IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2024.0429000*

# From Facial Expressions to User Experience (UX): How User Emotions Define the Design of Intelligent Systems

**ANTONIO DI TECCO**[1]
[1]Department of Excellence in Robotics and AI, Sant'Anna School of Advanced Studies

Corresponding Author: Antonio Di Tecco (e-mail: antonio.ditecco[at]outlook.com, antonio.ditecco[at]santannapisa.it).

This research work was supported by the Italian Ministry of University and Research (MUR) in the framework of the *AVATAR* Project (Department of Excellence).

This study involved humans in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Pisa Bioethical Committee (Authorization No. 8/2023, Protocol No. 12009/2023) and the Joint Bioethical Committee of Scuola Normale Superiore and Sant'Anna School of Advanced Studies of Pisa (Authorization Prot. No. 62/2024) for the *REBIO* Project (Department of Excellence).

**ABSTRACT** Emotions are a key determinant of User Experience (UX). This work investigates the relationship between facial emotions and UX. It presents *ALPHACA*, a predictive system that estimates user satisfaction from valence–arousal (VA) signals. *Dataset DT22*, a database of 112 participants, is used to evaluate models on the external test set. An *identity-free* processing pipeline that uses only VA features achieves ∼77% accuracy/weighted $F_1$-score. When contextual information is also available—Movie Clip ID and perceived *User Quality*—the *context-aware* pipeline increases performance to ∼85–92% in ablation analyses. This gain is obtained by ensembles with a *majority voting* decision rule. This indicates that contextual features provide complementary information to VA. To assess system robustness, Leave-One-Participant-Out (LOPO) and Leave-One-Movie-Out (LOMO) schemes are also investigated, as well as emotion-anchored evaluations (first- and last-emotion). The approach is modular and adheres to privacy-by-design principles by operating on *identity-free* descriptors rather than storing facial frames. Overall, the findings suggest that emotion signals can support adaptive UX. They can also enable dynamic content personalization in applications such as video streaming, customer-experience management, and e-learning. This research presents a practical study on integrating emotion-aware intelligence while respecting user privacy.

**INDEX TERMS** Artificial Intelligence, Cybernetics, Emotion Recognition, Emotion Prediction, Facial Expression Recognition (FER), Interactive Design, Machine Learning, User Experience (UX), User Satisfaction.

## I. INTRODUCTION

IN the digital age [1], where the consumption of products and services is increasingly intertwined with Human-Computer Interaction (HCI), the User Experience (UX) is a key factor in determining the success and adoption of new solutions in the market [2]. Indeed, the growing interest in UX is reflected in the search for a methodology to capture and translate user preferences and expectations into tangible metrics [3]. The fundamental challenge lies in identifying robust, reliable tools and processes for collecting and analyzing Human-Machine Interaction (HMI) data to predict UX. UX analysis is a promising approach for improving user products and services [2]. UX measurement encompasses contributions from engineering, cognitive psychology, ergonomics, and product design [4]–[8]. While UX merges various factors that influence user interaction, emotional states are a primary driver in shaping UX [2], [9]. In fact, human emotions act as a bridge between users and products or services, representing user perception, engagement, and satisfaction. Understanding user emotions provides insights into their motivations, preferences, and pain points, enabling UX designers to create products that resonate better with clients [10].

One of the most important UX metrics is user satisfaction, defined as the degree of users' comfort with a service or

product [3], [11]. In effect, users are asked about their overall satisfaction after using a product or service to indicate their final satisfaction with one or more aspects of that product or service. Hence, users' answers about their satisfaction are often expressed on a scale where higher scores indicate higher satisfaction [12].

Instead, users' emotional states change while using products or services through multiple stimulatory experiences [12]. These changes must be carefully considered. In fact, as shown in scientific literature, user emotions define how users will use the product the next time, whether they choose to use that product in the future, and whether they recommend it to someone else [3]. Therefore, predicting user satisfaction from the measured user emotions can be applied in many application domains.

Emotions may be measured in different ways. For instance, traditional approaches use surveys, interviews, and questionnaires [13], and they may be very slow or unsuitable in some circumstances, such as using recommender systems for online streaming of multimedia content.

This study presents a two-stage system in which a Facial Expression Recognition (FER) model extracts valence–arousal dynamics from facial video, and a User Experience Recognition (UXR) classifier maps these affective signals to user satisfaction levels. Thus, the work lies in moving beyond FER-only approaches by modeling the temporal evolution of affective states together with contextual information, such as clip identity and self-reported quality, to predict subjective user satisfaction. This integration addresses a gap in prior studies, which often relied on momentary signals or intrusive biosensors and could not generalize across different users and contents. Furthermore, this research employs a rigorous validation strategy based on external evaluation, as well as Leave-One-Participant-Out (LOPO), Leave-One-Movie-Out (LOMO), and ensemble schemes, among others, which demonstrate the robustness of the proposed models when applied to unseen data.

The obtained results show that models in *identity-free* mode achieve an accuracy of around 77%. While context-aware models with contextual features further improve performance, achieving over $\sim$85% accuracy across ablation studies.

The work advances the State-of-the-Art in several ways. It highlights the important role of contextual information in predicting user satisfaction. It also confirms the importance of modeling the temporal dynamics of emotions rather than relying on static signals. In addition, it establishes an effective pipeline for emotion-aware UX prediction that uses only camera input. This approach is less intrusive than physiological alternatives and can be more easily adapted to real-world applications.

This work is organized as follows: Section II provides an overview of the works related to this research study; Section III presents the background of this study; Section IV presents and describes the experimental protocol, scenario, tools, and data collection procedure; Section V presents and describes the system and classifier architecture, and data pipeline processing; Section VI presents the methodology for labeling data, signal feature extraction, and evaluation protocols; Section VII describes feature sets used for machine learning algorithms, their performance evaluation, and further analyses on machine learning algorithms; Section VIII presents a descriptive analysis of participant data; Section IX discusses finding results and the contribution to the State-of-the-Art; Sections X and XI comment on the limitations of this research work and future research; and finally, Section XII draws the article's conclusions.

## II. STATE-OF-THE-ART

Today, UX is crucial for the success of digital applications and their services; nevertheless, traditional techniques based on questionnaires, such as the Likert scale, are invasive, infrequent, and subject to low participation. To overcome these limitations, Facial Expression Recognition (FER) techniques have been explored.

The approach consists of analyzing users' emotional reactions while watching videos [14]–[17] so that human emotions may be observed. Emotions represent the human brain state associated with psychological changes, *e.g.*, heart rate, facial micro-expressions, brain activity, and so forth, that arise unconsciously and spontaneously. Emotions are important because they model and represent users' behavior and their decision-making process, including users' content selection, rating, and consumption [14], [18], [19].

However, until now, researchers have employed a combination of physiological measurements, subjective ratings, and machine learning algorithms to determine the user's emotional responses.

Physiological approaches using EEG and other biosignals have been employed to infer user emotions during video viewing. For instance, in [19], EEG, pupillary response, and gaze data were combined with SVM classifiers to estimate valence and arousal, achieving accuracies up to 76.4%. Similarly, [20] used EEG to classify emotions into Positive, Neutral, and Negative, confirming that emotional states can be reliably detected through neurophysiological data. However, these methods often require intrusive hardware setups, limiting scalability and usability in everyday digital environments—issues that this FER-based research work, camera-only approach aims to overcome.

Instead, [21] examines the application of real-time emotion recognition to enhance HCI by dynamically tailoring content to the user's emotional state. Using CNN and LSTM to analyze facial expressions and voice, the system showed significant improvements in engagement, learning, and satisfaction. They highlight both tangible benefits and technical and ethical challenges that need to be addressed for large-scale deployment.

Other researchers used FER machine learning models [27]. A FER algorithm is a tool that detects emotions by analyzing human facial expressions only [28]. These algorithms are often deep learning models developed, in general, on large

**TABLE 1. Comparative analysis of selected studies that use facial expression-based methods to measure User Experience (UX) or similar metrics. It shows methodological aspects in prior work and contrasts them with the contribution of this study.**

| Ref. | N. Part. | Method | Input | Feedback | Model | Output | Accuracy |
|---|---|---|---|---|---|---|---|
| [12] | 50+25 | UX Curve with ML Models | Momentary UX | Questionnaire | SVM, *k*-NN, etc. | UX | 93% |
| [22] | 60 | AV Emotion Recognition | Speech/Audio and Face Video | No | SVM | Customer Satisfaction | 95% |
| [23] | *N/A* | Optimized ML | Face Images | No | CNN and SVM | UX | 98.19% |
| [24] | 20 | FACS with ML | WebRTC-based Audio-Visual | Questionnaire | SVM, *k*-NN, ANN | QoE | 78% |
| [25] | 60 | UX Curve with ML Models | Task Sequence | Questionnaire | SVM, ANN, etc. | UX | 97% |
| [26] | 20 | ML-based Iterative Design | Contextual Data | Questionnaire | SVM, ANN, etc. | UX | 90% |
| **This Work** | 112 | FER and ML | Identity-free and Contextual-aware | Questionnaire | SVM, ANN, etc. | UX | **77%** **92.05%** |

datasets of photos [28], [29], and their output is based on the discrete or continuous emotion theory.

Several studies have explored FER as a tool for predicting user satisfaction. In [27], user facial expressions captured via webcam during video viewing were compared to the video's emotional profile, showing a positive correlation with user ratings, thus supporting FER as a source of implicit feedback. In [30], FER combined with heart rate data was used to train classifiers predicting binary user preferences (*Recommend* vs. *Do Not Recommend*), with *k*-NN achieving 68% accuracy and SVM reaching 86% after oversampling. These approaches demonstrate the feasibility of FER-based prediction, though often with limited accuracy and without modeling temporal evolution or rich context—limitations addressed by the proposed system in this work. When available, the accuracy of the FER can be increased by also utilizing emotion extracted from the audio data [22].

The dimensional theory represents each emotion as a point in a continuous multidimensional space where each dimension represents a quality of the emotion [31], [32]. Its main qualities are valence, which represents how negative or positive an emotion is, and arousal, which represents how calming or exciting the emotion is [29]. This theory is favored over the discrete theory because it represents the variety of emotions people exhibit in their everyday life [29], [33]. However, due to the lack of experimental data, these theories have not yet been used to predict user ratings.

This other work [23] proposes an optimization technique for emotion recognition from facial expressions, combining preprocessing filters with CNN and SVM. The algorithm selects the optimal filter parameters based on the CNN learning results. The method achieved an accuracy of 98.19%, showing high performance for embedded applications in HCI. Whereas [24] proposes *Quality of Experience* (QoE) estimation models for online videoconferencing based on facial expressions detected during audiovisual conversations. Using machine learning algorithms, the models estimate the QoE in real time without the need for explicit feedback. The results show an average accuracy of 0.78 with SVM and 0.70 with a fully connected network.

User ratings with facial emotions may have applications in various domains. For example, it was shown that user ratings had a positive correlation with user emotions [27]. This result confirmed that emotions can be valuable for implicit feedback rather than explicit techniques.

Another application field is represented by Advertising (AD), where attracting and retaining consumers' attention is important, thus limiting AD avoidance [34]. Indeed, one common problem in video ADs is that users often become bored watching ADs, even if they run for a very short time. Therefore, it is important to capture changes in user interests over a brief period, as in [35], to reduce ad avoidance or suggest other ADs. Instead, in [35], the authors utilize FER to find users' preferences dynamically. In particular, a CNN-based prediction model was developed to predict user ratings starting from a user's picture taken by a camera. A similar model was developed to search for users with similar preferences in real-time. The authors' idea was to predict the rating of video ADs using a CNN model while the user was watching the advertisement. When the predicted rating is below a certain threshold value, similar neighbors are found, and advertising videos that have been evaluated by the neighbors in the past are recommended to the target user. Moreover, predicting user satisfaction is also an application for Customer Relationship Management (CRM) systems, where customer satisfaction is evaluated over the full-service cycle using audio or speech analytics techniques [22], [36]. Recognizing user emotions and predicting their rating can also be applied to an e-commerce system, as demonstrated by [37].

A comparative summary of the most relevant studies is provided in Table 1. It highlights the number of participants, the method used, input types, feedback mechanisms, model output, and its performance.

In contrast to prior studies that rely on isolated emotion signals or limited context, the present work combines valence–arousal data extracted via FER with contextual information such as participant identity, video content, and perceived *User Quality* [38]. This integration enables improved prediction of user satisfaction, addressing key limitations in accuracy, personalization, and scalability observed in existing

**TABLE 2. Basic emotions and the neutral state with valence and arousal values normalized between the interval -1 and 1.**

| ID (↑) | Emotion | Valence | Arousal | Emoji |
|---|---|---|---|---|
| 1 | Sadness | -0.82 | -0.40 |  |
| 2 | Disgust | -0.67 | 0.49 |  |
| 3 | Contempt | -0.57 | 0.66 |  |
| 4 | Anger | -0.40 | 0.79 |  |
| 5 | Fear | -0.11 | 0.79 |  |
| 6 | Neutral | 0 | 0 |  |
| 7 | Surprise | 0.42 | 0.89 |  |
| 8 | Happiness | 0.90 | 0.16 |  |

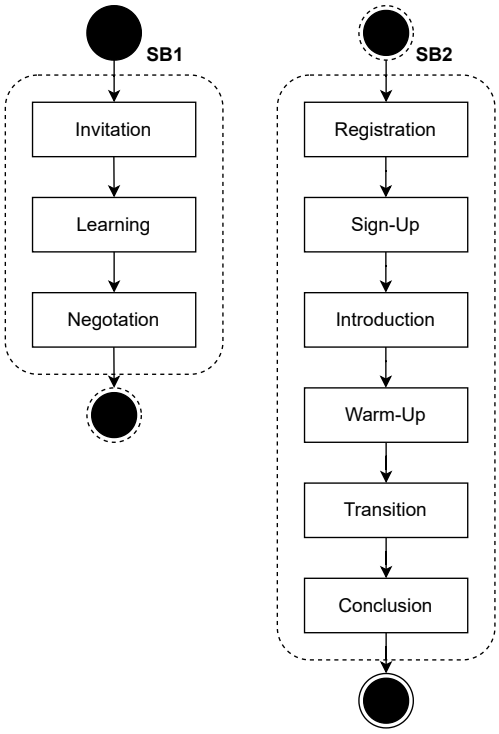


**FIGURE 1. Experimental procedure with the subprocedures Agreement (SB1) and Experimentation (SB2) and their phases.**

approaches.

## III. EMOTIONS AND USER EXPERIENCE (UX)

The section presents the nature of user emotions and explains their impact on UX, describing how emotions may influence user perceptions and human satisfaction.

### A. USER EMOTION

Every interaction with a product or service evokes emotional responses (emotions) in users [39], whether positive (happiness and surprise) or negative (sadness, disgust, anger, and fear). These emotions influence how users perceive the usability, value, and overall experience of a product or service [40]. One way to observe user emotions is to measure valence and arousal levels.

Valence indicates the degree of positivity or negativity of an emotion. In practice, an emotional state with high valence corresponds to pleasant or positive sensations, such as happiness or joy. In contrast, a low or negative valence corresponds to unpleasant sensations, such as sadness or disgust. Instead, arousal measures the level of activation or excitement associated with an emotional state. High arousal corresponds to very active and energetic emotions, such as surprise and anger, while low arousal describes calmer, more relaxed states, such as contentment, tranquility, and relaxation. Therefore, both valence and arousal are represented on a continuous scale from -1 (negative) to 1 (positive), where 0 indicates a relaxed emotional state or no activation (neutral), and -1 or 1 indicates maximal absolute psychophysiological activation. Valence and arousal are synthetic metrics that measure a person's emotional tone, allowing the model to capture both the emotion's polarity (positive vs. negative) and its activation level (calm vs. excited).

Table 2 presents the most frequent emotions considered in the scientific literature [29], [33], [38].

### B. UX AND USER SATISFACTION

The User Experience (UX) is the set of emotions that a user experiences when interacting with a product or service [40]. Therefore, while emotions reflect how people perceive products or services in a given moment, UX captures users' satisfaction during interaction. In addition, UX expresses the usability, efficiency, and satisfaction of used products or services [41]–[44]. Hence, this work uses human emotions to estimate user satisfaction.

The satisfaction metric is measured by using the Likert scale method [45]. The Likert scale is based on a question that measures user satisfaction using an odd number of response options, each representing a different level of satisfaction (*e.g.*, 1 to 5). In detail, a high level of user satisfaction indicates high satisfaction, whereas a low level indicates low satisfaction (dissatisfaction).

## IV. EXPERIMENTATION PROTOCOL

In the scientific literature, there is a lack of open data on emotions and UX. Thus, an experiment named *Sperimentazione DT22* is conducted to gather data and create a database named

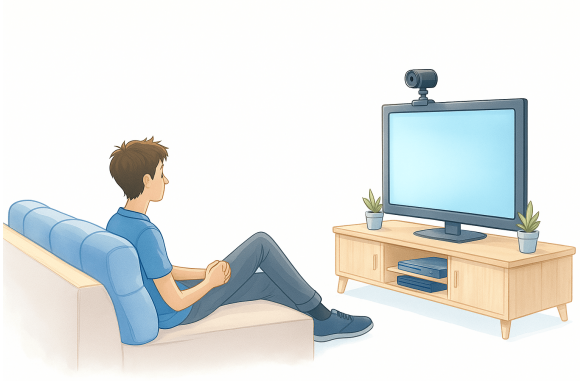

**FIGURE 2.** The experimental scenario involves a participant being recorded with a camera while watching a movie clip on a television monitor.

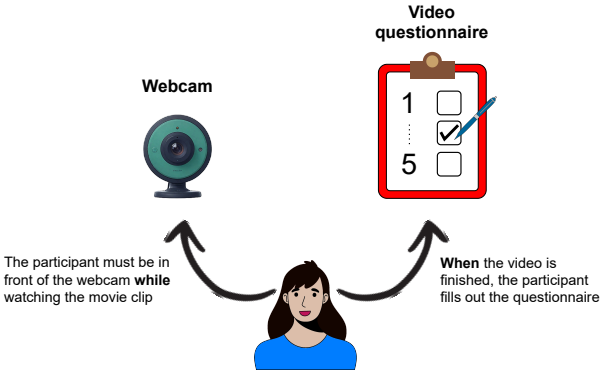


**FIGURE 3.** Tasks performed by the participant during the experiment.

*Dataset DT22* [38], [46]. The database contains signal data gathered from 112 users who participated in the above experimentation. In the database, emotion signals, such as valence-arousal (VA), are derived from participants' video sequences. These signal data are used to train and evaluate machine learning algorithms to predict user satisfaction. Hence, this section presents the experimental procedure and the data collected and used. Moreover, the experimental protocol is reviewed and approved by the *Bioethical Committee of the University of Pisa* (Authorization No. 8/2023, Protocol No. 12009/2023). In addition, it is also reviewed and approved by the *Joint Ethics Committee of the Scuola Normale Superiore and Sant'Anna School of Advanced Studies* of Pisa (Authorization Prot. No. 62/2024 on January 16*th*, 2025) for the *REBIO* Project.

### A. EXPERIMENTAL PROCEDURE

The experimentation procedure is divided into two subprocedures, namely *Agreement* and *Experimentation*, and each subprocedure has its phases, as shown in Fig. 1. The *Agreement* subprocedure is divided into *Invitation*, *Learning*, and *Negotiation* phases. Whereas the *Experimentation* subprocedure is divided into *Registration*, *Sign-Up*, *Introduction*, *Warm-up*, *Transition*, and *Conclusion* phases. In detail, the *Agreement* subprocedure involves the user in the experimentation. The user is informed of the experiment during the *Invitation* phase. Then, the user learns the experiment activities during the *Learning* phase. In the *Negotiation* phase, the user reads and, if they consent, signs the informed consent form, thereby confirming their participation in the experiment. By signing the form, the user agrees to participate in the experimental process. Then, they receive a private e-mail with a link to access the website for the experiment, as represented by the *Experimentation* subprocedure. It begins with the *Registration* phase, where the participant signs up for an account and provides personal data. They could log in on the platform home page to begin the experimentation. The *Introduction* phase follows. The participant watches an introduction clip that summarizes the experimental activities. Therefore, after watching the introduction clip, the participant may start the *Warm-Up* phase. In this phase, they watch two YouTube videos and apply the *learning-by-doing* method to learn their tasks. The *Transition* phase follows. During this phase, participants watch seven movie clips and fill out questionnaires. At the end, in the *Conclusion* phase, the participant is thanked.

### B. EXPERIMENTAL SCENARIO

The experimental scenario requires a participant and an electronic device connected to the internet and with a mounted webcam. This device could be a personal computer, laptop, tablet, smartphone, or simply a monitor display with a remote control. The experimentation could be carried out in a university setting or remotely, for instance, at home in a distraction-free environment without noise sources, as illustrated in Fig. 2. In this scenario, the participant watches nine video clips in total, pausing or resuming them at any moment. In particular, the first clips are used to teach participants about experimentation and as a warm-up. The other seven clips are movie clips that induced emotional stimuli in participants and from which data were gathered, as presented in Table 3. In the table, their title, scene description, start/end/length times, and dominant emotion (DE) are highlighted.

These movie clips were selected from validated film databases for emotion elicitation [47], [48]. The set was constructed to cover both positive (*e.g.*, happiness, surprise) and negative emotions (*e.g.*, sadness, anger, disgust, fear), in line with basic emotions (see Section III-A). Clip lengths ranged from 1:10 to 4:47 min (mean $\approx$ 2.6 min), so the overall viewing time per participant was comparable across clips. All participants observed the same set of clips. The order of the clips was fixed across participants.

In addition, the participant uses the webcam while watching these clips so their face can be recorded. The participant fills out zero or more *emotion questionnaires* and one *video questionnaire* for each clip. The participant completes the *emotion questionnaire* while they watch the clip, declaring their emotions at that moment and their valence and arousal levels between 1 and 9. Instead, the participant completes the *video questionnaire* at the end of each clip to indicate their level of satisfaction. Thus, these participant tasks are

**TABLE 3.** List of movie clips watched by participants during the experimentation that induced emotional stimuli.

| ID | Ref. | Title | Description | Start | End | Length | DE(s) |
|---|---|---|---|---|---|---|---|
| 3 | [47] | The Visitors (1993) | Jacquouille and Godfroid destroy the postman's car | 00:19:55 | 00:22:10 | 02:15 | Happiness |
| 4 | [47] | Schindler's List (1993) | The commander of a concentration camp wakes up and shoots the prisoners | 01:13:40 | 01:16:40 | 03:00 | Anger & Sadness |
| 5 | [47] | The Dead Poets Society (1989) | Todd commits suicide | 01:42:54 | 01:47:41 | 04:47 | Sadness |
| 6 | [47] | A Fish Called Wanda (1988) | Archie gets undressed, waiting for his girlfriend. Unexpectedly, the house owners discover him naked | 01:11:55 | 01:15:16 | 03:21 | Happiness |
| 7 | [47] | Trainspotting (1996) | A woman screams in an apartment, waking up the others. They find out that the woman's newborn baby is dead | 00:38:52 | 00:40:35 | 01:43 | Disgust & Sadness |
| 8 | [48] | Capricorn One (1977) | Men burst through the door unexpectedly | 01:32:51 | 01:34:01 | 01:10 | Surprise |
| 9 | [47] | Se7en (1995) | A man is found dead, tied to a bed. Unexpectedly, the man wakes up | 00:52:22 | 00:54:10 | 01:48 | Fear & Disgust |

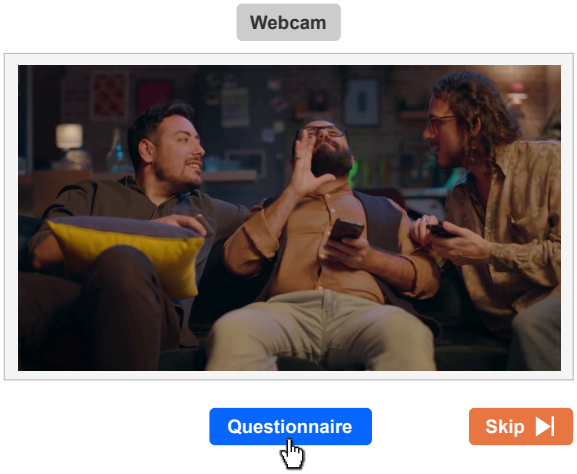


**FIGURE 4.** Guest User Interface (GUI) on the experimental web platform.

illustrated in Fig. 3. Moreover, technical support is guaranteed to the participant during experimentation to resolve platform issues, such as video playback.

Fig. 4 illustrates the online experimental platform's Guest User Interface (GUI) inspired by the VERA software platform [49]. The figure shows a frame where the clip is displayed and three buttons. On top of the frame, the gray button allows the participant to check that the webcam is working. Instead, at the bottom of the frame, the blue button allows participants to fill out the *emotion questionnaire*, and the orange button allows them to skip the clip if they do not like it.

### C. DATA SIGNAL

Participant data are collected via the online experimental platform from emotion and video questionnaires and webcam recordings.

In particular, after each movie clip ends and before the next one starts, the platform automatically displays a short *video questionnaire* in a pop-up window. Then, the participant answers the question *"How satisfied are you with watching this video?"* using a 5-point Likert scale, with the following textual anchors: 1 = *Very Dissatisfied*, 2 = *Dissatisfied*, 3 = *Neutral*, 4 = *Satisfied*, 5 = *Very Satisfied*. Each clip viewed by a participant is therefore associated with a satisfaction score $s \in S$, where $S = \{1, \ldots, 5\}$. These answers represent participants' UX levels, translating their subjective satisfaction into a quantifiable rating to assess the value of the movie clip.

Data is transmitted to the university server node, where it is processed and archived. Questionnaires' raw data are converted into a time series of time steps corresponding to the total number of *emotion questionnaires*. Instead, the webcam recordings are processed using the FER model [29]. The FER model takes a sequence of frames from the recording as input and outputs emotional data, including valence and arousal. The sequence of frames is converted into a time series with a duration equal to the participant's viewing time for the corresponding movie clip. Thus, the time series is labeled according to the participant's reported satisfaction level.

Each participant watches seven movie clips (see Section IV-B), resulting in a theoretical maximum of 784 *video questionnaires* (112 participants $\times$ 7 clips), as detailed in Section VIII-C.

## V. INTELLIGENT SYSTEM ARCHITECTURE

This section describes the architecture of the proposed intelligent system, named *ALPHACA* [50], and its core prediction module, the *UXR Engine* [51]. The architecture specifies the end-to-end processing pipeline that maps raw RGB frames captured from the participant's face to a discrete *User Satisfaction* score. A high-level block diagram of the architecture is shown in Fig. 5.

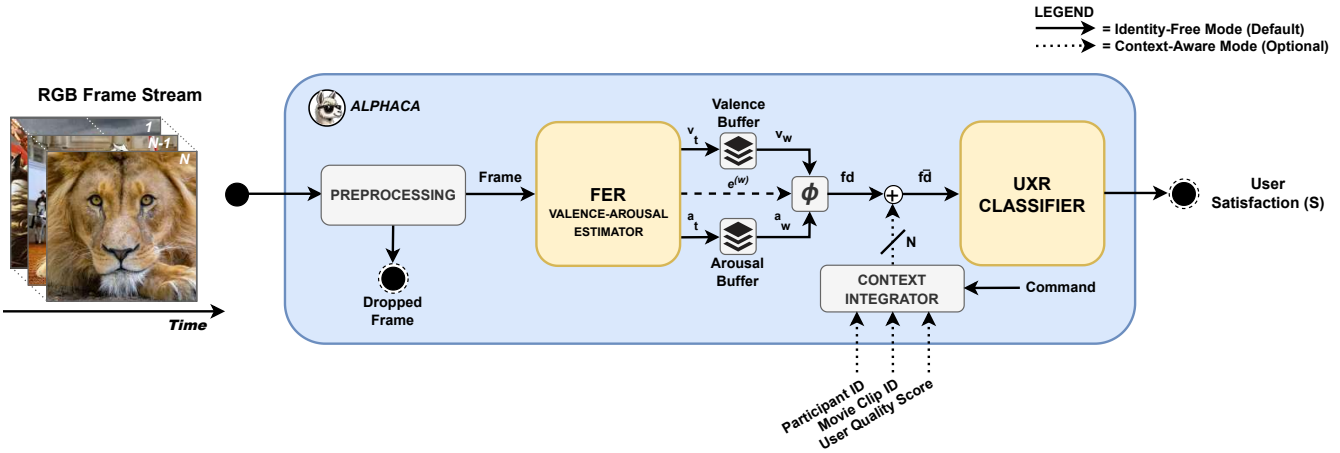


**FIGURE 5.** Block diagram of the ALPHACA intelligent system. Frames are preprocessed, passed to the FER model, and aggregated into window-level affect vectors $\mathbf{e}^{(w)}$. The feature extractor $\phi$ produces descriptors **fd**, optionally concatenated (+) with subsets of contextual features (Participant ID, Movie Clip ID, User Quality score). The downstream UXR classifier takes features as input and outputs the predicted User Satisfaction (*S*).

### A. ALPHACA SYSTEM

*ALPHACA* is a two-stage camera-only intelligent system that maps emotional dynamics extracted from the participant's face to estimate the *User Satisfaction* (UX) score. The architecture consists of a Facial Emotion Recognition (FER) classification model that produces valence–arousal time series from video frames, followed by a User Experience Recognition (UXR) classifier that infers the UX target label. Optional contextual features, such as Participant ID, Movie Clip ID, and *User Quality* score, can be given as input to the UXR classifier to improve robustness and personalization capability of the system.

**Inputs and Internal Variables.** The mandatory input of *ALPHACA* is a stream of RGB frames captured while the participant watches a movie clip. The frame is preprocessed through a preprocessing module that applies face detection, alignment, and quality checks (illumination/blur) algorithms. Then, the frame is passed as input to the FER model that outputs at each time step $t$ a continuous *affect vector* $\mathrm{e}_t$, as defined in 1:

$$\mathbf{e}_t = (v_t, a_t), \qquad v_t, a_t \in [-1, 1], \tag{1}$$

where $v_t$ and $a_t$ are the estimates of valence and arousal levels, at the time $t$, that constitute the system's *internal affective state* used downstream. These levels are sampled at a fixed rate and buffered into time windows of length $W$ (default $W = 60\,\mathrm{s}$) to stabilize the emotion estimate and capture its temporal dynamics. This representation of the emotion is provided to the downstream UXR model, which uses it to compute the system's output.

**Feature Building.** For a $W$-second window $w$, let $\mathbf{v}_w = [v_1, \ldots, v_{n_w}]$ and $\mathbf{a}_w = [a_1, \ldots, a_{n_w}]$ denote the frame-wise valence and arousal sequences within the time window, with $n_w = rW$ frames at frame rate $r = 1/\mathrm{fps}$, where $\mathrm{fps} = 30.30$. Define the window matrix $\mathbf{e}^{(w)}$ as in 2:

$$\mathbf{e}^{(w)} = \begin{bmatrix} \mathbf{v}_w \\ \mathbf{a}_w \end{bmatrix} \in \mathbb{R}^{2\times n_w}. \tag{2}$$

*ALPHACA* computes a window-level feature descriptor $\mathbf{fd}^{(w)}$ as in 3:

$$\mathbf{fd}^{(w)} = \phi\left(\mathbf{e}^{(w)}\right) = \phi(\mathbf{v}_w, \mathbf{a}_w) \in \mathbb{R}^N, \tag{3}$$

where $\phi : \mathbb{R}^{2\times n_w} \to \mathbb{R}^N$ aggregates features, for example, statistical, temporal, and shape features. The resulting $\mathbf{fd}^{(w)}$ is given as input to the UXR classifier. For simplicity and without loss of generality, the descriptor $\mathbf{fd}^{(w)}$ is denoted by **fd** throughout the remainder of the work.

**Optional Context.** If context information is present, three optional contextual features can be added to the feature descriptor **fd** as input for the UXR classifier: (i) *Participant ID* as $P$ (anonymized identifier), (ii) *Movie Clip ID* as $C$ (content identifier), and (iii) *User Quality* score as $Q$ (subjective 1–5 rating). These features are appended as scalars to **fd** to form the augmented descriptor $\bar{\mathbf{fd}}$, *i.e.*, $\mathbf{fd} \in \mathbb{R}^N \mapsto \bar{\mathbf{fd}} = [\mathbf{fd} \,|\, P \,|\, C \,|\, Q] \in \mathbb{R}^{N+3}$. They are not required for running the system, but can improve system robustness and personalization.

**Operating Modes and Privacy Regimes.** In all experiments, two pipelines of operating modes are distinguished.

*Identity-free Pipeline (Default).* In the *identity-free* mode, the UXR classifier receives only the VA-derived descriptor **fd** as input, without any participant identifier or contextual features. This mode can be deployed with on-device inference and without storing raw video or biometric templates, and it is therefore suitable for privacy-critical applications.

*Context-aware Pipeline.* When contextual information is available, **fd** can be augmented with one or more features $P$, $C$, $Q$ (Participant ID, Movie Clip ID, and *User Quality* score, respectively), resulting in the augmented descriptor

$\bar{\mathbf{fd}} = [\mathbf{fd} \,|\, P \,|\, C \,|\, Q]$. These context-aware variants trade a stronger trust assumption—the system has access to stable features and viewing metadata—for gains in robustness and personalization.

A discrete control signal, *Command*, specifies which subset of contextual features to use, *e.g.*, only `VA`, `VA+P`, `VA+C`, `VA+Q`, `VA+PC`, `VA+PQ`, `VA+CQ`, `VA+PCQ`. At inference time, the system follows a progressive fallback: it selects the feature subset consistent with *Command* value and the available inputs, defaulting to the *identity-free* pipeline whenever required contextual fields are missing or disabled.

**Interfaces and Deployment.** The pipeline exposes clean interfaces (see Fig. 5): (1) a FER API that ingests frames and outputs affect vectors $(v_t, a_t)$; (2) a Feature Builder that batches valence and arousal sequences $(\mathbf{v}_w, \mathbf{a}_w)$ and emits feature descriptors $\mathbf{fd}$; (3) an optional Context Integrator that adds $P, C, Q$ features to $\mathbf{fd}$; (4) a UXR classifier that takes the feature descriptors as input and returns their labels $s$. For privacy-by-design, raw frames do not need to be persisted; inference can run on-device, and only anonymized descriptors/metadata are stored when logging is enabled.

### B. UXR CLASSIFIER DESIGN

The UXR model is a supervised classifier, denoted $\mathrm{uxr}(\cdot)$, that maps a feature descriptor to a discrete *User Satisfaction* label. Let $\mathbf{fd} \in \mathbb{R}^N$ be the descriptor extracted from valence-arousal streams (see Section V), and let $\bar{\mathbf{fd}}$ be its augmented version when contextual features are provided. At inference time, the classifier outputs $s$ as in 4:

$$s = \mathrm{uxr}(\bar{\mathbf{fd}}), \qquad s \in \mathcal{S}, \tag{4}$$

and can optionally return a confidence score $p(s)$.

#### Baseline Models (Identity-Free Mode)

To accommodate different granularity requirements, three label-space configurations or policies are defined (see Table 4). Let $s \in \{1, 2, 3, 4, 5\}$ be the original Likert score. Each policy specifies a mapping $\pi(\cdot)$ from $s$ to a target label $s \in \mathcal{S}$, as in 5–7:

**Policy $P_{\text{2-3}}$ (binary).**

$$\mathcal{S} = \{1, 2\}, \quad \pi_{\text{2-3}}(s) = \begin{cases} 1, & s \in \{1, 2\}, \\ 2, & s \in \{3, 4, 5\}. \end{cases} \tag{5}$$

*Labels 1 and 2 correspond to Dissatisfied and Satisfied, respectively.*

**Policy $P_{\text{2-1-2}}$ (ternary).**

$$\mathcal{S} = \{1, 2, 3\}, \quad \pi_{\text{2-1-2}}(s) = \begin{cases} 1, & s \in \{1, 2\}, \\ 2, & s = 3, \\ 3, & s \in \{4, 5\}. \end{cases} \tag{6}$$

*Labels 1, 2, and 3 correspond to Dissatisfied, Neutral, and Satisfied, respectively.*

**Policy $P_{\text{1-to-5}}$ (five-class).**

$$\mathcal{S} = \{1, 2, 3, 4, 5\}, \quad \pi_{\text{1-to-5}}(s) = s. \tag{7}$$

*Labels 1–5 correspond to Very Dissatisfied, Dissatisfied, Neutral, Satisfied, and Very Satisfied, respectively.*

#### Ablation Models (Context-Aware Mode)

Starting from the optimal baseline policy, the *ablation* variants are derived. In detail, subsets of contextual features are added to the baseline models, and their contribution is measured.

The contextual features include Participant ID ($P$), Movie Clip ID ($C$), and *User Quality* score ($Q$) [38], [52]. Concretely, the classifier receives $\bar{\mathbf{fd}} = [\mathbf{fd} \,|\, P \,|\, C \,|\, Q]$ when context is available; otherwise, it operates in *identity-free* mode. In addition, ablation models are combined to create an ensemble of the top-performing classifiers aggregated with a *majority voting* rule (hard voting).

## VI. DATA PROCESSING AND EVALUATION METHODOLOGY

This section presents the algorithms employed to preprocess data [38], [46], label signals, extract signals' features, and create datasets for designing, training, and evaluating the UXR classification model.

### A. DATA PREPROCESSING AND SIGNALS LABELING

The system framework implements the FER and UXR models, as illustrated in Fig. 5. In the experiment, EmoNet is the FER model [29]. This model processes a sequence of frames as input, and it generates a sequence of emotions as output. In particular, the emotion considered in this study is defined with valence and arousal values, and a discrete value, as also represented in the *Plutchik's Wheel* (see Table 2). Valence and arousal are continuous values ranging from -1 to 1. Whereas the emotion is discrete, such as one for sadness, two for disgust, and so forth.

To a sequence of emotions corresponds a time series where a time step is a couple of values of valence and arousal. The time series has the same time length as the movie clip in seconds. However, the clip may have a different number of frames because the webcam device may differ for each user.

By analyzing the whole set of recordings, the median frame rate is 30.30 frames per second (fps). Thus, time series are resampled at the same sample rate of 30.30 fps. This resampling step is necessary because participants' devices and tools may differ (*e.g.*, webcam models, browsers, and operating systems). This approach mitigates the effects of device-related inconsistencies, ensuring that all time series are aligned and comparable.

Global and local features are extracted from time series. For local features, an exploratory analysis of fixed-length segmentation has been carried out with window sizes $W \in \{10, 20, 25, 30, 40, 45, 50, 60\}$ s to determine how $W$ may affect model performance. Shorter windows (*e.g.*, $W \leq 20$ s)

**TABLE 4. Labeling policies from Likert scales to the target class *S*.**

| Policy | Classes $S$ | Mapping $\pi(s)$ |
|---|---|---|
| $P_{2\text{-}3}$ (binary) | $\{1, 2\}$ | $1 : \{1, 2\}$ (Dissatisfied), $2 : \{3, 4, 5\}$ (Satisfied) |
| $P_{2\text{-}1\text{-}2}$ (ternary) | $\{1, 2, 3\}$ | $1 : \{1, 2\}$ (Dissatisfied), $2 : \{3\}$ (Neutral), $3 : \{4, 5\}$ (Satisfied) |
| $P_{1\text{-to-}5}$ (five-class) | $\{1, 2, 3, 4, 5\}$ | $1 : 1$ (Very Dissatisfied), $2 : 2$ (Dissatisfied), $3 : 3$ (Neutral), $4 : 4$ (Satisfied), $5 : 5$ (Very Satisfied) |

increase temporal resolution and the number of available segments. However, they produce fragmented and noisier emotional windows for users. Longer windows, instead, smooth out user reactions and reduce the number of segments. Across the exploratory evaluations, performance variations remained modest and within the variability of the estimates, and no single window size emerged as optimal. Consistent with prior observations [53], a time window length of $W = 60$ s is therefore adopted in this work as a pragmatic and interpretable default parameter, rather than as a definitive *optimum*.

Feature descriptors are defined as vectors of features extracted from a time series or from a specific time window. Each descriptor is associated with the label of the corresponding signal. These feature descriptors are used as input samples for the UXR model (see Fig. 5), while their labels serve as the classification targets.

Therefore, for each participant–clip pair, all descriptors extracted from the corresponding time series are labeled with the same clip-level satisfaction score $s$ (the 1–5 Likert rating described in Sections III-B and V-B). Therefore, all local descriptors extracted over 60-second windows inherit the clip's satisfaction score.

### B. FEATURE EXTRACTION

Statistical, temporal, and shape features are extracted from each emotion signal, such as valence and arousal, as input for the UXR classifier. The features are, as in [38], head value (first sample of the signal), tail value (last sample of the signal), maximum, minimum, ratio between the maximum and minimum values, difference between maximum and minimum value (range), root mean square, ratio between root mean square and median absolute deviation, mean absolute deviation ($MAD_0$), median absolute deviation ($MAD_1$), mean, median, mean of absolute values ($MAV$), mean square, power, geometric mean, harmonic mean, 10% trimmed mean, variance, standard deviation, 2*nd*-, 3*rd*, and 4*th*-order moments, 33*rd*-percentile, 1*st*-, and 3*rd*-quantile, interquartile, skewness, kurtosis, crest factor, clearance factor, impulse factor, peak value, and shape factor. Thus, VA features are concatenated to create a feature descriptor $\mathbf{fd}$ corresponding to a sample.

### C. DATASETS

Architectural models are trained and evaluated using specific datasets. Three datasets are defined and created, they are datasets $D_{2\text{-}3}$, $D_{2\text{-}1\text{-}2}$, and $D_{1\text{-to-}5}$. These datasets are then split into internal and external sets (90%/10%) with the same proportion of samples for each class. These sets are also named as training and evaluation sets in the literature.

To mitigate the class imbalance in satisfaction scores, in particular for classes with lower class frequencies (*e.g.*, scores 1 and 5), the ADASYN (Adaptive Synthetic Sampling) technique is applied to the internal set only. This oversampling method generates synthetic samples in the feature space, improving class balance before model training. In addition, stratified $k$-fold cross-validation is used to ensure that each fold preserves the original class distribution, enhancing robustness and fairness in the model evaluation process.

To provide a clearer overview of the dataset structure, Table 5 summarizes the types of data collected during the experimentation, their description, format, and the total number of global and segmented samples used in this research study. Moreover, Table 6 shows the number of global and segmented samples divided into internal (training) and external (test) sets without applying ADASYN.

### D. FEATURE SELECTION

Internal training sets are used to find the optimal feature sets using the *FeatureDS* software [54]. The Forward Sequential Feature Selection (FSFS) [55] finds the optimal feature set considering, as a criterion, the minimization of the cross-entropy of an ANN. The ANN is chosen because it can learn complex linear and nonlinear input-output relationships [56]. An ANN with a single hidden layer is chosen, with as many neurons in the output layer as the number of classes (two for $P_{2\text{-}3}$, three for $P_{2\text{-}1\text{-}2}$, and five for $P_{1\text{-to-}5}$). The number of hidden neurons is set according to 8, as in [57], such that:

$$H \approx \sqrt{N_i N_o}\,, \tag{8}$$

where $H$ is the number of hidden neurons, and $N_i$ and $N_o$ are the number of input and output, respectively. $N_i$ is set to 20 as a maximum of 20 features is considered for selection, whereas $N_o$ is equal to 2 for $P_{2\text{-}3}$, and 3 for $P_{2\text{-}1\text{-}2}$ and $P_{1\text{-to-}5}$. Then, $H = 6$ for $P_{2\text{-}3}$, $H = 8$ for $P_{2\text{-}1\text{-}2}$, and $H = 10$ for $P_{1\text{-to-}5}$.

A total of 100 executions of the FSFS are performed using stratified $k$-fold cross-validation on the training set, with $k$ equal to 10, thereby obtaining 100 candidate feature sets. Each fold contains randomly selected samples; of course, folds do not share any samples. A statistical evaluation leads to the most predictive feature set among the 100 candidates. In particular, each candidate feature set ($F_i$) is the input to train an ANN using a 30-run 10-fold cross-validation.

At each iteration, the performance of the ANN is assessed by measuring the cross-validation error ($E_{ik}$), representing the model's predictive accuracy, *i.e.*, the average of the mean cross entropies obtained for each of the 10 test folds at iteration $k$ using feature set $F_i$ as input.

**TABLE 5. Summary of dataset attributes used.**

| Data Type | Description | Format | Number of Samples |
|---|---|---|---|
| Facial Emotion Signals | Time series of valence and arousal per participant and video | Float32 (2D array) | 645 (global windows)<br>2,254 (segmented windows) |
| Categorical Emotions (FER) | Discrete emotion labels (*e.g.*, happiness, fear, surprise) | Integer (1-8) | ∼2,985,476 frame-level labels |
| Video Clip ID | Identifier for each movie clip shown to participants | Integer (3-9) | 7 unique values |
| Participant ID | Anonymized unique user identifier | Integer (1-112) | 112 participants |
| Self-reported Satisfaction | Participant's Likert score (1–5) on video satisfaction | Integer (1–5) | 645 entries |
| Self-reported Quality | Participant's Likert score (1–5) on video quality | Integer (1–5) | 645 entries |
| Features | Features extracted per window (mean, std, etc.) | Float Vector (size≈60) | 645 (global samples) and<br>2,254 (segmented samples) |

**TABLE 6. Number of global and segmented samples of each dataset $D_{2\text{-}3}$ (a), $D_{2\text{-}1\text{-}2}$ (b), and $D_{1\text{-}to\text{-}5}$ (c), in order.**

| Class | Internal Set | External Set |
|---|---|---|
| 1 | 261/882 | 29/97 |
| 2 | 320/1.147 | 35/128 |
| **Total** | 581/2.029 | 64/225 |

(a)

| Class | Internal Set | External Set |
|---|---|---|
| 1 | 261/881 | 29/98 |
| 2 | 198/674 | 21/74 |
| 3 | 122/474 | 14/53 |
| **Total** | 581/2.029 | 64/225 |

(b)

| Class | Internal Set | External Set |
|---|---|---|
| 1 | 62/188 | 7/21 |
| 2 | 199/692 | 22/78 |
| 3 | 197/673 | 22/75 |
| 4 | 102/395 | 11/43 |
| 5 | 21/81 | 2/8 |
| **Total** | 581/2.029 | 64/225 |

(c)

The Student's $t$-test with a significance level ($\alpha$) of 0.05 leads to finding the optimal feature set. This significance level represented the statistical significance threshold. In particular, Student's $t$-test is used to compare the means of the cross-validation errors generated by the ANN models using distinct candidate feature sets $F_i$ and $F_j$. The null hypothesis is as in 9:

$$H_0 : \text{the average mean cross-entropies of } F_i \text{ and } F_j \text{ do not show a statistically significant difference,} \tag{9}$$

suggesting that preferring one of the two feature sets over the other would not impact the model performance. The $t$-test produces a statistic ($t$) that measures the difference between the means, determined as in 10:

$$t = \frac{\overline{E}_i - \overline{E}_j}{s_p\sqrt{2/K}}, \tag{10}$$

where $\overline{E}_i$ and $\overline{E}_j$ are the means of the cross-validation errors for feature sets $F_i$ and $F_j$, respectively, $K$ is the number of iterations (*i.e.*, 30), and $s_p$ denotes the pooled standard deviation of errors, calculated as in 11:

$$s_p = \sqrt{\frac{(n-1)\sigma_i^2 + (m-1)\sigma_j^2}{n+m-2}}, \tag{11}$$

where $\sigma_i$ and $\sigma_j$ are the standard deviations of errors for feature sets $F_i$ and $F_j$, respectively, $n$ and $m$ are the sample sizes of the cross-validation errors for $F_i$ and $F_j$, both equal to 30.

The $t$-test also generates a $p$-value measuring the probability of observing a $t$-statistic as extreme as the one computed if the null hypothesis ($H_0$) is true. If the $p$-value is lower than $\alpha$, then the difference in the means of cross-validation errors is highly unlikely to have occurred by random chance. This indicates that the results are statistically significant, and the null hypothesis ($H_0$) is rejected as there is a statistically significant difference in the performance of candidate feature sets $F_i$ and $F_j$. The feature set that results in the highest number of rejections of the null hypothesis—indicating statistically significant differences in cross-validation errors compared to the others—is considered the most powerful for the classification task.

### E. PERFORMANCE ANALYSIS

The performance of the proposed models is evaluated through a structured protocol. A broad family of classifiers is considered, including optimized variants. Models are trained and validated by stratified cross-validation and, in addition, assessed on an external test set over 30 independent runs. Furthermore, Leave-One-Participant-Out (LOPO) and Leave-One-Movie-Out (LOMO) cross-validation schemes are employed. Predictions are aggregated either at the temporal level (window-to-clip) using LOMO or at the model level via *majority voting* in ensemble schemes. Finally, emotion-anchored analyses are performed.

#### 1) Classifier Families and Optimization Protocol

Many families of classifiers, such as Decision Trees, Discriminant Analysis, Naive Bayes, Logistic Regression, SVMs, ANNs, $k$-NNs, Kernels, and Ensembles, are considered and built for evaluating each architecture design, also by using *Bayesian search* to fine-tune classifiers' hyperparameters. Fine-tuned classifiers are named as *optimized* in this study.

**TABLE 7. Study, policies, and summary of the corresponding results. Each row links to where the models and metrics (accuracy and weighted $F_1$-score) are reported.**

| Study | Mode | Policy | Context | Model | Accuracy | wF1 | Ref. | Note |
|---|---|---|---|---|---|---|---|---|
| Baseline | Identity-free | $P_{2\text{-}3}$ | VA | Optimized $k$-NN | 0.7698 | **0.7615** | §VII-B | Only emotion signals. |
| Ablation | Context-aware | $P_{\text{CQ}}$ | VA + CQ | Fine $k$-NN | 0.8490 | 0.8436 | §VII-B | The optimal policy without P. |
| Ablation | Context-aware | $P_{\text{PCQ}}$ | VA + PCQ | Optimized $k$-NN | 0.9935 | 0.9935 | §VII-B | The optimal policy with P$^{\dagger}$. |
| Ablation | Context-aware | LOPO | VA + CQ | Fine $k$-NN | 0.8412 | 0.8397 | §VII-C | Cross-user robustness. |
| Ablation | Context-aware | LOMO | VA + CQ | Fine $k$-NN | 0.7709 | 0.7705 | §VII-D | Cross-clip robustness. |
| Ablation | Context-aware | Ensemble ($P_{\text{CQ}}$) | VA + CQ | Majority Voting | 0.9205 | **0.9184** | §VII-E | Top-3 $P_{\text{CQ}}$ ensemble. |
| Ablation | Context-aware | FE-only ($P_{\text{CQ}}$) | VA + CQ | Fine Tree | 0.7813 | 0.7804 | §VII-F | Compact temporal synthesis. |
| Ablation | Context-aware | LE-only ($P_{\text{CQ}}$) | VA + CQ | Medium ANN | 0.7812 | 0.7804 | §VII-F | Compact temporal synthesis. |

**Note.** VA: Valence–Arousal; P: Participant ID; C: Movie Clip ID; Q: User Quality score; $^{\dagger}$: Very high performance but likely leakage (use of P).

Standard and optimized classifiers are based on MATLAB R2024b Classification Learner App, using the *AutoML30X* software available on Zenodo [58].

For reproducibility, seeds and resulting datasets are generated in accordance with the MATLAB Classification Learner App's seeding rules, which drive cross-validation partitioning, model initialization [58], and Bayesian hyperparameter search before each run. In this case, classification algorithms are trained on the internal set, then evaluated on the external set for 30 independent iterations, as in [38], [53].

#### 2) Internal vs. External Evaluation Protocol

The internal evaluation assesses classification algorithms using a 30-run stratified 10-fold cross-validation. These algorithms are referred to as *internal models*, and they are used to demonstrate the classification algorithm's robustness against overfitting. However, internal evaluation results are ignored for simplicity.

These algorithms are then evaluated by performing external evaluation on the external set to demonstrate their performance on unseen samples, providing an unbiased estimate using real-world data. These models are named *external models*. The external models with the lowest mean loss are the optimal classifiers for the specific policy design considered in this study.

#### 3) Leave-One-Out Evaluation Protocol

The robustness of the optimal designs is evaluated using various evaluation protocol schemes based on the *leave-one-out* method: the Leave-One-Participant-Out (LOPO) and the Leave-One-Movie-Out (LOMO) schemes.

In the LOPO scheme, 112 classifiers are trained and evaluated, one classifier for each participant. For each run, the classifier is trained on samples from $N-1$ participants and then evaluated on the n*th* participant. This process is repeated until every participant has been a test subject once.

Similarly, the LOMO scheme trains and evaluates seven classifiers, one per movie clip. For each run, the classifier is trained on samples from $N-1$ movie clips and evaluated on the n*th* movie clip.

Both LOPO and LOMO schemes are evaluated using global and segmented samples to quantify the impact of temporal segmentation on generalization performance.

Moreover, in the case of segmented samples, the LOMO scheme employs the *majority voting* decision rule, based on a *temporal-level aggregation* strategy, to determine the final prediction for each movie clip. In detail, for a given clip and participant, the outputs produced by the same model on multiple temporal segments of that clip are aggregated. Since the number of segmented samples is identical to the number of global samples, this aggregation also enables a consistent and robust comparison between the global and segmented sample approaches. This way, if the predicted labels are tied with the segmented approach, the clip is considered incorrectly classified.

#### 4) Ensembles Evaluation Protocol

Ensembles of classifiers are an architectural design used to study and improve the performance compared to standard and optimized classification models. In detail, the optimal classifiers are combined into ensembles using a *majority voting* decision rule. A *model-level aggregation* strategy is used, where predictions from different models are aggregated to produce the final output.

#### 5) Emotion-Anchored Evaluation Protocol

In addition to the other evaluation protocols, specific analyses are performed anchored to the temporal position of the emotion signals. In particular, the advantage of using only the first or only the last emotion expressed by each participant during the interaction is investigated. To do this, two conditions are defined: the First Emotion-only (FE-only) analysis and the Last Emotion-only (LE-only) analysis, where the models are evaluated by considering the initial or the final emotional state, respectively.

#### 6) Feature Importance Analysis via Permutation

The predictive contribution of the input signals to the machine learning models is quantified using the *Feature Permutation Importance* (FPI) technique and by measuring the *Average Cumulative Importance* (ACI) score. For each feature, FPI measures the decrease in classification performance when the feature values are randomly permuted across samples, while keeping all other features unchanged. This procedure is repeated across the different evaluation runs, and the ACI score summarizes the resulting importance values. This evaluation

analysis estimates the relative importance of the selected features.

#### 7) Metrics and Hardware

The model performance is reported by measuring the mean accuracy and standard deviation, along with weighted precision, weighted recall, weighted $F_1$-score, and 95% confidence interval. Note that the word "weighted" is abbreviated as "w" when used within tables. In addition, if the standard deviation or confidence interval is approximately zero, this information is not presented in the tables.

Analyses are performed using a personal computer with an Intel CPU i7-14900K, 64 GB of DDR5 DRAM, and an NVIDIA GPU 4070 Ti.

### F. DATA ANALYSIS

The data from the *Dataset DT22*, including VA signals and categorical emotions obtained through FER, *User Quality* score, *User Satisfaction* score, and clip metadata, are used to present further statistical analyses. In detail, the distributions of emotions and satisfaction are represented for each movie clip and participant; in addition, correlation analysis is performed between emotion metrics and satisfaction, both for individual movie clips and for the whole set of clips. Finally, the demographic profile of the participants is represented for the sake of completeness and transparency.

## VII. MODEL PERFORMANCE AND RESULTS

This section presents the main findings of the present study, summarized in Table 7. The results are structured to illustrate the path from feature selection to model evaluation, highlighting the performance of different classification policies and algorithms. Attention is given to evaluation strategies across participants and multimedia contents, which permits assessing the robustness and generalization capability of the proposed *ALPHACA* system.

### A. SELECTED FEATURE SETS

The optimal feature sets $F_i$ were found, where $i$ is associated with each dataset ($D_{2\text{-}3}$, $D_{2\text{-}1\text{-}2}$, and $D_{1\text{-}to\text{-}5}$), including both global and local features. These feature sets were selected to capture key characteristics from the sample data and define the model inputs. In detail, feature sets $F_{2\text{-}3}$, $F_{2\text{-}1\text{-}2}$, and $F_{1\text{-}to\text{-}5}$ were chosen from datasets $D_{2\text{-}3}$, $D_{2\text{-}1\text{-}2}$, and $D_{1\text{-}to\text{-}5}$, respectively. However, feature sets containing local features are considered for simplicity only. In Appendix A, Table A1 presents these feature sets, divided by emotion signal types, such as valence and arousal.

### B. EXTERNAL EVALUATION

The models' performance is evaluated and compared by measuring performance metrics for external evaluations. In Appendix B, Table B1 presents models evaluated by using samples with global and local features. These models used design policies presented in Section V-B and their derivatives.

Table B1(a) presents models evaluated using global features. The optimal baseline model is the Wide ANN under *Policy* $P_{2\text{-}3}$, which achieves an accuracy of 58% and $F_1$-score of 0.5823. Thus, its derivatives Bagged Trees and Linear Discriminant under policies $P_{CQ}$ and $P_{PQ}$ reach an accuracy of ~80% and $F_1$-score of about 0.7954. Other models, such as the Cubic $k$-NN and Coarse Tree, achieved lower performance.

In contrast, Table B1(b) presents models using local features, showing a significant performance improvement. The Optimized $k$-NN models achieved the highest accuracy and $F_1$-score of over 0.97 under the $P_{PCQ}$ and $P_{PC}$ policies, demonstrating the highest performance when *User Quality* score ($Q$) and other features, such as Participant ID ($P$) and Movie Clip ID ($C$), are considered. Similarly, the $k$-NNs and Trees models under different policies performed well, with an accuracy and $F_1$-score exceeding 0.76. In Appendix C, the Figs. C1-C4 illustrate the confusion matrices of the segmented models with *Policy* $P_{2\text{-}3}$, $P_{2\text{-}1\text{-}2}$, $P_{1\text{-}to\text{-}5}$ and $P_{CQ}$.

Instead, Table B2 presents the hyperparameters of optimized models, listed as "*attribute: value*" pairs. Standard models' hyperparameters are not presented because they are already cited in scientific literature and also used in the MATLAB Classification Learner App.

In conclusion, policies $P_{CQ}$ and $P_{2\text{-}1\text{-}2}$ with global and segmented samples are considered for further analysis. The policies $P_{PCQ}$ and $P_{PC}$ are ignored for reasons discussed in Section IX.

### C. LEAVE-ONE-PARTICIPANT-OUT (LOPO) EVALUATION

The optimal baseline policies are considered for training and evaluating 112 classifiers using the Leave-One-Participant-Out (LOPO) scheme, where a classifier is dedicated to each participant. The used classification policies are *Policy* $P_{CQ}$ (2-class) and *Policy* $P_{2\text{-}1\text{-}2}$. Their performance is presented in Table B3 in Appendix B, considering both the global and segmented samples.

Looking at the table, results indicate that classifiers evaluated with segmented samples outperform those evaluated on global samples.

In detail, the segmented 2-class classifier achieved an average accuracy higher than the global 2-class classifier, by achieving an accuracy of 84.12% and an $F_1$-score of 83.97%, compared to the 69.34% of accuracy and 68.90% of $F_1$-score. In comparison, the segmented and global 3-class classifiers' performance decreased to an average ~45% of accuracy and ~44% of $F_1$-score. However, both segmented and global classifiers reached the minimum and maximum accuracies of 0% and 100%.

### D. LEAVE-ONE-MOVIE-OUT (LOMO) EVALUATION

The optimal classification policies are considered for training and evaluating seven classifiers using the Leave-One-Movie-Out (LOMO) scheme, with one classifier dedicated to each movie clip. The classification policies are the *Policy* $P_{CQ}$

**TABLE 8. Optimal three models that achieved the highest level of performance with the optimal Policy $P_{CQ}$.**

| Model | Accuracy | wF1 (↓) |
|---|---|---|
| Fine $k$-NN | 0.8490 | 0.8436 |
| Bagged Trees | 0.8180 | 0.8165 |
| Weighted $k$-NN | 0.8130 | 0.8128 |
| **Average** | 0.8267 | 0.8243 |

(2-class) and *Policy* ($P_{2\text{-}1\text{-}2}$). In addition, the *majority voting* decision rule is used to measure the performance when segmented samples are also considered. Their performance is presented in Table B4 in Appendix B, considering both global and segmented samples.

Looking at the Table B4(a), which does not use the *majority voting* rule, the classifiers achieve a high average accuracy of ~80% and $F_1$-score of 0.7604. In this case, the classifier evaluated on the clip with ID 3 is the optimal-performing model, with an accuracy of 83% and an $F_1$-score of 0.8325, whereas the classifier evaluated on Movie Clip ID 9 shows the lowest performance, with an accuracy of ~71% and an $F_1$-score of 0.7124.

Instead, Table B4(b) shows the results when the *majority voting* decision rule is used, yielding a light improvement in average performance with an accuracy of 77% and an $F_1$-score of 0.7705. The highest level of accuracy of ~84% is observed for the classifier evaluated with Movie Clip ID 3, while Clip 9 remains the lowest with an accuracy of 70%.

In addition, Table B4(c) presents the results for the 3-class classification setting, where performance drops considerably compared to the 2-class models. The average accuracy decreases to 59%, also in both precision and recall.

Whereas Table B4(d) reports the performance of the classifiers with a 3-class design using the *majority voting*.

The results show that average performance remains lower than the 2-class case (see Table B4(b)), with an average accuracy of ~60% and an $F_1$-score of 0.6059. This confirms the increased complexity introduced by the 3-class classification, which makes it more difficult to separate classes and increases the risk of misclassification.

### E. ENSEMBLES OF CLASSIFIERS EVALUATION

To further investigate whether model aggregation can improve the performance of the optimal classifier identified in Section VII-B, different ensembles of classifiers are evaluated. All ensembles are designed under *Policy* $P_{CQ}$ and combine the predictions of multiple models through a *majority voting* decision rule. Three complementary ensemble configurations are considered:

- *Three-Model Ensemble* (denoted $E_1$). After the analyses presented in Table 7, the ensemble is composed of the three top-performing classifiers under the *Policy* $P_{CQ}$, as listed in Table 8. Their predictions are aggregated by *majority voting* decision rule. The ensemble is evaluated statistically, and its performance is reported by measuring accuracy and weighted $F_1$-score.
- *Ensemble*$_{6/1}$ (denoted $E_2$). For each held-out movie clip $c \in C = \{3, \ldots, 9\}$, an ensemble of six classifiers is formed by excluding the *i*-classifier, *i.e.*, the classifier that achieved the optimal evaluation performance on clip $i$ (see Table B5 in Appendix B). The ensemble is trained on six clips $C \setminus \{c\}$ and evaluated on the clip $c$. Final predictions are obtained by *majority voting* decision rule. For example, given $c = 7$, the ensemble is trained on $\{3, 4, 5, 6, 8, 9\}$, and evaluated on clip 7. Per-clip results are reported in Table B6.
- *Ensemble*$_7$ (denoted $E_3$). For each held-out movie clip $c \in \{3, \ldots, 9\}$, an ensemble of seven classifiers is formed (see Table B5), training it on the union of the remaining six clips $C \setminus \{c\}$. Therefore, the ensemble is evaluated on the held-out clip $c$ and the final prediction is obtained by *majority voting* decision rule. For example, when $c = 7$, the ensemble is trained on $\{3, 4, 5, 6, 8, 9\}$ and their votes are aggregated to predict the samples on clip 7. Per-clip results for the *cth* held-out clip are reported in Table B7.

Starting with the *Three-Model Ensemble*, the accuracy of models, as probability metrics, is $P(A) = 0.8490$, $P(B) = 0.8180$, and $P(C) = 0.8130$. The ensemble accuracy $P(E_1)$ is measured statistically as in 12:

$$\begin{aligned} P(E_1) &= P(A) \times P(B) \times P(C) \\ &\quad + P(A) \times P(B) \times (1 - P(C)) \\ &\quad + P(A) \times P(C) \times (1 - P(B)) \\ &\quad + P(B) \times P(C) \times (1 - P(A)) \\ &= P(A \cap B) + P(A \cap C) + P(B \cap C) \\ &\quad - 2 \times P(A \cap B \cap C) \end{aligned} \tag{12}$$

Since probability independence is assumed between the predictions, the probability that two models correctly predict the output is:

$$\begin{aligned} P(A \cap B) &= P(A) \times P(B) = 0.8490 \times 0.8180 = 0.6945; \\ P(A \cap C) &= P(A) \times P(C) = 0.8490 \times 0.8130 = 0.6902; \\ P(B \cap C) &= P(B) \times P(C) = 0.8180 \times 0.8130 = 0.6650; \end{aligned}$$

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \times P(B) \times P(C) \\ &= 0.8490 \times 0.8180 \times 0.8130 = 0.5646. \end{aligned}$$

Thus, the accuracy of $P(E_1)$ is equal to:

$$\begin{aligned} P(E_1) &= 0.6945 + 0.6902 + 0.6650 \\ &\quad -2 \times 0.5646 \\ &\approx 0.9205 \ (92.05\%). \end{aligned}$$

Hence, the ensemble accuracy $P(E_1)$ is 92.05%, while $F_1$-score is 0.9184 (see Fig. C5 in Appendix C).

Whereas the mean ensemble accuracy $P(E_2)$ of the *Ensemble*$_{6/1}$ is 69.95%, as presented in Table B6. In particular, its results indicate an average accuracy of ~70% and an $F_1$-score of 0.6875, with individual clip accuracy ranging from

~60% (Clip 4) to ~78% (Clip 8). This suggests that some movie clips provide more consistent patterns for classification, whereas others introduce variability that affects accuracy. In addition, their precision values vary from the highest precision of 0.8133 for Clip 6 to the lowest of 0.5724 for Clip 4, reinforcing the idea that some clips are more informative for classification.

Instead, the $Ensemble_7$ mean accuracy $P(E_3)$ of 73% is represented in Table B7, which outperforms $Ensemble_{6/1}$, achieving also a higher average $F_1$-score of 0.7263. This improvement is observed across most clips, in particular in Clips 3 and 4 with an accuracy of ~73% and ~79%, respectively.

### F. EMOTION-ANCHORED EVALUATION

The method presented in Section V is reconsidered to build classifiers using segmented samples only. In this case, participants' first or last emotions are considered from each movie clip only. Therefore, other emotional states are ignored. Machine learning algorithms are trained and evaluated for First Emotion-only (FE-only) and Last Emotion-only (LE-only) policy designs based on *Policy* $P_{\text{CQ}}$. The models' performance is presented in Table B8 in Appendix B.

In particular, Table B8(a), which presents models able to predict the UX by first emotion, the Fine Tree model achieved the highest accuracy and $F_1$-score of over 0.78. Instead, the Kernel Naive Bayes and Bagged Trees models performed worse, with an accuracy and $F_1$-score of ~0.46, respectively.

Whereas Table B8(b) focuses on models able to use the last emotion. The Medium ANN model outperforms the other LE-only models, reaching an accuracy and $F_1$-score of ~0.78. Meanwhile, Subspace $k$-NN models under $P_{2\text{-}1\text{-}2}$ and $P_{1\text{-to-}5}$ policies achieved lower accuracy and $F_1$-score, indicating that they were not able to capture relevant emotion patterns.

These results indicate that the first and last emotions have an impact on model performance, compared to models presented in Table B1.

### G. EMOTIONAL AND CONTEXTUAL FEATURE IMPORTANCE

The predictive power of VA signals, along with contextual information including Movie Clip ID, and *User Quality* score, is assessed using the *Feature Permutation Importance* (FPI) technique [38]. This FPI technique is based on the out-of-bag random forest algorithm. It measures the impact of randomly shuffling feature values and evaluates model performance across multiple decision trees. Thus, features that reduce model performance influence model predictions.

In this study, an *importance score* is computed for each feature 30 times, and then their average and variance are measured. This permitted the computation of the *Average Cumulative Importance* (ACI) score for each signal as the sum of the average of the signal's features.

The dataset $D_{2\text{-}3}$ is used to measure the ACI score of the feature set $F_{2\text{-}3}$ with *Policy* $P_{\text{CQ}}$. The Fig. 9 shows the ACI score assigned to each signal. In addition, Table 17 presents the correlation coefficients between this feature set and the target class.

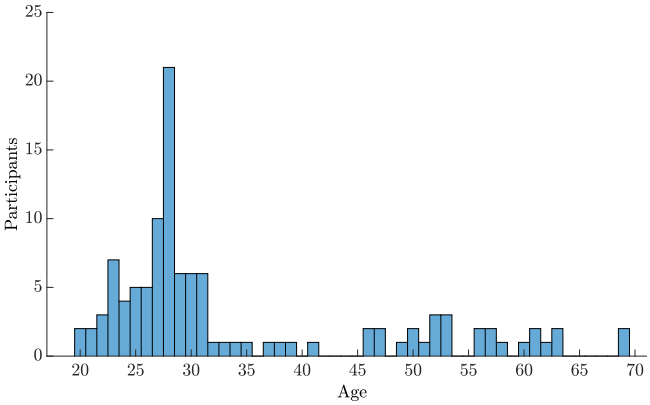


**FIGURE 6.** Age distribution of the participants involved in the study.

Consistent with the ACI scores in Fig. 9, features reveal their relative contribution: contextual information accounts for 48% of the total importance, emerging as the most informative group. In the $P_{\text{CQ}}$ policy this corresponds to an external accuracy/$F_1$-score of $\approx$ 85% (see Table B1(b)). Within the contextual feature subset, the *User Quality* score shows the strongest positive correlation with the target ($\approx$ 0.50), while the *Movie Clip ID* has negligible linear correlation ($\approx$ 0.03), suggesting a non-linear contribution. Valence and arousal contribute 33% and 19%, respectively; when used without contextual features, as in $P_{2\text{-}3}$, they reach $\approx$ 76% accuracy/$F_1$-score, confirming their relevance, though smaller than contextual information.

## VIII. DESCRIPTIVE ANALYSIS OF PARTICIPANT DATA

This section complements previous results evaluations by providing a descriptive and correlational analysis of the data within the *Dataset DT22*. Participants' characteristics are examined, self-reported emotions, perceived quality, and user satisfaction are distributed and interrelated.

### A. PARTICIPANTS AND DATA COLLECTION

A total of 112 participants (35 women and 77 men, mean age $\sim 34 \pm 13$ years, range = 20-69) were involved in the experimentation. They signed the informed consent and then completed the *registration questionnaire*. The questionnaire allowed for the collection of participants' age, highest educational level, computer expertise, and knowledge of video-on-demand services. Participants were Italian; recruitment included volunteers from the general population as well as students and employees of the University of Pisa. Most participants were younger than 30 years, with ages ranging from 20 to 69 years. These characteristics provide useful heterogeneity for the research findings, while noting that the sample reflects an Italian population predominantly of young adults.

The distribution of participants' ages is shown in Fig. 6. The participant population has an average age of 34.27 years and a standard deviation of 12.73 years. In addition, their age is also represented in Table 9, where they are divided into age

**TABLE 9. Distribution of participants grouped by age range.**

| Age Range | N. Participants |
|---|---|
| 20-29 | 65 (58.04%) |
| 30-39 | 19 (16.96%) |
| 40-49 | 6 (5.36%) |
| 50-59 | 14 (12.5%) |
| 60-69 | 8 (7.14%) |
| **Total** | 112 (100%) |

**TABLE 10. Distribution of participants by highest educational level attained.**

| Educational Level | N. Participants |
|---|---|
| Middle School Diploma | 3 (2.68%) |
| High School Diploma | 44 (39.29%) |
| Bachelor's Degree | 24 (21.43%) |
| Master's Degree | 36 (32.14%) |
| Doctor's Degree | 5 (4.46%) |
| **Total** | 112 (100%) |

**TABLE 11. Distribution of participants by level of computer experience.**

| Proficiency in Computer Systems | N. Participants |
|---|---|
| Beginner | 17 (15.18%) |
| Competent | 50 (44.64%) |
| Expert | 45 (40.18%) |
| **Total** | 112 (100%) |

**TABLE 12. Distribution of participants by familiarity with video-on-demand services.**

| Use of Streaming Platforms | N. Participants |
|---|---|
| Never | 11 (9.82%) |
| A little | 13 (11.60%) |
| Enough | 30 (26.79%) |
| Very | 42 (37.50%) |
| Very much | 16 (14.29%) |
| **Total** | 112 (100%) |

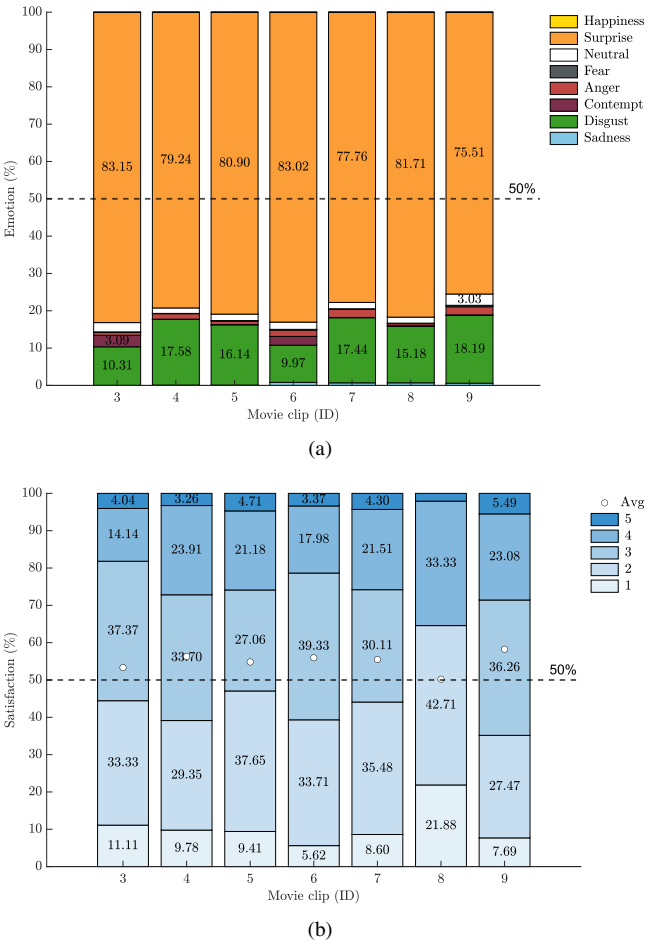


**FIGURE 7. User emotion (a) and User Satisfaction score (b) declarations for each movie clip.**

ranges. According to the table, 72.33% of the participants are between 20 and 35 years old, while 27.67% are older.

Instead, the distribution of the last educational level, computer expertise, and knowledge of video-on-demand services is presented in Tables 10–12, respectively. Looking at Table 10, 41.97% of participants hold only a Middle School Diploma or High School Diploma, whereas the remaining 58.03% have a University Degree. In Table 11, 15.18% declared themselves to be not proficient in computer usage, while 84.82% expressed their ability to use computers. In conclusion, as presented in Table 12, 9.82% declared that they have never used video-on-demand services, while 90.18% have used these services at least once.

### *B. DISTRIBUTIONS OF EMOTIONS AND SATISFACTIONS*

User emotions and *User Satisfaction* scores characterize the movie clips, describing the participant's affective responses elicited by the multimedia content.

To better understand the relationship between user emotion and the *User Satisfaction* score, both measures are analyzed for each video clip, and their relationship is illustrated in Fig. 7.

In particular, Fig. 7(a) represents how the FER model predicted the emotions for each movie clip. The figure shows that the FER model frequently detected surprise and negative emotions.

While Fig. 7(b) illustrates how the *User Satisfaction* score, which ranges from 1 to 5, is distributed across clips. In addition, the clip's average score is also highlighted. As illustrated, ratings of 2 and 3 are the most frequent (higher than 27%). Moreover, the average satisfaction score per clip is around 3, highlighting a neutral evaluation from participants.

Further insights are provided in Fig. 8, which highlights the relationships between the average *User Satisfaction* score and FER emotions, with a 95% confidence interval on the mean satisfaction score. The chart reveals a counterintuitive pattern: negative valence emotions, such as disgust, contempt, anger, and fear, are associated with higher satisfaction scores than positive emotions, including surprise and happiness. Among all emotions, sadness corresponds to the lowest satisfaction score, whereas contempt and fear show the highest. These findings suggest that intense, emotionally charged experiences—even when characterized by negative valence—may be perceived as more engaging and thus more satisfying by

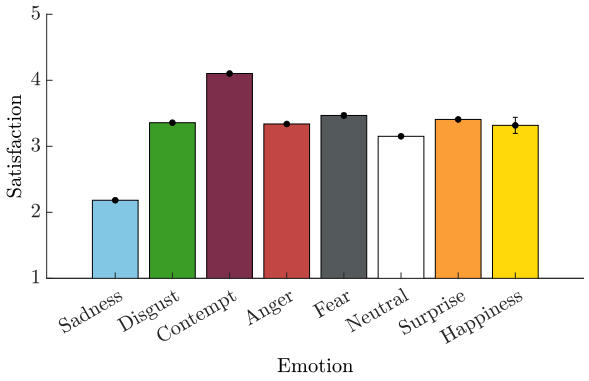


**FIGURE 8.** Average User Satisfaction score with a confidence interval of 95% for each emotion recognized by the FER model.

**TABLE 13.** Distribution of emotions recognized by the FER model across all participants.

| Emotion | Frequency (↓) |
|---|---|
| Surprise | 2,404,638 (80.55%) |
| Disgust | 439,265 (14.71%) |
| Neutral | 58,097 (1.95%) |
| Anger | 41,180 (1.38%) |
| Contempt | 26,964 (0.90%) |
| Sadness | 10,805 (0.36%) |
| Fear | 4,272 (0.14%) |
| Happiness | 255 (0.01%) |
| **Total** | 2,985,476 |

**TABLE 14.** Distribution of User Satisfaction score reported by participants via questionnaire.

| User Satisfaction score | N. Questionnaires (↓) |
|---|---|
| 2 | 221 (34.26%) |
| 3 | 219 (33.95%) |
| 4 | 113 (17.52%) |
| 1 | 69 (10.70%) |
| 5 | 23 (3.57%) |
| **Total** | 645 (100%) |

**TABLE 15.** Pearson Correlation Coefficient (PCC) between emotion signals (valence, arousal, and emotion) and User Satisfaction score, considering both individual clips (local) and the whole set of clips (global).

| Movie Clip ID | Valence | Arousal | Emotion | Satisfaction |
|---|---|---|---|---|
| 3 | 1 | | | |
| | -0.4049 | 1 | | |
| | 0.6289 | -0.5209 | 1 | |
| | 0.1578 | -0.0309 | 0.0972 | 1 |
| 4 | 1 | | | |
| | -0.5235 | 1 | | |
| | 0.6233 | -0.5815 | 1 | |
| | 0.0982 | -0.0053 | 0.0771 | 1 |
| 5 | 1 | | | |
| | -0.4324 | 1 | | |
| | 0.6511 | -0.5463 | 1 | |
| | 0.0331 | 0.0545 | 0.0821 | 1 |
| 6 | 1 | | | |
| | -0.3533 | 1 | | |
| | 0.6435 | -0.4079 | 1 | |
| | 0.2743 | -0.0755 | 0.1490 | 1 |
| 7 | 1 | | | |
| | -0.4805 | 1 | | |
| | 0.6893 | -0.5417 | 1 | |
| | -0.0475 | 0.0095 | 0.0121 | 1 |
| 8 | 1 | | | |
| | -0.4428 | 1 | | |
| | 0.6452 | -0.4954 | 1 | |
| | -0.0505 | 0.0199 | -0.0857 | 1 |
| 9 | 1 | | | |
| | -0.4409 | 1 | | |
| | 0.6823 | -0.5268 | 1 | |
| | 0.0361 | 0.1916 | -0.0116 | 1 |
| **Global** | 1 | | | |
| | -0.4314 | 1 | | |
| | 0.6525 | -0.5212 | 1 | |
| | 0.0140 | 0.0327 | 0.0142 | 1 |

**TABLE 16.** Correlation coefficients between User Quality and User Satisfaction scores.

| Coefficient | Value |
|---|---|
| Pearson | 0.5550 |
| Spearman | 0.5443 |
| Kendall | 0.4788 |

users.

### C. ANALYSIS OF EMOTION METRICS AND SATISFACTION

The analysis of participants' emotional responses and self-reported satisfaction provides a descriptive overview of the dataset.

Facial recordings were processed through the FER model, which extracted both dimensional signals (valence and arousal) and categorical emotions, *i.e.*, sadness, disgust, contempt, anger, fear, neutral, surprise, and happiness. Categorical emotions are chosen for descriptive analysis because they are more intuitive to interpret than continuous valence–arousal values. The distribution of these emotions is reported in Table 13, while participants' answers to the question *"How satisfied are you with watching this video?"* are summarized in Table 14.

The first table shows that the emotion *surprise* is the dominant emotion, representing 80.55% of detected participants' face expressions. However, as presented in Table 3, some clips must evoke positive emotions statistically. Moreover, 1.95% of the expressions are classified as neutral. This emotion represents moments in which participants did not show other specific emotions.

Whereas the second table shows that *User Satisfaction* score has a non-uniform distribution, with the prevalence of low ratings. Indeed, the most frequent score is 2, followed by 3, suggesting that most participants found the videos dissatisfying or neutral. Only 3.57% of participants declared the maximum score of 5, indicating a very low overall level of satisfaction. In total, 645 satisfaction questionnaires were completed out of 784 (112 participants $\times$ 7 clips), as some clips were skipped or not fully completed [38], as anticipated in Section IV-C.

### D. RELATIONSHIP BETWEEN EMOTION METRICS AND SATISFACTION

The relationship between emotion metrics—namely valence, arousal, and categorical emotions—and the *User Satisfaction*

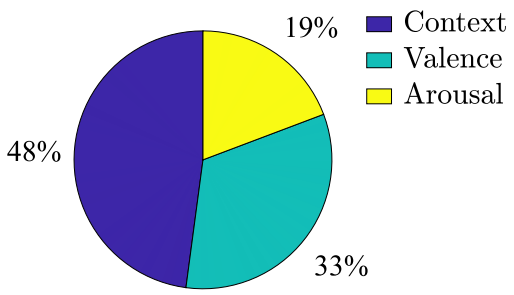


**FIGURE 9. Average Cumulative Importance (ACI) score with contextual information, valence, and arousal signals, using the optimal Policy $P_{CQ}$.**

score is examined through correlation analysis. The Pearson Correlation Coefficient (PCC) is computed both at the local level (per individual video clip) and at the global level (considering the entire dataset). In addition, to assess the statistical significance of these correlations, a Paired Sample Student's *t*-test is applied with a significance threshold of $\alpha = 0.05$. The null hypothesis ($H_0$) assumed that the observed correlation is not significantly different from zero. The full set of results is reported in Table 15.

The analysis shows a strong positive correlation between emotions and valence (*e.g.*, > 0.62 across clips), whereas the correlation between emotions and arousal is negative (*e.g.*, < –0.40). This confirms that the extracted emotion categories are more representative of the valence dimension than of arousal.

By contrast, the correlation between *User Satisfaction* scores and emotion metrics remains weak in all cases. None of the coefficients between satisfaction and valence, arousal, or categorical emotion exceeds 0.2, with many values close to zero. This indicates that while emotions and valence/arousal are interrelated, their direct linear relationship with *User Satisfaction* score is limited.

### E. RELATIONSHIP BETWEEN QUALITY AND SATISFACTION

The relationship between the perceived quality of a movie clip (*User Quality* score) and the corresponding *User Satisfaction* score represents an important aspect of audiovisual perception. Exploring this link is essential for refining predictive models and guiding the design of multimedia content toward more engaging and satisfying experiences.

The correlation analysis, summarized in Table 16, confirms a positive association between quality and satisfaction, though with varying strength depending on the statistical measure employed. In particular, the Pearson coefficient of ~0.56 indicates a moderate linear relationship, while the Spearman coefficient of ~0.54 supports the presence of a positive monotonic trend. The Kendall coefficient of ~0.48 is a little lower, suggesting that the connection between perceived quality and satisfaction may also be shaped by non-linear factors and subjective influences that traditional linear models do not fully capture.

## IX. DISCUSSION

This section interprets the findings by analyzing the contribution of emotion and contextual features, the robustness of the models across participants and video clips, and the role of signal temporal dynamics. The findings are compared with prior work, highlighting both strengths (*e.g.*, high accuracy with contextual features) and limitations (*e.g.*, reduced generalization). Implications for the design of more advanced systems are also discussed.

### A. KEY FINDINGS ON MODEL PERFORMANCE

The analysis of the selected features provides insights into the emotion signals and contextual information that contribute to predicting user satisfaction. Valence features are represented across all datasets as shown in Table A1 in Appendix A, highlighting their relationship with UX. In particular, the minimum valence value in both $F_{2\text{-}1\text{-}2}$ and $F_{1\text{-}to\text{-}5}$ indicates that the lowest emotional state experienced by participants is an important predictor of their final satisfaction. In the same way, the presence of peak valence across feature sets reinforces the role of emotional intensity in shaping user perceptions. On the contrary, arousal appears less frequent and with lower discriminative power, suggesting a weaker contribution to UX prediction compared to valence. Furthermore, the presence of moment- and quantile-based features suggests that non-linear patterns and asymmetries in the emotion signals are also relevant for classification.

The *Feature Permutation Importance* (FPI) analysis confirms that, although VA signals contribute to prediction, contextual information is the dominant factor shaping model performance. The ACI results show that contextual information, such as Movie Clip ID and *User Quality* score, accounts for 48% of the overall importance, compared with 52% from valence and arousal.

Moreover, valence proved more predictive than arousal [38], [52], a result that contrasts with studies where arousal better explains momentary engagement and related outcomes such as attentional capture, presence, memorability, and sharing [59]–[63].

Although prior works often assign arousal a stronger role, the target presented in this research study is to predict a final hedonic judgment (user satisfaction or UX), which is more aligned with valence. Moreover, in the 60 s time windows, arousal spikes tend to be noisier and less stable than valence trajectories; together with the data distribution, this explains why valence is a stronger predictor of satisfaction than arousal (see Table A1). This is consistent with the FPI analysis, where valence features yield larger average performance drops when permuted across all splits.

From a performance point of view, as presented in Section VII, some models achieved low accuracy and $F_1$-score, confirming the limited effectiveness when global signals are used. By contrast, other models performed better, demonstrating the advantage of using segmented signals for capturing the temporal dynamics of emotional responses.

**TABLE 17. Correlation between each feature in the optimal set $F_{2\text{-}3}$ selected under Policy $P_{CQ}$ and the target class.**

| | C | Q | F1 | F2 | F3 | F4 | F5 | Target |
|---|---|---|---|---|---|---|---|---|
| **Movie Clip ID (C)** | 1 | | | | | | | |
| **User Quality Score (Q)** | 0.0924 | 1 | | | | | | |
| **Maximum (F1)** | -0.0334 | 0.0210 | 1 | | | | | |
| **Mean Squared (F2)** | 0.0488 | 0.0037 | -0.5409 | 1 | | | | |
| **Impulse Factor (F3)** | 0.0724 | 0.0623 | -0.0027 | 0.6900 | 1 | | | |
| **Geometric Mean (F4)** | 0.0013 | -0.0001 | -0.2795 | 0.4494 | 0.2405 | 1 | | |
| **4*th*-order Momentum (F5)** | 0.0130 | 0.0210 | -0.2487 | 0.4559 | 0.2990 | 0.8196 | 1 | |
| **Target** | 0.0328 | 0.4969 | 0.1341 | 0.0211 | 0.1001 | -0.0466 | -0.0108 | 1 |

The highest performance is obtained by *ablation models* that utilize contextual features, including Participant ID (*P*), Movie Clip ID (*C*), and/or *User Quality* score (*Q*). In particular, the Optimized *k*-NNs under the $P_{PCQ}$ and $P_{PC}$ policies achieved an average $F_1$-score above 0.97, highlighting the improvement obtained when contextual features complement emotion signals. However, these two ablation designs may be influenced by data leakage because the Participant ID (*P*) is used. In addition, the (participant, clip) pair appears in both internal and external datasets.

Instead, the $P_{CQ}$ model provided a balanced trade-off between accuracy and robustness by combining valence–arousal signals with the Movie Clip ID and the *User Quality* score. In the external evaluation, this design achieved an $F_1$-score of approximately 0.84, further improved to about 0.92 by the Three-Model Ensemble under the $P_{CQ}$ policy (Table 7), while avoiding the nearly perfect yet likely overfitted performance of the $P_{PCQ}$ configuration (F1 $\approx$ 0.99). This design does not incur data leakage because the evaluation ensures that no (participant, clip) pair appears in both internal and external datasets. Hence, the model cannot identify the same user watching the same clip, and the Movie Clip ID and *User Quality* score cannot serve as a unique link to the label. This makes the $P_{CQ}$ design more reliable than *identity-free* mode models, which suffer from lower performance, and more generalizable than participant-dependent models, whose applicability in real-world scenarios with new users or unseen multimedia content is limited.

In conclusion, the correlation analysis provides insight into the relationship between *User Quality* score and UX. Pearson and Spearman coefficients of 0.5550 and 0.5443, respectively, suggest a moderate correlation between UX and *User Quality* score, whereas Kendall's coefficient of ~0.48 highlights a moderate non-linear dependency. These indicate that user satisfaction is a multidimensional metric influenced not only by emotion signals but also by cognitive and contextual factors [38], [52]. This reinforces the importance of integrating contextual features into UX models, moving beyond *identity-free* mode models that have often dominated prior studies.

### B. VALIDATION ACROSS SCHEME ANALYSIS

The performance analysis performed using the Leave-One-Participant-Out (LOPO) and Leave-One-Movie-Out (LOMO) schemes provided insights regarding system robustness and generalization of the proposed models (see Table 7).

LOPO results demonstrate that the model can generalize across different participants, especially when segmentation is applied. In this case, the segmented 2-class classifier achieved average accuracies above 84%, compared to the global 2-class classifier. This highlights the importance of capturing temporal information for predicting UX, enabling the model to be applied to new users without requiring retraining.

The LOMO scheme confirms the models' robustness across different movie clips, it also reveals the difficulty of multi-class prediction. In the 2-class policy, classifiers achieved average accuracy and $F_1$-score around 0.77 when using *majority voting* rule, demonstrating good performance across clips. By contrast, the introduction of a third class led to a drop in performance, with accuracy and $F_1$-score around 0.60 even with *majority voting* rule. These indicate that the additional class introduces model complexity, while the *majority voting* provides only marginal improvements. Therefore, the choice of classification policy (2-class vs. 3-class) is critical and should be guided by the specific application context.

These findings confirm that while the proposed models generalize well across both participants and video contents, especially in binary classification policy, their performance is challenged by more complex multi-class scenarios.

### C. TEMPORAL DYNAMICS AND EMOTION SIGNALS

The experiments on temporal dynamics highlight the importance of considering when emotions occur during the viewing of a clip. In particular, results show that the Fine Tree and Medium ANN are the most reliable models for both first and last emotion-based classifiers, with the latter yielding better performance (see Table B8). This suggests that both models are informative for predicting *User Satisfaction*.

Participant data revealed that the *surprise* emotion was the predominant emotion across all clips, with frequencies ranging from 75% to 83%. This suggests that the video stimuli elicited reactions of unexpectedness or wonder. Negative emotions such as disgust, anger, contempt, and sadness appeared less frequently. However, disgust emerged as the second most frequent emotion, due to the presence of unpleasant or provocative content in some clips. Instead, happiness was almost absent, suggesting limited reactivity among the participants.

Instead, the correlation analyses show the complexity of inferring *User Satisfaction* from emotions. The relationship

between valence and arousal was negative across clips, confirming affective models in which highly activated states are often associated with neutral or negative valence. However, the PCC metric between facial emotion signals and satisfaction scores was generally low (PCC < 0.2). This indicates that *User Satisfaction* is only weakly and linearly related to raw emotion signals. Such weak correlations can be attributed to the multidimensional nature of satisfaction, which also depends on cognitive engagement, context, and subjective preferences. In addition, measurement noise due to factors such as lighting conditions, camera angle, or participant posture may have further attenuated the correlations.

Despite these low linear correlations, the intelligent models developed in this study can leverage multivariate, nonlinear signal patterns by combining valence and arousal with contextual features such as Participant ID or Clip ID, and *User Quality* score. This integration allowed the models to maintain high predictive performance, even when simple correlations appeared weak.

### D. ENSEMBLE MODELS FOR ROBUSTNESS

Ensemble schemes prove the advantage of combining classifiers to improve robustness and accuracy compared to single models, as presented in Table 7. While the optimal ablation single classifier as Fine $k$-NN under $P_{\text{CQ}}$ achieved a performance of $F1 \approx 0.84$, ensemble schemes provided additional benefits in terms of stability and generalization across different movie clips.

The *Three-Model Ensemble*, built by aggregating the three optimal classifiers under the $P_{\text{CQ}}$ policy, reached an accuracy of 92.05% and an $F_1$-score of 91.84%. This result confirms that the *majority voting* decision rule increases robustness, reducing the influence of individual misclassification.

The $Ensemble_{6/1}$, which excluded the movie clip used as a test clip, achieved an average accuracy of ~70% ($F1 = 0.69$). This relatively lower performance highlights the dependence of the ensemble on the training data distribution and suggests that some clips (*e.g.*, Clip 8) provide more consistent predictive signals than others (*e.g.*, Clip 4).

By contrast, the $Ensemble_7$ improved upon the $Ensemble_{6/1}$, reaching an average accuracy of 73% ($F1 = 0.73$). The additional classifier trained on a separate clip improved the system's generalization, demonstrating that increasing model diversity within an ensemble can lead to more reliable performance.

These results indicate that ensembles can outperform single classifiers in terms of robustness when data variability is high across movie clips. Nevertheless, ensemble effectiveness strongly depends on the diversity and informativeness of the training data. Thus, ensembles with segmentation that use *majority voting* decision rule represent a solution for UX prediction. This solution may be particularly useful when models must generalize to unseen users and multimedia content.

### E. CONTEXTUAL FEATURES AND USER SATISFACTION

An important result of this study concerns the role of contextual features, such as Movie Clip ID and *User Quality* score, in shaping model performance. FPI analysis revealed that these contextual features accounted for 48% of the predictive power, whereas VA signals alone accounted for 52%. This indicates that, while emotion dynamics contribute to UX estimation, contextual features are the primary drivers of accurate predictions.

In addition, the relationship between *User Quality* score and *User Satisfaction* score reinforces findings about contextual features. Correlation analysis revealed moderate associations, with Pearson and Spearman coefficients of 0.5550 and 0.5443, respectively, and a Kendall coefficient of 0.4788, suggesting the presence of non-linear dependencies. These results suggest that a higher perceived quality corresponds to greater *User Satisfaction*; however, this relationship cannot be fully captured by simple linear models. Instead, additional subjective and contextual dimensions—such as cognitive engagement, expectations, and prior experiences—are likely to shape final satisfaction outcomes.

These findings have two significant implications for real-world predictive systems. First, they demonstrate that contextual features are important for building reliable models to predict the *User Satisfaction* score. Second, they highlight the need for advanced machine learning techniques able to capture non-linear interactions between quality, context, and emotion signals. Without such integration, systems based on affective cues only risk underestimating or misrepresenting the *User Satisfaction*.

### F. ETHICAL AND PRIVACY CONSIDERATIONS

Continuous inference of users' affective states can enable useful, adaptive experiences, yet it also raises ethical concerns around implicit profiling, transparency, informed consent, and the risk of misuse in commercial or other contexts, also using large-scale emotion recognition models.

In this work, these issues are addressed by aligning the technical claims and the evaluation protocol with privacy-preserving principles. Main results are based on *identity-free* mode: the primary models rely only on valence–arousal signals, while contextual features are based on ablation analyses (*context-aware* mode).

From a data protection standpoint, the processing pipeline follows a privacy-by-design approach. Inference can run on-device without requiring persistence of raw frames; when logging is enabled, only anonymized feature descriptors and minimal metadata are stored, and no biometric templates or identifiable video are needed for the reported results. This data minimization reduces re-identification risk and aligns the system's utility with proportionate data use. In practical deployments, these principles translate into clear disclosures about what is inferred and why, user control over activation and data retention, and default settings that prefer local processing when feasible.

Instead, considering the deployment, the *identity-free* and *context-aware* pipelines correspond to different privacy regimes. The *identity-free* mode is intended for privacy-critical scenarios in which storing persistent identifiers or detailed viewing histories is not acceptable (*e.g.*, consumer media platforms, classrooms, or clinical waiting rooms). In such cases, *ALPHACA* can be executed on-device, and logs can be restricted to short-lived by using anonymized feature descriptors.

By contrast, the *context-aware* mode requires trust assumptions, since it relies on features such as Movie Clip ID, *User Quality* scores, and pseudonymous Participant ID. This mode is appropriate for supervised environments where explicit consent, identity management, and clear data-governance policies are in place, such as UX laboratories, company-internal testbeds, or opt-in beta programs. In these contexts, linking multiple sessions for the same user can be justified by a clear benefit and transparent data policies.

### G. CONTRIBUTION TO THE STATE-OF-THE-ART

The results of this study advance the State-of-the-Art in UX prediction by demonstrating the benefits of integrating emotion signals with contextual features in machine learning models. Indeed, prior works relying solely on facial emotion recognition or multimodal biosignals achieved promising but limited performance, typically reporting accuracies in the range of 68–98% depending on the setting (see Table 1). For example, approaches based on EEG and physiological measures achieved accuracies around 76% [19], while FER-based models combined with additional biosignals or task ratings reported accuracies between 78% and 97% [23], [24]. However, these methods were either constrained by intrusive hardware setups, relied exclusively on momentary self-reports, or lacked the capacity to model temporal and contextual dynamics.

By contrast, the proposed ablation $P_{\mathrm{CQ}}$ model reached an accuracy and $F_1$-score above 0.84 by combining valence–arousal signals with contextual features such as Movie Clip ID and *User Quality* score. Moreover, ensemble schemes further improved robustness, with the *Three-Model Ensemble* system achieving an accuracy and $F_1$-score of 92.05%. These results demonstrate that contextual features complement affective signals and also act as a decisive factor in improving classification robustness and accuracy. Moreover, the use of LOPO and LOMO evaluation schemes confirmed that the models are able to generalize to unseen users and content, a requirement for practical deployment in intelligent systems.

Moreover, the results show that one can choose between a privacy-preserving *identity-free* mode, which already achieves competitive performance using only VA signals, and a higher-performing *context-aware* mode that exploits contextual features under trust assumptions.

This contribution challenges the common assumption in prior studies that emotional responses alone are sufficient to predict *User Satisfaction*. Instead, the results highlight that UX is a multidimensional metric shaped by affective, contextual, and quality-related factors. Integrating these elements yields predictive systems that are both more accurate and more adaptable to real-world variability.

At the same time, the very high performance achieved in this work raises important considerations about model generalization and information leakage from contextual features. Although cross-validation and external evaluations confirmed model robustness, the reliance on contextual IDs may limit applicability in scenarios with new users or unseen multimedia content. These limitations, along with issues such as emotion imbalance and satisfaction rating distribution, underscore the need for future research to broaden datasets, explore multimodal signals, and develop models less dependent on participant- or content-specific identifiers.

Thus, while this study establishes a clear advancement over the existing literature, it also opens avenues for further work on generalization, multimodality, and dataset diversity, which are discussed in the following section.

## X. LIMITATIONS

Despite the results presented in this study, different limitations must be acknowledged.

First, a certain degree of imbalance was observed in the distribution of both emotion labels and satisfaction ratings collected during the experiment. In particular, the emotion *surprise* accounted for more than 80% of the facial expressions detected by the FER model, while other emotions—such as happiness, sadness, anger, disgust, fear, and neutrality—were far less represented. Similarly, satisfaction scores tended to cluster around intermediate-to-lower values between scores 2 and 3, with fewer than 4% of responses corresponding to the highest level of satisfaction (score 5). While these distributions reflect the specific characteristics of the selected stimuli and participants' responses, they may affect the generalization of the proposed approach in broader contexts. Specifically:

- Reduced variability in emotion inputs may limit the classifier's ability to discriminate among satisfaction levels when other affective states are more prominent.
- The dominance of a single emotion may bias the learned associations between emotion and satisfaction, reducing model robustness when confronted with more diverse or balanced affective profiles.
- Applicability to real-world scenarios may be limited, as users typically express a broader spectrum of affective responses than those elicited in the present experimental setting.

Another limitation affects the risk of model overfitting. Although the *ablation models* achieved very high performance, for example, an accuracy and F1 $\approx 0.99$ with Policy $P_{\mathrm{PCQ}}$, such values may raise concerns regarding model generalization. To mitigate this, additional evaluations were carried out using external datasets, including Leave-One-Participant-Out (LOPO), Leave-One-Movie-Out (LOMO), and ensemble schemes, all of which showed consistent performance. Furthermore, ablation analyses demonstrated that using any

subset of contextual features (Participant ID, Movie Clip ID, and *User Quality* score) decreased performance from ~99% to ~66%. This confirms their importance but also highlights, in some cases, the information leakage between internal and external datasets, likely due to the use of Participant IDs. Moreover, while contextual features improve predictive accuracy, they may also limit generalization to new users or unseen multimedia content, especially when Participant ID is included.

A further limitation concerns the study's technological and methodological assumptions. The *ALPHACA* system was created and evaluated using a single FER model [29]. Therefore, the reported performance in Table 7 may reflect this model's biases and may not generalize to other FER implementations, camera setups, or lighting conditions. In addition, the experimentation was conducted on a controlled web-based platform with predefined movie clips, which does not capture the variability of real-world usage. Moreover, the *User Satisfaction* score was measured using the *video questionnaire*, which served as a subjective UX rating. These aspects highlight the need to validate *ALPHACA* in more realistic, real-world environments, such as production deployments of multimedia services or longitudinal field studies with diverse users and content.

## XI. FUTURE RESEARCH

While this study provides valuable contributions to the design of machine learning models for UX prediction, several avenues remain open for further investigation. For instance, the dataset used in this study is based on a predefined set of movie clips. Therefore, future analyses should extend to a wider range of content types, including updated multimedia content from contemporary platforms such as YouTube, Instagram, and TikTok, as well as new interaction modalities enabled by augmented and virtual reality devices. Such an expansion would increase the models' validity and applicability to real-world contexts.

In addition, although this work adopted standard machine learning algorithms (*e.g.*, Decision Trees, SVMs) to ensure transparency and rapid prototyping, deep learning techniques could be used in the future. For example, recurrent models such as LSTM networks could capture the temporal dynamics of valence and arousal sequences, while 3D CNNs and Transformers may enable end-to-end extraction of spatio-temporal features from face images. Hybrid architectures, such as CNN–LSTM models, can integrate visual and physiological signals to improve predictive accuracy.

Moreover, the integration of physiological and behavioral signals, such as eye tracking, heart rate, galvanic skin response, and body pose, should be explored. These multimodal data sources would help capture subtle or overlapping emotional states and reduce reliance on contextual identifiers.

Moreover, given the moderate but significant correlation observed between UX and *User Quality* score ($r \approx 0.5$), future investigations should also address the role of cognitive and subjective factors in shaping UX. Building on this, the present results suggest concrete implementation pathways for real-world systems. First, content recommendation can treat the predicted user satisfaction as a reward signal and dynamically substitute multimedia content. When the model forecasts user dissatisfaction, it can prioritize content whose valence–arousal trajectory has previously yielded higher satisfaction in similar contexts (*e.g.*, clip type, *User Quality* score). Second, *personalized user interfaces* can adapt pacing (*e.g.*, interaction time, notification frequency), presentation density, and tone (*e.g.*, copy, color palette) based on the current emotional state, for example, using arousal-informed controls for intensity and timing and valence-informed controls for hedonic framing. Third, *emotion-aware systems* can operationalize just-in-time adaptations across domains—entertainment (skip/replace scenes), education (scaffold or slow down when arousal is high and valence drops), and customer experience (route to different flows when dissatisfaction is predicted). Such advancements would extend the scope of *UX-aware systems*, enabling more engaging and context-sensitive interactions for users. These applications can be referred to as *smart applications*, *context-aware applications*, or simply *smart services*.

## XII. CONCLUSION

This study investigated the relationship between emotion signals and user satisfaction, using machine learning and ensemble models to improve classification performance. The results show that valence and arousal signals contribute to UX prediction. They also show that *identity-free* mode models achieve an accuracy of around 77%. Valence and arousal signals are sufficient to estimate the *User Satisfaction* score, but not enough to reach the highest levels of performance. Contextual features, such as Participant ID ($P$), Movie Clip ID ($C$), and *User Quality* score ($Q$), are important for improving classification accuracy. In ablation analyses, context-aware models achieve over 90% accuracy, even when excluding Participant ID ($P$), highlighting the complementary role of contextual features. Among ablation studies, the most significant results are achieved by the $P_{\mathrm{CQ}}$ policy with segmentation and the *Three-Model Ensemble* system with the *majority voting* decision rule.

In addition, the correlation analysis revealed that the relationship between perceived *User Quality* score and UX, as reported by participants, is moderate and non-linear. Therefore, alternative machine learning algorithms, such as deep learning techniques, may be required to capture complex interactions among input data.

These results have significant implications for the design of *UX-aware systems*, such as the *ALPHACA* system. At the same time, they enable the development of next-generation applications that can be integrated into frameworks or operating systems embedded in *smart environments*, such as smart mobility and transport, smart homes, and smart cities. In addition, these applications should follow privacy-by-design principles: no frame persistence or identity attributes are retained by the provider; the default configuration is set to *identity-*

*free* mode; and optional contextual features are enabled only when justified and consented to by the user.

In conclusion, future machine learning algorithms may better predict UX from emotion signals and contextual information, providing more engaging, personalized digital interactions via intelligent recommender systems.

## DATA AVAILABILITY STATEMENT

The *Dataset DT22* is available at the following link on Zenodo: 10.5281/zenodo.10086788.

## ACKNOWLEDGEMENTS

The Author thanks Prof. Pierfrancesco Foglia and Prof. Giovanni Stea of the University of Pisa for their support and suggestions. Thanks also to Prof. Claudio Loconsole of the University Mercatorum and Prof. Antonio Frisoli of the Sant'Anna School of Advanced Studies for providing the research funding. In conclusion, the Author extends special thanks to Prof. Beatrice Lazzerini and Prof. Cosimo Antonio Prete of the University of Pisa for providing the opportunity to publish this important research finding. Their contributions have been very important in shaping the direction and quality of this study.

## FUNDING SOURCES AND CONFLICTS OF INTEREST

This research study has been funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.1–Call for tender No. 104 published on February 2, 2022 by the Italian Ministry of University and Research (MUR), funded by the European Union–NextGenerationEU–Project Title *"AVATAR: Enhanced AI-enabled Avatar Robot for Remote Telepresence"*–CUP J53D23000860006, D53D23001490008–Grant Assignment Decree No. 960 adopted on June 30, 2023 by the Italian MUR.


## REFERENCES

[1] Antonio Di Tecco. Dataset DT22. Technical Report 10086789, Zenodo (CERN), Geneva, Switzerland, November 2023.
[2] Abdallah MH Abbas, Khairil Imran Ghauth, and Choo-Yee Ting. User Experience Design using Machine Learning: A Systematic Review. *IEEE Access*, 10:51501–51514, 2022.
[3] Bill Albert and Tom Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics*. Morgan Kaufmann, 2022.
[4] Tarannum Zaki and Muhammad Nazrul Islam. Neurological and Physiological Measures to Evaluate the Usability and User-Experience (UX) of Information Systems: A Systematic Literature Review. *Computer Science Review*, 40:100375, 2021.
[5] Catherine Harvey, Neville A Stanton, Carl A Pickering, Mike McDonald, and Pengjun Zheng. A Usability Evaluation Toolkit for In-Vehicle Information Systems (IVISs). *Applied Ergonomics*, 42(4):563–574, 2011.
[6] Christophe Morin. Neuromarketing: The New Science of Consumer Behavior. *Society*, 48(2):131–135, 2011.
[7] Yuri Borgianni, Erwin Rauch, Lorenzo Maccioni, and Benedikt Gregor Mark. User Experience Analysis in Industry 4.0-The Use of Biometric Devices in Engineering Design and Manufacturing. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 192–196. IEEE, 2018.
[8] Ainhoa Apraiz, Ganix Lasa, and Maitane Mazmela. Evaluation of User Experience in Human–Robot Interaction: A Systematic Literature Review. *International Journal of Social Robotics*, 15(2):187–210, 2023.
[9] Marc Hassenzahl and Noam Tractinsky. User experience-a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006.
[10] Jonathan Gutman. Means–end chains as goal hierarchies. *Psychology & marketing*, 14(6):545–560, 1997.
[11] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 345–352, 2000.
[12] Kitti Koonsanit and Nobuyuki Nishiuchi. Predicting Final User Satisfaction using Momentary UX Data and Machine Learning Techniques. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7):3136–3156, 2021.
[13] Sofien Gannouni, Kais Belwafi, Arwa Aledaily, Hatim Aboalsamh, and Abdelfettah Belghith. Software Usability Testing using EEG-based Emotion Detection and Deep Learning. *Sensors*, 23(11):5147, 2023.
[14] Dandan Wang and Xiaoming Zhao. Affective Video Recommender Systems: A Survey. *Frontiers in Neuroscience*, 16:984404, 2022.
[15] Toon De Pessemier, Ine Coppens, and Luc Martens. Evaluating Facial Recognition Services as Interaction Technique for Recommender Systems. *Multimedia Tools and Applications*, 79(31):23547–23570, 2020.
[16] Gustaf Bohlin, Kristoffer Linderman, Cecilia Alm, and Reynold Bailey. Considerations for Face-based Data Estimates: Affect Reactions to Videos. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications-HUCAPP*, volume 2, pages 188–194, 2019.
[17] Stanisław Saganowski, Bartosz Perz, Adam G Polak, and Przemysław Kazienko. Emotion Recognition for Everyday Life using Physiological Signals from Wearables: A Systematic Literature Review. *IEEE Transactions on Affective Computing*, 14(3):1876–1897, 2022.
[18] Mario GCA Cimino, Antonio Di Tecco, Pierfrancesco Foglia, and Cosimo A Prete. Using emotion recognition and temporary mobile social network in on-board services for car passengers. In *International Conference on Smart Cities and Green ICT Systems*, pages 158–171. Springer, 2022.
[19] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2011.
[20] Wei-Long Zheng and Bao-Liang Lu. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.
[21] Yuanyuan Xu, Yin-Shan Lin, Xiaofan Zhou, and Xinyang Shan. Utilizing emotion recognition technology to enhance user experience in real-time. *Computing and Artificial Intelligence*, 2(1):1388, Jun. 2024.
[22] Kah Phooi Seng and Li-Minn Ang. Video Analytics for Customer Emotion and Satisfaction at Contact Centers. *IEEE Transactions on Human-Machine Systems*, 48(3):266–278, 2017.
[23] Xiao Liu and Kiju Lee. Optimized facial emotion recognition technique for assessing user experience. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pages 1–9, 2018.
[24] Gülnaziye Bingöl, Simone Porcu, Alessandro Floris, and Luigi Atzori. Qoe estimation of webrtc-based audiovisual conversations from facial expressions. In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 577–584, 2022.
[25] Kitti Koonsanit, Daiki Hiruma, Vibol Yem, and Nobuyuki Nishiuchi. Using random ordering in user experience testing to predict final user satisfaction. *Informatics*, 9(4), 2022.
[26] Jingchen Cong, Pai Zheng, Yuan Bian, Chun-Hsien Chen, Jianmin Li, and Xinyu Li. A machine learning-based iterative design approach to automate user satisfaction degree prediction in smart product-service system. *Computers & Industrial Engineering*, 165:107939, 2022.
[27] Toon De Pessemier, Ine Coppens, and Luc Martens. Evaluating Facial Recognition Services as Interaction Technique for Recommender Systems. *Multimedia Tools and Applications*, 79(31):23547–23570, 2020.
[28] Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M Jose. Integrating Facial Expressions into User Profiling

for the Improvement of a Multimodal Recommender System. In *2009 IEEE International Conference on Multimedia and Expo*, pages 1440–1443. IEEE, 2009.

[29] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

[30] Yashowardhan Soni, Cecilia Ovesdotter Alm, and Reynold Bailey. Affective Video Recommender System. In *2019 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, pages 1–5. IEEE, 2019.

[31] Marko Tkalcic, Andrej Kosir, and Jurij Tasic. Affective Recommender Systems: The Role of Emotions in Recommender Systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems*, pages 9–13, 2011.

[32] James A Russell and James H Steiger. The Structure in Persons' Implicit Taxonomy of Emotions. *Journal of Research in Personality*, 16(4):447–469, 1982.

[33] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep Affect Prediction In-The-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.

[34] Thales Teixeira, Michel Wedel, and Rik Pieters. Emotion-Induced Engagement in Internet Video Advertisements. *Journal of Marketing Research*, 49(2):144–159, 2012.

[35] Gihwi Kim, Ilyoung Choi, Qinglong Li, and Jaekyeong Kim. A CNN-based Advertisement Recommendation through Real-Time User Face Recognition. *Applied Sciences*, 11(20):9705, 2021.

[36] Richard A Feinberg, Ik-Suk Kim, Leigh Hokama, Ko De Ruyter, and Cherie Keen. Operational Determinants of Caller Satisfaction in the Call Center. *International Journal of Service Industry Management*, 11(2):131–141, 2000.

[37] Bushra Alhijawi. Improving Collaborative Filtering Recommender System Results and Performance using Satisfaction Degree and Emotions of Users. *Web Intelligence*, 17(3):229–241, 2019.

[38] Antonio Di Tecco, Pierfrancesco Foglia, and Cosimo Antonio Prete. Video Quality Prediction: An Exploratory Study With Valence and Arousal Signals. *IEEE Access*, 2024.

[39] Karen A Machleit and Sevgin A Eroglu. Describing and Measuring Emotional Response to Shopping Experience. *Journal of Business Research*, 49(2):101–111, 2000.

[40] Jussi PP Jokinen. Emotional User Experience: Traits, Events, and States. *International Journal of Human-Computer Studies*, 76:67–77, 2015.

[41] Hala Magdy Hassan and Galal Hassan Galal-Edeen. From Usability to User Experience. In *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 216–222. IEEE, 2017.

[42] Simone Borsci, Stefano Federici, Silvia Bacci, Michela Gnaldi, and Francesco Bartolucci. Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction*, 31(8):484–495, 2015.

[43] Juan M Ferreira, Silvia T Acuña, Oscar Dieste, Sira Vegas, Adrián Santos, Francy Rodríguez, and Natalia Juristo. Impact of Usability Mechanisms: An Experiment on Efficiency, Effectiveness and User Satisfaction. *Information and Software Technology*, 117:106195, 2020.

[44] Cheng-Long Deng, Chen-Yu Tian, and Shu-Guang Kuai. A Combination of Eye-Gaze and Head-Gaze Interactions Improves Efficiency and User Experience in an Object Positioning Task in Virtual Environments. *Applied Ergonomics*, 103:103785, 2022.

[45] Rensis Likert. A Technique for the Measurement of Attitudes. *Archives of Psychology*, 1932.

[46] Antonio Di Tecco. Dataset DT22. *Zenodo*, 2024. 10.5281/zenodo.10086788.

[47] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the Effectiveness of a Large Database of Emotion-Eliciting Films: A New Tool for Emotion Researchers. *Cognition and Emotion*, 24(7):1153–1172, 2010.

[48] James J Gross and Robert W Levenson. Emotion Elicitation using Films. *Cognition and Emotion*, 9(1):87–108, 1995.

[49] Vanessa Lauro and Antonio Di Tecco. VERA: A Video Emotion Response Analysis Platform for Research Studies. *Zenodo*, November 2025. 10.5281/zenodo.15258012.

[50] Antonio Di Tecco. ALPHACA. *Zenodo*, 2025. 10.5281/zenodo.18222411.

[51] Antonio Di Tecco. UXR Engine. *Zenodo*, 2024. 10.5281/zenodo.12586529.

[52] Antonio Di Tecco. *Intelligent Systems for Human Health and Well-Being*. University of Florence, Sant'Anna School of Advanced Studies, 2024.

[53] Antonio Di Tecco, Francesco Pistolesi, and Beatrice Lazzerini. Elicitation of Anxiety Without Time Pressure and Its Detection Using Physiological Signals and Artificial Intelligence: A Proof of Concept. *IEEE Access*, 2024.

[54] Antonio Di Tecco. FeatureDS. *Zenodo*, 2024. 10.5281/zenodo.10086387.

[55] A. Jović, K. Brkić, and N. Bogunović. A Review of Feature Selection Methods with Applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015.

[56] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction*, pages 379–425. Springer, 2022.

[57] Timothy Masters. *Practical neural network recipes in C++*. Morgan Kaufmann, 1993.

[58] Antonio Di Tecco. AutoML30X. *Zenodo*, 2024. 10.5281/zenodo.10086765.

[59] Tae-Ho Lee, Michiko Sakaki, Ruth Cheng, Ricardo Velasco, and Mara Mather. Emotional Arousal Amplifies the Effects of Biased Competition in the Brain. *Social Cognitive and Affective Neuroscience*, 9(12):2067–2077, 2014.

[60] Crescent Jicol, Hoi Ying Cheng, Karin Petrini, and Eamonn O'Neill. A Predictive Model for Understanding the Role of Emotion for the Formation of Presence in Virtual Reality. *Plos one*, 18(3):e0280390, 2023.

[61] Chiara Mirandola and Enrico Toffalini. Arousal—but not Valence—reduces False Memories at Retrieval. *PloS one*, 11(3):e0148716, 2016.

[62] Davide Baldo, Vaidyanathan S Viswanathan, Richard J Timpone, and Vinod Venkatraman. The Heart, Brain, and Body of Marketing: Complementary Roles of Neurophysiological Measures in Tracking Emotions, Memory, and AD Effectiveness. *Psychology & Marketing*, 39(10):1979–1991, 2022.

[63] Zhe Wang, Yuqi Zhang, Björn B de Koning, Rachel Wong, and Shuangye Chen. Effects of Emotional Tones in Computer-based Learning: Insights from System-paced and Learner-paced Experiments. *Contemporary Educational Psychology*, 81:102368, 2025.

**ANTONIO DI TECCO** graduated *magna cum laude* with a master's degree in Computer Engineering (*curriculum* Computer Systems and Networks curriculum) from the University of Pisa, Pisa, Italy, and holds an M.B.A. from Valencian International University, Valencia, Spain. He was a Visiting Fellow at the University of Hertfordshire, Hatfield, U.K., where he conducted research on recommendation systems. He received the Ph.D. degree in Smart Computing from the Department of Information Engineering, University of Florence, Florence, Italy, and is currently a Research Fellow at the Department of Excellence in Robotics and AI, Sant'Anna School of Advanced Studies, Pisa, Italy. He is also a system architect with more than ten years of experience in platform and IT system design. His research interests include artificial intelligence, embedded systems, and the Internet of Things, with applications in usability, affective computing, and smart environments for health and well-being.

# APPENDIX A
# FEATURES SELECTED

**TABLE A1. Optimal feature sets selected with segmented samples for datasets $D_{2\text{-}3}$ (a), $D_{2\text{-}1\text{-}2}$ (b), and $D_{1\text{-}to\text{-}5}$ (c).**

| Signal | Feature | [Mean, Std] |
|---|---|---|
| Valence | Maximum | [0.14, 0.22] |
| | Mean Squared | [0.19, 0.12] |
| | Impulse Factor | [0.48, 0.16] |
| Arousal | Geometric Mean | [0.08, 0.10] |
| | 4*th*-order Momentum | [0.10, 0.12] |

(a)

| Signal | Feature | [Mean, Std] |
|---|---|---|
| Valence | Minimum | [-0.45, 0.16] |
| | Root Mean Square | [0.23, 0.11] |
| | Mean Absolute | [0.21, 0.12] |
| | Harmonic Mean | [0.14, 0.13] |
| | 3*rd*-Moment | $[7\times10^{-4}, 3\times10^{-3}]$ |
| | 2*nd*-Quantile | [-0.23, 0.16] |
| | 3*rd*-Quantile | [-0.12, 0.17] |
| | Skewness | [0.18, 0.68] |
| | Peak Value | [0.48, 0.15] |
| Arousal | Geometric Mean | [0.13, 0.10] |
| | Trimmed Mean | [0.13, 0.12] |

(b)

| Signal | Feature | [Mean, Std] |
|---|---|---|
| Valence | Maximum | [0.17, 0.22] |
| | Minimum | [-0.44, 0.16] |
| | Median | [-0.16, 0.16] |
| | 1*st*-Quantile | [-0.21, 0.15] |
| | Clearance Factor | [3.23, 1.57] |
| | Peak Value | [0.48, 0.15] |
| Arousal | Mean Absolute | [0.15, 0.09] |
| | Trimmed Mean | [0.12, 0.12] |
| | Variance | [0.01, 0.01] |
| | 1*st*-Quantile | [0.07, 0.11] |
| | Crest Factor | [2.79, 0.98] |
| | Peak Value | [0.41, 0.14] |

(c)

## APPENDIX B
## EXTERNAL EVALUATION PERFORMANCE AND PARAMETERS

**TABLE B1. External performance evaluations of different policy designs with global (a) and segmented (b) samples.**

| Study | Policy | Model | Mean | Std | wPrecision | wRecall | wF1 (↓) | 95%CI |
|---|---|---|---|---|---|---|---|---|
| Ablation | $P_{CQ}$ | Bagged Trees | 0.7970 | 0 | 0.7961 | 0.7962 | **0.7954** | 0 |
| Ablation | $P_{PQ}$ | Linear Discriminant | 0.7970 | 0 | 0.7952 | 0.7956 | **0.7952** | 0 |
| Ablation | $P_{PCQ}$ | Linear SVM | 0.7969 | 0 | 0.8080 | 0.7969 | 0.7921 | 0 |
| Baseline | $P_{2\text{-}3}$ | Wide ANN | 0.5828 | 0.0122 | 0.5937 | 0.5828 | **0.5823** | 0.0043 |
| Baseline | $P_{2\text{-}1\text{-}2}$ | Coarse Tree | 0.4688 | 0 | 0.4267 | 0.4688 | 0.4428 | 0 |
| Baseline | $P_{1\text{-to-}5}$ | Cubic $k$-NN | 0.3281 | 0 | 0.3516 | 0.3281 | 0.3350 | 0 |

(a)

| Study | Policy | Model | Mean | Std | wPrecision | wRecall | wF1 (↓) | 95%CI |
|---|---|---|---|---|---|---|---|---|
| Ablation | $P_{PCQ}$ | Optimized $k$-NN | 0.9935 | 0.0012 | 0.9935 | 0.9935 | **0.9935** | 0.0004 |
| Ablation | $P_{PC}$ | Optimized $k$-NN | 0.9739 | 0.0056 | 0.9751 | 0.9739 | 0.9738 | 0.0020 |
| Ablation | $P_{CQ}$ | Fine $k$-NN | 0.8490 | ~0 | 0.8534 | 0.8422 | **0.8436** | ~0 |
| Ablation | $P_{PQ}$ | Boosted Trees | 0.7780 | ~0 | 0.7830 | 0.7722 | **0.7736** | ~0 |
| Ablation | $P_{P}$ | Bagged Trees | 0.7665 | 0.0067 | 0.7700 | 0.7665 | 0.7671 | 0.0024 |
| Ablation | $P_{Q}$ | Bagged Trees | 0.7689 | 0.0046 | 0.7688 | 0.7689 | 0.7686 | 0.0017 |
| Baseline | $P_{2\text{-}3}$ | Optimized $k$-NN | 0.7698 | 0.0090 | 0.8475 | 0.7698 | **0.7615** | 0.0032 |
| Ablation | $P_{C}$ | Bagged Trees | 0.6600 | 0.0085 | 0.6686 | 0.6600 | 0.6615 | 0.0030 |
| Baseline | $P_{1\text{-to-}5}$ | Subspace $k$-NN | 0.6104 | 0.0047 | 0.6413 | 0.6104 | 0.6044 | 0.0017 |
| Baseline | $P_{2\text{-}1\text{-}2}$ | Bagged Trees | 0.5578 | 0.0081 | 0.5650 | 0.5578 | 0.5600 | 0.0029 |

(b)

**TABLE B2. Hyperparameters of optimized models.**

| Policy | Model | Hyperparameters |
|---|---|---|
| $P_{PCQ}$ | Optimized $k$-NN | No. Neighbors: 2, Distance: Jaccard, Distance Weight: Inverse. |
| $P_{PC}$ | Optimized $k$-NN | No. Neighbors: 2, Distance: Hamming, Distance Weight: Inverse. |
| $P_{2\text{-}3}$ | Optimized $k$-NN | No. Neighbors: 2, Distance: Jaccard, Distance Weight: Squared Inverse. |

**TABLE B3. LOPO performance evaluation analysis of 2- and 3-class designs evaluated for each participant at the experimentation.**

| Sample | Classifier | Mean | Std | wPrecision | wRecall | wF1 | 95%CI |
|---|---|---|---|---|---|---|---|
| Global | 2-class | 0.6934 | 0.2311 | 0.7560 | 0.6934 | 0.6890 | 0.0827 |
| Global | 3-class | 0.4298 | 0.2084 | 0.4797 | 0.4201 | 0.4070 | 0.0746 |
| Segmented | 2-class | 0.8412 | 0.1990 | 0.8609 | 0.8412 | 0.8397 | 0.0712 |
| Segmented | 3-class | 0.4857 | 0.2511 | 0.5393 | 0.4755 | 0.4686 | 0.0899 |

**TABLE B4.** LOMO performance for binary (2-class) and ternary (3-class) classifiers. (a) and (c) present the performance of 2-class classifiers, whereas (b) and (d) present the performance of 3-class classifiers. In addition, (a) and (c) classifiers use global samples, whereas (b) and (d) classifiers use segmented samples. Therefore, (b) and (d) use the majority voting decision rule.

| Movie Clip ID | Mean | Std | wPrecision | wRecall | wF1 | 95%CI |
|---|---|---|---|---|---|---|
| 3 | 0.8322 | ~0 | 0.8359 | 0.8322 | 0.8325 | ~0 |
| 4 | 0.6937 | 0.0070 | 0.7180 | 0.6937 | 0.6965 | 0.0025 |
| 5 | 0.7577 | 0.0011 | 0.7585 | 0.7577 | 0.7575 | 0.0004 |
| 6 | 0.7535 | ~0 | 0.8089 | 0.7535 | 0.7533 | ~0 |
| 7 | 0.7436 | 0.0077 | 0.7450 | 0.7436 | 0.7421 | 0.0028 |
| 8 | 0.8293 | 0.0029 | 0.8283 | 0.8293 | 0.8286 | 0.0010 |
| 9 | 0.7092 | 0.0118 | 0.7280 | 0.7092 | 0.7124 | 0.0042 |
| **Average** | 0.7599 | ~0 | 0.7683 | 0.7599 | 0.7604 | ~0 |

(a)

| Movie Clip ID | Mean | Std | wPrecision | wRecall | wF1 | 95%CI |
|---|---|---|---|---|---|---|
| 3 | 0.8384 | 0.0007 | 0.8422 | 0.8384 | 0.8384 | 0.0003 |
| 4 | 0.6916 | 0.0108 | 0.7125 | 0.6916 | 0.6932 | 0.0039 |
| 5 | 0.8153 | 0.0035 | 0.8154 | 0.8153 | 0.8152 | 0.0013 |
| 6 | 0.7704 | 0.0058 | 0.8234 | 0.7704 | 0.7659 | 0.0021 |
| 7 | 0.7479 | 0.0093 | 0.7528 | 0.7479 | 0.7431 | 0.0033 |
| 8 | 0.8293 | 0.0041 | 0.8284 | 0.8293 | 0.8286 | 0.0015 |
| 9 | 0.7036 | 0.0030 | 0.7271 | 0.7036 | 0.7091 | 0.0011 |
| **Average** | 0.7709 | 0.0053 | 0.7860 | 0.7709 | 0.7705 | 0.0019 |

(b)

| Movie Clip ID | Mean | Std | wPrecision | wRecall | wF1 | 95%CI |
|---|---|---|---|---|---|---|
| 3 | 0.7407 | 0.0064 | 0.7348 | 0.7407 | 0.7314 | 0.0023 |
| 4 | 0.4795 | 0.0043 | 0.4121 | 0.4795 | 0.4656 | 0.0015 |
| 5 | 0.5677 | 0.0023 | 0.6301 | 0.5677 | 0.5825 | 0.0008 |
| 6 | 0.6739 | 0.0045 | 0.6726 | 0.6739 | 0.6652 | 0.0016 |
| 7 | 0.5946 | 0.0102 | 0.6253 | 0.5946 | 0.5995 | 0.0036 |
| 8 | 0.5833 | 0.0096 | 0.7194 | 0.5833 | 0.6389 | 0.0034 |
| 9 | 0.5007 | 0.0026 | 0.5099 | 0.5007 | 0.4888 | 0.0009 |
| **Average** | 0.5914 | 0.0057 | 0.6149 | 0.5914 | 0.5960 | 0.0020 |

(c)

| Movie Clip ID | Mean | Std | wPrecision | wRecall | wF1 | 95%CI |
|---|---|---|---|---|---|---|
| 3 | 0.7458 | 0.0014 | 0.7466 | 0.7458 | 0.7261 | 0.0005 |
| 4 | 0.5051 | 0.0086 | 0.5439 | 0.5051 | 0.4922 | 0.0031 |
| 5 | 0.5977 | 0.0033 | 0.6503 | 0.5977 | 0.6109 | 0.0012 |
| 6 | 0.6940 | 0.0074 | 0.6967 | 0.6940 | 0.6828 | 0.0027 |
| 7 | 0.5624 | 0.0086 | 0.5953 | 0.5624 | 0.5701 | 0.0031 |
| 8 | 0.5892 | 0.0024 | 0.7173 | 0.5892 | 0.6352 | 0.0009 |
| 9 | 0.5377 | 0.0149 | 0.5356 | 0.5377 | 0.5239 | 0.0053 |
| **Average** | 0.6045 | 0.0067 | 0.6408 | 0.6045 | 0.6059 | 0.0024 |

(d)

**TABLE B5.** Optimal evaluation models based on LOMO analysis.

| Movie Clip ID | Model |
|---|---|
| 3 | RUSBoosted Trees |
| 4 | Subspace KNN |
| 5 | Fine Gaussian SVM |
| 6 | Gaussian Naive Bayes |
| 7 | Bagged Trees |
| 8 | Logistic Regression |
| 9 | Medium Gaussian SVM |

**Note.** See Table B4(b) for more details.

**TABLE B6.** Performance evaluation of the $\text{Ensemble}_{6/1}$ evaluated for each movie clip.

| Movie Clip ID | Mean | wPrecision | wRecall | wF1 |
|---|---|---|---|---|
| 3 | 0.6895 | 0.7876 | 0.6895 | 0.6687 |
| 4 | 0.5953 | 0.5724 | 0.5953 | 0.5530 |
| 5 | 0.7331 | 0.7429 | 0.7331 | 0.7300 |
| 6 | 0.7132 | 0.8133 | 0.7132 | 0.7048 |
| 7 | 0.6996 | 0.7240 | 0.6996 | 0.6798 |
| 8 | 0.7764 | 0.7967 | 0.7764 | 0.7810 |
| 9 | 0.6891 | 0.7209 | 0.6891 | 0.6955 |
| **Average** | 0.6995 | 0.7368 | 0.6995 | 0.6875 |

**TABLE B7.** Performance evaluation of the $\text{Ensemble}_{7}$ evaluated for each movie clip.

| Movie Clip ID | Mean | wPrecision | wRecall | wF1 |
|---|---|---|---|---|
| 3 | 0.7268 | 0.7793 | 0.7268 | 0.7201 |
| 4 | 0.7909 | 0.8074 | 0.7909 | 0.7890 |
| 5 | 0.7113 | 0.7607 | 0.7113 | 0.7044 |
| 6 | 0.7292 | 0.8141 | 0.7292 | 0.7236 |
| 7 | 0.6902 | 0.7340 | 0.6902 | 0.6604 |
| 8 | 0.7758 | 0.8030 | 0.7758 | 0.7809 |
| 9 | 0.7026 | 0.7129 | 0.7026 | 0.7062 |
| **Average** | 0.7324 | 0.7730 | 0.7324 | 0.7263 |

**TABLE B8.** External performance evaluations using segmented samples using first (a) or last (b) emotions for each clip only.

| Policy | Model | Mean | Std | wPrecision | wRecall | wF1 (↓) | 95%CI |
|---|---|---|---|---|---|---|---|
| $P_{\text{CQ}}$ | Fine Tree | 0.7813 | 0.0010 | 0.7811 | 0.7813 | 0.7804 | 0.0004 |
| $P_{\text{2-1-2}}$ | Kernel Naive Bayes | 0.4688 | 0 | 0.4630 | 0.4688 | 0.4652 | 0 |
| $P_{\text{1-to-5}}$ | Bagged Trees | 0.4531 | 0 | 0.4528 | 0.4531 | 0.4249 | 0 |

(a)

| Policy | Model | Mean | Std | wPrecision | wRecall | wF1 (↓) | 95%CI |
|---|---|---|---|---|---|---|---|
| $P_{\text{CQ}}$ | Medium ANN | 0.7812 | 0.0010 | 0.7811 | 0.7812 | 0.7804 | 0.0004 |
| $P_{\text{1-to-5}}$ | Subspace $k$-NN | 0.4688 | 0 | 0.5100 | 0.4688 | 0.4695 | 0 |
| $P_{\text{2-1-2}}$ | Subspace $k$-NN | 0.4615 | 0 | 0.4594 | 0.4615 | 0.4598 | 0 |

(b)

## APPENDIX C
## CONFUSION MATRICES

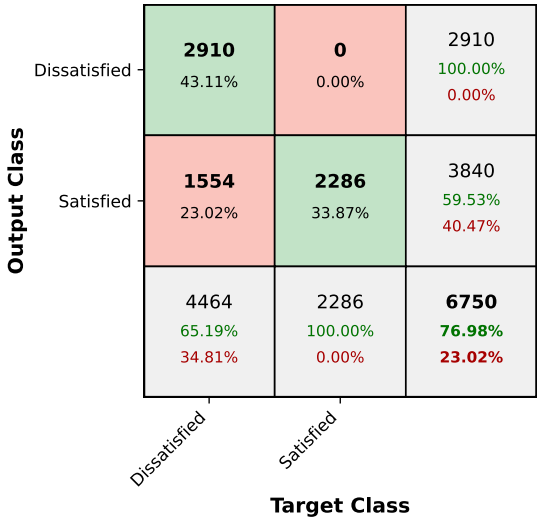


**FIGURE C1.** Confusion matrix of the model evaluated with segmented samples by Policy $P_{2\text{-}3}$.

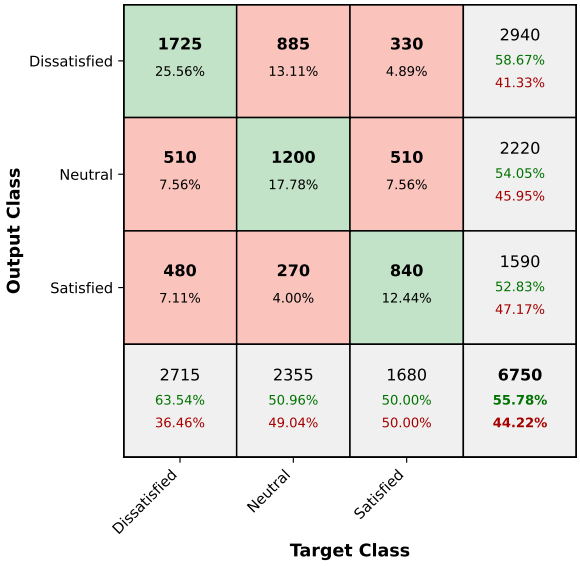


**FIGURE C2.** Confusion matrix of the model evaluated with segmented samples by Policy $P_{2\text{-}1\text{-}2}$.

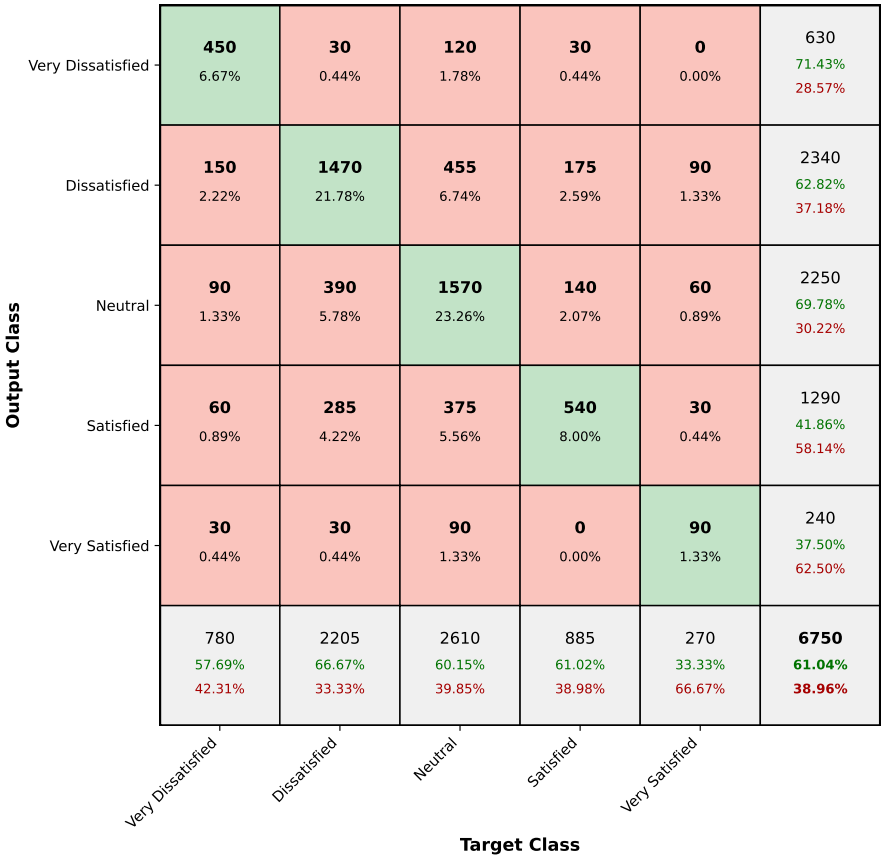


**FIGURE C3.** Confusion matrix of the model evaluated with segmented samples by Policy $P_{1\text{-to-}5}$.

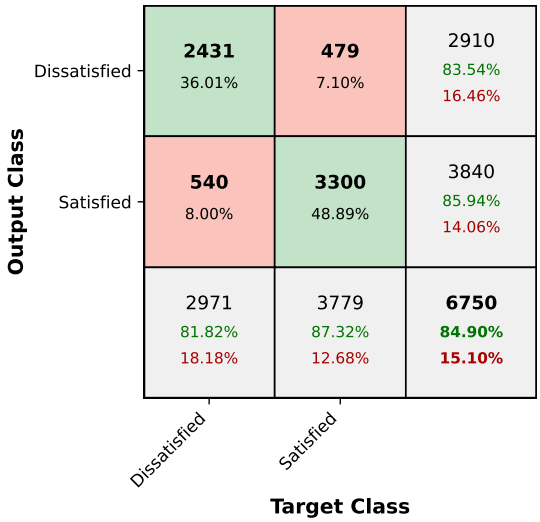


**FIGURE C4.** Confusion matrix of the model evaluated with segmented samples by Policy $P_{CQ}$.

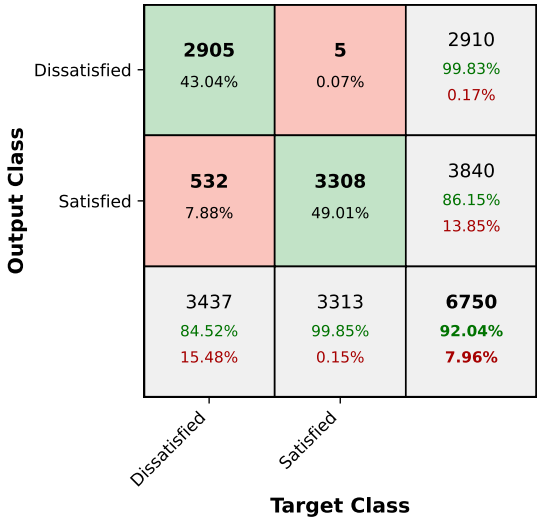


**FIGURE C5.** Confusion matrix of the ensemble composed by the three top-performing models with majority voting decision rule evaluated with segmented samples by Policy $P_{CQ}$.