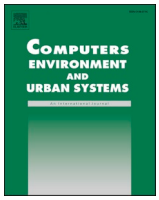
Computers, Environment and Urban Systems 110 (2024) 102112

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Exploring emergent soundscape profiles from crowdsourced audio data

Aura Kaarivuo [a,b,*], Jonas Oppenländer [c], Tommi Kärkkäinen [a], Tommi Mikkonen [a]

[a] Faculty of Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), Jyväskylän yliopisto FIN-40014, Finland
[b] School of Media, Design and Conservation, Metropolia University of Applied Sciences, P.O. Box 4072, Metropolia FIN-00079, Finland
[c] Elisa, Ratavartijankatu 5, Helsinki FIN-00520, Finland

ARTICLE INFO

Keywords:
Soundscapes
Mobile crowdsensing
Machine learning
Emotional information
Perception
Urban planning

ABSTRACT

The key component of designing sustainable, enriching, and inclusive cities is public participation. The soundscape is an integral part of an immersive environment in cities, and it should be considered as a resource that creates the acoustic image for an urban environment. For urban planning professionals, this requires an understanding of the constituents of citizens' emergent soundscape experience. The goal of this study is to present a systematic method for analyzing crowdsensed soundscape data with unsupervised machine learning methods. This study applies a crowdsensed sound- scape experience data collection method with low threshold for participation. The aim is to analyze the data using unsupervised machine learning methods to give insights into soundscape perception and quality.

For this purpose, qualitative and raw audio data were collected from 111 participants in Helsinki, Finland, and then clustered and further analyzed. We conclude that a machine learning analysis combined with accessible, mobile crowdsensing methods enable results that can be applied to track hidden experiential phenomena in the urban soundscape.

1. Introduction

Citizens' experience of the surrounding *soundscape* in rapidly growing, increasingly populous cities is strongly connected to well-being, comfort, and contentment (Kang, 2023; van Kamp, Leidelmeijer, Marsman, & de Hollander, 2003). Characterizing soundscapes of urban areas and defining when they are pleasing to the public has been a long-term goal of many soundscape research projects (Gontier, Aumond, Lagrande, Lavandier, & Petiot, 2018; Kang, 2023; Raimbault & Dubois, 2005; Xiao, Lavia, & Kang, 2018). The project of understanding and developing the quality of a soundscape dates back to R. Murray Schafer's "World Soundscape Project" in the 1960s (Schafer, 1977). In this international multidisciplinary project, Schafer aimed to find a sound- scape in which human society and the acoustic environment were in balance (Schafer, 1977). The term *acoustic environment* refers to the combination of sounds of a place or space that are modified by the environment (ISO, 2014) and can be heard (Brown, Gjestland, & Dubois, 2015). *A soundscape* is a person's perceptual concept (ISO, 2014) of the acoustic environment in question (Brown et al., 2015).

Urban soundscapes are living, multi-layered, and composed of an ongoing flow of events, (Arkette, 2004). As many of previous studies have concluded, it is difficult and highly problematic to describe the experience of a sound- scape using single words such as "eventful" or "pleasant" (Aletta et al., 2020; Axelsson, Guastavino, & Payne, 2019; Kang, 2023). This is due to the nature of sound and human perception. Sound is time bound and variable, and its percep- tion is dependent on individual and context-related judgment (Raimbault & Dubois, 2005; Schafer, 1977). Momentary changes in the soundscape can drastically change evaluation of it (Axelsson et al., 2019). It is also known that hedonistic judgment affects this evaluation and that individ- ual assessment is often based on semantic evaluation rather than solely on the perception of sound (Dubois, Guastavino, & Raimbault, 2006; Niessen, Cance, & Dubois, 2010). There- fore, due to the individual nature of the auditory perception, one person can evaluate the same sound sources differently than another (Guastavino, 2007; Mitchell, Aletta, & Kang, 2022). Citizen's needs, context, perceptions and experiences affect their evaluation of the soundscape (Yan, Meng, Yang, & Li, 2024). Soundscape experience is also affected by other sensations (smells, visuals, etc.), and the reporting of different perceptions might be conflated (Calleri et al., 2019; Engel, Paas, Schneider, Pfaffenbach, & Fels, 2018; Shao, Hao, Yin, Meng, & Xue, 2022; Wang, Zhang, Xie, Yang, & He, 2022). Several related studies have suggested that there should be

* Corresponding author at: Faculty of Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), Jyväskylän yliopisto FIN-40014, Finland.
E-mail address: aura.kaarivuo@metropolia.fi (A. Kaarivuo).

https://doi.org/10.1016/j.compenvurbsys.2024.102112
Received 12 September 2023; Received in revised form 24 March 2024; Accepted 25 March 2024
Available online 8 April 2024

more international and interdisciplinary collaboration in sound- scape research, as well as the development of new tools and methodological approaches, as traditional approaches and tools are not sufficient to holisti- cally represent and evaluate soundscapes (Axelsson et al., 2019; Aletta et al., 2020; Mitchell et al., 2022, Song, Meng, Kang, Yang, & Li, 2023). In particular, it would be necessary to develop collection methods and indicators that assess the health-related quality of soundscapes (Kang, 2023).

According to Potts (2020), the design of cities and their soundscapes should involve a wide range of interest groups and create consensus through social interaction, facilitated by city planners. In recent years, more inter- action, knowledge sharing, and debate between decision makers and inter- ested parties has resulted from the development of communication and mobile technologies (Potts, 2020). *Crowdsourcing* is an example of a method of fa- cilitating participation and democratic decision-making with a large group of dispersed people via the Internet (Brabham, 2013). Crowdsourcing pro- vides a means to gather extensive amounts of situated intelligence (Brabham, 2013) using smart and efficient methods (Liao et al., 2019). Urban crowdsourcing (Steils, Hanine, Rochdane, & Hamdani, 2021) can be used to inform the design of smart cities, using participatory design approaches (Mueller, Lu, Chirkin, Klein, & Schmitt, 2018) instead of traditional, expensive, labor-intensive methods, such as questionnaires or public hearings (Liao et al., 2019). Situated crowdsourcing has an enormous potential in soundscape design, as it allows participants to provide both qual- itative and quantitative information in-situ, in everyday life situations, and in larger groups (Craig, Moore, & Knox, 2017). *Crowdsensing* here refers to collabo- rations with citizens in which both people and their mobile devices act as sensors (Brambilla & Pedrielli, 2020; Cardone et al., 2013; Lefevre, Agarwal, Issarny, & Mallet, 2021). Crowdsensing (or mobile crowdsensing) has been utilized especially for noise monitoring and mapping soundscape quality (Craig et al., 2017; Li, Liu, & Haklay, 2018; Orio, De Carolis, & Liotard, 2021).

Crowdsensed data can provide more diverse information for soundscape research (Brambilla & Pedrielli, 2020; Gontier et al., 2018; Nieto-Mora, Rodríguez-Buritica, Rodríguez-Marín, Martínez- Vargaz, & Isaza-Narváez, 2023; Zappatore, Longo, & Bochicchio, 2017). According to recent studies, the most common analysis methods consist of manual labeling of data by listening to record- ings or visually inspecting spectrograms, summarizing variations in acoustic energy, or automatically recognizing sound sources or insides using machine learning algorithms (Benocci, Afify, Potenza, Roman, & Zambon, 2023; Nieto-Mora et al., 2023). However, big audio data cannot be manually labeled and analyzed, due to its time- consuming nature (Benocci et al., 2023; Nieto-Mora et al., 2023). Automatic recognition of acoustic insides and sound sources is sensitive to noise and the sound sources may vary depending on the specific environment being studied. Machine learning methods have been used to identify geographic patterns (Quinn et al., 2022), to evaluate urban spaces (Yu & Kang, 2009), and to classify species and other acoustic features (Dias, Ponti, & Minghim, 2022). Both supervised and unsupervised techniques have offered promising results, but again supervised machine learning is labor intensive and time consuming (Nieto-Mora et al., 2023).

The goal of this paper is to present a systematic method for analyzing crowdsensed soundscape data with unsupervised machine learning methods. We will apply unsupervised machine learning methods to the results of man- ual qualitative data analysis of soundscapes, and observe the resulting clus- ters to obtain information about the perceived quality of the soundscape.

These aims are addressed through the following research questions:

RQ1. How can crowdsensed soundscape data be analyzed using unsupervised machine learning methods?

RQ2. What kind of soundscape profiles emerge from the analysis and how could their interpretation be linked to improve our understanding of urban soundscape experiences?

The rest of this paper is structured as follows. We will present an analy- sis that employs manual labeling, qualitative analysis, and machine learning methods (see Fig. 1) for soundscape data which is collected with participa- tory crowdsensing method. We use methodological triangulation to augment the findings of different analysis methods (Denzin, 1970). First, in Section 2, we describe the data collection, manual labeling and automated analysis of soundscape data, which was based on a combination of unsupervised ma- chine learning and feature selection methods and the results of the qualitative analysis. Second, in Section 3, we provide details of the identified clusters and analyze the groups and profiles of the soundscape experience from the crowdsensed audio data and manual qualitative analysis. We compare the results of the manual qualitative analysis with the results of the unsupervised machine learning approach and, finally, present the general characterization of the emerging soundscape experience. In Section 4, we discuss the inter- pretation and key findings of the research. Finally, in Section 5, we draw conclusions and suggest implications and ideas for future work.

2. Material and methods

Various research and methodological approaches, solutions, and frame- works for soundscape data collection and analysis have been presented over the past five decades at an accelerating pace (Aletta, Kang, & Axelsson, 2016; Guastavino, 2007; Jiang et al., 2022; Kang, 2010, 2023; Kang & Aletta, 2018; Schafer, 1977). The current standardized method is presented in ISO standard 12,913 parts 1–3, which contain a definition of soundscape and a conceptual frame- work, data collection, reporting, and analysis requirements (ISO, 2014, 2018, 2019) for research. According to this ISO standard, a soundscape study should be holistic and contain several investigative methods to ensure that the study considers different viewpoints, such as the human perception, the acoustic environment, and the context in question. The standard does not give a single answer or a clear research approach but recommends a collection of methods because a consensus could not be reached regarding a protocol (Mitchell et al., 2022). Qualitative data analysis is recommended to be done with a chosen coding method to generalize the observations. Quantitative analysis is recommended but is considered less important, especially in cases of qualitative and explorative methods. The analysis of responses about the perceived quality of a soundscape is presented in the following dimension (ISO, 2019):

- pleasant – unpleasant
- calm – chaotic
- vibrant – monotonous
- eventful – uneventful

The following data collection and analysis method loosely follows the ISO standards. With the ISO standard, the fundamental question is that the definitions for dimensions are presented in English, and as Aletta et al. (2020) state in their article, sounds are described in a different way in different languages (Guastavino, 2007). According to Axelsson et al. (2019) context and person-related factors create great variance, which leads to difficulties in interpretation of the results. These and other limitations and perspectives of the critique toward the ISO standards (Aletta et al., 2020; Jo, Seo, & Jeon, 2020; Mitchell et al., 2022) were considered when designing this method. According to the ISO standard the choice of indicators depends on the people, acoustic environment and context.

The data set contained 111 one-minute-long raw audio files and question- naire answers related to them. The data collection method used here follows a method developed and tested by Kaarivuo, Salo, and Mikkonen (2021). The aim of this method was to develop an accessible, mobile, and participatory method that would produce live recordings of a soundscape in addition to traditional written descriptions and questionnaires. The purpose of this approach was to observe emerging pleasant soundscapes that citizens pass through in their everyday lives.

Fig. 1. Data collection and analysis method.

According to the evaluation of technical procedure and the functionality of the mobile data collection method it seemed that the recording with mobile tools and sharing the audio is easy and does not require any specific applications or even technical instructions. The evaluation also showed that this particular method identifies pleasant and easily accessible places in the city in which the participants enjoy in their surroundings. The study concluded that with this method it would be possible to collect training data for machine learning. (Kaarivuo et al., 2021).

2.1. Participants, context, and data collection

The research participants were first-year university media production stu- dents at a university of applied sciences located in Helsinki. The experiential soundscape data was collected in three workshops in August and September of 2020–2022 in the greater Helsinki area, which is the home environment for the participants. There were 111 participants in total, 35 to 38 participants per year. Most of the participants (68.5%) were 18 to 25 years old, 28.8% 26 to 35 years old, and 1.8% 36 to 45 years old.

The motivation of the media students to complete the assignment was most likely higher than average due to their motivation and interest in audio and sound design, but their technical competencies or listening and analyzing skills at the beginning of the studies were quite diverse. Most of the students were not familiar with soundscapes, urban planning, or analytical listening.

Participants received a short introduction lesson about the surrounding acoustic environment and a listening and soundscape recording assignment. To strengthen the engagement of the participants, the assignment was designed so that it connected to the participants' personal experiences about the urban soundscape (Neuvonen, 2019). In the assignment, the participants were asked to choose a location in the city in which they found the sound- scape pleasant and comfortable. They were asked to focus and listen to the soundscape for 20 min and record it using any kind of recording device and application for one minute. Next, they were asked to share the recording via an online form and answer questions concerning the soundscape. The questions in the online form were as follows:

- What is the name of the location?
- List the sounds you heard.
- What sounds would you add to the soundscape to make it more pleasant?
- What sounds would you remove or reduce?
- In your own words, describe how the soundscape feels and sounds and justify why. What in the soundscape evokes these feelings?

The online questionnaire was designed to be a combination of a questionnaire and an interview, both of which are mentioned as data collection methods in the ISO 12913-2 standard (ISO, 2018). As the participants were not describing the same locations, it was necessary to collect more detailed information about the soundscape, such as sounds

heard in-situ. The list of sounds provided a reference point for comparing the recordings, and the question about emotion provided information about the emotions and fea- tures experiences, such as pleasantness, calmness, vibrancy, eventfulness, and loudness.

As our approach aimed to lower the threshold of participation, self-reporting was made easy. We aimed to design the questions so that they were easy to answer and would produce detailed data about the physical and psycho-acoustic features of the soundscape. The idea was to lead the par- ticipator to first observe their surrounding soundscape in a focused manner, to recognize the elements in the soundscape, and then to create associations between emerging emotions and sounds and feelings. The aim was to create a procedure that can be repeated with any group of people, regardless of their age, education, prior knowledge, or sonological competence.

2.2. Manual qualitative analysis method

The self-reported emotional perceptions of the participants and locations of the recorded soundscapes were manually coded and labeled, drawing on categorizations from the related literature.

The emotional answers were coded under naturally emerging categories, following a grounded theory approach (Glaser & Strauss, 1967), rather than strictly applying the ISO standard labels. The qualitative analysis of the questionnaire was conducted in the following steps:

Step 1. Open coding: recognizing key terms concerning the emotions associated with sounds.

Step 2. Eliminating unnecessary and irrelevant information that is not directly related to the soundscape in question.

Step 3. Identifying repeated words and expressions.

Step 4. Identifying concepts: comparing the emerging terms and expressions to the ISO 12913-3:2019 standard for perceived soundscape affective quality.

Step 5. Generating categories: grouping similar expressions and concepts.

Step 6. Generating subcategories: modifying the chosen framework to illustrate the emerging phenomena.

Step 7. Drawing conclusions from the results.

The testing of the manual analysis indicated that the labeling of freely written Finnish answers with the original ISO standard English dimensions is problematic. The free-form lyrics did not distinguish between, e.g., vibrancy and eventfulness because there is no Finnish translation which would trans- late similarly. Also, the clustering of a small sample requires that the number of evaluation axes is reasonably small. We decided to test the analysis methods on the basis of what emerges from the data. Therefore, the dimensions were narrowed down to three:

pleasant – unpleasant calm – chaotic vibrant – monotonous

According to the reported locations of the recordings, we identified the recording locations and categorized them. The seven identified location cat- egories are as follows:

1. Sports/activity,
2. Street,
3. Social activity,
4. Neighborhood,
5. Station,
6. Park,
7. Miscellaneous.

2.3. Manual qualitative analysis results

In all three rounds, the participants chose locations mainly in the Helsinki metropolitan area in Finland. It seems that the selected locations are close to the places where the students live, commute between home and university, or spend their free time.[1]

The participants recorded mainly street locations, such as bus stops and other places where it is convenient to stay for a while to listen. About one quarter of the participants (23%) selected a park to represent a comfortable soundscape. Residential areas, train and metro stations, sports venues, and cafe terraces were mentioned <10 times each. The miscellaneous category contained recordings that did not meet the requirements of the assignments, and were recorded in indoor spaces such as shopping centers, vehicles, and indoor metro stations. The distribution of the created categories is presented in Fig. 2.

The answers to the question *"In your own words, describe how the sound- scape feels and sounds and justify why. What in the soundscape evokes these feelings?"* produced a variety of thoughts and opinions about the soundscape and the participants' memories, associative thoughts, and emotions and rela- tion toward the sounds and the place. It is well known that people describe their experience of an environment affectively (ISO, 2019). However, the an- swers contained expressions of the pleasantness, calmness, and vibrancy of the places in question, or the opposite.

The pleasant and unpleasant soundscapes were described, for exam- ple, as "homelike," "safe," "cozy," "comfortable," or with words like "gloomy," "restless," "disturbing," and "inharmonious." As the precondition of the task was to go to a place where the soundscape was comfortable, 77.5% of the soundscapes were labeled as pleasant and 22.5% unpleasant.

The calmness and chaos of the places could also be characterized as quiet and loud. As the task concerned urban environments, the word "quietness" did not appear in the answers. These impressions were expressed with words such as "relaxing," "carefree," "serene" or "smooth" and "noisy," "hectic," "busy," or "stentorian." The distribution was fairly even, with 51.4% of the soundscapes being described as calm.

The vibrancy and monotony of the soundscapes were expressed within various contexts. A monotonous soundscape was a place in which partici- pants could pick up "quiet sounds" and "be with your own thoughts" and a vibrant one was "multi-layered" or "eventful" with "continuous stimuli." A soundscape was "morning-like, with only small sounds" or "ordinary and bor- ing." In contrast the soundscapes were "speedy" and had "sounds of life" and "there [was] a lot going on around". Over half (58.6%) of the places were described as vibrant and 41.4% as monotonous.

The self-reported written descriptions of the emotions related to the soundscapes were categorized under three label pairs (see Table 1).

[1] In 2020, the Covid-19 pandemic affected our lives, including social behavior. However, in August–September 2020, the Covid-19 situation in Finland was fairly stable, allowing students to study on campus, use public transportation, and freely move outdoors. Restau- rants and other leisure activities were available, with certain limitations (Ministry of Social Affairs and Health and the National Institute for Health and Welfare, 2020). The circumstances might have affected the participants, choices of recording locations. As the main aim of our study was to develop a method for deriving insights from recorded locations, the circumstances in 2020 did not compromise the collected data and the development of the method.

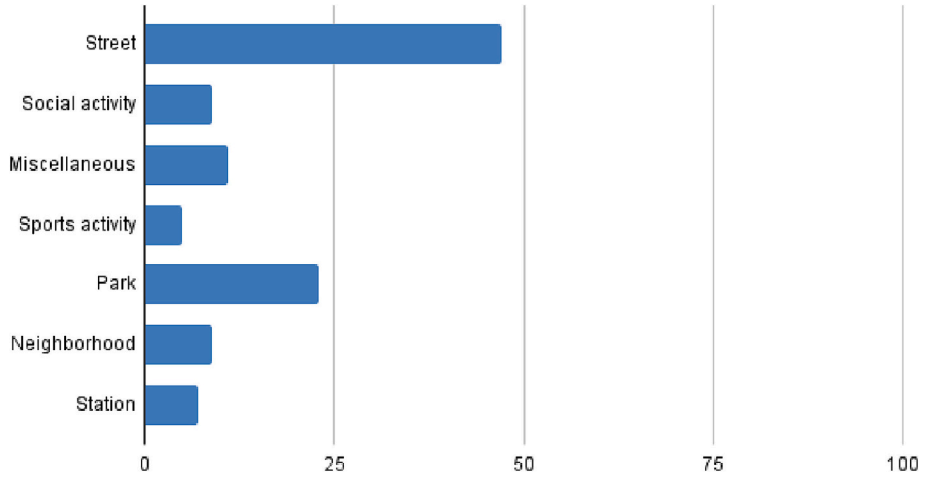


Fig. 2. Distribution of the created location categories per 111 audio samples.

Table 1
Categorization of the 111 soundscapes into three categories according to labeling of the self-reported expressions.

Pleasant		Unpleasant		Quiet		Loud		Monotonous		Vibrant	
86	(77.5%)	25	(22.5%)	57	(51.4%)	54	(48.6%)	46	(41.4%)	65	(58.6%)

2.4. Unsupervised machine learning based analysis

For automatic profiling of the soundscape experience, we applied a four- step procedure (see below Section 2.4) to identify the most important audio features based on the manual qualitative categorization. These most important features are used to link the manually produced knowledge to the raw audio recordings, thereby indicating which features are primarily related to different categories and which features contribute the most to the classification.

The recorded audio files were preprocessed as follows. The audio files were first converted to 16-bit with a sampling rate of 44.2 kHz and two channels with normalized volume, using ffmpeg. Each audio file was then truncated to the median length of all audio files (61.69 s). Files below this minimum length were padded with silence at the end of the audio file. Audio features were extracted using OpenSMILE 3.0.1 (Eyben, Wöllmer, & Schuller, 2010). In total, we extracted 988 functional features using the specification file *emobase.conf*. As summarized in Appendix A, this set of acoustic features that are com- monly used in emotion recognition research (Schuller, Steidl, & Batliner, 2009) contains statistical transformations (e.g., maximum, minimum, range, mean, stddev, skewness, kurtosis, and quartiles) as well as first- and second-order deriva- tives of the following basic groups of audio descriptors: intensity, loudness, spectral envelope, zero crossing, speech probability, fundamental frequency, pitch, and Mel-frequency cepstral coefficients (MFCC). While many of these feature sets relate to the paralinguistic analysis of a voiced speech, emobase has been applied in various other contexts of affective computing, including soundscape analysis (Lionello, Aletta, & Kang, 2020).

We followed a four-step procedure to use the extracted audio features to identify a small set of similar groups of soundscape experiences based on the audio recordings and their qualitative analysis. The first three steps perform a filter-type feature selection (Linja, Hämäläinen, Nieminen, & Kärkkäinen, 2023), and the last step establishes the division into soundscape clusters (Niemelä, Äyrämö, & Kärkkäinen, 2021).

Step 1. The range, *Rng*, of the original 988 emobase features varied in 0–2.14e+4. A range of zero means a constant, noninformative feature. Therefore, features whose range is close to zero are treated as non-informative. There were slightly >100 features with ranges of around 1e-3 or less, so we decided to drop the 102 features whose range was below this threshold. The basis for this decision is illus- trated in Fig. 3 (left).

Step 2. As defined in Cord, Ambroise, and Cocquerez (2006) and applied in, for example, Saarela, Hämäläinen, and Kärkkäinen (2017) and Jääskelä, Heilala, Kärkkäinen, and Häkkinen (2021), the H statistics of the non- parametric Kruskal-Wallis (or Mann-Whitney U for binary labelling) test (Kruskal & Wallis, 1952) can be used to evaluate how well a certain feature signifies a given classification. We computed these values with respect to the three soundscape categorizations that were derived in Section 2.2 (see Table 1). To unify the scale of statistics, all three sets were individually normalized by division of the largest value, resulting in the uniform range [0,1].

Step 3. To ensure that a feature can separate all three of the qualitative categories, we computed the minimum H statistics value over the normalized sets and sorted this vector into decreasing order. These values were then given to the knee point detection algorithm (Kaplan, 2023), which estimated the location where the curve "turns" (the "knee," see Thorndike (1953)). This point provided us the in- dex (351) and the tolerance level (0.05) that identified the point at which additional features signified less correspondence to the three manual classifications. Therefore, these 536 non-strongly separating features on the tail were removed, and we ended up with 350 features that were used in the consequent clustering step. This selection is illustrated in Fig. 3 (right).

Step 4. Because of the non-Gaussian distribution of the features to be ana- lyzed, the robust k-spatmeds++ clustering algorithm (Hämäläinen, Jauhiainen, & Kärkkäinen, 2017) with 1000 repetitions for the number of clusters ranging from $k = 2 \ldots 10$ was applied using the toolbox given in in the study by Niemelä et al. (2021).

The Wemmert-Gancarski (WG) cluster validation index, which was the best performing one in the comparisons of large-dimensional datasets with hundreds of features performed in Niemelä et al. (2021), was applied to es- timate the number of clusters. As depicted in Fig. 4, the best division into nondisjoint clusters is given with three or five clusters. These results are analyzed next.

3. Results

This section first presents the results of the machine learning based