qualitative characteristics of the soundscape (**Qualitative Audio Features**). Subsequently, computational metrics describing the acoustic properties of each audio track were extracted (**Quantitative Audio Features**).

3.1.1 Qualitative Features

Subjective evaluations were collected via a questionnaire that included the following types of questions:

- **Likert Scales (1-5):** to quantify the presence of sources (e.g., “How much do you hear traffic noise? (cars, motorcycles, buses, trains, ships, planes, etc.)”, “How much do you hear nature sounds? (animals, wind, water, rain, trees, etc.)”).
- **Multiple Choice:** to assess general perception (e.g., “What do you think of the sound context of the place you are currently in?” [Very comfortable - Very uncomfortable], “The soundscape is...” [Monotonous - Vibrant]).
- **Open-ended Questions:** to identify sources of comfort or discomfort (e.g., “Identify a sound source (if present) from the general sound context that conveys a sense of comfort”)

For the construction of the training dataset, open-ended and multiple-choice questions with optional missing answers were excluded. The final set includes 19 Qualitative Audio Features, which have been grouped according to the following categories:

Location & Setting:

- *What kind of environment are you in right now?* [Park - Monument]
- *Do you think the sound context is relevant to the appreciation of this place?* [No - I don’t know - Yes]
- *Does the overall sound context of this place match the expectations you had made of it before arriving here?* [No - I don’t know - Yes]
- *Do you think the overall sound context is unique to this place?* [No - I don’t know - Yes]
- *Do you think this place needs noise control measures?* [No - I don’t know - Yes]
- *Do you think there are sounds to protect in this place?* [No - I don’t know -

Yes]

– *What do you think of the sound context of where you are right now?* [It makes me very uncomfortable - It makes me uncomfortable - Neutral - Comfortable - Very comfortable]

– *Do you feel that your presence impacts the overall soundscape?* [No - I don't know - Yes]

Perceptual Impact:

– *How would you define the acoustic environment in which you are currently?* [Irritating - Neither one nor the other - Pleasant]

– *You think the soundscape is...* [Chaotic - Neither one nor the other - Calm]

– *Evaluate your degree of participation in the overall soundscape according to the following scale of values:* [1 – 5]

– *Does the soundscape have any kind of event?* [Eventless (daily life) - Neither one nor the other - With events (extraordinary event)]

Noise Sources:

– *How much does traffic noise sound? (car, motorcycle, bus, train, ship, plane, etc.)* [1 – 5]

– *How much do you hear the sound of construction, industry, electronic equipment, household appliances, etc.* [1 – 5]

– *How much do you hear the sound of music, radio, amplified music, musical instruments, disco, etc.* [1 – 5]

– *How many other sounds are heard (such as bells, whistles, fireworks, sirens, etc.)* [1 – 5]

– *How much do you hear the sounds of nature (animals, wind, water, rain, trees, etc.)* [1 – 5]

- *How much human sounds are heard (voices, conversations, laughter, singing, neighbors, footsteps, etc.)* [1 – 5].

Additionally, context features related to the time and day of registration have been added: – *hour* [0 – 24];

– *minutes* [1- 60];

– *day of the week* [Monday – Sunday];

– *time of the day* [morning – afternoon - evening - night].

3.1.2 Audio Labeling

Each participant has been asked to listen to the audio and label them from very unpleasant to very pleasant. The results of the study were accumulated to find the most agreed upon label for each of the samples using a simple majority voting strategy. In cases of a tie, the corresponding audio recording was discarded. As a result, one recording was discarded for each type of acquisition sensor, yielding a final dataset of 68 recordings per device.

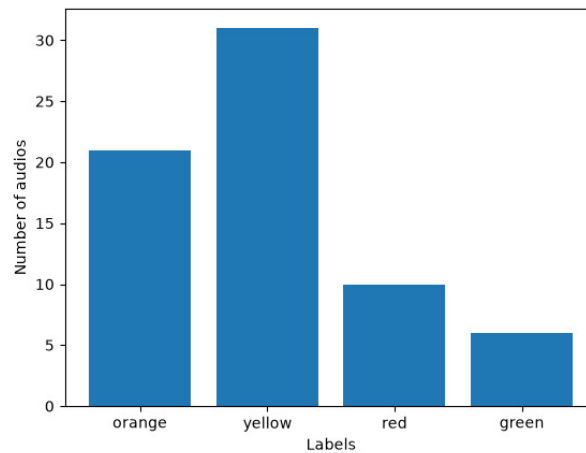


Figure 3.1: Number of audio recordings per label. Red indicates "very unpleasant", orange indicates "unpleasant", yellow indicates "pleasant", and Green indicates "very pleasant".

The audio recordings were evaluated as follows: 21 were labeled orange, 31 yellow, 10 red, and 6 green, as shown here above in figure 3.1

As part of the classification process, since machine learning algorithms require numerical inputs, we assigned numerical values to these labels as follows: Green = 0, Yellow = 1, Orange = 2, and Red = 3. This mapping is appropriate for multiclass classification; however, for binary classification, we took an additional step: the Green and Yellow labels were mapped to 0 (quiet), and the Orange and Red labels were mapped to 1 (noisy).

3.1.3 Quantitative Features

Acoustic features were calculated for each device (Mobile and Zoom) by unitizing the recordings into temporal segments of 90, 60 and 30 seconds. The final set includes 38 quantitative features inspired by previous work in soundscape analysis and classification (see [17; 18; 20; 21]). The features are grouped in the following categories.

- **Energy / Loudness features:** Root Mean Square (RMS); A-weighted average sound-pressure level (L_{Aeq}); 99th percentile of the A-weighted sound-pressure level (L_{A99}); minimum A-weighted sound-pressure level recorded (L_{Amin}); Crest Factor (CR) – the ratio between the peak value and the RMS value
- **Zero Crossing Rate (ZCR):** the frequency of changes in the signal’s sign.
- **Mel-Frequency Cepstral Coefficients (MFCC):** the first 13 cepstral coefficients on the Mel scale.
- **Chroma:** the temporal mean of the Chroma Short-Time Fourier Transform projecting the energy of the audio spectrum onto the 12 pitch classes of the chromatic scale.

- **Spectral Contrast:** the difference (in dBs) between peaks and valleys of the spectrum in different frequency bands.

3.1.4 Hybrid Features

The hybrid features were constructed by concatenating the quantitative and qualitative feature sets described above, resulting in a single feature vector for each split of recording that combines objective acoustic measurements with participants’ subjective perceptions of the environment. This combination allows the classification models to leverage both the physical properties of the soundscape and the human perceptual dimension, enabling an assessment of whether incorporating subjective evaluations improves classification performance beyond what is achievable using quantitative acoustic features alone.

Throughout our work, we use a concise notation to refer to quantitative features and to these combinations. For instance, **QuantM30**, **QuantM60**, and **QuantM90** refer, respectively, to models trained on quantitative features extracted from 30-second, 60-second, and 90-second Mobile audio recordings, while **QuantZ30**, **QuantZ60**, and **QuantZ90** refer, respectively, to models trained on quantitative features extracted from 30-second, 60-second, and 90-second Zoom audio recordings.

Similarly, **HM30**, **HM60**, and **HM90** refer, respectively, to models trained on both quantitative and qualitative features (Hybrid) extracted from 30-second, 60-second, and 90-second Mobile audio recordings, while **HZ30**, **HZ60**, and **HZ90** refer, respectively, to models trained on both quantitative and qualitative features (Hybrid) extracted from 30-second, 60-second, and 90-second Zoom audio recordings.

3.2 Model Selection

To classify urban soundscapes as noisy or quiet, six models were considered: five supervised machine learning classifiers, spanning different algorithmic families and levels of complexity, and one baseline model. This diversity allows us to compare simple, interpretable decision rules against increasingly flexible learning algorithms, and to assess whether the added complexity of machine learning translates into a genuine performance gain. All machine learning models were implemented using the scikit-learn library [24].

Support Vector Machine with RBF kernel (RBF_SVC) : . SVMs construct a decision boundary that maximizes the margin between classes. When classes are not linearly separable in the original feature space, the Radial Basis Function (RBF) kernel implicitly projects the data into a higher-dimensional space where a linear separation becomes possible, without explicitly computing that transformation (the so-called "kernel trick"). This makes the RBF_SVC particularly suited to capturing non-linear relationships between acoustic (and, where applicable, perceptual) features and the noisy/quiet classification, at the cost of reduced interpretability compared to linear models.

KERNEL LINEAR SVC (LIN_SVC) : As a counterpart to the RBF variant, the linear-kernel SVC seeks a maximum-margin hyperplane directly in the original feature space, without any non-linear transformation. Including this model alongside its RBF counterpart allows us to isolate the effect of non-linearity: if the linear and RBF variants perform comparably, this suggests that the relationship between the acoustic or perceptual features and soundscape classification is largely linear, and that the additional complexity of the RBF kernel is not warranted.

LINEAR SVC WITH L1 NORM (LIN_SVC_L1) : This model shares the linear decision boundary of the previous classifier, but replaces the standard L2 regularization with an L1 penalty. Unlike the two previous models, this one was not implemented using an SVC with a linear kernel; instead, we used LinearSVC, which differs in that it is implemented using LIBLINEAR rather than LIBSVM. The L1 norm encourages sparsity in the model’s coefficients, effectively performing implicit feature selection by shrinking the weights of less informative features to zero. This is particularly valuable given the relatively high dimensionality of the quantitative and qualitative feature sets used in this work, as it helps identify which acoustic and perceptual descriptors contribute most to distinguishing noisy from quiet environments.

LOGISTIC REGRESSION (LOG_REG) : Logistic Regression models the probability of a location being classified as noisy (versus quiet) as a logistic function of a linear combination of the input features. Despite its simplicity, it remains a strong and widely used baseline in binary classification tasks, offering directly interpretable coefficients that indicate the direction and relative magnitude of each feature’s contribution. In this work, an L1 penalty was applied, for the same feature-selection rationale described above.

RANDOM FOREST (RF) : Random Forest is an ensemble method that aggregates the predictions of a large number of decision trees, each trained on a bootstrap-resampled subset of the data and a random subset of features. This dual randomization reduces overfitting and typically yields robust performance even in the presence of noisy or correlated features, while also providing a natural measure of feature importance. Unlike the SVM-based and linear models above, Random Forest can capture complex, non-linear, and non-monotonic interactions between features without requiring an explicit kernel or transformation, making

it a valuable point of comparison in terms of both accuracy and model interpretability.

In this work, the Random Forest has been developed to do 2 different tasks. The first one is the multiclass classification task, assigning each environment to one of four classes (3 - Very unpleasant, 2 - Unpleasant, 1 - Pleasant, 0 - Very pleasant). The second is the binary classification task, categorizing each environment as either noisy or quiet.

BASELINE MODEL : The Baseline Model was built by exploiting the interpretability property of Random Forest. Specifically, we examined the permutation feature importance, which identifies the features on which the Random Forest relies most heavily when making its binary classification decisions in the Audios cross-validation scheme.

L_{Aeq} emerged repeatedly, and with considerable values, when computing the permutation importance for the different Random Forest models previously trained on the quantitative datasets extracted from recordings split into the various window lengths (QuantM30, QuantM60, QuantM90, QuantZ30, QuantZ60, QuantZ90); see the box plots in figs. 3.2 to 3.4 and table 3.1.

Based on this observation, we computed the average L_{Aeq} value over the training and validation sets of each dataset and used it as a discriminator: each element of the test set was classified as quiet if its L_{Aeq} value fell below this discriminator, and as noisy otherwise.

3.2 Model Selection

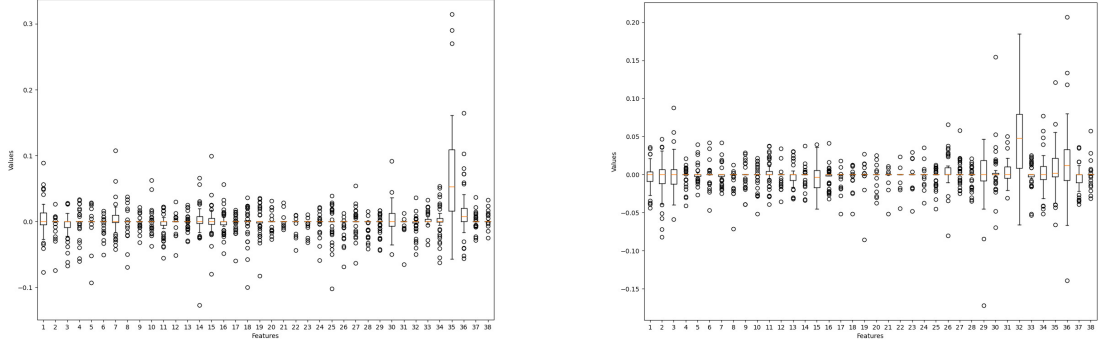


Figure 3.2: Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 30-second windows. Left: Mobile30 dataset (smart-phone recordings). Right: Zoom30 dataset (digital recorder recordings)

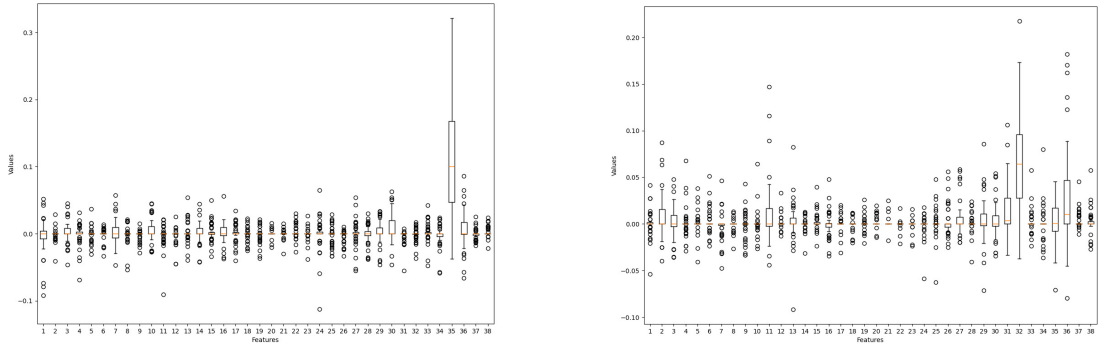


Figure 3.3: Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 60-second windows. Left: Mobile60 dataset (smart-phone recordings). Right: Zoom60 dataset (digital recorder recordings)

3.2 Model Selection

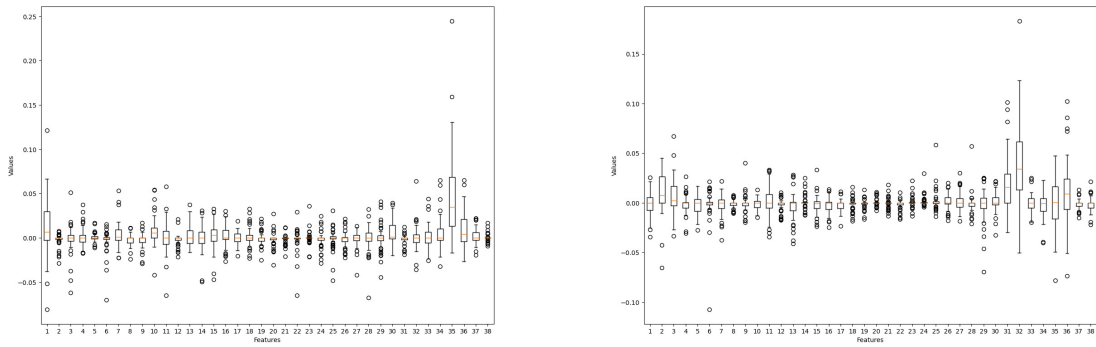


Figure 3.4: Box plots of permutation feature importance obtained by training a Random Forest for binary classification on the quantitative datasets, extracted from recordings split into 30-second windows. Left: Mobile90 dataset (smart-phone recordings). Right: Zoom90 dataset (digital recorder recordings)

3.2 Model Selection

index feature	name feature
1	MFCC_1_mean
2	MFCC_2_mean
3	MFCC_3_mean
4	MFCC_4_mean
5	MFCC_5_mean
6	MFCC_6_mean
7	MFCC_7_mean
8	MFCC_8_mean
9	MFCC_9_mean
10	MFCC_10_mean
11	MFCC_11_mean
12	MFCC_12_mean
13	MFCC_13_mean
14	Chroma_1_mean
15	Chroma_2_mean
16	Chroma_3_mean
17	Chroma_4_mean
18	Chroma_5_mean
19	Chroma_6_mean
20	Chroma_7_mean
21	Chroma_8_mean
22	Chroma_9_mean
23	Chroma_10_mean
24	Chroma_11_mean
25	Chroma_12_mean
26	Spectral_Contrast_1_mean
27	Spectral_Contrast_2_mean
28	Spectral_Contrast_3_mean
29	Spectral_Contrast_4_mean
30	Spectral_Contrast_5_mean
31	Spectral_Contrast_6_mean
32	Spectral_Contrast_7_mean
33	Zero_Crossing_Rate_mean
34	RMS_Energy_mean
35	LAeq
36	LA1
37	LAImin
38	Crest_Factor

Table 3.1: Mapping of the indices appearing in the feature names shown in figs. 3.2 to 3.4

3.2.1 Hyperparameter Selection

This subsection describes how the hyperparameters of each classifier were selected.

RBF_SVC The performance of the RBF SVC depends critically on two hyperparameters: the regularization parameter C , which controls the trade-off between margin width and classification error, and γ , which governs the influence radius of each training point. Both were tuned by grid search (10-fold cross-validation, accuracy as the scoring metric), exploring ten logarithmically spaced values between 10^{-4} and 10^3 for each parameter. The best-performing combination was retained for evaluation on the test set.

LIN_SVC The same procedure was followed for the LIN_SVC: a 10-fold grid search over the regularization parameter C , again spanning ten logarithmically spaced values between 10^{-4} and 10^3 , with accuracy as the scoring metric.

LIN_SVC_L1 The regularization strength C was tuned by 10-fold grid search over the same logarithmic range (10^{-4} to 10^3 , ten values). The configuration that maximized the cross-validated accuracy was selected.

LOG_REG An L1 penalty was applied, for the same feature-selection rationale described above. The regularization parameter C was tuned by 10-fold grid search over ten logarithmically spaced values between 10^{-4} and 10^3 , with accuracy as the scoring metric.

RF Given the larger number of hyperparameters involved (due to its structure), tuning was performed by random search rather than by an exhaustive grid. For a fixed number of iterations, the following hyperparameters were sampled independently:

3.3 Model Training And Evaluation

- number of estimators (50 to 990, in steps of 10);
- maximum tree depth (unrestricted, or between 1 and 20);
- minimum number of samples required to split an internal node (2 to 20);
- minimum number of samples required at a leaf node (1 to 20);
- maximum number of features considered at each split (unrestricted, or a fraction between 0.1 and 0.9);
- use of bootstrap sampling (yes or no)

The classifier was additionally configured with *class_weight* = “*balanced_subsample*” to mitigate any class imbalance, both in the binary and in the multiclass task.

Unlike the other models, hyperparameter selection for Random Forest combined random search with 3-fold cross-validation. A total of 100 configurations were sampled from the distributions above, and each was evaluated by 3-fold cross-validation. This yielded three F1-scores per configuration (macro-averaged for the multiclass task, or $F1_1$, computed on the noisy class, for the binary task), and therefore 300 F1-scores in total. The configuration associated with the single highest F1-score among these 300 values was selected to build the final model.

3.3 Model Training And Evaluation

Having defined the datasets and the set of candidate models, this section describes how each model was trained, tuned, and evaluated. We first outline the training procedure and the hyperparameter optimization strategy adopted for these classifiers, followed by the evaluation metrics used to assess and compare their performance.

3.3.1 Model Training

To evaluate the robustness of the models, we designed three testing protocols by splitting the training and test sets according to three stratification strategies to simulate different levels of generalization:

1. **Unseen users (Users):** the test set contains users absent in the training set, but the evaluated audio recordings might be the same (i.e., already “heard” by the model during training via other users). This situation corresponds to the recommendation system predicting how a new user will evaluate a soundscape, which is already part of an archive of soundscapes whose acoustic features were previously computed and processed.
2. **Unseen audios (Audios):** the test set contains audio recordings absent in the training set, but potentially has access to subjective evaluations from a user who has uploaded multiple recordings. This situation corresponds to the recommendation system predicting how a user who has already used the system and recorded her preferences (i.e., the Qualitative Audio Features for other soundscapes) would evaluate a new soundscape, for example, one she is capturing at that time with her smartphone.
3. **Unseen users and audios (Users $\times$ Audios):** the most complex condition, where both users and recordings in the test set are completely novel. This situation corresponds to the recommendation system predicting how a new user would evaluate a new soundscape, e.g., a user installing the recommendation system’s app and capturing the soundscape at that time to get immediate information about its quality.

3.3.2 Model Evaluation

To assess and compare the performance of the trained models, a set of evaluation metrics was computed for each model’s predictions on the validation and test set. Since two distinct classification tasks were addressed in this work, binary and multiclass, the choice of metrics was adapted accordingly, as detailed below. For both tasks overall accuracy were computed as a general summary of classification performance.

- **Binary Classification** For the binary classification task (noisy vs. quiet), performance was evaluated using precision, recall, and F1-score computed separately for each class, rather than averaged, in order to assess the model’s ability to correctly identify both noisy and quiet locations independently. In addition, balanced accuracy was computed to account for potential class imbalance between the two categories, providing a more reliable summary metric than overall accuracy alone in such cases.
- **Multiclass Classification** For the multiclass classification task (Green, Yellow, Orange, Red), precision, recall, and F1-score were computed using macro-averaging, meaning that the metric is calculated independently for each class and then averaged, giving equal weight to each class regardless of its frequency in the dataset. As in the binary case, balanced accuracy was also computed to mitigate the effect of class imbalance across the four categories.

To test the models, we retained approximately 20% of the data, enforcing one of the aforementioned protocols, while the remaining data was used for hyperparameter tuning, as described above. Moreover, we repeated the process above 50 times to sample different test sets using different seeds.

3.4 Model Comparison

Before proceeding with the model comparison (RQ1, RQ2, RQ3, RQ4), it is necessary to recall the structural properties of the experimental setup, since they motivate every methodological choice adopted in the analysis that follows.

Our work span 12 experimental conditions, obtained from the factorial combination of:

- **Feature type:** Quantitative, Hybrid (2 levels)
- **Recording device:** Mobile, Zoom (2 levels)
- **Window length:** 30s, 60s, 90s (3 levels)

For each condition, six models were evaluated: the Baseline and five machine learning classifiers (RBF SVC, Kernel Linear SVC, Logistic Regression, LinearSVC with L1 norm, and Random Forest). Each model was evaluated on the same 50 test splits for each cross-validation protocol, yielding a distribution of 50 balanced accuracy scores per model, per condition, and per cross-validation protocol.

For all these models, for each cross-validation protocol and for each feature type, it has been reported mean and standard deviation of the balanced accuracy.

3.4.1 Preliminary Considerations for Statistical Testing

This shared set of 50 splits, mentioned above, constitutes the most important structural property of the evaluation design: since all models are evaluated on identical test data, any pairwise comparison between two models within the same condition involves paired observations rather than independent samples. This

property requires the use of paired statistical tests throughout the analysis and substantially increases the statistical power of each comparison.

- **Normality Check**

Parametric tests, such as the paired t-test, assume that the differences between paired observations are approximately normally distributed; violating this assumption can inflate the Type I error rate or reduce statistical power. Although the Central Limit Theorem offers partial protection at $n = 50$, balanced accuracy scores are bounded between 0 and 1 and can exhibit skew, particularly when a model performs near chance (≈ 0.5) or near ceiling (≈ 0.9). For this reason, normality was verified rather than assumed.

For every pair of models compared within a given condition, the 50 pairwise differences (score of model A - score of model B, computed for each split) were tested for normality using the Shapiro-Wilk test, which is well suited to $n = 50$ due to its statistical power at this sample size and its established use in the range $n = 3-50$. It is important to note that the test was applied to the differences between paired scores, rather than to the raw scores themselves, as this is the correct assumption to verify for a paired test. Under the null hypothesis, the differences are normally distributed; a significant result ($p < 0.05$) indicates that normality cannot be assumed.

Based on the outcome of the Shapiro-Wilk test, the following decision rule was applied uniformly throughout the analysis: if the differences were normally distributed ($p \geq 0.05$), the paired t-test was used; otherwise ($p < 0.05$), the Wilcoxon signed-rank test was used. In practice, results were mixed across conditions. Given that a substantial proportion of pairs were found to be non-normal, the Wilcoxon signed-rank test was adopted throughout the analysis, as a pragmatic and conservative choice; t-test re-

sults were nonetheless consistent with the Wilcoxon results wherever normality held. Although the Wilcoxon test is slightly less powerful than the paired t-test under normality, it remains valid regardless of the distribution of the differences, making its universal application a safe methodological choice.

- **Multiple Comparisons and Error Rate Control** Since more than one statistical test was performed, the multiple comparisons problem was addressed: the probability of obtaining at least one false positive by chance increases with the number of tests performed. With five machine learning models compared against the Baseline across 12 conditions, 60 tests were run for this comparison alone; without correction, approximately 3 false positives would be expected at $\alpha = 0.05$ purely by chance.

To control for this, the False Discovery Rate (FDR) was controlled using the Benjamini-Hochberg (BH) procedure, rather than the more conservative Bonferroni correction. This choice was motivated by the fact that the Bonferroni correction, which controls the Family-Wise Error Rate, is excessively strict when the number of tests is large and the tests are correlated, as is the case here, since the same 50 splits underlie all comparisons. FDR control instead allows a specified proportion of significant findings to be false positives, offering a more appropriate trade-off for this exploratory and confirmatory evaluation.

The BH procedure is applied as follows: all tests within a given family were run and their raw p-values collected; the p-values were then ranked from smallest (rank 1) to largest (rank m , where m is the total number of tests in the family); the largest rank k satisfying $p(k) \leq (k/m) \times \alpha$ was identified; and all tests with rank $\leq k$ were declared significant.

Actually, we did not implement the BH as above, we used the one present in the statsmodels library [25].

3.4.2 RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?

This research question asks whether any of the five machine learning models provides a statistically significant improvement over the Baseline, and whether the answer differs depending on the feature type used (Quantitative vs. Hybrid). The procedure described above was applied to confirm or disconfirm this pattern rigorously.

The analysis was separated by feature type, since the Quantitative and Hybrid settings represent fundamentally different conditions: the Hybrid setting introduces questionnaire-derived features that the Baseline Model cannot exploit in the same way as the machine learning classifiers, given that it relies solely on a single quantitative acoustic feature as its discriminator. Treating both settings as a single family would therefore obscure this distinction, which is central to interpreting the results.

For each condition within the Quantitative setting (6 conditions, obtained from the combination of 2 devices and 3 window lengths), 5 paired tests were performed, one per machine learning model against the Baseline. The Benjamini-Hochberg FDR correction was then applied across all 30 tests (5 models $\times$ 6 conditions) within the Quantitative family. The same procedure was repeated independently for the Hybrid setting, yielding a second family of 30 tests corrected separately.

3.4.3 RQ2 — Does the Hybrid Setting Improve Performance Over Quantitative Alone?

Unlike RQ1, this research question does not concern whether machine learning models outperform the Baseline; rather, it asks whether adding questionnaire-derived features (the Hybrid setting) benefits each model relative to using acoustic features alone (the Quantitative setting). This is therefore a within-model comparison across feature types, with device and window length held constant.

For each of the six models (including the Baseline) and for each of the six device–window length combinations, the 50 paired differences in balanced accuracy between the Hybrid and Quantitative settings were computed for each test split. This yielded a total of 36 paired comparisons (6 models $\times$ 6 device–window combinations). For each comparison, the normality check described above was applied to the corresponding set of paired differences, and the appropriate test (paired t-test or Wilcoxon signed-rank test) was selected accordingly; the Benjamini-Hochberg FDR correction was then applied jointly across all 36 tests, treated as a single family.

For each comparison, the following quantities were reported: the mean balanced accuracy in the Quantitative setting, the mean balanced accuracy in the Hybrid setting, their difference ($\Delta = \text{Hybrid} - \text{Quantitative}$), the raw p-value, the BH-corrected p-value, and the corresponding significance flag.

3.4.4 RQ3 — Is There a Consistently Best ML Model, or Are They Statistically Equivalent?

When comparing more than two models, performing all possible pairwise tests, without a prior omnibus test is methodologically problematic: pairwise tests do not account for the dependency structure among the comparisons, and the

likelihood of false positives increases rapidly with their number. For this reason, a two-stage procedure was adopted: first, an omnibus test was used to assess whether any difference existed among the models; pairwise post-hoc comparisons were then performed only if this omnibus test proved significant.

- **Friedman test (Omnibus):** We choose to use the Friedman test because it is the non-parametric equivalent of a one-way repeated-measures ANOVA. This choice is appropriate for three reasons: first, the same 50 test splits were used to evaluate all models, so the resulting measurements are related (i.e., the 50 splits are the "subjects" and the models are the "conditions"); second, balanced accuracy is ordinal-compatible, meaning that observations can be meaningfully ranked; and third, normality of the 50 scores obtained for each model cannot be assumed.

The Friedman test operates by ranking the balanced accuracy scores within each test split across the six models considered, and then testing whether the resulting average ranks differ significantly across models. Under the null hypothesis that all models share the same performance distribution, the corresponding test statistic follows a chi-squared distribution with $k - 1$ degrees of freedom.

- **Post-hoc test: Nemenyi test** Nemenyi test is performed only when the Friedman test is significant for a given condition. It has specifically been designed as a post-hoc procedure following the Friedman test and compares average ranks between pairs of models and uses a studentized range distribution to control the family-wise error rate. Because it is based on ranks rather than raw scores, it does not require any distributional assumption. For each significant condition, the full matrix of pairwise p-values across all model comparisons was reported, with pairs satisfying $p < 0.05$ highlighted.

No additional Benjamini-Hochberg correction was applied at this stage, as the Nemenyi test already corrects internally for multiple comparisons.

3.4.5 RQ4 — Do Device and Window Length Affect Performance?

This research question is composed of two parts as shown below

- Device effect (Mobile vs Zoom)** For each model–feature type combination, three paired score distributions were available for Mobile recordings (one per window length) and three for Zoom recordings. Rather than concatenating the 50 scores across the three window lengths for each device, which would introduce within-sample dependence due to each test split appearing three times, performance was first averaged across window lengths for each split and device, yielding a single mean score per split and per device (50 values per device, per model). Then, a paired Wilcoxon signed-rank test or t-test was applied to these averaged values, preserving the paired structure of the data while avoiding the dependence problem. The Benjamini-Hochberg FDR correction was applied across the 12 resulting device comparisons ($6 \text{ models} \times 2 \text{ feature types}$), and the corresponding effect sizes were reported.
- Window length effect (30s vs 60s vs 90s)** An analogous procedure was followed to assess the effect of window length. For each model–feature type combination, two paired score distributions were available for each window length (one per device). Performance was first averaged across devices for each split and window length, yielding a single mean score per split and per window length (50 values per window length, per model), thereby avoiding the within-sample dependence that would otherwise result from pooling

scores across devices. Since this is inherently a multi-level factor (3 levels) rather than a binary comparison, a pairwise test is not appropriate as a first step. Instead we used A two-stage procedure: a Friedman omnibus test, and the Nemenyi post-hoc test only when the Friedman test was significant (see 3.4.5 for more details).

Chapter 4

Experimental Results

4.1 Model Evaluation

This section reports the results obtained by each model on both the quantitative and hybrid datasets, across the six device–window length conditions (Mobile and Zoom, at 30, 60, and 90 seconds). It should be noted that both quantitative and hybrid results are available for the audios cross-validation, whereas only hybrid results are available for the Users and Audio & User cross-validations: since these two cross-validation schemes split the data based on user identity, and the quantitative datasets do not retain this information, the corresponding models cannot be trained in the quantitative setting. For each model, accuracy, F1-score, precision, and recall are reported separately for the quiet (0) and noisy (1) classes, together with balanced accuracy as the primary summary metric. Full numeric results are reported in the corresponding tables.

4.1.1 Audios cross-validation

We begin with the multiclass Random Forest, the only model in this work trained to perform multiclass rather than binary classification and have trained only for

the quantitative datasets and for the audios cross-validation. We then proceed to the binary Random Forest and to the Baseline Model built, before turning to the remaining binary classifiers.

RF Model for Multiclass Classification (Quantitative) A Random Forest model was trained to perform multiclass classification, assigning each recording to one of the four original pleasantness classes (Green, Yellow, Orange, Red). Unlike every other model discussed in this section, RF is the only classifier that was also trained and evaluated in its multiclass form; there is no corresponding binary-versus-multiclass comparison for the other models. Table 4.1 reports the results obtained on the quantitative datasets. Overall performance is considerably lower than that observed for the binary classifiers presented later in this section, as expected given the increased number of classes: accuracy ranged from 36.71% (QuantZ60) to 43.37% (QuantM30), and balanced accuracy ranged from 31.86% (QuantZ60) to 40.12% (QuantM30). As will also be observed for the binary classifiers, Mobile conditions outperformed Zoom conditions across every window length; however, performance here declined as window length increased for Mobile recordings, from 40.12% (QuantM30) down to 34.24% (QuantM90), a trend opposite to the one that will emerge for the binary Random Forest model.

Condition	Accuracy	F1	Precision	Recall	Balanced Accuracy
QuantM30	43.37%	42.33%	36.81%	40.69%	40.12%
QuantM60	41.76%	38.15%	33.64%	37.62%	35.75%
QuantM90	40.98%	36.42%	31.71%	34.66%	34.24%
QuantZ30	37.99%	36.46%	30.72%	34.56%	34.51%
QuantZ60	36.71%	34.00%	28.29%	31.14%	31.86%
QuantZ90	37.77%	35.39%	29.72%	33.24%	33.30%

Table 4.1: Results for RF for multiclass classification trained & tested on quantitative datasets, obtained from audios cross-validation.

4.1 Model Evaluation

RF Model (Quantitative, hybrid) Having examined the multiclass results, we now turn to the binary RF classifier, presented ahead of the remaining binary models since its interpretability property (i.e., feature importance) was leveraged to identify the most consistently influential quantitative feature, LAeq, which was used to build our Baseline. On the quantitative datasets (Table 4.2), balanced accuracy ranged from 68.11% (QuantZ60) to 74.63% (QuantM60), with Mobile conditions consistently outperforming Zoom conditions across all three window lengths.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	70.82%	71.97%	67.90%	73.35%	69.25%	72.83%	69.38%	71.10%
QuantM60	74.53%	75.48%	71.73%	77.02%	73.50%	76.21%	73.06%	74.63%
QuantM90	73.43%	74.48%	70.82%	76.55%	72.21%	75.13%	72.67%	73.90%
QuantZ30	69.63%	69.72%	67.27%	74.23%	66.90%	68.44%	70.68%	69.56%
QuantZ60	68.43%	68.75%	65.32%	72.77%	65.62%	67.97%	68.25%	68.11%
QuantZ90	69.03%	69.94%	65.33%	73.79%	66.70%	69.73%	68.10%	68.91%

Table 4.2: Results for RF trained & tested on quantitative datasets, obtained from audios cross-validation.

On the hybrid datasets (Table 4.3), Random Forest achieves the highest and most consistent results among the binary classifiers considered in this section, with balanced accuracy ranging from 86.15% (HM60) to 87.58% (HZ30). Notably, the clear device gap present in the quantitative setting disappears almost entirely here, with Mobile and Zoom conditions performing within roughly one point of each other across all window lengths.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	87.35%	87.72%	85.62%	88.29%	86.15%	88.27%	86.51%	87.39%
HM60	86.12%	86.39%	84.20%	88.17%	83.68%	85.91%	86.39%	86.15%
HM90	86.43%	86.64%	84.78%	88.26%	84.05%	86.10%	86.89%	86.49%
HZ30	87.43%	87.69%	85.99%	89.08%	85.57%	87.42%	87.74%	87.58%
HZ60	86.86%	87.24%	85.03%	89.08%	84.45%	86.70%	86.93%	86.82%
HZ90	87.00%	87.07%	85.64%	89.34%	84.72%	86.13%	88.01%	87.07%

Table 4.3: Results for RF trained & tested on hybrid datasets, obtained from audios cross-validation.

4.1 Model Evaluation

Baseline (Quantitative, hybrid) Having identified LAeq as the most consistently important feature for RF, we now turn to our Baseline model, which uses this feature directly as a discriminator between noisy and quiet locations. On the quantitative datasets (Table 4.4), balanced accuracy ranged from 63.11% (QuantZ30) to 70.40% (QuantM90), with Mobile recordings consistently outperforming Zoom recordings across all three window lengths. A further asymmetry is visible between classes: precision was systematically higher for the quiet class than for the noisy class, while recall followed the opposite trend.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	69.97%	69.19%	68.99%	75.10%	65.61%	65.75%	74.61%	70.18%
QuantM60	69.98%	68.81%	69.29%	75.44%	65.49%	64.95%	75.45%	70.20%
QuantM90	70.21%	68.97%	69.48%	76.51%	65.42%	64.66%	76.15%	70.40%
QuantZ30	63.27%	61.20%	62.72%	68.94%	58.47%	56.71%	69.52%	63.11%
QuantZ60	64.95%	63.53%	63.77%	69.82%	60.66%	60.12%	69.28%	64.70%
QuantZ90	66.26%	63.86%	65.62%	72.97%	61.27%	59.08%	73.31%	66.20%

Table 4.4: Results for BASELINE trained & tested on quantitative datasets, obtained from audios cross-validation.

Moving to the hybrid setting (Table 4.5), the same overall structure persists, with comparable performances to the quantitative setting: balanced accuracy ranges from 63.11% (HZ30) to 70.40% (HM90), and Mobile recordings continue to outperform Zoom recordings throughout. The asymmetry between classes persists in the hybrid setting as well: precision is again systematically higher for the quiet class than for the noisy class, while recall again shows the opposite trend.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	69.97%	69.19%	68.99%	75.10%	65.61%	65.75%	74.61%	70.18%
HM60	69.98%	68.81%	69.29%	75.44%	65.49%	64.95%	75.45%	70.20%
HM90	70.21%	68.97%	69.48%	76.51%	65.42%	64.66%	76.15%	70.40%
HZ30	63.27%	61.20%	62.72%	68.94%	58.47%	56.71%	69.52%	63.11%
HZ60	64.95%	63.53%	63.77%	69.82%	60.66%	60.12%	69.28%	64.70%
HZ90	66.26%	63.86%	65.62%	72.97%	61.27%	59.08%	73.31%	66.20%

Table 4.5: Results for baseline trained & tested on hybrid datasets, obtained from audios cross-validation.

4.1 Model Evaluation

RBF_SVC Model (Quantitative, hybrid) Turning to the machine learning classifiers, we begin with the `rbf_svc`. On the quantitative datasets (Table 4.6), balanced accuracy ranged from 69.89% (QuantM30) to 74.39% (QuantZ60); notably, the device ranking is reversed compared to the baseline model, with Zoom recordings now outperforming Mobile recordings across every window length. As with the baseline, however, F1-score remained consistently higher for the quiet class than for the noisy class.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	69.35%	71.68%	65.33%	71.12%	69.79%	74.98%	64.81%	69.89%
QuantM60	70.40%	72.54%	66.69%	72.99%	70.31%	75.23%	67.04%	71.13%
QuantM90	70.18%	72.56%	65.63%	72.33%	72.10%	76.79%	64.60%	70.70%
QuantZ30	73.60%	75.29%	70.07%	75.11%	72.62%	77.26%	69.79%	73.52%
QuantZ60	74.71%	76.64%	70.64%	75.25%	74.90%	79.96%	68.82%	74.39%
QuantZ90	74.38%	76.91%	69.40%	73.66%	76.66%	82.37%	65.53%	73.95%

Table 4.6: Results for `rbf_svc` trained & tested on quantitative datasets, obtained from audios cross-validation.

Introducing qualitative features produces a marked shift for this model: on the hybrid datasets (Table 4.7), balanced accuracy climbs to between 80.29% (HM30) and 85.22% (HZ60), well above anything observed in the quantitative setting. The device ranking from the quantitative case is preserved, with Zoom still ahead of Mobile, and the quiet class continues to be classified more reliably than the noisy class.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	80.21%	81.06%	77.34%	82.17%	78.35%	81.61%	78.98%	80.29%
HM60	81.91%	82.51%	79.31%	84.13%	79.27%	82.21%	81.21%	81.71%
HM90	81.17%	82.03%	78.19%	82.10%	79.80%	83.38%	78.52%	80.95%
HZ30	84.53%	84.90%	82.43%	85.68%	83.12%	85.30%	83.17%	84.24%
HZ60	85.63%	86.18%	83.34%	86.23%	84.88%	87.22%	83.21%	85.22%
HZ90	83.62%	84.42%	81.00%	84.60%	82.48%	85.56%	81.19%	83.38%

Table 4.7: Results for `rbf_svc` trained & tested on hybrid datasets, obtained from audios cross-validation.

4.1 Model Evaluation

LIN_SVC Model (Quantitative, hybrid) A similar picture emerges for the linear-kernel counterpart of the SVC. On the quantitative datasets (Table 4.8), balanced accuracy ranged from 70.83% (QuantM90) to 75.91% (QuantZ60), with Zoom conditions again outperforming Mobile conditions throughout, and F1-score consistently favoring the quiet class.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	71.41%	73.37%	68.01%	72.95%	71.57%	76.22%	67.40%	71.81%
QuantM60	70.89%	73.20%	67.19%	72.74%	73.00%	77.16%	66.49%	71.83%
QuantM90	70.42%	73.24%	65.66%	71.47%	72.94%	78.35%	63.31%	70.83%
QuantZ30	74.46%	76.02%	71.15%	75.52%	73.74%	78.17%	70.65%	74.41%
QuantZ60	76.12%	77.88%	72.42%	76.58%	76.53%	81.01%	70.80%	75.91%
QuantZ90	73.77%	76.20%	68.97%	73.69%	75.22%	80.89%	65.99%	73.44%

Table 4.8: Results for LIN_SVC trained & tested on quantitative datasets, obtained from audios cross-validation.

On the hybrid datasets (Table 4.9), balanced accuracy rises to between 82.08% (HM90) and 83.13% (HM60); interestingly, this gain comes with a notable change in structure, as the device gap seen in the quantitative setting largely disappears, with Mobile and Zoom conditions now performing within a narrow one-point range of each other.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	82.97%	83.78%	80.42%	84.31%	81.49%	84.65%	81.41%	83.03%
HM60	83.29%	84.19%	80.35%	84.22%	81.84%	85.37%	80.89%	83.13%
HM90	82.27%	82.57%	79.88%	82.89%	81.75%	83.81%	80.35%	82.08%
HZ30	82.89%	83.68%	80.08%	84.05%	82.00%	84.91%	80.22%	82.56%
HZ60	82.78%	83.95%	79.47%	83.56%	82.25%	85.79%	79.04%	82.42%
HZ90	83.03%	83.85%	80.37%	83.91%	82.01%	85.16%	80.60%	82.88%

Table 4.9: Results for LIN_SVC trained & tested on hybrid datasets, obtained from audios cross-validation.

LOG_REG Model (Quantitative, hybrid) Logistic Regression follows a pattern broadly consistent with the two SVM-based models. On the quantitative datasets (Table 4.10), balanced accuracy ranged from 71.25% (QuantM30) to 76.45% (QuantZ30), with Zoom conditions again ahead of Mobile conditions,

4.1 Model Evaluation

and precision once more higher for the quiet class across every condition.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	70.61%	72.75%	66.90%	72.45%	70.85%	75.92%	66.59%	71.25%
QuantM60	72.12%	73.70%	69.04%	74.32%	72.11%	75.85%	69.49%	72.67%
QuantM90	71.04%	72.72%	67.65%	73.50%	70.64%	74.97%	67.99%	71.48%
QuantZ30	76.42%	77.81%	73.28%	78.51%	74.33%	78.73%	74.16%	76.45%
QuantZ60	75.49%	76.80%	72.42%	78.11%	73.05%	77.18%	73.91%	75.54%
QuantZ90	75.46%	76.48%	72.50%	78.49%	73.08%	76.60%	74.49%	75.54%

Table 4.10: Results for LOG_REG trained & tested on quantitative datasets, obtained from audios cross-validation.

The hybrid setting (Table 4.11) brings not only a higher but also a markedly more stable performance: balanced accuracy ranges narrowly from 84.15% (HM30) to 85.65% (HZ90), with Mobile and Zoom conditions performing comparably and a mild upward trend as window length increases for both devices. Of the models considered so far, Logistic Regression is the one for which the hybrid setting appears to flatten device- and window-related variability the most.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	83.81%	84.40%	81.74%	85.21%	83.35%	85.59%	82.70%	84.15%
HM60	84.24%	84.69%	82.35%	85.67%	83.52%	85.41%	83.44%	84.43%
HM90	84.85%	85.00%	83.19%	85.58%	84.95%	86.10%	83.76%	84.93%
HZ30	84.84%	85.42%	82.56%	85.27%	84.25%	86.71%	82.55%	84.63%
HZ60	85.37%	85.65%	83.60%	86.29%	84.42%	86.24%	84.39%	85.31%
HZ90	85.52%	85.92%	83.73%	86.04%	84.74%	86.94%	84.36%	85.65%

Table 4.11: Results for LOG_REG trained & tested on hybrid datasets, obtained from audios cross-validation.

LIN_SVC_L1 Model (Quantitative, hybrid) The L1-regularized LinearSVC stands out among the quantitative results (Table 4.12): balanced accuracy ranges from 70.10% (QuantM90) to 77.66% (QuantZ90), the highest value recorded by any model in this setting. As with the other linear and kernel-based classifiers, Zoom conditions outperform Mobile conditions, but here the gap between the quiet and noisy classes is noticeably narrower across all metrics than for the models discussed so far.

4.1 Model Evaluation

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
QuantM30	71.93%	72.55%	70.04%	75.95%	69.38%	71.90%	73.39%	72.64%
QuantM60	69.73%	69.84%	68.30%	74.67%	67.17%	68.56%	72.54%	70.55%
QuantM90	69.54%	69.86%	67.45%	74.46%	67.03%	68.96%	71.23%	70.10%
QuantZ30	75.35%	76.30%	72.74%	78.54%	72.20%	75.64%	75.19%	75.41%
QuantZ60	76.00%	76.61%	73.61%	79.09%	73.31%	76.07%	75.98%	76.03%
QuantZ90	77.47%	78.12%	75.14%	80.96%	74.59%	77.37%	77.94%	77.66%

Table 4.12: Results for LIN_SVC_L1 trained & tested on quantitative datasets, obtained from audios cross-validation.

The hybrid setting (Table 4.13) brings a further improvement, with balanced accuracy reaching between 83.51% (HZ90) and 86.44% (HM60), the highest value obtained by this model across both settings. Here, however, the device ranking reverses relative to the quantitative case: Mobile conditions now outperform Zoom conditions, a reversal not observed for any of the models discussed previously.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	85.15%	85.40%	83.63%	87.80%	82.72%	84.57%	86.19%	85.38%
HM60	86.37%	86.41%	85.03%	88.95%	83.76%	85.21%	87.68%	86.44%
HM90	85.88%	85.80%	84.56%	88.70%	82.84%	84.31%	87.90%	86.11%
HZ30	84.24%	84.71%	82.26%	86.33%	82.20%	84.55%	84.02%	84.29%
HZ60	83.85%	84.51%	81.53%	85.56%	81.90%	84.74%	82.77%	83.76%
HZ90	83.44%	83.96%	81.34%	85.73%	81.23%	83.70%	83.32%	83.51%

Table 4.13: Results for LIN_SVC_L1 trained & tested on hybrid datasets, obtained from audios cross-validation.

4.1.2 Users cross-validation

Having presented the results obtained under the audios cross-validation scheme, we now turn to the alternative scheme in which models are built using the Users cross-validation. As mentioned above, this scheme was applied only to the Hybrid datasets, since user identity is captured solely through the qualitative features collected via the questionnaire; the Quantitative datasets, lacking this information, cannot support a user-based split. Accordingly, only Hybrid results are reported below for the baseline model and for the five binary classifiers. Across all five models, performance under this scheme is markedly higher than that observed

4.1 Model Evaluation

under the audios cross-validation scheme, with balanced accuracy exceeding 88% under most conditions.

RF Model (Hybrid) We begin with Random Forest (Table 4.14), which delivers the strongest and most stable performance among the five models considered in this scheme: balanced accuracy ranges narrowly from 90.33% (HM90) to 92.05% (HZ30), with Zoom conditions holding a slight but consistent edge over Mobile conditions across all window lengths.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	91,11%	91,82%	90,09%	92,96%	89,85%	91,22%	91,13%	91,18%
HM60	90,50%	91,26%	89,46%	92,38%	89,20%	90,70%	90,53%	90,61%
HM90	90,31%	91,21%	89,06%	91,46%	89,89%	91,50%	89,17%	90,33%
HZ30	91,83%	92,43%	91,02%	93,98%	90,11%	91,42%	92,69%	92,05%
HZ60	91,56%	92,09%	90,85%	95,19%	88,27%	89,62%	94,21%	91,92%
HZ90	90,89%	91,43%	90,13%	94,64%	87,77%	88,99%	93,44%	91,21%

Table 4.14: Results for RF trained & tested on hybrid datasets, obtained from Users cross-validation.

Baseline Model (Hybrid) By contrast, our baseline Model (Table 4.15) performs considerably below RF under this scheme, with balanced accuracy ranging from 66.57% (HZ30) to 77.29% (HM90). Here, however, the device ranking reverses: a clear gap favors Mobile over Zoom conditions, and performance increases steadily with window length for both devices, from 74.07% to 77.29% for Mobile and from 66.57% to 69.20% for Zoom.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	73,72%	74,50%	72,24%	79,89%	69,19%	70,93%	77,21%	74,07%
HM60	75,40%	76,12%	73,96%	81,44%	71,15%	72,67%	78,81%	75,74%
HM90	76,90%	77,72%	75,43%	83,29%	72,29%	73,94%	80,64%	77,29%
HZ30	65,48%	64,16%	66,04%	75,51%	59,34%	57,26%	75,87%	66,57%
HZ60	67,49%	66,86%	67,46%	76,70%	61,51%	60,70%	76,13%	68,42%
HZ90	68,22%	67,53%	68,14%	77,78%	62,36%	61,42%	76,98%	69,20%

Table 4.15: Results for baseline trained & tested on hybrid datasets, obtained from Users cross-validation.

4.1 Model Evaluation

RBF_SVC Model (HYbrid) Returning to machine learning classifiers, the `rbf_svc` (Table 4.16) achieves the highest balanced accuracy of all five models under this scheme, ranging from 90.41% (HM90) to 92.48% (HZ30), slightly surpassing even Random Forest. As with RF, and unlike our baseline model, Zoom conditions consistently outperform Mobile conditions, with the best results for both devices obtained at the shortest window length (30s) and performance declining slightly as window length increases.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	91,56%	92,21%	90,68%	93,04%	90,53%	91,83%	91,46%	91,65%
HM60	90,44%	91,25%	89,32%	91,53%	90,10%	91,54%	89,38%	90,46%
HM90	90,39%	91,23%	89,19%	91,95%	89,72%	91,16%	89,67%	90,41%
HZ30	92,42%	93,06%	91,53%	94,14%	91,00%	92,36%	92,60%	92,48%
HZ60	91,96%	92,72%	90,83%	93,28%	91,48%	92,76%	91,20%	91,98%
HZ90	90,52%	91,33%	89,32%	92,62%	89,43%	90,83%	90,51%	90,67%

Table 4.16: Results for `rbf_svc` trained & tested on hybrid datasets, obtained from Users cross-validation.

LIN_SVC Model (Hybrid) The linear-kernel counterpart of the SVC tells a different story, however: Table 4.17 shows balanced accuracy clustering tightly between 88.08% (HZ30) and 89.79% (HM30), noticeably lower and narrower than its RBF-kernel counterpart. Unlike both the RBF_SVC and Random Forest, this model shows no clear device advantage, with Mobile and Zoom conditions remaining within roughly one percentage point of each other throughout, and balanced accuracy essentially flat across window lengths for both devices.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	89,68%	90,48%	88,57%	91,66%	88,40%	89,91%	89,66%	89,79%
HM60	89,31%	90,29%	87,89%	90,49%	89,38%	90,88%	87,67%	89,28%
HM90	89,54%	90,53%	88,06%	90,66%	89,88%	91,24%	87,72%	89,48%
HZ30	88,04%	89,13%	86,46%	90,17%	87,20%	88,98%	87,18%	88,08%
HZ60	88,98%	90,05%	87,43%	90,69%	88,49%	90,25%	87,82%	89,03%
HZ90	89,20%	90,06%	87,98%	91,94%	87,42%	89,04%	89,81%	89,43%

Table 4.17: Results for LIN_SVC trained & tested on hybrid datasets, obtained from Users cross-validation.

4.1 Model Evaluation

LOG_REG Model (Hybrid) in Table 4.18, performs similarly to LIN_SVC, with a balanced accuracy ranging from 88.18% (HM60) to 89.72% (HZ90). As with the LIN_SVC model, performance remains fairly stable across Mobile conditions; on the Zoom side, however, a mild upward trend emerges with increasing window length, culminating in the highest score recorded for this model at HZ90.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	88,07%	89,04%	86,76%	90,62%	86,26%	88,15%	88,23%	88,19%
HM60	88,02%	88,95%	86,76%	90,91%	85,70%	87,62%	88,74%	88,18%
HM90	88,44%	89,39%	87,16%	90,86%	86,49%	88,49%	88,67%	88,58%
HZ30	88,37%	89,46%	86,82%	90,25%	87,46%	89,41%	87,33%	88,37%
HZ60	89,23%	90,15%	87,92%	91,89%	87,43%	89,15%	89,54%	89,35%
HZ90	89,64%	90,58%	88,29%	91,89%	88,30%	90,02%	89,42%	89,72%

Table 4.18: Results for LOG_REG trained & tested on hybrid datasets, obtained from Users cross-validation.

LIN_SVC_L1 Model (Hybrid) We conclude with LIN_SVC_L1 (Table 4.19), whose balanced accuracy, ranging from 88.24% (HZ30) to 90.05% (HZ90), sits slightly above the two models just discussed. Unlike the consistent device or window-length trends observed for the previous classifiers, the pattern here is mixed: Zoom conditions yield both the lowest score (HZ30) and the highest score (HZ90) observed for this model, while Mobile conditions increase steadily with window length, from 88.91% to 89.84%, closing out this scheme on a result broadly comparable to LOG_REG and the LIN_SVC, but still well below RF and the rbf_svc.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	88,58%	89,31%	87,62%	92,54%	85,17%	86,81%	91,01%	88,91%
HM60	88,86%	89,61%	87,87%	92,37%	85,70%	87,46%	90,77%	89,11%
HM90	89,55%	90,20%	88,64%	93,48%	86,10%	87,61%	92,06%	89,84%
HZ30	88,07%	89,02%	86,75%	91,14%	85,84%	87,71%	88,77%	88,24%
HZ60	88,21%	89,14%	86,84%	91,16%	86,51%	88,09%	88,62%	88,36%
HZ90	89,67%	90,33%	88,75%	94,00%	86,12%	87,57%	92,54%	90,05%

Table 4.19: Results for LIN_SVC_L1 trained & tested on hybrid datasets, obtained from Users cross-validation.

4.1.3 Audio & User

We now turn to the third and final cross-validation scheme, the Audio & User cross-validation. As with the Users cross-validation scheme, this configuration was applied only to the hybrid datasets, for the same reason: the Quantitative datasets do not retain user identity and therefore cannot support a split that depends on it. Accordingly, only Hybrid results are reported below, for the baseline model and for the five binary classifiers. Overall, performance under this scheme sits between the two previously discussed, higher than the audio cross-validation scheme but somewhat lower than the Users cross-validation scheme, with balanced accuracy mostly falling between 82% and 87%.

RF (Hybrid) We begin with RF (Table 4.20), which achieves balanced accuracy ranging from 84.92% (HZ90) to 86.77% (HM60). Mobile conditions outperform Zoom conditions across all three window lengths, a pattern also reflected in the noisy-class metrics: $F1_1$ is higher for every Mobile condition (83.91%–84.86%) than for any Zoom condition (82.66%–83.95%).

Condition	Accuracy	$F1_0$	$F1_1$	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	86.52%	88.03%	84.04%	87.45%	87.51%	89.83%	82.53%	86.18%
HM60	87.18%	88.53%	84.86%	88.38%	87.81%	89.83%	83.71%	86.77%
HM90	86.38%	87.80%	83.91%	87.93%	87.12%	89.13%	82.98%	86.05%
HZ30	85.32%	86.86%	82.67%	87.67%	84.04%	87.22%	82.82%	85.02%
HZ60	86.09%	87.37%	83.95%	89.03%	84.48%	86.88%	84.83%	85.86%
HZ90	85.18%	86.67%	82.66%	87.42%	84.34%	87.15%	82.69%	84.92%

Table 4.20: Results for RF trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

Baseline Model (Hybrid) The Baseline model (Table 4.21), by contrast, performs well below RF, with balanced accuracy ranging from 65.24% (HZ30) to 74.32% (HM90). As with RF, Mobile conditions outperform Zoom conditions, but here the gap is considerably wider, and performance increases steadily with

4.1 Model Evaluation

window length for both devices, mirroring the trend already observed for this model under the Users cross-validation scheme.

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	72.20%	73.47%	69.03%	78.81%	68.26%	71.01%	73.45%	72.23%
HM60	72.92%	74.19%	69.71%	79.38%	69.65%	72.10%	73.73%	72.91%
HM90	74.40%	75.74%	70.85%	81.14%	71.05%	73.37%	75.26%	74.32%
HZ30	64.12%	61.52%	63.79%	77.17%	58.01%	55.09%	75.39%	65.24%
HZ60	65.89%	63.62%	65.22%	78.14%	60.24%	57.88%	75.74%	66.81%
HZ90	66.54%	64.28%	65.83%	80.08%	60.52%	58.10%	77.06%	67.58%

Table 4.21: Results for baseline trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

RBF_SVC Model (Hybrid) Returning to machine learning classifiers, the `rbf_svc` (Table 4.22) achieves balanced accuracy ranging from 84.31% (HM60) to 86.44% (HZ60), the highest value recorded by any model in this scheme. Unlike RF and the baseline model, Zoom conditions hold a slight edge here, supported by a consistently higher precision for the noisy class on Zoom (83.96%–85.66%) compared to Mobile (83.17%–85.24%).

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	84.46%	86.02%	82.09%	86.94%	83.17%	86.32%	82.92%	84.62%
HM60	84.34%	86.13%	81.56%	86.62%	83.21%	86.69%	81.93%	84.31%
HM90	85.44%	87.08%	82.68%	86.95%	85.24%	88.34%	82.07%	85.20%
HZ30	85.43%	87.03%	82.83%	87.59%	83.96%	87.33%	83.13%	85.23%
HZ60	86.48%	87.97%	84.09%	88.62%	85.12%	88.19%	84.69%	86.44%
HZ90	86.45%	88.03%	83.95%	87.99%	85.66%	88.89%	83.78%	86.34%

Table 4.22: Results for `rbf_svc` trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

LIN_SVC Model (Hybrid) The linear-kernel counterpart of the SVC (Table 4.23) tells a more balanced story: balanced accuracy ranges from 84.16% (HM30) to 85.62% (HZ90), with Mobile and Zoom conditions remaining close throughout and no consistent device advantage emerging, in line with the lack of a clear device effect already observed for this model under the two previous cross-validation schemes.

4.1 Model Evaluation

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	84.19%	85.83%	81.60%	86.23%	83.56%	86.76%	81.55%	84.16%
HM60	85.38%	87.11%	82.62%	86.92%	85.42%	88.41%	81.91%	85.16%
HM90	84.51%	86.08%	82.02%	86.48%	84.63%	87.11%	81.82%	84.46%
HZ30	84.69%	86.32%	81.96%	87.00%	83.43%	86.75%	82.29%	84.52%
HZ60	84.12%	85.70%	81.73%	87.20%	82.13%	85.36%	83.30%	84.33%
HZ90	85.55%	87.15%	83.04%	88.19%	83.95%	87.20%	84.04%	85.62%

Table 4.23: Results for LIN_SVC trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

LOG_REG Model (Hybrid) Logistic Regression (Table 4.24) achieves balanced accuracy ranging from 82.67% (HM30) to 85.64% (HZ60), the widest range among the five models in this scheme. Zoom conditions edge ahead of Mobile conditions overall, and this is reflected in the quiet-class F1-score, which is consistently higher for Zoom (86.57%–87.18%) than for Mobile (84.54%–86.28%).

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	82.68%	84.54%	79.91%	84.91%	81.60%	85.31%	80.02%	82.67%
HM60	83.51%	85.43%	80.46%	85.52%	82.25%	86.30%	80.23%	83.27%
HM90	84.45%	86.28%	81.58%	86.26%	83.77%	87.36%	81.16%	84.26%
HZ30	85.13%	86.78%	82.45%	86.78%	84.20%	87.70%	81.99%	84.85%
HZ60	85.71%	87.18%	83.40%	88.16%	83.66%	86.99%	84.30%	85.64%
HZ90	84.89%	86.57%	82.24%	86.80%	84.12%	87.40%	82.13%	84.77%

Table 4.24: Results for LOG_REG trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

LIN_SVC_L1 Model (Hybrid) We conclude with LinearSVC with L1 norm (Table 4.25), whose balanced accuracy ranges from 84.38% (HZ30) to 86.19% (HM90), the second-highest value observed in this scheme after the rbf_svc. Here, Mobile conditions take a slight lead over Zoom conditions, driven in large part by the strong HM90 result, where both precision for the quiet class (89.59%) and recall for the noisy class (86.04%) reach their highest values of any condition for this model, closing out this scheme with performance broadly comparable to, though marginally more variable than, the rbf_svc.

4.2 Model Comparison

Condition	Accuracy	F1 ₀	F1 ₁	Precision ₀	Precision ₁	Recall ₀	Recall ₁	Balanced Accuracy
HM30	84.29%	85.82%	82.00%	87.50%	81.70%	85.12%	83.73%	84.43%
HM60	84.60%	86.10%	82.33%	87.59%	82.13%	85.50%	83.84%	84.67%
HM90	86.11%	87.44%	83.99%	89.59%	83.38%	86.34%	86.04%	86.19%
HZ30	84.44%	85.97%	82.00%	87.07%	82.36%	85.87%	82.90%	84.38%
HZ60	85.18%	86.72%	82.76%	87.53%	83.27%	86.72%	83.40%	85.06%
HZ90	84.96%	86.45%	82.56%	88.07%	82.71%	85.94%	83.90%	84.92%

Table 4.25: Results for LOG.REG trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

4.2 Model Comparison

It is now time to present the results concerning our Research Questions (RQs), which are cross-cutting questions in the sense that they could, in principle, be answered under each of the three cross-validation schemes considered in this work (Audios, Users, and Audios $\times$ Users). However, as mentioned in Section 4.1, the Users and Audios $\times$ Users cross-validation schemes rely on user identity to construct the data splits; since the quantitative datasets do not retain this information, no models could be trained under these two schemes using quantitative features. As a consequence, RQ2, which concerns the impact of adding qualitative features to the quantitative ones, cannot be addressed under the Users and Audios $\times$ Users cross-validation schemes, since answering it requires comparing models built on quantitative and hybrid datasets directly. In contrast, RQ1, RQ3, and RQ4 can still be addressed under these two schemes, though only partially, using models built on hybrid datasets alone.

To answer these research questions, a series of statistical tests was performed on the models, for which an evaluation metric first had to be selected. Balanced accuracy was chosen for this purpose, as it is more robust than accuracy to potential class imbalance between the noisy and quiet categories. We therefore begin by collecting, for each cross-validation scheme in turn, the balanced accuracy

4.2 Model Comparison

achieved by every model on every dataset.

We start with the **audios cross-validation** scheme, for which Table 4.26 reports the balanced accuracy of all six models built on the quantitative datasets. RF achieves the highest balanced accuracy on four of the six Mobile and Zoom conditions, peaking at 74.63% (QuantM60), while LIN_SVC_L1 achieves the highest value overall, at 77.66% (QuantZ90). Our Baseline model consistently ranks lowest among the six models across all conditions. Table 4.27 instead reports the corresponding results on the hybrid datasets. Every machine learning model improves substantially over its quantitative counterpart in every condition, and the ranking among models changes accordingly: RF now achieves the highest balanced accuracy in five of the six conditions, peaking at 87.58% (HZ30), while our Baseline model remains the weakest model throughout, its balanced accuracy barely changing between the quantitative and hybrid settings.

Condition	Baseline	RBF_SVC	LIN_SVC	LOG_REG	LIN_SVC_L1	RF
QuantM30	70.18%	69.89%	71.81%	71.25%	72.64%	71.10%
QuantM60	70.20%	71.13%	71.83%	72.67%	70.55%	74.63%
QuantM90	70.40%	70.70%	70.83%	71.48%	70.10%	73.90%
QuantZ30	63.11%	73.52%	74.41%	76.45%	75.41%	69.56%
QuantZ60	64.70%	74.39%	75.91%	75.54%	76.03%	68.11%
QuantZ90	66.20%	73.95%	73.44%	75.54%	77.66%	68.91%

Table 4.26: Balanced accuracy of each model trained & tested on quantitative datasets, obtained from audios cross-validation.

Condition	Baseline	RBF_SVC	LIN_SVC	LOG_REG	LIN_SVC_L1	RF
HM30	70,18%	80,29%	83,03%	84,15%	85,38%	87,39%
HM60	70,20%	81,71%	83,13%	84,43%	86,44%	86,15%
HM90	70,40%	80,95%	82,08%	84,93%	86,10%	86,49%
HZ30	63,11%	84,24%	82,56%	84,63%	84,29%	87,58%
HZ60	64,70%	85,22%	82,42%	85,31%	83,76%	86,82%
HZ90	66,20%	83,38%	82,88%	85,65%	83,51%	87,07%

Table 4.27: Balanced accuracy of each model trained & tested on hybrid datasets, obtained from audios cross-validation.

4.2 Model Comparison

Then we turn to the **Users cross-validation** scheme, for which, as explained above, only results from models built on hybrid datasets are available. Table 4.28 reports the balanced accuracy of all six models under this scheme. Balanced accuracy is considerably higher here than under the Audios cross-validation scheme for every model except our Baseline model, whose values remain comparable to those observed previously. The RBF_SVC achieves the highest balanced accuracy in four of the six conditions, peaking at 92.48% (HZ30), closely followed by Random Forest, which peaks at 92.05% under the same condition.

Condition	Baseline	RBF_SVC	LIN_SVC	LOG_REG	LIN_SVC_L1	RF
HM30	74.07%	91.65%	89.79%	88.19%	88.91%	91.18%
HM60	75.74%	90.46%	89.28%	88.18%	89.11%	90.61%
HM90	77.29%	90.41%	89.48%	88.58%	89.84%	90.33%
HZ30	66.57%	92.48%	88.08%	88.37%	88.24%	92.05%
HZ60	68.42%	91.98%	89.03%	89.35%	88.36%	91.92%
HZ90	69.20%	90.67%	89.43%	89.72%	90.05%	91.21%

Table 4.28: Balanced accuracy of each model trained & tested on hybrid datasets, obtained from Users cross-validation.

We end with the **Audios $\times$ Users cross-validation** scheme, for which, as with the Users scheme, only results from models built on hybrid datasets are available. Table 4.29 reports the balanced accuracy of all six models under this scheme. Balanced accuracy here sits between the values observed for the Audios and Users cross-validation schemes for every machine learning model, while the Baseline remains broadly in line with its performance under the other two schemes. RF achieves the highest balanced accuracy in four of the six conditions, peaking at 86.77% (HM60), with the RBF_SVC close behind, peaking at 86.44% (HZ60).

4.2 Model Comparison

Condition	Baseline	RBF_SVC	LIN_SVC	LOG_REG	LIN_SVC_L1	RF
HM30	72.23%	84.62%	84.16%	82.67%	84.43%	86.18%
HM60	72.91%	84.31%	85.16%	83.27%	84.67%	86.77%
HM90	74.32%	85.20%	84.46%	84.26%	86.19%	86.05%
HZ30	65.24%	85.23%	84.52%	84.85%	84.38%	85.02%
HZ60	66.81%	86.44%	84.33%	85.64%	85.06%	85.86%
HZ90	67.58%	86.34%	85.62%	84.77%	84.92%	84.92%

Table 4.29: Balanced accuracy of each model trained & tested on hybrid datasets, obtained from Audio & User cross-validation.

4.2.1 RQ1 - Is the Baseline Sufficient or Do ML Models Improve Performance

Following the procedure described in 3.4.2, each of the five machine learning models was compared against the Baseline, separately for each of the six device-window length conditions (HM30, HM60, HM90, HZ30, HZ60, HZ90), yielding 30 paired comparisons in total. For each comparison, the 50 pairwise differences in balanced accuracy were first tested for normality using the Shapiro-Wilk test; the paired t-test was then applied where normality could be assumed ($p \geq 0.05$), and the Wilcoxon signed-rank test otherwise. The resulting p-values were corrected for multiple comparisons across all 30 tests using the Benjamini-Hochberg FDR procedure.

1. Audios cross-validation

Using models built on the **quantitative datasets**, the picture is considerably more mixed (see Table 4.30) than under the other schemes that will be examined in this RQ. Of the 30 comparisons, only 16 reach significance after correction. No machine learning model significantly outperforms the Baseline on the Mobile conditions at 60 and 90 seconds, nor at 30 seconds with the exception of LIN_SVC_L1 (72.64%, $\Delta = 2.46$, $pval_corrected \approx 0.033$); RF, by contrast, becomes significant precisely at the two longer

4.2 Model Comparison

Mobile windows (M60: $\Delta = 4.43$, $pval_corrected \approx 0.0018$; M90: $\Delta = 3.49$, $pval_corrected \approx 0.0072$) but not at M30. On the Zoom conditions, the pattern reverses: every model except RF reaches significance across all three window lengths, with improvements ranging from roughly 7 to 13 percentage points, while RF itself is significant only at Z30 ($\Delta = 6.45$, $pval_corrected \approx 1.7 \times 10^{-4}$). It is important to note two additional points. First, in a few cases the Baseline actually outperformed the corresponding machine learning model (RBF_SVC versus Baseline on QuantM30 $\Delta = -0.29$ and LIN_SVC_L1 versus Baseline on QuantM90 $\Delta = -0.31$). Second, the paired t-test was used for the vast majority of comparisons, with the Wilcoxon signed-rank test applied in only two cases (RBF_SVC versus Baseline on QuantM30, and LOG_REG versus Baseline on QuantZ60).

Comparison	Condition	Test	Mean Bal. Acc. Baseline	Mean Bal. Acc. ML	Δ Mean Bal. Acc. (Baseline - ML)	pval.raw	pval.corrected
RBF_SVC VS Baseline	QuantM30	wilcoxon	70.18%	69.89%	-0.29%	0.8384	0.8384
RBF_SVC VS Baseline	QuantM60	ttest	70.20%	71.13%	0.93%	0.4946	0.5936
RBF_SVC VS Baseline	QuantM90	ttest	70.40%	70.70%	0.29%	0.8293	0.8384
RBF_SVC VS Baseline	QuantZ30	ttest	63.11%	73.52%	10.41%	7.6e-09	3.3e-08*
RBF_SVC VS Baseline	QuantZ60	ttest	64.70%	74.39%	9.69%	2.4e-08	8.9e-08*
RBF_SVC VS Baseline	QuantZ90	ttest	66.20%	73.95%	7.75%	3.7e-06	1.0e-05*
LIN_SVC VS Baseline	QuantM30	ttest	70.18%	71.81%	1.63%	0.1509	0.2263
LIN_SVC VS Baseline	QuantM60	ttest	70.20%	71.83%	1.62%	0.2520	0.3600
LIN_SVC VS Baseline	QuantM90	ttest	70.40%	70.83%	0.43%	0.7421	0.8357
LIN_SVC VS Baseline	QuantZ30	ttest	63.11%	74.41%	11.30%	3.1e-10	1.9e-09*
LIN_SVC VS Baseline	QuantZ60	ttest	64.70%	75.91%	11.21%	2.1e-10	1.9e-09*
LIN_SVC VS Baseline	QuantZ90	ttest	66.20%	73.44%	7.25%	1.1e-05	2.7e-05*
LOG_REG VS Baseline	QuantM30	ttest	70.18%	71.25%	1.07%	0.3358	0.4381
LOG_REG VS Baseline	QuantM60	ttest	70.20%	72.67%	2.47%	0.0579	0.0914
LOG_REG VS Baseline	QuantM90	ttest	70.40%	71.48%	1.08%	0.3294	0.4381
LOG_REG VS Baseline	QuantZ30	ttest	63.11%	76.45%	13.33%	4.5e-12	1.4e-10*
LOG_REG VS Baseline	QuantZ60	wilcoxon	64.70%	75.54%	10.84%	1.3e-07	3.8e-07*
LOG_REG VS Baseline	QuantZ90	ttest	66.20%	75.54%	9.35%	3.5e-08	1.2e-07*
LIN_SVC_L1 VS Baseline	QuantM30	ttest	70.18%	72.64%	2.46%	0.0174	0.0326*
LIN_SVC_L1 VS Baseline	QuantM60	ttest	70.20%	70.55%	0.35%	0.7802	0.8359
LIN_SVC_L1 VS Baseline	QuantM90	ttest	70.40%	70.10%	-0.31%	0.7521	0.8357
LIN_SVC_L1 VS Baseline	QuantZ30	ttest	63.11%	75.41%	12.30%	3.4e-11	5.1e-10*
LIN_SVC_L1 VS Baseline	QuantZ60	ttest	64.70%	76.03%	11.33%	5.7e-09	2.9e-08*
LIN_SVC_L1 VS Baseline	QuantZ90	ttest	66.20%	77.66%	11.46%	2.8e-10	1.9e-09*
RF VS Baseline	QuantM30	ttest	70.18%	71.10%	0.92%	0.3879	0.4849
RF VS Baseline	QuantM60	ttest	70.20%	74.63%	4.43%	8.6e-04	0.0018*
RF VS Baseline	QuantM90	ttest	70.40%	73.90%	3.49%	0.0036	0.0072*
RF VS Baseline	QuantZ30	ttest	63.11%	69.56%	6.45%	7.4e-05	1.7e-04*
RF VS Baseline	QuantZ60	ttest	64.70%	68.11%	3.42%	0.0516	0.0860
RF VS Baseline	QuantZ90	ttest	66.20%	68.91%	2.72%	0.0365	0.0644

Table 4.30: Results RQ1 tests of models built on quantitative datasets in the audios cross-validation scheme. * indicates that a significant difference has been found.

4.2 Model Comparison

Turning to the models built on the **hybrid datasets** under the same scheme (see Table 4.31), the picture changes markedly: all 30 comparisons reach statistical significance after correction, mirroring the pattern that will also be observed under the Users and Audios_×_Users schemes. Improvements range from 10.11 percentage points (RBF_SVC, HM30) to 24.47 percentage points (RF, HZ30), with RF and the RBF_SVC again producing some of the largest gains on the Zoom conditions, while the smallest improvements are concentrated on the Mobile conditions, consistent with the device pattern already noted in the descriptive results.

Comparison	Condition	Test	Mean Bal. Acc. Baseline	Mean Bal. Acc. ML	Δ Mean Bal. Acc. (Baseline - ML)	pval_raw	pval_corrected
RBF_SVC VS Baseline	HM30	ttest	70.18%	80.29%	10.11%	7.9e-09	7.9e-09*
RBF_SVC VS Baseline	HM60	ttest	70.20%	81.71%	11.51%	4.8e-11	5.2e-11*
RBF_SVC VS Baseline	HM90	ttest	70.40%	80.95%	10.54%	5.2e-09	5.4e-09*
RBF_SVC VS Baseline	HZ30	ttest	63.11%	84.24%	21.12%	2.1e-16	1.6e-15*
RBF_SVC VS Baseline	HZ60	ttest	64.70%	85.22%	20.52%	2.8e-16	1.7e-15*
RBF_SVC VS Baseline	HZ90	ttest	66.20%	83.38%	17.18%	7.1e-14	1.3e-13*
LIN_SVC VS Baseline	HM30	ttest	70.18%	83.03%	12.85%	2.9e-12	3.6e-12*
LIN_SVC VS Baseline	HM60	ttest	70.20%	83.13%	12.93%	1.4e-12	2.0e-12*
LIN_SVC VS Baseline	HM90	ttest	70.40%	82.08%	11.68%	2.0e-11	2.3e-11*
LIN_SVC VS Baseline	HZ30	ttest	63.11%	82.56%	19.45%	5.9e-16	2.9e-15*
LIN_SVC VS Baseline	HZ60	ttest	64.70%	82.42%	17.72%	1.9e-14	4.2e-14*
LIN_SVC VS Baseline	HZ90	ttest	66.20%	82.88%	16.68%	1.8e-12	2.5e-12*
LOG_REG VS Baseline	HM30	ttest	70.18%	84.15%	13.96%	7.0e-12	8.0e-12*
LOG_REG VS Baseline	HM60	ttest	70.20%	84.43%	14.23%	1.7e-14	3.9e-14*
LOG_REG VS Baseline	HM90	ttest	70.40%	84.93%	14.53%	2.1e-12	2.7e-12*
LOG_REG VS Baseline	HZ30	ttest	63.11%	84.63%	21.52%	4.9e-15	1.3e-14*
LOG_REG VS Baseline	HZ60	ttest	64.70%	85.31%	20.62%	9.8e-16	3.7e-15*
LOG_REG VS Baseline	HZ90	ttest	66.20%	85.65%	19.45%	1.0e-13	1.7e-13*
LIN_SVC.L1 VS Baseline	HM30	ttest	70.18%	85.38%	15.20%	1.7e-14	3.9e-14*
LIN_SVC.L1 VS Baseline	HM60	ttest	70.20%	86.44%	16.24%	3.1e-15	9.2e-15*
LIN_SVC.L1 VS Baseline	HM90	ttest	70.40%	86.10%	15.70%	7.1e-13	1.1e-12*
LIN_SVC.L1 VS Baseline	HZ30	ttest	63.11%	84.29%	21.17%	1.1e-15	3.7e-15*
LIN_SVC.L1 VS Baseline	HZ60	ttest	64.70%	83.76%	19.06%	1.1e-13	1.7e-13*
LIN_SVC.L1 VS Baseline	HZ90	ttest	66.20%	83.51%	17.32%	5.6e-12	6.8e-12*
RF VS Baseline	HM30	ttest	70.18%	87.39%	17.21%	2.9e-17	2.9e-16*
RF VS Baseline	HM60	ttest	70.20%	86.15%	15.95%	3.6e-14	7.2e-14*
RF VS Baseline	HM90	ttest	70.40%	86.49%	16.09%	5.0e-14	9.4e-14*
RF VS Baseline	HZ30	ttest	63.11%	87.58%	24.47%	2.4e-18	7.2e-17*
RF VS Baseline	HZ60	ttest	64.70%	86.82%	22.12%	1.3e-17	1.9e-16*
RF VS Baseline	HZ90	ttest	66.20%	87.07%	20.87%	1.0e-15	3.7e-15*

Table 4.31: Results RQ1 tests of models built on hybrid datasets in the audios cross-validation scheme. * indicates that a significant difference has been found.

2. Users cross-validation

The second cross-validation scheme that we examined is the **Users cross-validation**, for which, as noted above, are only available models built on hybrid datasets.

Table 4.32 summarizes the results. Each of the 30 comparisons reached statistical significance after correction ($pval_corrected < 0.05$), indicating that all five machine learning models significantly outperform our Baseline across every condition under this scheme. The t-test was used for the majority of comparisons; the Wilcoxon signed-rank test was used instead for three comparisons involving non-normal differences (LOG_REG at HZ60 and HZ90, and LIN_SVC.L1 at HZ90).

The magnitude of the improvement varies both by model and by condition. The largest gains are observed on the Zoom conditions, particularly HZ30, where the Baseline’s balanced accuracy of 66.57% is improved by over 25 percentage points by both the RBF_SVC (92.48%, $\Delta = 25.91$) and RF (92.05%, $\Delta = 25.49$). The smallest gains, by contrast, are observed on the Mobile conditions at longer window lengths, most notably HM90, where the Baseline already performs comparatively well (77.29%); here, the improvement brought by LOG_REG is the smallest in the entire table (88.58%, $\Delta = 11.29$), though still highly significant ($pval_corrected \approx 1.3 \times 10^{-12}$). Across all 30 comparisons, RF and the RBF_SVC tend to produce the largest improvements over the Baseline, while LOG_REG tends to produce the smallest, though the difference between models is modest compared to the difference between conditions.

4.2 Model Comparison

Comparison	Condition	Test	Mean Bal. Acc. Baseline	Mean Bal. Acc. ML	Δ Mean Bal. Acc. (Baseline - ML)	pval_raw	pval_corrected
RBF_SVC VS Baseline	HM30	ttest	74.07%	91.65%	17.58%	6.8e-23	1.7e-22*
RBF_SVC VS Baseline	HM60	ttest	75.74%	90.46%	14.72%	2.8e-21	6.0e-21*
RBF_SVC VS Baseline	HM90	ttest	77.29%	90.41%	13.12%	3.1e-18	4.9e-18*
RBF_SVC VS Baseline	HZ30	ttest	66.57%	92.48%	25.91%	3.0e-31	9.1e-30*
RBF_SVC VS Baseline	HZ60	ttest	68.42%	91.98%	23.56%	7.9e-26	3.0e-25*
RBF_SVC VS Baseline	HZ90	ttest	69.20%	90.67%	21.47%	9.5e-23	2.2e-22*
LIN_SVC VS Baseline	HM30	ttest	74.07%	89.79%	15.72%	3.2e-21	6.3e-21*
LIN_SVC VS Baseline	HM60	ttest	75.74%	89.28%	13.54%	2.0e-19	3.5e-19*
LIN_SVC VS Baseline	HM90	ttest	77.29%	89.48%	12.19%	1.0e-15	1.4e-15*
LIN_SVC VS Baseline	HZ30	ttest	66.57%	88.08%	21.52%	5.7e-27	3.4e-26*
LIN_SVC VS Baseline	HZ60	ttest	68.42%	89.03%	20.62%	1.8e-25	5.9e-25*
LIN_SVC VS Baseline	HZ90	ttest	69.20%	89.43%	20.23%	3.8e-23	1.0e-22*
LOG_REG VS Baseline	HM30	ttest	74.07%	88.19%	14.12%	1.0e-16	1.5e-16*
LOG_REG VS Baseline	HM60	ttest	75.74%	88.18%	12.44%	2.1e-15	2.4e-15*
LOG_REG VS Baseline	HM90	ttest	77.29%	88.58%	11.29%	1.3e-12	1.3e-12*
LOG_REG VS Baseline	HZ30	ttest	66.57%	88.37%	21.80%	1.5e-26	6.7e-26*
LOG_REG VS Baseline	HZ60	wilcoxon	68.42%	89.35%	20.93%	1.8e-15	2.0e-15*
LOG_REG VS Baseline	HZ90	wilcoxon	69.20%	89.72%	20.52%	1.8e-15	2.0e-15*
LIN_SVC.L1 VS Baseline	HM30	ttest	74.07%	88.91%	14.84%	4.1e-18	6.2e-18*
LIN_SVC.L1 VS Baseline	HM60	ttest	75.74%	89.11%	13.38%	2.6e-16	3.5e-16*
LIN_SVC.L1 VS Baseline	HM90	ttest	77.29%	89.84%	12.55%	1.2e-14	1.3e-14*
LIN_SVC.L1 VS Baseline	HZ30	ttest	66.57%	88.24%	21.67%	7.6e-28	5.7e-27*
LIN_SVC.L1 VS Baseline	HZ60	ttest	68.42%	88.36%	19.94%	9.8e-24	2.9e-23*
LIN_SVC.L1 VS Baseline	HZ90	wilcoxon	69.20%	90.05%	20.85%	1.8e-15	2.0e-15*
RF VS Baseline	HM30	ttest	74.07%	91.18%	17.11%	6.9e-21	1.3e-20*
RF VS Baseline	HM60	ttest	75.74%	90.61%	14.88%	2.9e-18	4.8e-18*
RF VS Baseline	HM90	ttest	77.29%	90.33%	13.04%	2.9e-15	3.1e-15*
RF VS Baseline	HZ30	ttest	66.57%	92.05%	25.49%	9.5e-31	1.4e-29*
RF VS Baseline	HZ60	ttest	68.42%	91.92%	23.50%	9.5e-29	9.5e-28*
RF VS Baseline	HZ90	ttest	69.20%	91.21%	22.01%	1.6e-26	6.7e-26*

Table 4.32: Results RQ1 tests of models built on hybrid datasets in the Users cross-validation scheme. * indicates that a significant difference has been found.

3. Audio & User cross-validation

The last cross-validation scheme that we examined in this scheme is the Audio & User cross-validation, for which, as in the Users cross-validation scheme, are only available models built on hybrid datasets.

Here in Table 4.33, as in the previous scheme, it can be observed that each of the 30 comparisons reached statistical significance after correction ($pval_corrected < 0.05$), but unlike the Users cross-validation scheme, all them relied on the paired t-test, with no instance of non-normal differences requiring the Wilcoxon signed-rank test.

The same pattern observed in the previous scheme emerges here as well: the size of the improvement again depends more strongly on device than on model. The largest gains are concentrated on the Zoom conditions, peaking at HZ30, where the Baseline’s balanced accuracy of 65.24% is improved

4.2 Model Comparison

by nearly 20 percentage points by Random Forest (85.02%, $\Delta = 19.78$) and the RBF_SVC (85.23%, $\Delta = 19.99$), the latter being the largest single improvement recorded under this scheme. The smallest gains, conversely, occur on the Mobile conditions at longer window lengths: at HM90, LOG_REG again yields the smallest improvement of the six conditions (84.26%, $\Delta = 9.95$), the smallest margin observed across all three cross-validation schemes so far, though still associated with a highly significant result ($pval_corrected \approx 1.6 \times 10^{-8}$).

Comparison	Condition	Test	Mean Bal. Acc. Baseline	Mean Bal. Acc. ML	Δ Mean Bal. Acc. (Baseline - ML)	pval_raw	pval_corrected
RBF_SVC VS Baseline	HM30	ttest	72.23%	84.62%	12.39%	3.1e-13	4.9e-13*
RBF_SVC VS Baseline	HM60	ttest	72.91%	84.31%	11.40%	4.2e-12	5.7e-12*
RBF_SVC VS Baseline	HM90	ttest	74.32%	85.20%	10.89%	7.2e-12	9.0e-12*
RBF_SVC VS Baseline	HZ30	ttest	65.24%	85.23%	19.99%	2.4e-23	7.3e-22*
RBF_SVC VS Baseline	HZ60	ttest	66.81%	86.44%	19.63%	3.1e-19	1.2e-18*
RBF_SVC VS Baseline	HZ90	ttest	67.58%	86.34%	18.76%	3.1e-20	2.3e-19*
LIN_SVC VS Baseline	HM30	ttest	72.23%	84.16%	11.93%	2.2e-11	2.5e-11*
LIN_SVC VS Baseline	HM60	ttest	72.91%	85.16%	12.24%	3.7e-14	6.2e-14*
LIN_SVC VS Baseline	HM90	ttest	74.32%	84.46%	10.15%	6.9e-12	8.9e-12*
LIN_SVC VS Baseline	HZ30	ttest	65.24%	84.52%	19.28%	3.3e-22	3.5e-21*
LIN_SVC VS Baseline	HZ60	ttest	66.81%	84.33%	17.52%	5.5e-17	1.0e-16*
LIN_SVC VS Baseline	HZ90	ttest	67.58%	85.62%	18.04%	4.8e-20	2.9e-19*
LOG_REG VS Baseline	HM30	ttest	72.23%	82.67%	10.44%	2.2e-11	2.5e-11*
LOG_REG VS Baseline	HM60	ttest	72.91%	83.27%	10.35%	4.2e-11	4.3e-11*
LOG_REG VS Baseline	HM90	ttest	74.32%	84.26%	9.95%	1.6e-08	1.6e-08*
LOG_REG VS Baseline	HZ30	ttest	65.24%	84.85%	19.61%	6.0e-20	3.0e-19*
LOG_REG VS Baseline	HZ60	ttest	66.81%	85.64%	18.84%	4.6e-18	1.2e-17*
LOG_REG VS Baseline	HZ90	ttest	67.58%	84.77%	17.19%	4.2e-18	1.2e-17*
LIN_SVC.L1 VS Baseline	HM30	ttest	72.23%	84.43%	12.20%	2.8e-12	4.0e-12*
LIN_SVC.L1 VS Baseline	HM60	ttest	72.91%	84.67%	11.76%	3.2e-11	3.6e-11*
LIN_SVC.L1 VS Baseline	HM90	ttest	74.32%	86.19%	11.87%	3.6e-11	3.9e-11*
LIN_SVC.L1 VS Baseline	HZ30	ttest	65.24%	84.38%	19.14%	2.3e-18	7.7e-18*
LIN_SVC.L1 VS Baseline	HZ60	ttest	66.81%	85.06%	18.26%	1.1e-16	1.9e-16*
LIN_SVC.L1 VS Baseline	HZ90	ttest	67.58%	84.92%	17.34%	1.6e-17	3.6e-17*
RF VS Baseline	HM30	ttest	72.23%	86.18%	13.95%	1.7e-17	3.6e-17*
RF VS Baseline	HM60	ttest	72.91%	86.77%	13.85%	2.5e-19	1.1e-18*
RF VS Baseline	HM90	ttest	74.32%	86.05%	11.74%	2.8e-12	4.0e-12*
RF VS Baseline	HZ30	ttest	65.24%	85.02%	19.78%	3.5e-22	3.5e-21*
RF VS Baseline	HZ60	ttest	66.81%	85.86%	19.05%	2.9e-18	8.8e-18*
RF VS Baseline	HZ90	ttest	67.58%	84.92%	17.34%	2.1e-17	4.3e-17*

Table 4.33: Results RQ1 tests of models built on hybrid datasets in the Audio & User cross-validation scheme. * indicates that a significant difference has been found.

4.2.2 RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?

To answer to RQ2, where it is asked whether adding the qualitative features (Hybrid setting) improves each model’s performance over the acoustic features alone (Quantitative setting), we computed the 50 paired differences in balanced accuracy between the hybrid and quantitative versions ($\Delta = \textit{Hybrid} - \textit{Quantitative}$) of each of the six models (the Baseline Model and the five machine learning classifiers) and each of the six device-window length conditions. This gives 36 paired comparisons in total. As mentioned in , the Shapiro-Wilk test was used to choose between the paired t-test and the Wilcoxon signed-rank test, and Benjamini-Hochberg FDR correction was applied jointly across the 36 tests as a single family.

1. Audios cross-validation

For the Baseline Model, the hybrid and quantitative balanced accuracies are identical in all six conditions (70.18%, 70.20%, 70.40%, 63.11%, 64.70%, 66.20%), so every paired difference is exactly zero. The Shapiro-Wilk test cannot be computed on a constant sample, and no significance test could be applied (reported as "No-test"). These six comparisons were assigned $p = 1$ and are not significant.

All 30 comparisons involving the machine learning models were significant after correction ($pval_corrected < 0.05$). The largest corrected p-value was more or less 1.68×10^{-3} (LIN_SVC_L1, Z90), and most were below 10^{-7} . Every machine learning model improved in every condition, with Δ between 5.85 percentage points (LIN_SVC_L1, Z90) and 18.70 percentage points (RF, Z60). The paired t-test was used in 29 of these 30 comparisons, and the Wilcoxon signed-rank test only for the LIN_SVC at M60.

The size of the improvement varied by model:

- **RBF_SVC:** Δ was nearly uniform across conditions, from 9.43 (Z90) to 10.83 (Z60).
- **LIN_SVC:** Δ was between 11.22 and 11.30 on the Mobile conditions and between 6.51 and 9.44 on the Zoom conditions.
- **LOG_REG:** Δ was between 11.76 and 13.45 on Mobile and between 8.18 and 10.10 on Zoom.
- **LIN_SVC_L1:** gains were larger on Mobile (12.74 to 16.01) than on Zoom (5.85 to 8.87), and decreased with window length on Zoom.
- **RF:** gains were largest on Zoom (18.02 to 18.70). On Mobile they ranged from 11.52 (M60) to 16.29 (M30).

In the hybrid setting, the machine learning models reached between 80.29% and 87.58% balanced accuracy. In the Quantitative setting they reached between 68.11% and 77.66%.

4.2 Model Comparison

Model	Comparison	Test	Mean Bal. Acc. Hyb	Mean Bal. Acc. Quant	Δ Mean Bal. Acc. (Hyb - Quant)	pval_raw	pval_corrected
Baseline	QuantM30 VS HM30	No-test	70.18%	70.18%	0.00%	1.0	1.0
Baseline	QuantM60 VS HM60	No-test	70.20%	70.20%	0.00%	1.0	1.0
Baseline	QuantM90 VS HM90	No-test	70.40%	70.40%	0.00%	1.0	1.0
Baseline	QuantZ30 VS HZ30	No-test	63.11%	63.11%	0.00%	1.0	1.0
Baseline	QuantZ60 VS HZ60	No-test	64.70%	64.70%	0.00%	1.0	1.0
Baseline	QuantZ90 VS HZ90	No-test	66.20%	66.20%	0.00%	1.0	1.0
RBF_SVC	QuantM30 VS HM30	ttest	80.29%	69.89%	10.40%	5.7e-11	1.4e-10*
RBF_SVC	QuantM60 VS HM60	ttest	81.71%	71.13%	10.58%	8.6e-11	1.9e-10*
RBF_SVC	QuantM90 VS HM90	ttest	80.95%	70.70%	10.25%	4.7e-10	9.4e-10*
RBF_SVC	QuantZ30 VS HZ30	ttest	84.24%	73.52%	10.72%	2.8e-12	1.1e-11*
RBF_SVC	QuantZ60 VS HZ60	ttest	85.22%	74.39%	10.83%	9.4e-10	1.7e-09*
RBF_SVC	QuantZ90 VS HZ90	ttest	83.38%	73.95%	9.43%	2.1e-07	3.4e-07*
LIN_SVC	QuantM30 VS HM30	ttest	83.03%	71.81%	11.22%	1.6e-13	9.4e-13*
LIN_SVC	QuantM60 VS HM60	wilcoxon	83.13%	71.83%	11.30%	2.8e-10	6.0e-10*
LIN_SVC	QuantM90 VS HM90	ttest	82.08%	70.83%	11.25%	2.5e-11	7.4e-11*
LIN_SVC	QuantZ30 VS HZ30	ttest	82.56%	74.41%	8.15%	8.0e-09	1.4e-08*
LIN_SVC	QuantZ60 VS HZ60	ttest	82.42%	75.91%	6.51%	2.0e-04	2.5e-04*
LIN_SVC	QuantZ90 VS HZ90	ttest	82.88%	73.44%	9.44%	2.2e-07	3.4e-07*
LOG_REG	QuantM30 VS HM30	ttest	84.15%	71.25%	12.89%	3.3e-11	9.1e-11*
LOG_REG	QuantM60 VS HM60	ttest	84.43%	72.67%	11.76%	6.4e-10	1.2e-09*
LOG_REG	QuantM90 VS HM90	ttest	84.93%	71.48%	13.45%	1.5e-12	6.6e-12*
LOG_REG	QuantZ30 VS HZ30	ttest	84.63%	76.45%	8.18%	2.0e-06	2.7e-06*
LOG_REG	QuantZ60 VS HZ60	ttest	85.31%	75.54%	9.77%	5.5e-08	9.0e-08*
LOG_REG	QuantZ90 VS HZ90	ttest	85.65%	75.54%	10.10%	5.3e-07	7.3e-07*
LIN_SVC.L1	QuantM30 VS HM30	ttest	85.38%	72.64%	12.74%	9.6e-12	3.4e-11*
LIN_SVC.L1	QuantM60 VS HM60	ttest	86.44%	70.55%	15.89%	1.1e-13	8.1e-13*
LIN_SVC.L1	QuantM90 VS HM90	ttest	86.10%	70.10%	16.01%	2.7e-13	1.4e-12*
LIN_SVC.L1	QuantZ30 VS HZ30	ttest	84.29%	75.41%	8.87%	3.7e-07	5.4e-07*
LIN_SVC.L1	QuantZ60 VS HZ60	ttest	83.76%	76.03%	7.73%	4.2e-05	5.4e-05*
LIN_SVC.L1	QuantZ90 VS HZ90	ttest	83.51%	77.66%	5.85%	0.0014	0.0017*
RF	QuantM30 VS HM30	ttest	87.39%	71.10%	16.29%	3.0e-16	2.7e-15*
RF	QuantM60 VS HM60	ttest	86.15%	74.63%	11.52%	4.7e-11	1.2e-10*
RF	QuantM90 VS HM90	ttest	86.49%	73.90%	12.59%	2.2e-11	7.3e-11*
RF	QuantZ30 VS HZ30	ttest	87.58%	69.56%	18.02%	4.6e-18	8.3e-17*
RF	QuantZ60 VS HZ60	ttest	86.82%	68.11%	18.70%	5.9e-19	2.1e-17*
RF	QuantZ90 VS HZ90	ttest	87.07%	68.91%	18.15%	7.3e-18	8.7e-17*

Table 4.34: Results RQ2 tests of models built on hybrid and on quantitative datasets (Hyb. Vs. Quant) in the audios cross-validation scheme. * indicates that a significant difference has been found.

4.2.3 RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent

Following the two-stage procedure described 3.4.4, the Friedman omnibus test was followed by the Nemenyi post-hoc test, applied jointly to the 36 model-condition combinations (e.g., Baseline_M30, RBF_SVC_Z60, RF_Z90), if a significant Friedman p-value is found. The Nemenyi internally controls the family-wise error rate and therefore requires no additional Benjamini-Hochberg correction. A pairwise comparison is considered significant when its Nemenyi p-value falls below 0.05; by convention, p-values smaller than 10^{-6} are reported as exactly zero. This has been done for each cross-validation.

1. Audios cross-validation

Starting with the models built on the quantitative datasets, the Friedman test returned a significant p-value ($p \approx 8.90 \times 10^{-53}$), which justified proceeding to the Nemenyi post-hoc test. The resulting matrix of pairwise p-values is reported in 4.1. From that matrix three patterns emerge.

- **Baseline combinations remain mutually indistinguishable.** All pairwise comparisons among the six Baseline combinations are non-significant. The smallest *pvalue* ≈ 0.15
- **The Baseline is separated from the machine learning models.** Of the 180 comparisons between a Baseline combination and a machine learning combination, only 77 (43%) are significant. The significant differences are concentrated on specific model-device combinations:
 - The three Mobile Baseline combinations are never significantly different from any RBF_SVC or RF combination ($p \geq 0.12$). They differ significantly only from the Zoom combinations of three models: LOG_REG (all three window lengths, p from 2.3×10^{-4} to 0.031), LIN_SVC_L1 (all three, p from 5×10^{-6} to 0.006) and the LIN_SVC at Z60 ($p \leq 0.006$).
 - The Zoom Baseline combinations significantly differ from the Zoom combinations of the machine learning models. The exception is RF: its Zoom combinations are statistically indistinguishable from every Zoom Baseline combination ($p \geq 0.23$)
- **Differences among machine learning models are few but structured by device.** Only 60 of the 360 comparisons between different machine learning models are significant, and they are not scattered. Among the nine Mobile combinations of the machine learning models, a single pair differs significantly (RBF_SVC_M30 vs. RF_M60, $p =$

0.041), so on Mobile recordings the five classifiers are practically equivalent. On Zoom recordings, the RBF_SVC, LIN_SVC, LOG_REG and LIN_SVC_L1 are mutually indistinguishable. All 26 significant Zoom-Zoom differences involve a Random Forest combination, RF_Z60 being the most frequent (10 pairs). For example, RF_Z60 is significantly lower than LOG_REG_Z30 and LIN_SVC_L1_Z90 (both $p < 10^{-6}$) and than LIN_SVC_Z60 ($p = 3 \times 10^{-6}$). The remaining significant pairs mix devices, and the most distinct combinations are LIN_SVC_L1_Z90 (77.66%, significantly different from 20 of the other 35 combinations) and LOG_REG_Z30 (76.45%, 18 of 35).

Turning to models built on **hybrid datasets**, we found again a significant Friedman p-value ($p \approx 1.96 \times 10^{-117}$), so we proceeded again to the Nemenyi post-hoc test. The resulting matrix is shown in 4.2.

Here too, three patterns can be observed

- As in the quantitative setting, **Baseline combinations remain mutually indistinguishable**: all pairwise comparisons among the six Baseline combinations are non-significant.
- **Baseline combinations are significantly outperformed by nearly every machine learning combination**. Of the 180 comparisons between a Baseline combination and a machine learning combination, the overwhelming majority are significant, with p-values effectively equal to zero. The few exceptions involve RBF_SVC_M30, which does not differ significantly from four of the six Baseline combinations (Baseline_M30: $p \approx 0.380$; Baseline_M60: $p \approx 0.232$; Baseline_M90: $p \approx 0.369$; Baseline_Z90: $p \approx 0.143$).

- **Among machine learning combinations, differences are largely non-significant, though not uniformly so.** The majority of pairwise comparisons between machine learning combinations, regardless of model or condition, yield high p-values (often close to 1). Here too, there are some exceptions. The first is again RBF_SVC_M30, which differs significantly from RBF_SVC_Z60, LOG_REG_M90, LOG_REG_Z60, LOG_REG_Z90, all Mobile combinations of LIN_SVC_L1, and all Mobile and Zoom combinations of RF. The second exception concerns RF_M30 and RF_Z30, which differ significantly from RBF_SVC_M60, RBF_SVC_M90, LIN_SVC_Z30, and LIN_SVC_Z60.

	Baseline_M3	Baseline_M6	Baseline_M9	Baseline_Z3	Baseline_Z6	Baseline_Z9	RBF_SVC_M3	RBF_SVC_M6	RBF_SVC_M9	RBF_SVC_Z3	RBF_SVC_Z6	RBF_SVC_Z9	LN_SVC_M3	LN_SVC_M6	LN_SVC_M9	LN_SVC_Z3	LN_SVC_Z6	LN_SVC_Z9	LOG_REG_M3	LOG_REG_M6	LOG_REG_M9	LOG_REG_Z3	LOG_REG_Z6	LOG_REG_Z9	SVC_L1_M3	SVC_L1_M6	SVC_L1_M9	SVC_L1_Z3	SVC_L1_Z6	SVC_L1_Z9	RF_M30	RF_M60	RF_M90	RF_Z30	RF_Z60	RF_Z90
Baseline_M30	1.0	1.0	1.0	0.251511	0.952857	0.998729	1.0	1.0	1.0	0.881232	0.125927	0.67255	0.999999	0.999978	1.0	0.10923	0.0028	0.927863	1.0	0.954096	1.0	0.00023	0.007562	0.014919	0.955311	1.0	1.0	0.002514	0.001271	5e-06	1.0	0.144555	0.323959	1.0	0.999999	1.0
Baseline_M60	1.0	1.0	1.0	0.154626	0.88358	0.992842	1.0	1.0	1.0	0.951593	0.210334	0.810113	1.0	0.999999	1.0	0.185614	0.00631	0.974907	1.0	0.986093	1.0	0.000583	0.016083	0.030327	0.98657	1.0	1.0	0.0057	0.002985	1.5e-05	1.0	0.23724	0.468294	1.0	0.999972	1.0
Baseline_M90	1.0	1.0	1.0	0.152571	0.881232	0.992558	1.0	1.0	1.0	0.952857	0.212926	0.813194	1.0	0.999999	1.0	0.187988	0.006439	0.975672	1.0	0.98657	1.0	0.000597	0.016387	0.030862	0.987034	1.0	1.0	0.005818	0.003049	1.6e-05	1.0	0.240051	0.472251	1.0	0.99997	1.0
Baseline_Z30	0.251511	0.154626	0.152571	1.0	1.0	0.999974	0.552667	0.019361	0.098632	6e-06	0.0	1e-06	0.002461	0.000992	0.04724	0.0	0.0	1.1e-05	0.01464	1.8e-05	0.00787	0.0	0.0	0.0	1.8e-05	0.014366	0.254431	0.0	0.0	0.0	0.012101	0.0	0.0	0.23724	0.993118	0.629285
Baseline_Z60	0.952857	0.88358	0.881232	1.0	1.0	1.0	0.997517	0.417782	0.794324	0.00139	4e-06	0.000308	0.125927	0.068464	0.6213	3e-06	0.0	0.002306	0.362279	0.003391	0.257371	0.0	0.0	0.0	0.003463	0.358706	0.954096	0.0	0.0	0.0	0.32735	5e-06	3.2e-05	0.946288	1.0	0.999013
Baseline_Z90	0.998729	0.992842	0.992558	0.999974	1.0	1.0	0.999994	0.79109	0.977858	0.010571	5.4e-05	0.0028	0.395174	0.257371	0.920846	4.1e-05	0.0	0.016387	0.739812	0.022812	0.617297	0.0	1e-06	2e-06	0.023228	0.736223	0.998791	0.0	0.0	0.0	0.703091	6.9e-05	0.000364	0.996375	1.0	0.999999
RBF_SVC_M30	1.0	1.0	1.0	0.552667	0.997517	0.999994	1.0	1.0	1.0	0.601224	0.034253	0.348091	0.99965	0.997156	1.0	0.02877	0.00043	0.691746	1.0	0.757455	0.999993	2.8e-05	0.0013	0.0028	0.760921	1.0	1.0	0.000382	0.00018	0.0	0.999999	0.040634	0.115684	1.0	1.0	1.0
RBF_SVC_M60	1.0	1.0	1.0	0.019361	0.417782	0.79109	1.0	1.0	1.0	0.999782	0.680269	0.994364	1.0	1.0	1.0	0.641203	0.066343	0.999951	1.0	0.999987	1.0	0.009588	0.134993	0.212926	0.999988	1.0	1.0	0.061284	0.036691	0.00042	1.0	0.717992	0.911437	1.0	0.980528	0.999999
RBF_SVC_M90	1.0	1.0	1.0	0.098632	0.794324	0.977858	1.0	1.0	1.0	0.98154	0.304017	0.894879	1.0	1.0	1.0	0.272396	0.01071	0.991648	1.0	0.995962	1.0	0.001215	0.02877	0.052143	0.996131	1.0	1.0	0.010778	0.005818	3.6e-05	1.0	0.337638	0.597193	1.0	0.999769	1.0
RBF_SVC_Z30	0.881232	0.951593	0.952857	6e-06	0.00139	0.010571	0.601224	0.999782	0.981154	1.0	1.0	1.0	1.0	0.99689	1.0	0.981764	1.0	0.999916	1.0	0.999992	0.774561	0.997156	0.999508	1.0	0.999921	0.878855	0.978551	0.947652	0.242883	0.999958	1.0	1.0	0.892678	0.064279	0.524329	
RBF_SVC_Z60	0.125927	0.210334	0.212926	0.0	4e-06	5.4e-05	0.034253	0.680269	0.304017	1.0	1.0	1.0	0.952857	0.98657	0.476217	1.0	1.0	1.0	0.736223	0.999999	0.839715	0.99996	1.0	1.0	0.999999	0.739812	0.124172	1.0	1.0	0.965195	0.771185	1.0	1.0	0.134993	0.000656	0.024516
RBF_SVC_Z90	0.67255	0.810113	0.813194	1e-06	0.000308	0.0028	0.348091	0.994364	0.894879	1.0	1.0	1.0	0.999984	1.0	0.967146	1.0	0.998729	1.0	0.997025	1.0	0.999387	0.93601	0.999921	0.999994	1.0	0.997156	0.668671	0.998375	0.993642	0.468294	0.998123	1.0	1.0	0.691746	0.021216	0.284797
LN_SVC_M30	0.999999	1.0	1.0	0.002461	0.125927	0.395174	0.99965	1.0	1.0	1.0	0.952857	0.999984	1.0	1.0	1.0	0.939079	0.269349	1.0	1.0	1.0	1.0	0.060312	0.436985	0.576983	1.0	1.0	0.999999	0.254431	0.174068	0.004183	1.0	0.964187	0.996749	1.0	0.787832	0.999013
LN_SVC_M60	0.999978	0.999999	0.999999	0.000992	0.068464	0.257371	0.997156	1.0	1.0	1.0	0.98657	1.0	1.0	1.0	1.0	0.981154	0.41019	1.0	1.0	1.0	1.0	0.11242	0.601224	0.736223	1.0	1.0	0.999976	0.391456	0.284797	0.009588	1.0	0.990647	0.999585	0.999985	0.637239	0.936462
LN_SVC_M90	1.0	1.0	1.0	0.04724	0.6213	0.920846	1.0	1.0	1.0	0.99689	0.476217	0.967146	1.0	1.0	1.0	0.436985	0.028267	0.998961	1.0	0.999608	1.0	0.003391	0.063267	0.107663	0.999629	1.0	1.0	0.025867	0.01464	0.000121	1.0	0.516257	0.774561	1.0	0.997281	1.0
LN_SVC_Z30	0.10923	0.185614	0.187988	0.0	3e-06	4.1e-05	0.02877	0.641203	0.272396	1.0	1.0	1.0	0.939079	0.981154	0.436985	1.0	1.0	1.0	0.699325	0.999998	0.810113	0.999981	1.0	1.0	0.999998	0.703091	0.107663	1.0	1.0	0.974123	0.736223	1.0	1.0	0.117344	0.000519	0.020455
LN_SVC_Z60	0.0028	0.00631	0.006439	0.0	0.0	0.0	0.00043	0.066343	0.011871	0.981764	1.0	0.998729	0.269349	0.41019	0.028267	1.0	1.0	0.963158	0.083675	0.940573	0.13314	1.0	1.0	1.0	0.939079	0.084952	0.00274	1.0	1.0	1.0	0.097188	1.0	0.999997	0.003115	3e-06	0.000273
LN_SVC_Z90	0.927863	0.974907	0.975672	1.1e-05	0.002306	0.016387	0.691746	0.999951	0.991648	1.0	1.0	1.0	1.0	1.0	0.998961	1.0	0.963158	1.0	0.999984	1.0	0.999999	0.691746	0.992558	0.998375	1.0	0.999985	0.926151	0.957669	0.909469	0.183262	0.999993	1.0	1.0	0.93601	0.091584	0.617297
LOG_REG_M30	1.0	1.0	1.0	0.01464	0.362279	0.739812	1.0	1.0	1.0	0.999916	0.736223	0.997025	1.0	1.0	1.0	0.699325	0.083675	0.999984	1.0	0.999996	1.0	0.012817	0.165216	0.254431	0.999996	1.0	1.0	0.077522	0.04724	0.000597	1.0	0.771185	0.937558	1.0	0.969012	0.999997
LOG_REG_M60	0.954096	0.986093	0.98657	1.8e-05	0.003391	0.022812	0.757455	0.999987	0.995962	1.0	0.999999	1.0	1.0	1.0	0.999608	0.999998	0.940573	1.0	0.999996	1.0	1.0	0.6213	0.985603	0.996294	1.0	0.999996	0.952857	0.932834	0.869053	0.144555	0.999998	1.0	1.0	0.959934	0.119022	0.687935
LOG_REG_M90	1.0	1.0	1.0	0.00787	0.257371	0.617297	0.999993	1.0	1.0	0.999992	0.839715	0.999387	1.0	1.0	1.0	0.810113	0.13314	0.999999	1.0	1.0	1.0	0.023228	0.245738	0.358706	1.0	1.0	1.0	0.124172	0.078722	0.001243	1.0	0.86653	0.974123	1.0	0.927863	0.999968
LOG_REG_Z30	0.00023	0.000583	0.000597	0.0	0.0	0.0	2.8e-05	0.009588	0.001215	0.774561	0.99996	0.93601	0.060312	0.11242	0.003391	0.999981	1.0	0.691746	0.012817	0.6213	0.023228	1.0	1.0	1.0	0.617297	0.013064	0.000224	1.0	1.0	1.0	0.015491	0.999921	0.99689	0.00026	0.0	1.7e-05
LOG_REG_Z60	0.007562	0.016083	0.016387	0.0	0.0	1e-06	0.0013	0.134993	0.02877	0.997156	1.0	0.999921	0.436985	0.601224	0.063267	1.0	1.0	0.992558	0.165216	0.985603	0.245738	1.0	1.0	1.0	0.985098	0.167397	0.007412	1.0	1.0	0.999991	0.187988	1.0	1.0	0.008353	1e-05	0.000845
LOG_REG_Z90	0.014919	0.030327	0.030862	0.0	0.0	2e-06	0.0028	0.212926	0.052143	0.999508	1.0	0.999994	0.576983	0.736223	0.107663	1.0	1.0	0.998375	0.254431	0.996294	0.358706	1.0	1.0	1.0	0.996131	0.257371	0.01464	1.0	1.0	0.999891	0.284797	1.0	1.0	0.016387	2.6e-05	0.001854
LN_SVC_L1_M30	0.955311	0.98657	0.987034	1.8e-05	0.003463	0.023228	0.760921	0.999988	0.996131	1.0	0.999999	1.0	1.0	1.0	0.999629	0.999998	0.930079	1.0	0.999996	1.0	1.0	0.617297	0.985098	0.996131	1.0	0.999997	0.954096	0.931204	0.86653	0.142603	0.999999	1.0	1.0	0.961031	0.12072	0.691746
LN_SVC_L1_M60	1.0	1.0	1.0	0.014366	0.358706	0.736223	1.0	1.0	1.0	0.999921	0.739812	0.997156	1.0	1.0	1.0	0.703091	0.084952	0.999985	1.0	0.999996	1.0	0.013064	0.167397	0.257371	0.999997	1.0	1.0	0.078722	0.048028	0.000611	1.0	0.774561	0.939079	1.0	0.968089	0.999997
LN_SVC_L1_M90	1.0	1.0	1.0	0.254431	0.954096	0.998791	1.0	1.0	1.0	0.878855	0.124172	0.668671	0.999999	0.999976	1.0	0.107663	0.00274	0.926151	1.0	0.952857	1.0	0.000224	0.007412	0.01464	0.954096	1.0	1.0	0.002461	0.001243	5e-06	1.0	0.142603	0.320586	1.0	0.999999	1.0
LN_SVC_L1_Z30	0.002514	0.0057	0.005818	0.0	0.0	0.0	0.000382	0.061284	0.010778	0.785551	1.0	0.998375	0.254431	0.391456	0.025867	1.0	1.0	0.957669	0.077522	0.932834	0.124172	1.0	1.0	1.0	0.931204	0.078722	0.002461	1.0	1.0	1.0	0.090224	1.0	0.999995	0.0028	2e-06	0.000241
LN_SVC_L1_Z60	0.001271	0.002985	0.003049	0.0	0.0	0.0	0.00018	0.036691	0.005818	0.947652	1.0	0.993642	0.174068	0.284797	0.01464	1.0	1.0	0.909469	0.04724	0.869053	0.078722	1.0	1.0	1.0	0.86653	0.048028	0.001243	1.0	1.0	1.0	0.055648	1.0	0.999948	0.001422	1e-06	0.000112

	Baseline_M3	Baseline_M6	Baseline_M9	Baseline_Z3	Baseline_Z6	Baseline_Z9	RBF_SVC_M3	RBF_SVC_M6	RBF_SVC_M9	RBF_SVC_Z3	RBF_SVC_Z6	RBF_SVC_Z9	LIN_SVC_M3	LIN_SVC_M6	LIN_SVC_M9	LIN_SVC_Z3	LIN_SVC_Z6	LIN_SVC_Z9	LOG_REG_M3	LOG_REG_M6	LOG_REG_M9	LOG_REG_Z3	LOG_REG_Z6	LOG_REG_Z9	SVC_L1_M3	SVC_L1_M6	SVC_L1_M9	SVC_L1_Z3	SVC_L1_Z6	SVC_L1_Z9	RF_M30	RF_M60	RF_M90	RF_Z30	RF_Z60	RF_Z90						
Baseline_M30	1.0	1.0	1.0	0.999199	1.0	1.0	0.380395	0.000382	0.000373	0.0	0.0	2e-06	1.1e-05	0.0	4.8e-05	0.000484	0.000286	3.1e-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2e-06	1.8e-05	0.0	0.0	0.0	0.0	0.0	0.0						
Baseline_M60	1.0	1.0	1.0	0.999951	1.0	1.0	0.231684	0.000124	0.000121	0.0	0.0	1e-06	3e-06	0.0	1.4e-05	0.000159	9.2e-05	9e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	5e-06	0.0	0.0	0.0	0.0	0.0	0.0						
Baseline_M90	1.0	1.0	1.0	0.999317	1.0	1.0	0.369476	0.000355	0.000347	0.0	0.0	2e-06	1e-05	0.0	4.5e-05	0.000451	0.000266	2.9e-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2e-06	1.7e-05	0.0	0.0	0.0	0.0	0.0	0.0						
Baseline_Z30	0.999199	0.999951	0.999317	1.0	1.0	0.999997	0.000656	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0					
Baseline_Z60	1.0	1.0	1.0	1.0	1.0	1.0	0.014366	1e-06	1e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	1e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0					
Baseline_Z90	1.0	1.0	1.0	0.999997	1.0	1.0	0.142603	4.7e-05	4.6e-05	0.0	0.0	0.0	1e-06	0.0	5e-06	6.1e-05	3.4e-05	3e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0					
RBF_SVC_M30	0.380395	0.231684	0.369476	0.000656	0.014366	0.142603	1.0	0.997517	0.997401	0.226215	0.028267	0.637239	0.842524	0.32735	0.952857	0.998452	0.995786	0.927863	0.35515	0.11242	0.010778	0.272396	0.01464	0.006183	0.006183	8.3e-05	9.6e-05	0.330761	0.637239	0.888187	2e-06	0.000597	4.6e-05	4e-06	4.8e-05	2.9e-05	0.0					
RBF_SVC_M60	0.000382	0.000124	0.000355	0.0	1e-06	4.7e-05	0.997517	1.0	1.0	0.999982	0.978551	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	0.999999	0.999451	0.913376	0.999995	0.940573	0.848059	0.848059	0.190384	0.205215	0.999999	1.0	1.0	0.01972	0.452558	0.14067	0.032519	0.144555	0.10923	0.0					
RBF_SVC_M90	0.000373	0.000121	0.000347	0.0	1e-06	4.6e-05	0.997401	1.0	1.0	0.999984	0.979227	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	0.99948	0.915287	0.999995	0.942041	0.850783	0.850783	0.192801	0.207763	0.999999	1.0	1.0	0.020085	0.456477	0.142603	0.033088	0.146529	0.110816	0.0						
RBF_SVC_Z30	0.0	0.0	0.0	0.0	0.0	0.0	0.226215	0.999982	0.999984	1.0	1.0	1.0	1.0	1.0	1.0	0.999963	0.999993	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.777914	0.999994	0.98876	0.861396	0.989549	0.979227	0.0	0.0						
RBF_SVC_Z60	0.0	0.0	0.0	0.0	0.0	0.0	0.028267	0.978551	0.979227	1.0	1.0	1.0	0.999981	1.0	0.999353	0.970795	0.985603	0.999754	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999931	0.994132	1.0	0.999996	0.998294	0.999997	0.999984	0.0					
RBF_SVC_Z90	2e-06	1e-06	2e-06	0.0	0.0	0.0	0.637239	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999989	1.0	0.999997	0.999916	0.999916	0.871547	0.885898	1.0	1.0	1.0	1.0	0.344589	0.98458	0.807006	0.448648	0.813194	0.746931	0.0	0.0					
LIN_SVC_M30	1.1e-05	3e-06	1e-05	0.0	0.0	1e-06	0.842524	1.0	1.0	1.0	0.999981	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999451	1.0	0.999782	0.997736	0.997736	0.680269	0.703091	1.0	1.0	1.0	1.0	0.174068	0.919022	0.589119	0.245738	0.597193	0.516257	0.0	0.0					
LIN_SVC_M60	0.0	0.0	0.0	0.0	0.0	0.0	0.32735	0.999999	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	0.999997	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.656963	0.99956	0.967146	0.760921	0.969012	0.946288	0.0	0.0					
LIN_SVC_M90	4.8e-05	1.4e-05	4.5e-05	0.0	0.0	5e-06	0.952857	1.0	1.0	1.0	0.999353	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999999	0.992842	1.0	0.996294	0.980528	0.980528	0.460407	0.484172	1.0	1.0	1.0	0.079937	0.774561	0.373099	0.12072	0.380395	0.310586	0.0	0.0					
LIN_SVC_Z30	0.000484	0.000159	0.000451	0.0	1e-06	6.1e-05	0.998452	1.0	1.0	0.999963	0.970795	1.0	1.0	0.999997	1.0	1.0	1.0	1.0	0.999999	0.999063	0.892678	0.999989	0.924411	0.819277	0.819277	0.167397	0.180931	0.999997	1.0	1.0	0.016387	0.413979	0.122436	0.027284	0.125927	0.094352	0.0	0.0				
LIN_SVC_Z60	0.000286	9.2e-05	0.000266	0.0	1e-06	3.4e-05	0.995786	1.0	1.0	0.999993	0.985603	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999723	0.934435	0.999998	0.956502	0.878855	0.878855	0.220833	0.23724	1.0	1.0	1.0	0.024516	0.500168	0.165216	0.039953	0.1696	0.129494	0.0	0.0					
LIN_SVC_Z90	3.1e-05	9e-06	2.9e-05	0.0	0.0	3e-06	0.927863	1.0	1.0	1.0	0.999754	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.996452	1.0	0.998294	0.98916	0.98916	0.52837	0.552667	1.0	1.0	1.0	1.0	0.103068	0.8282	0.436985	0.152571	0.444749	0.369476	0.0	0.0					
LOG_REG_M30	0.0	0.0	0.0	0.0	0.0	0.0	0.35515	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.979227	0.982936	1.0	1.0	1.0	0.625296	0.99317	0.958813	0.732614	0.961031	0.934435	0.0	0.0
LOG_REG_M60	0.0	0.0	0.0	0.0	0.0	0.0	0.11242	0.999451	0.99948	1.0	1.0	1.0	1.0	1.0	0.999999	0.999063	0.999723	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99967	0.999769	1.0	1.0	1.0	0.917169	0.999999	0.99885	0.958813	0.998961	0.997281	0.0	0.0	
LOG_REG_M90	0.0	0.0	0.0	0.0	0.0	0.0	0.010778	0.913376	0.915287	1.0	1.0	0.999989	0.999451	1.0	0.992842	0.892678	0.934435	0.996452	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.991325	0.993118	1.0	1.0	1.0	0.721675	0.99984	0.980528	0.816249	0.981764	0.966181	0.0	0.0
LOG_REG_Z30	0.0	0.0	0.0	0.0	0.0	0.0	0.272396	0.999995	0.999995	1.0	1.0	1.0	1.0	1.0	1.0	0.999989	0.999998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.991325	0.993118	1.0	1.0	1.0	0.721675	0.99984	0.980528	0.816249	0.981764	0.966181	0.0	0.0
LOG_REG_Z60	0.0	0.0	0.0	0.0	0.0	0.0	0.01464	0.940573	0.942041	1.0	1.0	0.999997	0.999782	1.0	0.996294	0.924411	0.956502	0.998294	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999997	0.999387	0.99885	1.0	1.0	0.999754	1.0	0.999999	0.0	0.0			
LOG_REG_Z90	0.0	0.0	0.0	0.0	0.0	0.0	0.006183	0.848059	0.850783	1.0	1.0	0.999916	0.997736	1.0	0.980528	0.819277	0.878855	0.98916	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999916	0.995017	0.99991	1.0	1.0	0.999988	1.0	1.0	0.0	0.0			
LIN_SVC_L1_M30	0.0	0.0	0.0	0.0	0.0	0.0	0.006183	0.848059	0.850783	1.0	1.0	0.999916	0.997736	1.0	0.980528	0.819277	0.878855	0.98916	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999916	0.995017	0.99991	1.0	1.0	0.999988	1.0	1.0	0.0	0.0			
LIN_SVC_L1_M60	0.0	0.0	0.0	0.0	0.0	0.0	8.3e-05	0.190384	0.192801	0.995415	1.0	0.871547	0.680269	0.984047	0.460407	0.167397	0.220833	0.52837	0.979227	0.99967	1.0	0.991325	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.983499	0.871547	0.609272	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0		
LIN_SVC_L1_M90	0.0	0.0	0.0	0.0	0.0	0.0	9.6e-05	0.205215	0.207763	0.996452	1.0	0.885898	0.703091	0.987034	0.484172	0.180931	0.23724	0.552667	0.982936	0.999769	1.0	0.993118	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.98657	0.885898	0.633266	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0		
LIN_SVC_L1_Z30	0.0	0.0	0.0	0.0	0.0	0.0	0.330761	0.999999	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	0.999997	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.983499	0.98657	1.0	1.0	1.0	0.653038	0.999535	0.966181	0.757455	0.968089	0.944898	0.0	0.0
LIN_SVC_L1_Z60	2e-06	1e-06	2e-06	0.0	0.0	0.0	0.637239	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999989	1.0	0.999997	0.999916	0.999916	0.871547	0.885898	1.0	1.0	1.0	1.0	0.344589	0.98458	0.807006	0.448648	0.813194	0.746931	0.0	0.0	0.0	0.0			
LIN_SVC_L1_Z90	1.8e-05	5e-06	1.7e-05	0.0	0.0	2e-06	0.888187	1.0	1.0	1.0	0.999931	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.998596	1.0	0.999387	0.995017	0.995017	0.609272	0.633266	1.0	1.0	1.0	1.0	0.136865	0.881232	0.516257	0.197701	0.524329	0.444749	0.0	0.0	0.0	0.0			
RF_M30	0.0	0.0	0.0	0.0	0.0	0.0	2e-06	0.01972	0.020085	0.777914	0.994132	0.344589	0.174068	0.656																												

2. Users cross-validation

We now turn to the Users cross-validation scheme, for which, as for the Audios Hybrid case, the Friedman test returned a significant p-value ($p \approx 9.42 \times 10^{-146}$). We therefore proceeded to the Nemenyi post-hoc test, applied jointly to the 36 model-condition combinations. The models here are built only on hybrid datasets. The resulting matrix is shown in 4.3.

Three patterns can be observed, and the third is more structured than in the Audios scheme:

- As in the previous settings, **baseline combinations remain mutually indistinguishable**: all pairwise comparisons among the six Baseline combinations are non-significant (smallest $p \approx 0.684$, between Baseline_M90 and Baseline_Z30).
- **Baseline combinations are significantly outperformed by every machine learning combination**. All 180 comparisons between a Baseline combination and a machine learning combination are significant, with p-values effectively equal to zero for the Zoom Baseline combinations. The largest p-value is $p = 0.0026$ (Baseline_M90 vs. LOG_REG_M60), and the Mobile Baseline combinations are the ones closest to the machine learning models. Unlike the Audios Hybrid case, where RBF_SVC_M30 was not significantly different from four Baseline combinations, there is no exception here.
- **Among machine learning combinations, differences are largely non-significant, but the few significant ones are concentrated**. Only 27 of the 360 comparisons between different machine learning models are significant, and no comparison between combinations of the same model is. All 27 involve at least one of four combinations:

RBFSVC_Z30, RBFSVC_Z60, RF_Z30, and RF_Z60.

- RBFSVC_Z30 and RBFSVC_Z60 each differ significantly from eight combinations: LINSVC_Z30, LOGREG_M30, LOGREG_M60, LOGREG_M90, LOGREG_Z30, LINSVC_L1_M30, LINSVC_L1_Z30, and LINSVC_L1_Z60.
- RF_Z30 differs significantly from seven of the same combinations (all but LOGREG_M90). RF_Z60 differs from four: LINSVC_Z30, LOGREG_M30, LOGREG_M60, and LINSVC_L1_Z30.

	Baseline_M3	Baseline_M6	Baseline_M9	Baseline_Z3	Baseline_Z6	Baseline_Z9	RBF_SVC_M3	RBF_SVC_M6	RBF_SVC_M9	RBF_SVC_Z3	RBF_SVC_Z6	RBF_SVC_Z9	LN_SVC_M3	LN_SVC_M6	LN_SVC_M9	LN_SVC_Z3	LN_SVC_Z6	LN_SVC_Z9	LOG_REG_M3	LOG_REG_M6	LOG_REG_M9	LOG_REG_Z3	LOG_REG_Z6	LOG_REG_Z9	SVC_L1_M3	SVC_L1_M6	SVC_L1_M9	SVC_L1_Z3	SVC_L1_Z6	SVC_L1_Z9	RF_M30	RF_M60	RF_M90	RF_Z30	RF_Z60	RF_Z90		
Baseline_M30	1.0	1.0	0.999997	0.999996	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	0.0	0.0	1e-06	2e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_M60	1.0	1.0	1.0	0.995415	0.999997	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8e-05	0.0	0.0	2.9e-05	4.4e-05	1e-06	3e-06	0.0	0.0	1e-06	0.0	0.0	3.6e-05	3e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_M90	0.999997	1.0	1.0	0.684109	0.969914	0.98916	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	0.0	0.001215	2e-06	1e-06	0.001854	0.002569	7.1e-05	0.000308	1e-06	0.0	0.000109	5e-06	0.0	0.002208	0.000273	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_Z30	0.999996	0.995415	0.684109	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_Z60	1.0	0.999997	0.969914	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_Z90	1.0	1.0	0.98916	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RBF_SVC_M30	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.965195	0.942041	0.990991	0.094352	0.863978	0.929548	0.070641	0.055648	0.398906	0.210334	0.943483	0.993642	0.337638	0.78455	0.999891	0.062269	0.223513	0.999963	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
RBF_SVC_M60	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.974907	0.995219	1.0	1.0	1.0	1.0	0.858786	0.999999	1.0	0.803874	0.753968	0.995415	0.964187	1.0	1.0	0.990991	0.999986	1.0	0.777914	0.969012	1.0	1.0	1.0	1.0	0.99689	0.999951	1.0	1.0	
RBF_SVC_M90	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.977858	0.995962	1.0	1.0	1.0	1.0	0.848059	0.999998	1.0	0.79109	0.739812	0.994589	0.959934	1.0	1.0	0.989549	0.999981	1.0	0.764364	0.965195	1.0	1.0	1.0	1.0	0.997401	0.999963	1.0	1.0	
RBF_SVC_Z30	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.974907	0.977858	1.0	1.0	0.998211	0.307291	0.248614	0.456477	0.001854	0.152571	0.226215	0.001215	0.000865	0.020085	0.006439	0.251511	0.492157	0.01464	0.104581	0.794324	0.001014	0.007121	0.842524	0.999795	0.951593	0.996603	1.0	1.0	0.999993	1.0	
RBF_SVC_Z60	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.995219	0.995962	1.0	1.0	0.999859	0.476217	0.402653	0.641203	0.004938	0.269349	0.373099	0.00332	0.002408	0.044942	0.015785	0.406415	0.676416	0.033666	0.195241	0.913376	0.0028	0.017329	0.940573	0.999992	0.988347	0.99967	1.0	1.0	1.0	1.0	
RBF_SVC_Z90	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.998211	0.999859	1.0	0.999996	0.999984	1.0	0.625296	0.999769	0.99997	0.548614	0.488161	0.954096	0.834013	0.999985	1.0	0.929548	0.998961	1.0	0.516257	0.848059	1.0	1.0	1.0	1.0	1.0	0.999926	1.0	1.0	1.0
LN_SVC_M30	0.0	0.0	0.0	0.0	0.0	0.0	0.965195	1.0	1.0	0.307291	0.999996	1.0	1.0	1.0	0.99999	1.0	1.0	0.999958	0.999883	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999926	1.0	0.99999	0.516257	0.797532	0.998791	0.99999	0.516257	0.797532	0.998791	0.998791	
LN_SVC_M60	0.0	0.0	1e-06	0.0	0.0	0.0	0.942041	1.0	1.0	0.248614	0.402653	0.999984	1.0	1.0	1.0	0.999998	1.0	1.0	0.99999	0.999968	1.0	1.0	1.0	1.0	1.0	1.0	0.999981	1.0	1.0	0.999754	1.0	0.999995	0.440862	0.732614	0.997025	0.99999	0.440862	0.732614
LN_SVC_M90	0.0	0.0	0.0	0.0	0.0	0.0	0.990991	1.0	1.0	0.456477	0.641203	1.0	1.0	1.0	1.0	0.999819	1.0	1.0	0.999451	0.998791	1.0	0.999998	1.0	1.0	1.0	1.0	0.999156	0.999999	1.0	0.999997	1.0	1.0	0.680269	0.905446	0.999883	0.99999	0.680269	0.905446
LN_SVC_Z30	1e-06	1.8e-05	0.001215	0.0	0.0	0.0	0.094352	0.858786	0.848059	0.001854	0.004938	0.625296	0.99999	0.999998	0.999819	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	0.999998	0.999689	1.0	1.0	0.987922	1.0	1.0	0.979227	0.464346	0.911437	0.680269	0.006059	0.024959	0.291122	0.99999
LN_SVC_Z60	0.0	0.0	2e-06	0.0	0.0	0.0	0.863978	0.999999	0.999998	0.152571	0.269349	0.999769	1.0	1.0	1.0	1.0	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.998032	1.0	0.999904	0.300763	0.581031	0.986093	0.99999	0.300763	0.581031	0.986093	0.99999
LN_SVC_Z90	0.0	0.0	1e-06	0.0	0.0	0.0	0.929548	1.0	1.0	0.226215	0.373099	0.99997	1.0	1.0	1.0	0.999999	1.0	1.0	0.999995	0.999982	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99999	1.0	0.99999	1.0	0.99999	0.41019	0.703091	0.995786	0.99999	0.41019	0.703091
LOG_REG_M30	1e-06	2.9e-05	0.001854	0.0	0.0	0.0	0.070641	0.803874	0.79109	0.001215	0.00332	0.548614	0.999958	0.99999	0.999451	1.0	1.0	0.999995	1.0	1.0	1.0	1.0	1.0	0.999989	0.99911	1.0	1.0	0.977147	1.0	1.0	0.963158	0.391456	0.869053	0.605251	0.004097	0.017653	0.234451	0.99999
LOG_REG_M60	2e-06	4.4e-05	0.002569	0.0	0.0	0.0	0.055648	0.753968	0.739812	0.000865	0.002408	0.488161	0.999883	0.999968	0.998791	1.0	0.999999	0.999982	1.0	1.0	1.0	1.0	1.0	0.999966	0.998123	1.0	1.0	0.964187	1.0	1.0	0.944898	0.337638	0.8282	0.544562	0.002985	0.013316	0.195241	0.99999
LOG_REG_M90	0.0	1e-06	7.1e-05	0.0	0.0	0.0	0.398906	0.995415	0.994589	0.020085	0.044942	0.054096	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999976	1.0	1.0	0.999926	0.88358	0.998375	0.968012	0.053001	0.158799	0.739812	0.99999	
LOG_REG_Z30	0.0	3e-06	0.000308	0.0	0.0	0.0	0.210334	0.964187	0.959934	0.006439	0.015785	0.834013	1.0	1.0	0.999998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999995	1.0	1.0	0.999063	1.0	1.0	0.997937	0.699325	0.982358	0.871547	0.019008	0.067397	0.508202	0.99999	
LOG_REG_Z60	0.0	0.0	1e-06	0.0	0.0	0.0	0.943483	1.0	1.0	0.251511	0.406415	0.999985	1.0	1.0	1.0	0.999998	1.0	1.0	0.999989	0.999966	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999979	1.0	1.0	0.999769	1.0	0.999995	0.444749	0.736223	0.997156	0.99999	
LOG_REG_Z90	0.0	0.0	0.0	0.0	0.0	0.0	0.993642	1.0	1.0	0.492157	0.676416	1.0	1.0	1.0	1.0	0.999689	1.0	1.0	0.99911	0.998123	1.0	0.999995	1.0	1.0	1.0	1.0	1.0	0.998664	0.999997	1.0	0.999999	1.0	1.0	0.714292	0.922643	0.999936	0.99999	
LN_SVC_L1_M30	0.0	1e-06	0.000109	0.0	0.0	0.0	0.337638	0.990991	0.989549	0.01464	0.033666	0.929548	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999921	1.0	1.0	0.999782	0.839715	0.996452	0.950304	0.039953	0.125927	0.676416	0.99999	
LN_SVC_L1_M60	0.0	0.0	5e-06	0.0	0.0	0.0	0.78455	0.999986	0.999981	0.104581	0.195241	0.998961	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.993642	0.999998	0.9999508	0.220833	0.472251	0.967146	0.99999	
LN_SVC_L1_M90	0.0	0.0	0.0	0.0	0.0	0.0	0.999891	1.0	1.0	0.794324	0.913376	1.0	1.0	1.0	1.0	0.987922	1.0	1.0	0.977147	0.964187	0.999976	0.999063	1.0	1.0	0.999921	1.0	1.0	0.970795	0.99928	1.0	1.0	1.0	1.0	0.931204	0.993118	1.0	1.0	
LN_SVC_L1_Z30	2e-06	3.6e-05	0.002208	0.0	0.0	0.0	0.062269	0.777914	0.764364	0.001014	0.0028	0.516257	0.999926	0.999981	0.999156	1.0	0.999999	0.99999	1.0	1.0	1.0	1.0	1.0	0.999979	0.998664	1.0	1.0	0.970795	1.0	1.0	0.954096	0.362279	0.848059	0.572933	0.003463	0.015203	0.212926	0.99999
LN_SVC_L1_Z60	0.0	3e-06	0.000273	0.0	0.0	0.0	0.223513	0.969012	0.965195	0.007121	0.017329	0.848059	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999997	1.0	1.0	0.99928	1.0	1.0	0.998375	0.717992	0.985098	0.88358	0.020832	0.027875	0.52837	0.99999	
LN_SVC_L1_Z90	0.0	0.0	0.0	0.0	0.0	0.0	0.999963	1.0	1.0	0.842524	0.940573	1.0	1.0	1.0	1.0	0.979227	1.0	1.0	0.963158	0.944898	0.999926	0.997937	1.0	1.0	0.999782	1.0	1.0	0.954096	0.998375	1.0	1.0	1.0	1.0	0.954096	0.996452	1.0	1.0	
RF_M30	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.999795	0.999992	1.0	0.999926	0.999754	0.999997	0.464346	0.998032	0.999608	0.391456	0.337638	0.86358	0.699325																

3. Audios_x_Users cross-validation

Finally, we consider the Audios_x_Users cross-validation scheme, again using only the hybrid datasets. Here too the Friedman test returned a significant p-value ($p \approx 8.68 \times 10^{-104}$), so we proceeded to the Nemenyi post-hoc test, applied jointly to the 36 model-condition combinations. The resulting matrix is shown in

Three patterns can be observed, and the third is the most extreme of the three schemes:

- **Baseline combinations remain mutually indistinguishable.** All pairwise comparisons among the six Baseline combinations are non-significant (smallest $p \approx 0.520$, between Baseline_M90 and Baseline_Z30).
- **Baseline combinations are significantly outperformed by every machine learning combination.** All 180 comparisons between a Baseline combination and a machine learning combination are significant. For the Zoom Baseline combinations and for Baseline_M30 and Baseline_M60, the p-values are below 10^{-4} in nearly all cases, and mostly effectively equal to zero. The Mobile Baseline combinations at 90 seconds are the closest to the machine learning models, with the largest p-value being $p \approx 0.016$ (Baseline_M90 vs. LOG_REG_M30). As under the Users scheme, and unlike the Audios scheme, there is no exception.
- **Among machine learning combinations, no difference is significant.** None of the 360 comparisons between different machine learning models, and none of the comparisons between combinations of the same model, reaches significance. The smallest p-value among all machine learning combinations is $p \approx 0.183$ (RBF_SVC_Z60 vs.

LOG_REG_M30), and the large majority of p-values are above 0.9. Under this scheme the five classifiers are therefore statistically equivalent across all device and window-length conditions.

	Baseline_M3	Baseline_M60	Baseline_M90	Baseline_Z30	Baseline_Z60	Baseline_Z90	RBF_SVC_M3	RBF_SVC_M60	RBF_SVC_M90	RBF_SVC_Z30	RBF_SVC_Z60	RBF_SVC_Z90	LN_SVC_M3	LN_SVC_M60	LN_SVC_M90	LN_SVC_Z30	LN_SVC_Z60	LN_SVC_Z90	LOG_REG_M3	LOG_REG_M60	LOG_REG_M90	LOG_REG_Z30	LOG_REG_Z60	LOG_REG_Z90	SVC_L1_M3	SVC_L1_M60	SVC_L1_M90	SVC_L1_Z30	SVC_L1_Z60	SVC_L1_Z90	RF_M30	RF_M60	RF_M90	RF_Z30	RF_Z60	RF_Z90			
Baseline_M30	1.0	1.0	0.999998	0.999898	1.0	1.0	0.0	1e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	0.0	2.5e-05	1e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1e-06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Baseline_M60	1.0	1.0	1.0	0.986093	0.999904	0.999999	0.0	1e-05	0.0	0.0	0.0	0.0	1e-06	0.0	1e-06	7e-06	1.3e-05	0.0	0.000323	1.7e-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7e-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Baseline_M90	0.999998	1.0	1.0	0.520291	0.878855	0.976419	1e-05	0.000992	0.0	4e-06	0.0	0.0	0.000155	5e-06	9.4e-05	0.000703	0.001215	0.0	0.016387	0.001554	1.9e-05	1.3e-05	0.0	2.7e-05	3.3e-05	1.7e-05	0.0	0.001554	8e-06	3e-06	0.0	0.0	0.0	0.0	3e-06	2e-06	3.2e-05	0.0	
Baseline_Z30	0.999898	0.986093	0.520291	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Baseline_Z60	1.0	0.999904	0.878855	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Baseline_Z90	1.0	0.999999	0.976419	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
RBF_SVC_M30	0.0	0.0	1e-05	0.0	0.0	0.0	1.0	1.0	1.0	0.996294	0.99999	1.0	1.0	1.0	1.0	1.0	1.0	0.999968	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999063	1.0	1.0	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	1.0		
RBF_SVC_M60	1e-06	1e-05	0.000992	0.0	0.0	0.0	1.0	1.0	0.999994	1.0	0.656963	0.939079	1.0	1.0	1.0	1.0	1.0	0.999994	1.0	1.0	1.0	1.0	0.999951	1.0	1.0	1.0	0.764364	1.0	1.0	1.0	0.999921	0.974123	0.999996	1.0	1.0	1.0	1.0		
RBF_SVC_M90	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.999994	1.0	1.0	0.999997	1.0	1.0	1.0	1.0	0.999998	1.0	0.988347	0.999972	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999972	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
RBF_SVC_Z30	0.0	0.0	4e-06	0.0	0.0	0.0	1.0	1.0	1.0	0.999387	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99965	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999891	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
RBF_SVC_Z60	0.0	0.0	0.0	0.0	0.0	0.0	0.996294	0.656963	0.999997	0.999387	1.0	1.0	0.901307	0.998907	0.937558	0.714292	0.6213	1.0	0.183262	0.576983	0.990647	0.994589	1.0	0.98458	0.979886	0.991961	1.0	0.576983	0.997281	0.999689	1.0	1.0	0.999995	0.999508	0.99984	0.981154	0.0		
RBF_SVC_Z90	0.0	0.0	0.0	0.0	0.0	0.0	0.99999	0.939079	1.0	1.0	1.0	1.0	0.995219	0.999999	0.998032	0.958813	0.924411	1.0	0.512227	0.903391	0.999944	0.999979	1.0	0.999859	0.999769	0.999958	1.0	0.903391	0.999994	1.0	1.0	1.0	1.0	1.0	1.0	0.999795	0.0		
LN_SVC_M30	0.0	1e-06	0.000155	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.901307	0.995219	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.950304	1.0	1.0	1.0	1.0	1.0	0.99885	1.0	1.0	1.0	1.0	1.0		
LN_SVC_M60	0.0	0.0	5e-06	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.998907	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	0.999819	1.0	1.0	1.0	1.0	1.0	1.0	0.999782	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_M90	0.0	1e-06	9.4e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.937558	0.998032	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.971656	1.0	1.0	1.0	1.0	0.999608	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_Z30	0.0	7e-06	0.000703	0.0	0.0	0.0	1.0	1.0	0.999998	1.0	0.714292	0.958813	1.0	1.0	1.0	1.0	1.0	0.999979	1.0	1.0	1.0	1.0	0.999984	1.0	1.0	1.0	0.813194	1.0	1.0	1.0	0.999972	0.984047	0.999999	1.0	1.0	1.0	1.0		
LN_SVC_Z60	1e-06	1.3e-05	0.001215	0.0	0.0	0.0	1.0	1.0	0.999988	1.0	0.6213	0.924411	1.0	1.0	1.0	1.0	1.0	0.999891	1.0	1.0	1.0	1.0	0.99991	1.0	1.0	1.0	0.732614	1.0	1.0	1.0	0.999859	0.966181	0.999992	1.0	1.0	1.0	1.0		
LN_SVC_Z90	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.99994	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999979	0.999891	1.0	0.968089	0.999782	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999782	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
LOG_REG_M30	2.5e-05	0.000323	0.016387	0.0	0.0	0.0	0.999968	1.0	0.988347	0.99965	0.183262	0.512227	1.0	0.999819	1.0	1.0	1.0	0.968089	1.0	1.0	0.999995	0.999984	0.970795	0.999998	0.999999	0.999993	0.260334	1.0	0.999948	0.999317	0.964187	0.637239	0.990292	0.99956	0.998791	0.999999	0.0		
LOG_REG_M60	1e-06	1.7e-05	0.001554	0.0	0.0	0.0	1.0	1.0	0.999972	1.0	0.576983	0.903391	1.0	1.0	1.0	1.0	1.0	0.999782	1.0	1.0	1.0	1.0	0.999819	1.0	1.0	1.0	0.691746	1.0	1.0	1.0	0.999723	0.954096	0.999981	1.0	1.0	1.0	1.0		
LOG_REG_M90	0.0	0.0	1.9e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.990647	0.999944	1.0	1.0	1.0	1.0	1.0	0.999995	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.997156	1.0	1.0	1.0	1.0	0.999995	1.0	1.0	1.0	1.0	1.0	1.0		
LOG_REG_Z30	0.0	0.0	1.3e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.994589	0.999979	1.0	1.0	1.0	1.0	1.0	0.999984	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.998526	1.0	1.0	1.0	1.0	0.999998	1.0	1.0	1.0	1.0	1.0	1.0		
LOG_REG_Z60	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.999951	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999984	0.99991	1.0	0.970795	0.999819	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
LOG_REG_Z90	0.0	0.0	2.7e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.98458	0.999859	1.0	1.0	1.0	1.0	1.0	0.999998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.994807	1.0	1.0	1.0	1.0	0.999984	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_L1_M30	0.0	0.0	3.3e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.979886	0.999769	1.0	1.0	1.0	1.0	1.0	0.999999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.992842	1.0	1.0	1.0	1.0	0.99997	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_L1_M60	0.0	0.0	1.7e-05	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.991961	0.999958	1.0	1.0	1.0	1.0	1.0	0.999993	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.997629	1.0	1.0	1.0	1.0	0.999996	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_L1_M90	0.0	0.0	0.0	0.0	0.0	0.0	0.999063	0.764364	1.0	0.999891	1.0	1.0	0.950304	0.999782	0.971656	0.813194	0.732614	1.0	0.260334	0.691746	0.997156	0.998526	1.0	0.994807	0.992842	0.997629	1.0	0.691746	0.999353	0.999951	1.0	1.0	1.0	0.999916	0.999978	0.993384	0.0		
LN_SVC_L1_Z30	1e-06	1.7e-05	0.001554	0.0	0.0	0.0	1.0	1.0	0.999972	1.0	0.576983	0.903391	1.0	1.0	1.0	1.0	1.0	0.999782	1.0	1.0	1.0	1.0	0.999819	1.0	1.0	1.0	0.691746	1.0	1.0	1.0	1.0	0.999723	0.954096	0.999981	1.0	1.0	1.0	1.0	
LN_SVC_L1_Z60	0.0	0.0	8e-06	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.997281	0.999994	1.0	1.0	1.0	1.0	1.0	0.999948	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999353	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
LN_SVC_L1_Z90	0.0	0.0	3e-06	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.999689	1.0	1.0	1.0	1.0	1.0	1.0	0.999317	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999951	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
RF_M30	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.999921	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999972	0.999859	1.0	0.964187	0.999723	1.0	1.0	1.0	1.0	1.0	1.0	0.999723	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
RF_M60	0.0	0.0	0.0	0.0	0.0	0.0	0.999999	0.974123	1.0	1.0	1.0	1.0	0.99885	1.0	0.999608	0.984047	0.966181	1.0	0.637239	0.954096	0.999995	0.999998	1.0	0.999984	0.999997	0.999996	1.0	0.954096	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.999974	0.0		
RF_M90	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.999996	1.0	1.0	0.999995	1.0	1.0	1.0	1.0	0.999999	0.999992	1.0	0.990292	0.999981	1.0	1.0	1.0	1.0	1.0	1.0	0.999981	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
RF_Z30	0.0	0.0	3e-06	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.999508	1.0	1.0	1.0	1.0	1.0	1.0	0.9995																					

4.2.4 RQ4 - Do Device and Window Length Affect Performance?

This research question is divided into two parts: the effect of the recording device and the effect of the window length.

4.2.4.1 Device effect (mobile vs zoom)

As explained in 3.4.5, for each model and feature type, balanced accuracy was first averaged across the three window lengths for each test split and device, giving 50 paired values per device. This preserves the paired structure and avoids counting the same split three times. The Shapiro-Wilk test was applied to the 50 differences ($\Delta = \text{Mobile} - \text{Zoom}$) to choose between the paired t-test and the Wilcoxon signed-rank test. Benjamini-Hochberg FDR correction was applied separately within each feature type, over the six model comparisons.

1. Audios cross-validation

The analysis of the device effect begin with models built on the **quantitative datasets** of the Audios cross-validation. All six comparisons (see Table 4.35) were significant after correction ($pval_corrected < 0.05$), and the paired t-test was used in every case. The direction of the effect depended on the model. The Baseline performed better on Mobile recordings (70.26% vs. 64.67%, $\Delta = 5.59$, $pval_corrected = 8.20 \times 10^{-4}$). RF showed the same direction, with a smaller gap (73.21% vs. 68.86%, $\Delta = 4.35$, $pval_corrected = 3.28 \times 10^{-3}$).

The other four classifiers showed the opposite pattern, with Zoom significantly better than Mobile:

- **RBF_SVC**: $\Delta = -3.38$, $pval_corrected \approx 9.66 \times 10^{-3}$.

4.2 Model Comparison

- **LIN_SVC**: $\Delta = -3.10$, $pval_corrected \approx 0.016$.
- **LOG_REG**: $\Delta = -4.04$, $pval_corrected \approx 4.48 \times 10^{-3}$.
- **LIN_SVC_L1**: $\Delta = -5.27$, $pval_corrected \approx 1.07 \times 10^{-4}$, the largest Zoom advantage and the smallest corrected p-value of the table

Model	Test	Mean Bal. Acc. (Mobile)	Mean Bal. Acc. (Zoom)	$\Delta = Mobile - Zoom$	pval_raw	pval_corrected	Significant	Best device
Baseline	t-test	70.26%	64.67%	5.59%	0.000273	0.000820	True	MOBILE
RBF_SVC	t-test	70.57%	73.95%	-3.38%	0.008045	0.009655	True	ZOOM
LIN_SVC	t-test	71.49%	74.59%	-3.10%	0.016338	0.016338	True	ZOOM
LOG_REG	t-test	71.80%	75.84%	-4.04%	0.002985	0.004478	True	ZOOM
LIN_SVC_L1	t-test	71.10%	76.37%	-5.27%	0.000018	0.000107	True	ZOOM
RF	t-test	73.21%	68.86%	4.35%	0.001641	0.003282	True	MOBILE

Table 4.35: device effect on quantitative datasets, obtained from audios cross-validation.

Table 4.36 instead reports the same analysis for the models built on the **hybrid datasets**. The picture is clearly different: only three of the six comparisons were significant, and the Wilcoxon signed-rank test was needed for three models (RBF_SVC, LIN_SVC_L1 and RF). The Baseline is unchanged, since it cannot use the qualitative features ($\Delta = 5.59$, $pval_corrected \approx 1.64 \times 10^{-3}$, in favour of Mobile). Among the classifiers, two significant effects remain, and they go in opposite directions. The RBF_SVC performed better on Zoom (84.28% vs. 80.98%, $\Delta = -3.29$, $pval_corrected = 6.37 \times 10^{-3}$). LIN_SVC_L1 performed better on Mobile (85.98% vs. 83.85%, $\Delta = 2.13$, $pval_corrected = 0.019$).

For the other three classifiers, no significant device difference was found. This holds for the LIN_SVC ($\Delta = 0.13$, $pval_corrected = 0.88$), LOG_REG ($\Delta = -0.70$, $pval_corrected = 0.46$) and RF ($\Delta = -0.48$, $pval_corrected = 0.46$). In particular, RF, which showed a Mobile advantage of 4.35 points with quantitative features, shows no significant device effect with hybrid features.

4.2 Model Comparison

Model	Test	Mean Bal. Acc. (Mobile)	Mean Bal. Acc. (Zoom)	$\Delta = \text{Mobile} - \text{Zoom}$	pval_raw	pval_corrected	Significant	Best device
Baseline	t-test	70.26%	64.67%	5.59%	0.000273	0.001639	True	MOBILE
RBF_SVC	wilcoxon	80.98%	84.28%	-3.29%	0.002122	0.006367	True	ZOOM
LIN_SVC	t-test	82.75%	82.62%	0.13%	0.884259	0.884259	False	N.S.E.D
LOG_REG	t-test	84.50%	85.20%	-0.70%	0.382098	0.458518	False	N.S.E.D
LIN_SVC_L1	wilcoxon	85.98%	83.85%	2.13%	0.009256	0.018512	True	MOBILE
RF	wilcoxon	86.68%	87.16%	-0.48%	0.324807	0.458518	False	N.S.E.D

Table 4.36: device effect on hybrid datasets, obtained from audios cross-validation. N.S.E.D. stands for No Strong Evidence of Difference

2. Users cross-validation

Moving on the Users cross-validation scheme, for which only models built on hybrid datasets are available, the device effect is considerably weaker than under the Audios scheme. There, with hybrid features, three of the six comparisons were significant, two of them involving a machine learning model. Here, the same procedure yields a significant device difference only for the Baseline (Table 4.37).

The Baseline performed significantly better on Mobile recordings (75.70% vs. 68.06%, $\Delta = 7.64$, $pval_corrected = 4.39 \times 10^{-9}$), the largest device gap observed in any RQ4-1 comparison so far. For the five machine learning models, no significant difference was found, and the paired t-test was used in all six comparisons. Zoom was numerically ahead for the RBF_SVC ($\Delta = -0.87$, $pval_corrected = 0.197$), LOG_REG ($\Delta = -0.83$, $pval_corrected = 0.197$) and RF ($\Delta = -1.02$, $pval_corrected = 0.093$). Mobile was numerically ahead for the LIN_SVC ($\Delta = 0.67$, $pval_corrected = 0.237$) and LIN_SVC_L1 ($\Delta = 0.41$, $pval_corrected = 0.365$). The RF comparison is the closest to significance. Its raw p-value (0.031) falls below 0.05 but does not survive the correction.

4.2 Model Comparison

Model	Test	Mean Bal. Acc. (Mobile)	Mean Bal. Acc. (Zoom)	$\Delta = \text{Mobile} - \text{Zoom}$	pval_raw	pval_corrected	Significant	Best device
Baseline	t-test	75.70%	68.06%	7.64%	7.31e-10	4.39e-9	True	MOBILE
RBF_SVC	t-test	90.84%	91.71%	-0.87%	0.13114987	0.19672480	False	N.S.E.D
LIN_SVC	t-test	89.51%	88.85%	0.67%	0.19725570	0.23670684	False	N.S.E.D
LOG_REG	t-test	88.32%	89.15%	-0.83%	0.12675023	0.19672480	False	N.S.E.D
LIN_SVC.L1	t-test	89.29%	88.88%	0.41%	0.36545108	0.36545108	False	N.S.E.D
RF	t-test	90.71%	91.73%	-1.02%	0.03089838	0.09269514	False	N.S.E.D

Table 4.37: device effect on hybrid datasets, obtained from Users cross-validation. N.S.E.D. stands for No Strong Evidence of Difference

3. Audio & User cross-validation

We now consider the Audio & User cross-validation scheme, again with only models built on hybrid datasets. The picture is the same as under the Users scheme: a significant device difference is found only for the Baseline (Table 4.38). The same procedure was applied.

The Baseline performed significantly better on Mobile recordings (73.15% vs. 66.54%, $\Delta = 6.61$, $pval_corrected = 7.00 \times 10^{-5}$). This gap lies between those observed under the Audios scheme (5.59) and the Users scheme (7.64), so the Baseline’s Mobile advantage appears under all three schemes. For the five machine learning models, no significant difference was found. The Wilcoxon signed-rank test was used for three of them (RBF_SVC, LOG_REG and RF) and the paired t-test for the other two (LIN_SVC and LIN_SVC.L1). Zoom was numerically ahead for the RBF_SVC ($\Delta = -1.29$, $pval_corrected \approx 0.054$), Logistic Regression ($\Delta = -1.69$, $pval_corrected \approx 0.290$) and the LIN_SVC ($\Delta = -0.23$, $pval_corrected \approx 0.774$). Mobile was numerically ahead for LIN_SVC.L1 ($\Delta = 0.31$, $pval_corrected \approx 0.774$) and Random Forest ($\Delta = 1.07$, $pval_corrected \approx 0.590$). The RBF_SVC comparison is the closest to significance. Its raw p-value (0.018) falls below 0.05 but does not survive the correction.

4.2 Model Comparison

Model	Test	Mean Bal. Acc. (Mobile)	Mean Bal. Acc. (Zoom)	$\Delta = \text{Mobile} - \text{Zoom}$	pval_raw	pval_corrected	Significant	Best device
Baseline	wilcoxon	73.15%	66.54%	6.61%	0.000012	0.000070	True	MOBILE
RBF_SVC	wilcoxon	84.71%	86.00%	-1.29%	0.018061	0.054183	False	N.S.E.D
LIN_SVC	t-test	84.59%	84.82%	-0.23%	0.774408	0.774408	False	N.S.E.D
LOG_REG	wilcoxon	83.40%	85.09%	-1.69%	0.145039	0.290079	False	N.S.E.D
LIN_SVC_L1	t-test	85.09%	84.79%	0.31%	0.649636	0.774408	False	N.S.E.D
RF	wilcoxon	86.33%	85.27%	1.07%	0.393495	0.590242	False	N.S.E.D

Table 4.38: device effect on hybrid datasets, obtained from Audio & User cross-validation. N.S.E.D. stands for No Strong Evidence of Difference

4.2.4.2 Window length effect (30s vs 60s vs 90s)

We now turn to RQ4-2, the effect of the window length. Since three window lengths (30, 60 and 90 seconds) are compared, pairwise tests are no longer appropriate. As for RQ3, we used the two-stage procedure: a Friedman omnibus test, and the Nemenyi post-hoc test only when the Friedman test was significant. For each model and feature type, balanced accuracy was first averaged across the two devices for each test split, giving 50 values per window length, with the 50 splits acting as the "subjects" and the three window lengths as the "conditions". As the Nemenyi test corrects internally for multiple comparisons, no further correction was applied

1. Audios cross-validation

Starting with Audios cross-validation, we first applied the procedure using models built **quantitative datasets** and Table 4.39 reports the results that we obtained. The Friedman test was significant for one model only, the Baseline ($p \approx 5.17 \times 10^{-4}$). No significant effect of the window length was found for the five machine learning models i.e., RBF_SVC ($p \approx 0.099$), LIN_SVC ($p \approx 0.127$), LOG_REG ($p \approx 0.868$), LIN_SVC_L1 ($p \approx 0.995$), and RF ($p \approx 0.245$).

For the Baseline, balanced accuracy increased steadily with window length (66.65%, 67.45%, 68.30%). The Nemenyi test showed that 30 seconds differed significantly from both 60 seconds ($p \approx 0.00549$) and 90 seconds

4.2 Model Comparison

($p \approx 0.0014$), while 60 and 90 seconds were not distinguishable ($p \approx 0.916$). The gain therefore appears between 30 and 60 seconds, then levels off. For the machine learning models, the differences between window lengths were small, at most about 1.7 points (LIN_SVC: 73.87% at 60 seconds vs. 72.14% at 90 seconds). The LIN_SVC and the RBF_SVC showed their lowest values at 90 and 30 seconds respectively, but these differences are not significant.

Model	Mean Bal. Acc. (win30)	Mean Bal. Acc. (win60)	Mean Bal. Acc. (win90)	pval_friedman	pval_nemenyi 30-60	pval_nemenyi 30-90	pval_nemenyi 60-90
Baseline	66.65%	67.45%	68.30%	0.000517	5.49e-03	1.35e-03	0.916
RBF_SVC	71.71%	72.76%	72.32%	0.099140			
LIN_SVC	73.11%	73.87%	72.14%	0.127454			
LOG_REG	73.85%	74.10%	73.51%	0.868130			
LIN_SVC_L1	74.03%	73.29%	73.88%	0.994987			
RF	70.33%	71.37%	71.41%	0.245099			

Table 4.39: window length effect on quantitative datasets, obtained from audios cross-validation.

Subsequently, the same procedure was performed using models built on **hybrid datasets**. Table 4.40 reports the same analysis for the hybrid datasets. The result is the same as in the Quantitative setting: the Friedman test was significant only for the Baseline, whose values and Nemenyi p-values are identical to those above, as expected since it cannot use the qualitative features. Here also no significant effect of the window length was found for the five machine learning models i.e., RBF_SVC ($p \approx 0.144$), LIN_SVC ($p \approx 0.350$), LOG_REG ($p \approx 0.080$), LIN_SVC_L1 ($p \approx 0.913$), and RF ($p \approx 0.439$).

LOG_REG is the closest to significance. Its balanced accuracy increased monotonically with window length (84.39%, 84.87%, 85.29%), but the Friedman p-value (0.080) does not reach 0.05, so this trend cannot be confirmed. RF showed the opposite tendency, with its highest value at 30 seconds (87.49%), again without significance. For all models the spread across window lengths is at most about 1.3 points (RF: 87.49% vs. 86.48%).

4.2 Model Comparison

Model	Mean Bal. Acc. (win30)	Mean Bal. Acc. (win60)	Mean Bal. Acc. (win90)	pval_friedman	pval_nemenyi 30-60	pval_nemenyi 30-90	pval_nemenyi 60-90
Baseline	66.65%	67.45%	68.30%	0.000517	5.49e-03	1.35e-03	0.916
RBF_SVC	82.27%	83.46%	82.16%	0.143704			
LIN_SVC	82.80%	82.77%	82.48%	0.349761			
LOG_REG	84.39%	84.87%	85.29%	0.080393			
LIN_SVC_L1	84.83%	85.10%	84.81%	0.912627			
RF	87.49%	86.48%	86.78%	0.439454			

Table 4.40: window length effect on hybrid datasets, obtained from audios cross-validation.

2. Users cross-validation

Shifting our attention to the Users cross-validation scheme, for which only models built on hybrid datasets are available, the window length has a stronger effect than under the Audios scheme. There, only the Baseline was affected, and the gain levelled off after 60 seconds. Here the same procedure shows a significant effect for two models: the Baseline and, for the first time in RQ4-2, one machine learning classifier (Table 4.41).

The Baseline showed a significant effect ($p = 1.81 \times 10^{-13}$), with balanced accuracy increasing steadily with window length (70.32%, 72.08%, 73.25%). Unlike under the Audios scheme, all three pairwise differences were significant: 30 vs. 60 seconds ($p = 1.02 \times 10^{-6}$), 30 vs. 90 seconds ($p = 1.91 \times 10^{-13}$) and 60 vs. 90 seconds ($p = 0.043$). The gain therefore does not stop at 60 seconds, although the 60-to-90 step is the smallest and the weakest in significance.

Among the classifiers, LIN_SVC_L1 was the only one with a significant Friedman test ($p = 2.24 \times 10^{-3}$). Its balanced accuracy was 88.58%, 88.74% and 89.94% at 30, 60 and 90 seconds. The Nemenyi test showed that 90 seconds differed significantly from both 30 seconds ($p \approx 0.0055$) and 60 seconds ($p = 0.016$), while 30 and 60 seconds were not distinguishable ($p = 0.935$). The advantage of the longest window is about 1.2 to 1.4 points.

No significant effect was found for the other four classifiers: RBF_SVC

($p = 0.078$), LIN_SVC ($p = 0.192$), LOG_REG ($p = 0.199$), and RF ($p = 0.344$).

The tendencies were not uniform. LIN_SVC and LOG_REG increased slightly with window length (by 0.52 and 0.87 points between 30 and 90 seconds). The RBF_SVC and RF showed the opposite tendency, with their highest values at 30 seconds (92.06% and 91.61%). The decline was larger for the RBF_SVC (1.52 points), whose p-value (0.078) is the closest to significance among these four.

Model	Mean Bal. Acc. (win30)	Mean Bal. Acc. (win60)	Mean Bal. Acc. (win90)	pval_friedman	pval_nemenyi 30-60	pval_nemenyi 30-90	pval_nemenyi 60-90
Baseline	70.32%	72.08%	73.25%	1.81e-13	1.02e-06	1.91e-13	4.33e-02
RBF_SVC	92.06%	91.22%	90.54%	0.07825843236870701			
LIN_SVC	88.93%	89.15%	89.45%	0.19175914424239393			
LOG_REG	88.28%	88.76%	89.15%	0.19920873948932147			
LIN_SVC.L1	88.58%	88.74%	89.94%	0.0022416813308758503	0.935	5.49e-03	0.016
RF	91.61%	91.27%	90.77%	0.34415378686540515			

Table 4.41: window length effect on hybrid datasets, obtained from Users cross-validation.

3. Audio & User cross-validation

We conclude RQ4-2 and this chapter with the Audio & User cross-validation scheme, again with only models built on hybrid datasets. The window length has its broadest effect here. Under the Audios scheme only the Baseline was affected, and under the Users scheme the Baseline and one classifier were. Here the same procedure gives a significant effect for the Baseline and for two classifiers (Table 4.42).

The Baseline showed a significant effect ($p = 2 \times 10^{-6}$), with balanced accuracy increasing steadily with window length (68.74%, 69.86%, 70.95%). As under the Users scheme, all three pairwise differences were significant: 30 vs. 60 seconds ($p \approx 0.019$), 30 vs. 90 seconds ($p = 1 \times 10^{-6}$) and 60 vs. 90 seconds ($p \approx 0.043$). The gain from 30 to 90 seconds (2.21 points) lies between those observed under the Audios (1.65) and Users (2.93) schemes.

4.2 Model Comparison

Among the classifiers, two had a significant Friedman test, and their patterns differ:

- **LIN_SVC_L1** ($p \approx 0.011$): balanced accuracy was 84.40%, 84.87% and 85.55% at 30, 60 and 90 seconds. The Nemenyi test showed that only 30 vs. 90 seconds differed significantly ($p \approx 0.014$), while 30 vs. 60 seconds ($p \approx 0.703$) and 60 vs. 90 seconds ($p \approx 0.112$) did not. The gain is gradual, about 1.2 points over the full range, and it is the same direction as under the Users scheme.
- **RF** ($p = 0.042$): balanced accuracy was 85.60%, 86.31% and 85.49%, so the best window was 60 seconds. The only significant pair was 60 vs. 90 seconds ($p \approx 0.033$), with a difference of about 0.8 points. The pairs 30 vs. 60 seconds ($p = 0.514$) and 30 vs. 90 seconds ($p = 0.341$) were not significant.

No significant effect was found for the remaining three classifiers: RBF_SVC ($p \approx 0.965$), LIN_SVC ($p \approx 0.554$), and LOG_REG ($p \approx 0.123$). Nevertheless, RBF_SVC and LIN_SVC showed slight increases in performance as the window length increased, with gains of 0.85 and 0.70 percentage points, respectively, from 30 s to 90 s. LOG_REG exhibited a similar trend and was the classifier closest to reaching statistical significance, with an improvement of 0.75 percentage points.

Model	Mean Bal. Acc. (win30)	Mean Bal. Acc. (win60)	Mean Bal. Acc. (win90)	pval_friedman	pval_nemenyi 30-60	pval_nemenyi 30-90	pval_nemenyi 60-90
Baseline	68.74%	69.86%	70.95%	2.00e-6	0.019004	1.00e-6	0.043289
RBF_SVC	84.92%	85.38%	85.77%	0.965436			
LIN_SVC	84.34%	84.75%	85.04%	0.553824			
LOG_REG	83.76%	84.46%	84.51%	0.123164			
LIN_SVC_L1	84.40%	84.87%	85.55%	0.010870	0.703015	0.014150	0.112183
RF	85.60%	86.31%	85.49%	0.041932	0.514054	0.340875	0.033242

Table 4.42: window length effect on hybrid datasets, obtained from Audio & User cross-validation.

Chapter 5

Discussion

This chapter interprets the results of the experimental campaign in light of the research questions introduced earlier (RQ1-RQ4) and of the three usage scenarios represented by the cross-validation schemes: a returning user capturing a new soundscape (Audios), a new user evaluating an archived soundscape (Users), and a new user capturing a new soundscape (Audios $\times$ Users).

A preliminary remark concerns the choice of the classification task. Throughout this work we adopted binary classification (noisy vs. quiet). Although binary and multiclass classification were not systematically and statistically compared across all models, the results of the RF models built on the quantitative datasets under the Audios cross-validation scheme to do these tasks (see Table 4.1 for Multiclass classification and 4.2 for Binary one) show that the multiclass performance was considerably lower than the binary one. This discrepancy may be due to the limited amount of data available, which is probably insufficient to support a four-class problem. The binary formulation was therefore preferred, and the discussion that follows refers to it exclusively.

We start with RQ1, which asks whether the Baseline Model is sufficient or whether machine learning models improve performance.

5.1 RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?

1. Audios cross-validation

The answer to RQ1 under the Audios scheme depends on the type of features used.

On the **quantitative datasets**, 16 of the 30 comparisons against the Baseline were significant after correction, so the machine learning models do not improve on the Baseline uniformly. The improvement is concentrated on the Zoom recordings, where the Baseline is weakest (63.11%-66.20% balanced accuracy). There, four of the five models (RBF_SVC, LIN_SVC, LOG_REG, LIN_SVC_L1) are significantly better at all three window lengths, with gains of roughly 7 to 13 percentage points. On the Mobile recordings the Baseline (70.18%-70.40%) is statistically indistinguishable from most classifiers. In two conditions it is even numerically ahead (RBF_SVC at QuantM30, $\Delta = -0.29$, and LIN_SVC_L1 at QuantM90, $\Delta = -0.31$), although these differences are far from significant. This pattern is consistent with how the Baseline was built. It relies on a single feature, L_{Aeq} , which the permutation importance of RF repeatedly ranked as the most influential. A plausible reading is that on Mobile recordings this one level-based descriptor already captures most of the information that the other quantitative features can add. The more flexible models then have little room to improve. On Zoom recordings, the Baseline's threshold may transfer less well, whereas models that combine several descriptors (e.g., spectral features) can compensate. The data do not isolate the cause, so this explanation remains a hypothesis. RF departs from the other classifiers. It is the only model whose

5.1 RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?

gains are significant on Mobile (M60 and M90) but largely not on Zoom (significant only at Z30). This resembles the device pattern of the Baseline and is plausible given that the Baseline was derived from RF’s feature importance. However, in the quantitative setting the baseline cannot be considered clearly insufficient. Its sufficiency depends on the device, and on Mobile recordings it is a competitive reference.

As noted above, another pattern emerges: unlike the other four machine learning models, RF, and consequently the Baseline, appears to perform better when built on Mobile recordings than on Zoom recordings. This can be explained by the fact that Mobiles have a built-in preprocessing system that already acts as a filter, removing small fluctuations and retaining only the larger ones. When classification is performed on such data, the decision threshold is somewhat higher, which allows the Mobile recordings to be classified well. The Zoom recorders, by contrast, capture the full signal variations without any filtering. Given the limited amount of data, the models struggle to learn a suitable decision boundary and therefore classify less accurately.

With the **hybrid datasets** the picture changes completely. All 30 comparisons are significant, with improvements between about 10 and 24 percentage point. The pattern observed for RF and the Baseline on the quantitative datasets persisted only for the Baseline when the hybrid datasets were used. Once the questionnaire features are available, the Baseline Model is clearly insufficient. A single acoustic threshold cannot exploit perceptual and contextual information, such as the type of environment or the perceived relevance of the sound context, which the machine learning models can use.

2. Users cross-validation

5.1 RQ1 - Is the Baseline Model Sufficient, or Do Machine Learning Models Improve Performance?

Moving on to Users cross-validation, where only models built on **hybrid datasets** are available, the answer to RQ1 is unambiguous. All 30 comparisons are significant after correction, with improvements from 11.29 (LOG_REG, HM90) to 25.91 percentage points (RBF_SVC, HZ30). The machine learning models reach about 88%-92% balanced accuracy, whereas the Baseline remains between 66.57% and 77.29%. As in the Audios scheme, the gains are largest on Zoom, where the Baseline is weakest, and smallest on longer Mobile windows, where it is strongest. The Baseline's weakness on Zoom recordings therefore does not seem to depend on how the data are split.

The machine learning models also reach their highest performance in this scheme. This is consistent with the scenario it represents: since recordings are shared between training and test sets, the models can exploit what they have learned about each soundscape from other users' evaluations, whereas the Baseline, which uses only a L_{Aeq} threshold, cannot. The results should therefore be read as an estimate of performance for new users on archived soundscapes, not for new soundscapes. In this archive-based scenario, the baseline is clearly insufficient.

3. Audios_X_Users cross-validation

Even in this cold-start scenario, all 30 comparisons are significant, with improvements from 9.95 to 19.99 percentage points and machine learning models at about 82%-87% balanced accuracy. The advantage over the Baseline therefore does not depend on prior knowledge of either the user or the recording.

The device pattern is again the same: gains are largest on Zoom (up to $\Delta = 19.99$ for the RBF_SVC at HZ30) and smallest on longer Mobile win-

5.2 RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?

dows. All comparisons relied on the paired t-test, so, unlike in the Users scheme, no comparison involved non-normal differences. The margins are smaller than under the Users scheme, and the smallest improvement of the whole study (LOG_REG at HM90, $\Delta = 9.95$) occurs here. This latter is compatible with the overlap hypothesis above, since removing shared recordings should reduce the achievable performance. It remains compatible rather than demonstrated.

5.2 RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?

Shifting our attention to RQ2, the tests needed to answer it could be performed only under the Audios scheme, since the Users and Audios_X_Users schemes split the data by participant, and the quantitative datasets do not contain user identity. The discussion therefore focuses on the Audios scheme and its usage scenario.

1. Audios cross-validation

The answer is clear for the machine learning models. All 30 comparisons were significant after correction, with gains between 5.85 and 18.70 percentage points. The hybrid machine learning models reached 80.29%-87.58% balanced accuracy, against 68.11%-77.66% for their quantitative counterparts. Adding the qualitative features therefore improves every classifier in every device and window condition. The Baseline Model is the only exception, with $\Delta = 0$ in all six conditions. This is true by construction, because the Baseline uses a L_{Aeq} threshold and cannot exploit the questionnaire. It is nonetheless informative. The gain appears only in models that can use the qualitative variables.

5.2 RQ2 - Does the Hybrid Setting Improve Performance Over Quantitative Alone?

The size of the gain differs across models, and the differences follow the quantitative performance of each model. RF shows the largest gains on Zoom (18.02-18.70 points), which is where its quantitative results are weakest (68.11%-69.56%). LIN_SVC_L1 shows the opposite pattern: its gains are larger on Mobile (12.74-16.01) than on Zoom (5.85-8.87), and on Zoom it already had its best quantitative results (up to 77.66%). The smallest gain among the LIN_SVC conditions (6.51 at Z60) also occurs where its quantitative score is highest (75.91%). A plausible reading is that the hybrid features act as a common ceiling, so the models with the least acoustic information to start from gain the most. This is supported by the hybrid scores, which are far less dispersed than the quantitative ones. It remains an interpretation, because the analysis does not separate the contribution of each qualitative variable.

The device gap also changes. In the quantitative setting, device strongly affects the results, and its direction depends on the model. In the hybrid setting, RF and the linear models give very similar results on Mobile and Zoom, whereas the RBF_SVC retains a modest advantage on Zoom over Mobile (about 3 points). The qualitative features therefore reduce the dependence on the recording device, which is relevant for an application used on smartphones. This is consistent with the reading proposed for RQ1, that the models built on quantitative datasets are limited mainly by how well the acoustic descriptors transfer across devices. For the returning-user scenario, the practical conclusion is that storing the user's previous evaluations is clearly worthwhile. The system can gain 6 to 19 points of balanced accuracy over a purely acoustic one, and the gain is largest where acoustic information alone performs worst.

5.3 RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent?

To answer the RQ3 a two-phase Friedman-Nemenyi posthoc was applied to the 36 model-condition combinations, so each significant difference mixes a model effect and a condition effect. The discussion therefore interprets patterns of combinations (for example, "RF at Z60") and not model rankings in the abstract. It is organised by usage scenario, as for RQ1 and RQ2.

1. Audios cross-validation

Using models built on **quantitative datasets**, the answer is that no model is consistently best, and that the structure of the differences is dictated by the device. The Baseline is separated from the machine learning combinations in only 77 of 180 comparisons. The three Mobile Baseline combinations are never significantly different from any RBF_SVC or RF combination. On Mobile recordings the five classifiers are also practically equivalent to each other (a single significant pair among the nine Mobile combinations). Together with RQ1, this indicates that with acoustic information alone, extracted on smartphone recordings, neither the choice of classifier nor the use of a classifier instead of a L_{Aeq} threshold has a reliable effect. This is the setting most relevant to an application running on smartphones.

On Zoom recordings, the picture changes in a specific way. The RBF_SVC, LIN_SVC, LOG_REG and LIN_SVC_L1 are mutually indistinguishable. All the significant Zoom-Zoom differences involve RF, whose Zoom combinations are statistically indistinguishable from the Zoom Baseline. This agrees with the RQ1 result that RF was significant only at Z30. It also fits the reading given for RQ1: RF was the model from which the Baseline was derived, and it shares its dependence on L_{Aeq} . The best combinations (LIN_SVC_L1

5.3 RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent?

at Z90, 77.66%; LOG_REG at Z30, 76.45%) are the ones that differ from the largest number of other combinations, but the same LIN_SVC_L1 is almost identical to the Baseline at M90 (70.10% vs. 70.40%). A model that is best in one condition and indistinguishable from the baseline in another cannot be called consistently superior.

Once the qualitative features are available and combined with the acoustic features (**hybrid features**), the Baseline is significantly outperformed by nearly every combination (176 of 180). The only exception is RBF_SVC at M30, which is not distinguishable from four Baseline combinations. This is consistent with its numerically close balanced accuracy to the Baseline in that condition in the quantitative setting, and it is the weakest hybrid RBF_SVC condition (80.29%). Differences among the classifiers are few and attached to specific combinations. RBF_SVC at M30 differs from some LOG_REG, LIN_SVC_L1 and RF combinations. RF at M30 and Z30 differ from RBF_SVC at M60 and M90 and from LIN_SVC at Z30 and Z60. RF has the highest balanced accuracy in most hybrid conditions (up to 87.58%), but it is not significantly better than most other classifiers. Its numerical lead over the others is therefore not enough to claim consistent superiority. For the returning-user scenario, the practical consequence is that, when the stored evaluations are available, the classifier can be chosen on secondary criteria, such as interpretability, training cost or robustness, without a detectable loss of performance.

2. Users cross-validation

In this scheme, the Baseline is separated from all machine learning combinations (180 of 180, with no exception). This fits the reading given for RQ1: all classifiers outperformed the Baseline. Among the classifiers, only

5.3 RQ3 - Is There a Consistently Best ML Model, or Are They Statistically Equivalent?

27 of the 360 cross-model comparisons are significant, and they are concentrated on four combinations: RBF_SVC and RF at Z30 and Z60. These four are the best-performing combinations (91.92%-92.48%) and they differ from the lowest-scoring linear combinations (88.08%-88.91%), mostly LOG_REG and LIN_SVC_L1 on Mobile and Zoom. RBF_SVC and RF are never significantly different from each other, and neither differs significantly from its own other conditions.

Two readings are possible. The first is that a modest non-linear advantage exists in this scenario. It would be consistent with the fact that the same recordings appear in training and test, so flexible models can exploit what they have learned about each soundscape from other users. The second is that the significant pairs reflect only the spread of balanced accuracy values, since the matrix compares combinations across devices and windows. The present data do not separate the two. The conservative conclusion is that the difference is small (about 3-4 points), confined to specific Zoom conditions, and not general. For the archive scenario, a non-linear model is a reasonable default. The system would not lose much with a linear one.

3. Audio & User cross-validation

We conclude this section answering the RQ also in the Audio & User cross-validation.

Here again, the Baseline is fully separated from all machine learning combinations (180 of 180). The largest p-value (0.016) involves Baseline at M90 and the lowest-scoring classifier combination (LOG_REG at M30, 82.67%). Among the classifiers, no comparison is significant, either between models or between conditions of the same model. Under the most demanding scenario, the five classifiers are therefore statistically equivalent in every de-

5.4 RQ4 - Do Device and Window Length Affect Performance?

vice and window condition. RF at M60 has the highest balanced accuracy (86.77%), but this cannot be distinguished from the others. This result needs to be read carefully. Not finding a significant difference is not proof of equivalence, especially in a test with 36 groups, where the Nemenyi test has limited power. The correct statement is that the data give no evidence of a difference. Even so, the numerical spread among the classifiers is small (roughly 82.7%-86.8%), and much smaller than the gap to the Baseline (at most 74.3%). For a deployed application that must serve new users on new soundscapes, this suggests that the Baseline is the worst choice among the six models, which confirms the answer given in RQ1, and that the choice of the classifier is not the critical design decision.

5.4 RQ4 - Do Device and Window Length Affect Performance?

The answer to this question was obtained by first comparing the devices and then comparing the window lengths, in order to assess the effect of each on performance. The discussion is organised by usage scenario, as for RQ1-RQ3.

5.4.1 Device effect (mobile vs zoom)

1. Audios cross-validation

Quantitative datasets: with acoustic information alone, the device has a significant effect for every model, but its direction depends on the model. The Baseline performs better on Mobile recordings ($\Delta = 5.59$), and so does RF ($\Delta = 4.35$). The other four classifiers perform significantly better on Zoom, by 3.10 to 5.27 points. These results are consistent with those for RQ1 and RQ3. In both, RF behaved like the Baseline, with an advantage

5.4 RQ4 - Do Device and Window Length Affect Performance?

on Mobile and no clear separation from the Baseline on Zoom. This supports the earlier reading that RF and the Baseline share their dependence on L_{Aeq} . The classifiers that favor Zoom combine several descriptors. A plausible reading is that the digital recorder provides a cleaner signal for spectral and timbral features, which these models can exploit. On smartphone recordings, level information is relatively more useful.

The practical implication is mixed. In a smartphone application, which is the intended setting for this scenario, using only acoustic features offers no robust advantage from a more flexible model. Most classifiers would work better on the better recorder than on the device the system targets.

Hybrid datasets: The device effect is much weaker once the qualitative features are included. Only the Baseline, which cannot use them, keeps its Mobile advantage unchanged ($\Delta = 5.59$). Among the classifiers, LIN_SVC, LOG_REG and RF show no significant difference, and two small effects remain in opposite directions: the RBF_SVC favors Zoom ($\Delta = -3.29$) and LIN_SVC_L1 favours Mobile ($\Delta = 2.13$). Because the two effects have opposite signs, there is no evidence of a general device effect for the machine learning models. Both should be treated with caution, since they are the smallest significant differences in the analysis and each is one of six comparisons. The main result is the disappearance of the quantitative device gaps, in particular the Mobile advantage of RF. This suggests that the qualitative information compensates for the differences between devices. For the returning-user scenario, a smartphone user would not be penalised by the device, provided the stored evaluations are available.

2. Users cross-validation

5.4 RQ4 - Do Device and Window Length Affect Performance?

The Baseline is again the only model with a significant device difference, in favor of Mobile, and the gap ($\Delta = 7.64$) is the largest in all of RQ4-1. No machine learning model shows a significant effect, and the differences are small (0.41 to 1.02 points). RF, whose Zoom advantage of 1.02 points has a raw p-value of 0.031 that does not survive the correction, is the closest to significance. This should not be read as an effect, since it is not significant after correction and the sign is the opposite of its quantitative result. The larger Baseline gap is probably linked to the different composition of the splits, in which all recordings can appear in training. However, the Baseline does not use that information, so its sensitivity to the device is a property of the L_{Aeq} threshold. In the archive scenario, the recording device is therefore not a relevant design constraint for the classifiers. This matters because the archive’s acoustic features were computed beforehand, and recordings from both devices can be stored without a measurable change in performance.

3. Audio & User cross-validation

Here, the pattern is the same as under the Users scheme. The Baseline has a significant Mobile advantage ($\Delta = 6.61$), and none of the five classifiers shows a significant difference (0.23 to 1.69 points). The RBF_SVC is the closest to significance ($\Delta = -1.29$, corrected $p = 0.054$), with a Zoom advantage that may be real but cannot be confirmed here. Even in the most demanding scenario, the machine learning models do not depend measurably on the device, which is the most useful result for an application used by new people on their own phones. As in RQ3, a non-significant result is not proof of equivalence: with 50 splits and a Benjamini-Hochberg correction over six tests, small differences of one to two points could remain undetected. The conservative conclusion is that any device effect for the classifiers is at most small.

5.4.2 Window length effect (mobile vs zoom)

1. Audios cross-validation

With both **quantitative** and **hybrid features**, the window length has no significant effect on any of the five machine learning models. The differences are small (at most about 1.7 points for quantitative features and 1.3 points for hybrid features), and their direction is not consistent: LOG_REG tends to improve with longer windows (84.39% to 85.29%, Friedman $p = 0.080$), while RF tends to do the opposite (87.49% at 30 seconds, 86.78% at 90 seconds). As for RQ3 and RQ4-1, a non-significant result is not proof of no effect. The conservative conclusion is that, in this scenario, any effect of the window length on the classifiers is small.

The Baseline is the only model affected. Its balanced accuracy increases from 30 to 60 seconds and then levels off, with no difference between 60 and 90 seconds. A plausible reading is that L_{Aeq} , averaged over a longer period, is a more stable estimate of the sound level of the place, and that the stabilization is already reached at about 60 seconds. The multi-feature classifiers are less affected, possibly because they have more descriptors to compensate for a noisier level estimate. This remains a hypothesis. The practical implication is favorable for a smartphone application: a 30-second recording is sufficient for the classifiers, which allows faster feedback to the user. The longer window is useful only for a threshold-based approach.

2. Users cross-validation

The Baseline benefits from longer windows more clearly than under the Audios scheme, with all three pairwise differences significant (a gain of 2.93 points from 30 to 90 seconds). The gain does not stop at 60 seconds, although the step from 60 to 90 seconds is the weakest. The sensitivity of

5.4 RQ4 - Do Device and Window Length Affect Performance?

the Baseline to the window length is therefore consistent across schemes, which supports the interpretation that it comes from the stability of the L_{Aeq} estimate, and not from the data split. Among the classifiers, only LIN_SVC_L1 shows a significant effect ($p = 0.0022$), in favor of the 90-second window by about 1.2 to 1.4 points. It seems that LIN_SVC_L1 could provide more stable values on longer windows. The RBF_SVC and RF even show the opposite tendency, with their best values at 30 seconds, and neither is significant. For the archive scenario, the choice of window length is not a critical parameter for the classifiers, although the longest window is not harmful.

3. Audio & User cross-validation

The Baseline is again significantly affected, with the three pairwise differences significant and a gain of 2.21 points between 30 and 90 seconds. Two classifiers are also significant: LIN_SVC_L1, with the same direction as under the Users scheme (the 90-second window is better than the 30-second one, by about 1.2 points), and RF, whose best window is 60 seconds. The RF effect is small (about 0.8 points between 60 and 90 seconds), borderline ($p = 0.042$), and not a gain from longer windows. The other three classifiers show no significant effect, with a slight tendency to improve with longer windows. The increase in the number of significant classifiers across the schemes (none, one, two) should not be over-interpreted. The effects are small and not consistent across models. A possible reading is that, with new users and new recordings, the models are less able to compensate for the variability of the shorter windows, but the data do not support more than a suggestion. In the cold-start scenario the practical consequence is modest: windows of 60 to 90 seconds may give a slight advantage to some models, but the 30-second window loses at most about one point.

Chapter 6

Conclusions

Although technology has evolved continuously over the last decade, only a few researchers and practitioners are exploiting it to build automatic classification systems for the soundscapes in which we are immersed every day and at every moment. Such systems could nonetheless serve as a guide: for individual users, to find healthier and more peaceful environments, and for urban planners, to design less noisy ones. Existing research on this task, however, leaves three gaps open. The first concerns the contribution of perceptual information when it is combined with signal-derived features: the two research directions have largely been developed in isolation, and only [19; 22] have partially addressed the question, even then separately from the classification model itself. The second, and the least examined, concerns how window length affects classification performance on longer, real-world recordings: only [20] investigated it, and within a narrow range (from 4 seconds down to 125 milliseconds). The third concerns model selection: apart from [17; 21; 22], who each compare a few algorithms, the reviewed studies generally rely on a single classifier, which gives little insight into how the choice of model interacts with feature type or window length.

This thesis set out to address these gaps by asking whether urban soundscapes can be classified as noisy or quiet and, above all, whether this task requires

machine learning or can be handled by a simple acoustic rule. To this end, we compared a Baseline, based on the equivalent sound level (L_{Aeq}) identified through the permutation importance of a Random Forest, with five classifiers: RBF_SVC, LIN_SVC, LOG_REG, LIN_SVC_L1, and RF. The comparison was carried out on recordings from two devices (Mobile and Zoom), at three window lengths (30, 60 and 90 seconds), using either acoustic features alone (Quantitative) or acoustic features combined with questionnaire responses (Hybrid). Three cross-validation schemes (Audios, Users, and Audio & User) allowed us to evaluate the models in three realistic usage scenarios of a soundscape recommendation system: a returning user capturing a new soundscape, a new user facing an archived soundscape, and a new user facing a new soundscape.

The first result concerns the sufficiency of the Baseline. When only acoustic information is available, a simple threshold on L_{Aeq} proves to be a surprisingly strong competitor, particularly on Mobile recordings, where it is statistically indistinguishable from most classifiers; only on the Zoom recordings is it clearly outperformed. The picture changes completely once questionnaire information is added: in every condition of every scheme, including the most demanding cold-start scenario, all five classifiers significantly outperform the Baseline.

This difference speaks directly to the first gap and points to the central finding of the work: the decisive factor is the information supplied to the models, rather than the models themselves. Adding the qualitative features improved every classifier in every condition, by between 5.85 and 18.70 points of balanced accuracy, while the Baseline, which cannot exploit them, remained unchanged. Consistently with this, and in answer to the third gap, the choice of algorithm had a much smaller effect. The five classifiers were largely statistically indistinguishable, with the few significant differences tied to specific combinations of device and window length, and with none at all in the most demanding scheme.

RF and the RBF_SVC were numerically the strongest, but never by a margin that held across conditions.

As for the second gap, the analysis of device and window length showed that these factors matter mainly for the Baseline, which performs better on Mobile recordings and benefits from longer windows. With hybrid features, the classifiers were largely insensitive to both, apart from small and inconsistent effects. In practical terms, this supports the use of short recordings captured with smartphones.

These conclusions must be read in light of several limitations. The dataset is small, which explain our choice for a classical machine learning models instead of deep learning ones and also explains why the multiclass formulation performed poorly and why the work was restricted to binary classification. The test of the contribution of qualitative features (RQ2) was possible only under the Audios scheme, since the quantitative datasets contain no user identity. In the analysis of device and window length a non-significant result should be interpreted as absence of evidence, not as evidence of equivalence. Moreover, some of the explanations proposed, such as the role of L_{Aeq} across devices or the effect of smartphone pre-processing, are plausible hypotheses that the data do not test.

Future work should therefore start from a larger and more varied dataset, which would also make multiclass classification feasible. It should also identify which qualitative variables drive the improvement, since this determines what a new user would have to provide in a real application. Finally, a prospective evaluation of the system in the field would show whether these results hold outside the experimental setting.

References

- [1] K. M. De Paiva Vianna, M. R. Alves Cardoso, and R. M. Rodrigues, “Noise pollution and annoyance: An urban soundscapes study,” *Noise & Health* 17,76(2015), p. 125–133, 2015. 1
- [2] K. Sun, B. De Coensel, K. Filipan, F. Aletta, T. Van Renterghem, T. De Pessemier, J. Wout, and D. Botteldooren, “Classification of soundscapes of urban public open spaces,” *Landscape and Urban Planning* 189(2019), pp. 139–155, 2019. 1
- [3] M. S. Hammer, T. K. Swinburn, and R. L. Neitzel, “Environmental Noise Pollution in the United States: Developing an Effective Public Health Response,” *Environmental Health Perspectives* 122, 2(2014), p. 115–119, 2014. 1
- [4] E. Ratcliffe, “Sound and soundscape in restorative natural environments: A narrative literature review,” *Frontiers in Psychology* 12(2021), 570563, 2021. 1
- [5] E. Ratcliffe, B. Gatersleben, and P. T. Sowden, “Bird sounds and their contributions to perceived attention restoration and stress recovery,” *Journal of Environmental Psychology* 36(2013), p. 221–228, 2013. 1
- [6] J. J. Alvarsson, S. Wiens, and M. E. Nilsson, “Stress recovery during ex-

REFERENCES

- posure to nature sound and environmental noise,” *International Journal of Environmental Research and Public Health* 7,3(2010), p. 1036–1046, 2010. 1
- [7] M. G. Berman, J. Jonides, and S. Kaplan, “The cognitive benefits of interacting with nature,” *Psychological Science* 19,12(2008), p. 1207–1212, 2008. 1
- [8] S. C. Van Hedger, H. C. Nusbaum, L. Clohisy, S. M. Jaeggi, M. Buschkuehl, and M. G. Berman, “Of cricket chirps and car horns: The effect of nature sounds on cognitive performance,” *Psychonomic Bulletin & Review* 26,2(2019), p. 522–530, 2019. 1
- [9] C. Twohig-Bennett and A. Jones, “The health benefits of the great outdoors: A systematic review and meta-analysis of green space exposure and health outcomes,” *Environmental Research* 166(2018), p. 628–637, 2018. 1
- [10] M. P. White, I. Alcock, J. Grellier, B. W. Wheeler, T. Hartig, S. L. Warber, A. Bone, M. H. Depledge, and L. E. Fleming, “Spending at least 120 minutes a week in nature is associated with good health and wellbeing,” *Scientific Reports* 9, 1(2019), p. 1–11, 2019. 1
- [11] Y. Zhang, S. Mavoa, J. Zhao, D. Raphael, and M. Smith, “he association between green space and adolescents’ mental well-being: A systematic review,” *International Journal of Environmental Research and Public Health* 17,18(2020),6642, 2020. 1
- [12] V. Bountourakis, L. Vrysis, and G. Papanikolaou, “Machine learning algorithms for environmental sound recognition: Towards soundscape semantics,” *In Proceedings of the Audio Mostly 2015 on Interaction With Sound (Thessaloniki, Greece)(AM’15)*, 2015. Association for Computing Machinery, NewYork, NY, USA, Article5, 7pages. 2

REFERENCES

- [13] S. Chachada and C.-C. JayKuo, “Environmental sound recognition: a survey,” *APSIPA Transactions on Signal and Information Processing* 3(2014), e14, 2014. 2
- [14] S. Oramas, V. C. Ostuni, T. Di Noia, X. Serra, and E. Di Sciascio, “Sound and music recommendation with knowledge graphs,” *ACM Trans. Intell. Syst. Technol.* 8, 2, Article 21(Oct.2016), 21pages, 2016. 2
- [15] Z. Zeng, Y. Wang, Y. Zhao, and W. Shi, “A survey of music recommendation systems,” *In Proceedings of the 5th International Conference on Computer Information and Big Data Applications(Wuhan,China)(CIBDA’24)*, p. 507–519, 2024. Association for Computing Machinery, New York, NY, USA. 2
- [16] F. Viola, A. Stolfi, A. Milo, M. Ceriani, M. Barthet, and G. Fazekas, “Playsound.space: enhancing a live music performance tool with semantic recommendations (saam’18),” p. 46–53, 2018. Association for Computing Machinery, NewYork, NY, USA. 2
- [17] K. M Bellisario, T. Broadhead, D. Savage, Z. Zhao, H. Omrani, S. Zhang, J. Springer, and B. C Pijanowski, “Contributions of mir to soundscape ecology. part3: Tagging and classifying audio features using a multi-labeling k-nearest neighbor approach,” *Ecological Informatics* 51(2019), p. 103–111, 2010. 2, 7, 14, 94
- [18] I. McGregor, G. Leplâtre, A. Crerar, and D. Benyon, “Sound and soundscape classification: Establishing key auditory dimensions and their relative importance,” 2006. Georgia Institute of Technology. 2, 5, 8, 14
- [19] C. A. Quinn, P. Burns, G. Gill, S. Baligar, R. L. Snyder, L. Salas, S. J. Goetz, and M. L. Clark, “Soundscape classification with convolutional neural

REFERENCES

- networks reveals temporal and geographic patterns in ecoacoustic data,” *Ecological Indicators* 138(2022), 108831, 2022. 2, 5, 7, 8, 94
- [20] M. Thorogood, J. Fan, and P. Pasquier, “Bf-classifier: Background/foreground classification and segmentation of soundscape recordings,” *In Proceedings of the Audio Mostly 2015 on Interaction With Sound*, p. 1–6, 2015. 2, 6, 14, 94
- [21] O. Kafui Toffa and M. Mignotte, “Environmental sound classification using local binary pattern and audio features collaboration,” *IEEE Transactions on Multimedia* 23(2020), p. 3978–3985, 2020. 2, 7, 14, 94
- [22] A. J. Torija, D. P. Ruiz, and A. F. Ramos-Ridao, “A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model,” *Science of the Total Environment* 482-483(2014), pp. 440–451, 2013. 2, 6, 9, 94
- [23] A. J. Torija, D. P. Ruiz, and Ramos-Ridao, “Application of a methodology for categorizing and differentiating urban soundscapes using acoustical descriptors and semantic-differential attributes,” *J. Acoust. Soc. Am.* 134, 791–802 (2013), p. 791–802, 2013. 6
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 16
- [25] S. Seabold and J. Perktold, “statsmodels: Econometric and statistical modeling with python,” in *9th Python in Science Conference*, 2010. 29