Human Engagement and Multimedia Content: A Predictive Study Based on Self-Reported Affective and Experiential Variables

Antonio Di Tecco
Department of AI and Robotics
Sant'Anna School of Advanced Studies
Pisa, Italy
0000-0003-0126-8079

Nambobi Mutwalibi
Department of Information Technology
Islamic University in Uganda
Mbale, Uganda
0000-0001-6822-616X

Chongomweru Halimu
Computer Science Department
Islamic University in Uganda
Mbale, Uganda
0000-0003-0374-8287

***Abstract*—This study investigates whether user engagement during multimedia viewing can be predicted from self-reported affective and experiential variables. The analysis is based on data collected from 112 participants who watched seven movie clips and completed a four-item post-viewing Likert questionnaire after each clip, reporting prior knowledge of the clip, user experience (UX), engagement (EN), and emotional experience (EX). In addition, a dominant-emotion label derived from the clip-level emotional characterization was included as one of the input variables. Using stratified 10-fold cross-validation repeated 30 times, we evaluated 40 machine-learning classifiers under 5-class, 3-class, and 2-class engagement settings. The best-performing model was a 2-class Wide Artificial Neural Network, which achieved an average accuracy above 87% with a weighted F1-score above 0.87. Models with fewer output classes consistently outperformed the 5-class setting, highlighting a trade-off between granularity and predictive reliability, although these results should be interpreted in light of the lower difficulty of coarser classification tasks. Correlation analysis, Somers' D, and feature permutation importance converged in showing that emotional experience is the strongest predictor of engagement, followed by user experience. These findings should be interpreted as a benchmark based on subjective post-viewing measures rather than as a real-time multimodal sensing system. Nevertheless, the study offers a useful baseline for future work integrating physiological, behavioral, or audiovisual signals into adaptive context-aware systems.**

***Index Terms*—Human Engagement, Multimedia Content, Affective Computing, Machine Learning, Human-Computer Interaction.**

I. Introduction

The multimedia content domain is expanding continuously, becoming an integral part of our daily interactions, entertainment, and information consumption [15]. Moreover, with the proliferation of social media and online platforms, multimedia has evolved beyond mere text-based communication, integrating visuals and other sensory modalities to convey information and evoke emotions [2]. So, understanding and predicting human engagement with multimedia content is crucial for content creators, marketers, and educators alike [18].

979-8-3195-3447-7/$31.00 ©2026 IEEE

Artificial intelligence and affective computing can provide useful tools for understanding how users perceive, process, and respond to multimedia stimuli [3], [10]. Such understanding may support content adaptation, user-centered design, and the development of more effective multimedia applications. At the same time, emotional experience remains difficult to model because emotions are subjective and characterized by complex nuances [25], [27]. For this reason, robust and interpretable methods for analyzing affective responses are still needed [25], [27].

Affective computing is relevant in several application domains, including robotics, education, marketing, and entertainment, because it may support more adaptive and user-centered forms of interaction [4], [8]. More generally, this field lies at the intersection of psychology, biology, and computer science, and investigates how technology can account for users' affective states in a responsible and interpretable way [13], [26]. In the broader literature, affective states may be studied through several modalities; however, the present work focuses on a narrower setting and considers only self-reported post-viewing variables.

Accordingly, this paper focuses on a narrower and more clearly defined problem than full multimodal real-time emotion recognition. Rather than collecting physiological, facial, or vocal signals, it investigates whether post-viewing self-reported variables can provide a useful predictive signal for engagement during multimedia consumption. More specifically, it analyses the predictive value of user experience (UX), prior knowledge of the clip (KN), emotional experience (EX), and a dominant-emotion label (EM) in relation to self-reported engagement (EN). The contribution of the paper is threefold: (i) it provides an empirical benchmark across multiple machine-learning classifiers and output granularities; (ii) it compares 5-class, 3-class, and 2-class engagement prediction settings; and (iii) it analyses variable importance to clarify which subjective factors are most strongly associated with engagement. The study should therefore be read as a baseline for future context-aware systems rather than as a complete multimodal sensing framework.

This paper is structured as follows: Section II represents the state-of-the-art on recent HCI approaches and affective computing; Section III presents the experimental procedure, the data gathered, and used for developing machine learning algorithms; Section IV presents and discusses the results based on statistical data analysis and performance for machine learning algorithms; whereas Section VII concludes this work and proposes future research studies and analyses.

II. State-of-the-Art

Research on engagement and affect in multimedia systems spans several methodological approaches, including behavioral analysis, affective computing, user modeling, and multimodal signal processing. In the broader literature, engagement may be studied through visual, auditory, physiological, or interaction-based data. However, these approaches differ substantially in terms of data requirements, interpretability, and deployment conditions. The present study addresses a narrower problem by focusing exclusively on self-reported post-viewing variables derived from questionnaires.

Previous HCI interactions can reveal patterns and preferences that can be used to personalize user content recommendations, optimize content presentation, and improve user satisfaction. Emotion recognition has attracted interest from researchers from diverse fields [28]. It is a fundamental aspect of human cognition, influencing decision-making, perception, and interaction [24]. Specifically, human emotions are complex and multifaceted based on subjective experiences [7], [17], [22]. In addition, emotions can be expressed through various channels, such as facial expressions, vocal cues, body language, and physiological signals. It is important to recognize and interpret human emotions to develop affective computing systems. It is essential to address biases in emotion recognition algorithms and ensure these technologies are used responsibly and comply with international frameworks/ standards, such as the EU AI Act [16], particularly in sensitive contexts, such as healthcare and education.

In human communication and multimedia content, nonverbal cues, such as facial expressions, gestures, and body language, often convey information, sometimes exceeding the verbal content [14]. In fact, studies show that only 7% of a message's impact comes from the words themselves, while 38% is attributed to vocal cues, and a striking 55% stems from facial expressions [9], [11]. Physiological signals, such as heart rate, skin conductance, and brain activity, can provide, statistically speaking, objective measures of emotional arousal and valence. Electroencephalogram signals are valuable, as they respond sensitively and in real-time to fluctuations in affective states, providing useful features for emotion recognition [28].

The broader literature has shown that engagement and affect may be investigated through multiple sources of information, including verbal reports and, in other studies, multimodal signals [20], [28]. Multimodal approaches are often motivated by the idea that different data sources may capture complementary aspects of users' responses [6], [28]. At the same time, these approaches usually require richer acquisition settings and more complex data-processing pipelines [28]. For this reason, not all engagement studies rely on physiological, vocal, or visual signals [5]. In the present work, the analysis is intentionally restricted to self-reported post-viewing variables, to provide an interpretable benchmark based on questionnaire-derived measures.

Accordingly, the present study should not be interpreted as a computer-vision-based or multimodal sensing framework. Rather, it focuses on user experience, prior knowledge, emotional experience, and a dominant-emotion label derived at the clip level, all used to predict self-reported engagement [1], [7]. In this sense, the proposed approach provides an interpretable benchmark that may complement future studies based on richer multimodal sensing settings [19].

III. Method

In the scientific literature, there is a lack of open data on emotions and UX. An experiment named *Sperimentazione DT22* was conducted to gather data and create a database named *Dataset DT22* [7]. The database contains data gathered from 112 users who participated in the above experimentation. Although the broader database includes facially derived emotional data in addition to questionnaire data, the present study relies exclusively on questionnaire-based self-reported variables for statistical analysis and for training machine-learning models to predict participants' level of engagement while watching movie clips. However, emotional data are not considered in this research study. Whereas questionnaire data is used for statistical analysis and to develop machine learning algorithms to predict the participant's *level of engagement* while watching movie clips. Hence, this section summarizes the experimentation procedure, presenting how questionnaire data is gathered and used. This experimental protocol was reviewed and approved by the *Bioethical Committee of the University of Pisa* (Authorization No. 8/2023, Protocol No. 12009/2023). However, it was 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 sub-procedures, *Agreement* and *Experimentation*, and each sub-procedure has its steps, as shown in Fig. 1. The *Agreement* sub-procedure was divided into *Invitation*, *Learning*, and *Negotiation* phases. However the *Experimentation* sub-procedure was divided up into *Registration*, *Sign-Up*, *Introduction*, *Warm-up*, *Transition*, and *Conclusion* phases. The *Agreement* sub-procedure involved the user in the experimentation. Indeed, the participants were informed about this experimental activity during the *Invitation* phase. Then, during the *Learning* phase, the user learned the experiment activities. In the *Negotiation* phase, they read and, eventually, signed the informed consent form highlighting their participation in this experiment. Next, if participants agreed to participate, they received a private

TABLE I
LIST OF MOVIE CLIPS WATCHED BY PARTICIPANTS DURING THE EXPERIMENTAL TRANSITION PHASE.

ID	Ref.	Title	Description	Start	End	Length	Emotion
1	[21]	The Visitors (1993)	Jacquouille and Godfroid destroy the postman's car	00:19:55	00:22:10	02:15	Happiness
2	[21]	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
3	[21]	The Dead Poets Society (1989)	Todd commits suicide	01:42:54	01:47:41	04:47	Sadness
4	[21]	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
5	[21]	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
6	[12]	Capricorn One (1977)	Men burst through the door unexpectedly	01:32:51	01:34:01	01:10	Surprise
7	[21]	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

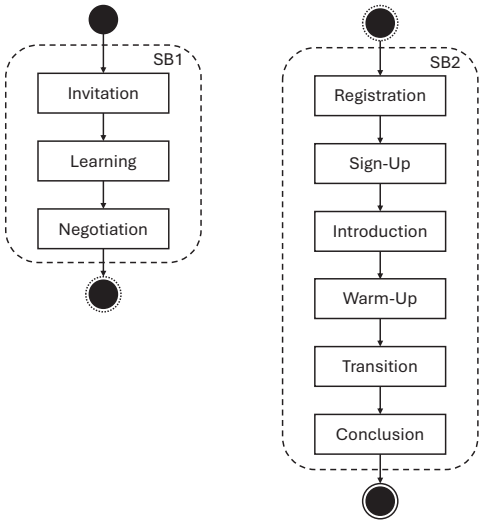


Fig. 1. Overview of phases with their sub-procedures which include Agreement ($SB1$) for informed consent and Experimentation ($SB2$) involving video viewing.

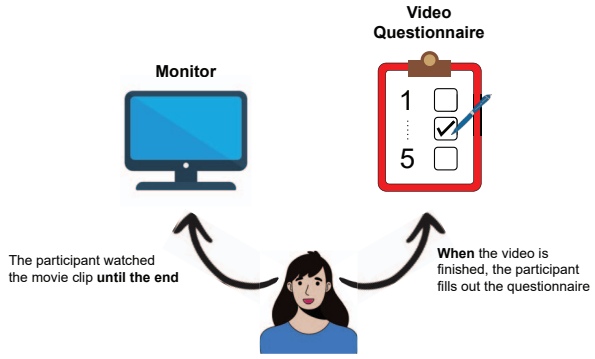


Fig. 2. Experimental scenario with a participant who watches the movie clip on the monitor and then completes the post-viewing questionnaire.

Answer	UX	EN	KN	EX
1	86	55	567	62
2	253	192	42	128
3	247	231	33	173
4	124	182	32	202
5	26	76	62	171
Total				**736**

TABLE II
PARTICIPANT ANSWERS GATHERED VIA POST-VIEWING QUESTIONNAIRE. IN TOTAL, 112 PARTICIPANTS WATCHED 7 MOVIE CLIPS, COMPLETING 736 QUESTIONNAIRES.

e-mail with the link to access the website to participate in the experiment, as represented by the *Experimentation* sub-procedure. They began with the *Registration* phase. After the user registered their account specifying their personal data, the *Sign-Up* phase started. In fact, they could log in to the home page to begin the experimentation. Then, the *Introduction* phase followed. The participant watched an introduction video that summarized the experimental activities. After watching this video, the participant started the *Warm-Up* phase. In this phase, they watched two YouTube videos. In this manner, the *learning by doing* approach was applied to teach participants their tasks. Next is the *Transition* phase.

During this phase, participants watched seven movie clips (video segments), as indicated in Table I, and completed a post-viewing questionnaire after each clip, as shown in Fig. 2. This questionnaire collected self-reported post-viewing responses for each specific clip; further details are presented in the next subsection. Then, the *Conclusion* phase followed, where the participant was thanked.

B. Questionnaire Data

Various data were collected by administering questionnaires to participants. In this study, the variables derived from

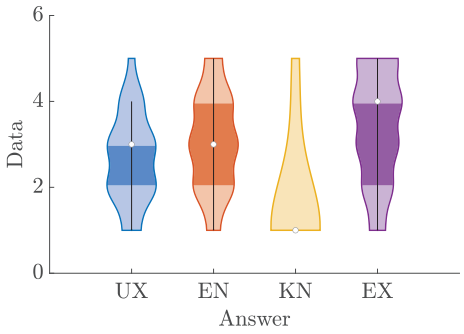


Fig. 3. Participant answer data distributions for User Experience (UX), Engagement (EN), Knowledge (KN), and Emotional Experience (EX) between 1 to 5.

the post-viewing questionnaire are treated as self-reported variables. This questionnaire was completed after each clip. Participants were asked to select an answer on a Likert scale from 1 (minimum value) to 5 (maximum value) for each of the following questions:

1) *"How satisfied are you with watching the video?"*;
2) *"How much did the video engage you?"*;
3) *"Do you already know the video?"*;
4) *"How do you rate the video's ability to evoke an emotion?"*.

Each item provided a self-reported post-viewing measure of user experience, engagement, prior knowledge of the clip, and emotional experience, respectively.

C. Dataset

The answers to the questions described in the previous subsection were used to construct the dataset employed in this study. Each participant could watch up to seven movie clips and complete up to seven post-viewing video questionnaires. As a result, the complete experimental design would have yielded 784 questionnaires; however, 48 were missing because participants were free to withdraw from the experiment at any time. The final dataset contains 736 valid questionnaire instances.

The dataset includes four input variables and one target variable. The input variables are User Experience (UX), Prior Knowledge of the clip (KN), Emotional Experience (EX), and Dominant Emotion (EM). The target variable is Engagement (EN), which was obtained directly from the participant's response to the engagement item in the post-viewing questionnaire. Accordingly, the present study models engagement from subjective post-viewing variables and does not claim real-time measurement.

Table II reports the distribution of questionnaire answers for UX, EN, KN, and EX, each measured on a Likert scale from 1 to 5. The distributions of these variables are further illustrated in Fig. 3 through a violin plot, which highlights medians, quartiles, density, and possible asymmetries across the variables. From Table II and Fig. 3, EX and EN show higher mean and standard deviation values ($\mu_{EX} = 3.40$, $\mu_{EN} = 3.04$, $\sigma_{EX} = 1.25$, $\sigma_{EN} = 1.10$) than UX and KN ($\mu_{UX} = 2.66$, $\mu_{KN} = 1.61$, $\sigma_{UX} = 1.00$, $\sigma_{KN} = 1.27$). This suggests that participants' emotional experience and engagement varied more across clips than their prior knowledge of the multimedia content.

In addition to the questionnaire-based variables, the dataset includes Dominant Emotion (EM). EM was not collected through an additional questionnaire; rather, it was derived from the dominant emotion associated with each movie clip on the basis of the emotional characterization reported in the literature and summarized in Table I [7]. For the purposes of this study, EM was reduced to a binary valence label: value 1 indicates negative valence, whereas value 2 indicates non-negative valence. Overall, 420 instances were labeled with value 1 and 316 with value 2.

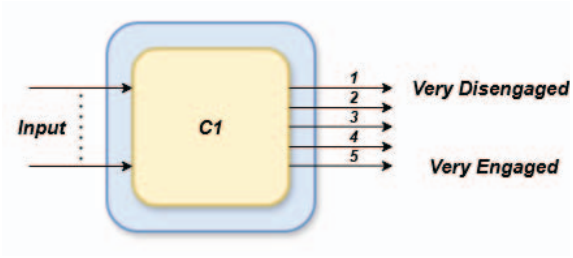
The final dataset is composed of 736 samples described by four input variables, namely UX, KN, EX, and EM, and labeled with the corresponding participant's level of engagement (EN) on a scale from 1 to 5. No additional label transformation was required, since EN was directly obtained from the participant's self-reported answer to the second question of the post-viewing questionnaire.

D. System Architecture Design

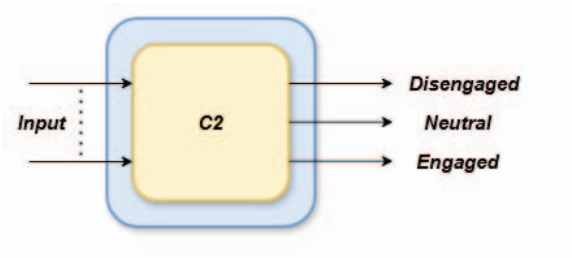
To address the challenge of user engagement prediction presented in this study, an intelligent classifier system called *Engagement Recognizer* (ER) is designed and developed. It is represented as a function $er(\cdot)$ that takes as input the four input variables UX, KN, EX, and EM, and returns a predicted value of the target variable EN. Accordingly, the system can be expressed as $EN = er(UX, KN, EX, EM)$.

The system architecture is developed using three different classification designs, such as C_1, C_2, and C_3, and their derivatives as illustrated in Fig. 4. Because these designs involve different numbers of output classes, they also differ in task granularity and classification difficulty. Therefore, their performance values should not be interpreted as directly equivalent, but rather as indicative of the trade-off between predictive reliability and output resolution. Each design aimed to explore a different level of engagement classification as system output, as follows:

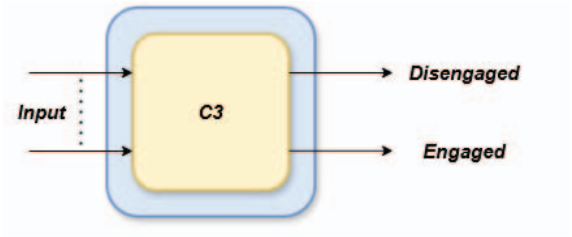
- *Classifier* C_1: This design considers each engagement level from 1 (*Very Disengaged*) to 5 (*Very Engaged*) as a distinct class. It represents the most refined and informative design, but also the most complex in learning, useful for scenarios where it is necessary to distinguish emotional nuances in user engagement;
- *Classifier* C_2: In this design, the classes are grouped to distinguish three classes: scores 1 and 2 indicate *Disengaged* users, score 3 represents a *Neutral* state, and scores 4 and 5 correspond to *Engaged* users. This design reduces the model complexity while preserving its ability to understand important information trends in engagement, it is suited to adaptive systems with multiple outputs;



(a)



(b)



(c)

Fig. 4. Architecture design of a 5-class (a), 3-class (b), and a 2-class (c) classifier.

- *Classifier* C_3: This design is explored by two configurations. The configuration design C_{3a} grouped classes 1, 2, and 3 as class 1 (*Disengaged*) and classes 4 and 5 as class 2 (*Engaged*), and the configuration C_{3b} applied the inverse logic; specifically, C_{3b} grouped classes 1 and 2 as class 1 (*Disengaged*) and classes 3, 4 and 5 as class 2 (*Engaged*).

In addition, this design C_1 is studied, also adding an *optimization* layer as shown in Fig. 5. This layer represents a logic system that takes C_1 output as input and changes the system output (number of classes) via the *command* input. In this design, the system function takes the variable *opt* (optimization) as optional input, such that $EN = er(UX, KN, EX, EM, [opt])$. It is set to zero by default, which means that the layer does not change the layer input. Higher values mean that the layer changes its output. The *command* input may be changed by the user or dynamically using another software. The layer output becomes C_2, C_{3a}, and C_{3b} through this input. Thence, alternative C_1 designs are C_{1_2}, C_{1_a}, C_{1_b} respectively.

These architecture designs allow both a performance comparative evaluation of resulting models, as explained in the next subsection, and greater flexibility in adapting the system to different application needs.

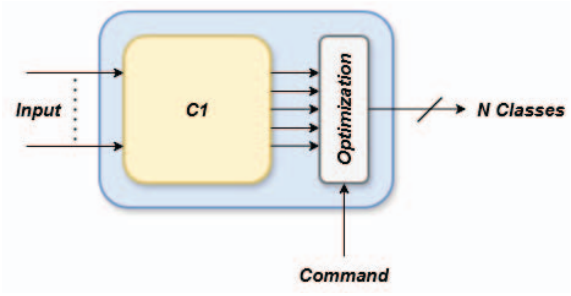


Fig. 5. Architecture design of the N-class system with the optimization layer.

	UX	EN	KN	EX	EM
UX	1				
EN	0.6004	1			
KN	0.2994	0.2807	1		
EX	0.5113	0.7942	0.2749	1	
EM	-0.1070	-0.3813	-0.2922	-0.4082	1

(a)

	UX	EN	KN	EX	EM
UX	1				
EN	0.5470	1			
KN	0.2655	0.2484	1		
EX	0.4503	0.7169	0.2404	1	
EM	-0.0981	-0.3461	-0.2804	-0.3679	1

(b)

TABLE III
SPEARMAN (A) AND KENDALL (B) CORRELATION COEFFICIENTS AMONG VARIABLES.

E. System Training and Evaluation

Three main classification designs (C_1, C_2, and C_3) were used to train and evaluate 40 machine learning algorithms using stratified k-fold cross-validation with k equal to 10 for 30 times. These machine learning algorithms belonged to various algorithm families already defined in scientific literature and present on MATLAB software, such as Decision Trees, Discriminant Analysis, Naive Bayes, Logistic Regression, Support Vector Machines (SVMs), ANNs, k-Nearest Neighbors (k-NNs), Kernels, and Ensembles of classifiers. Then, various performance metrics are measured. Models were compared using the Student t-test with a significance level (α) of 0.01, where the Null Hypothesis (H_0) assumes there is no significant difference in mean performance of the models and the Alternative Hypothesis (H_1) assumes that models have significant performance differences. The best model was the one that surpassed the others with the Student t-test and achieved the lowest mean loss. Analyses were performed using a personal computer equipped with an Intel CPU i7-14900K, 64 GB of DDR5 DRAM, and an NVIDIA GPU 4070 Ti.

IV. RESULTS

This section presents the experimental results by analyzing the collected data and developing machine learning algorithms.

Design	Model	Mean	Std	Precision	Recall	F1 (↑)
C_1	Fine Tree	0.6457	0.0026	0.6549	0.6550	0.6550
C_{1_2}	Fine Tree	0.7473	0.0023	0.7524	0.7474	0.7489
C_2	Wide ANN	0.7512	0.0017	0.7587	0.7590	0.7588
C_{1a}	Logistic Regression	0.8640	0.0014	0.8640	0.8652	0.8645
C_{3a}	Naive Bayes	0.8653	0.0008	0.8697	0.8702	0.8700
C_{1b}	Ensemble Trees	0.8745	0.0013	0.8733	0.8545	0.8728
C_{3b}	Wide ANN	0.8810	0.0020	0.8803	0.8810	**0.8789**

TABLE IV
PERFORMANCE EVALUATION OF THE BEST MODELS FOR EACH CLASSIFICATION DESIGN.

A. Statistical Variables Analysis

Statistical analysis was conducted on the dataset to better understand data relationships that may influence users' *level of engagement*. Spearman (ρ) and Kendall (τ) correlation coefficients were measured among User Experience (UX), Engagement (EN), Knowledge (KN), Emotional Experience (EX), and Dominant Emotion (EM). Pearson correlation was not measured because the data were not parametric but qualitative ordinal variables. Correlations between data distribution (variable) pairs were then estimated with 99% confidence intervals and a significance level (α) of 0.005 to provide a rigorous statistical assessment correlation. Data correlations are presented in Table III. Looking at the table, the Spearman correlation shows a strong correlation between EN and EX ($\rho = 0.7942$) and a moderate correlation between UX and EX ($\rho = 0.5113$). Also, the correlation between EN and UX is high ($\rho = 0.6004$). Whereas the Kendall coefficient detected consistent correlations, such as a strong correlation between EN and EX ($\tau = 0.7169$) and a moderate between UX and EX ($\tau = 0.4503$). In addition, the correlation between EN and UX is significant ($\tau = 0.5470$), although less than Spearman. Moreover, the Knowledge correlations are weaker but statistically significant. Very interesting is that the correlation between KN and EN is lower than Spearman ($\rho = 0.2807$) and Kendall ($\tau = 0.2484$) as well, suggesting that while knowledge may facilitate comprehension or contextual framing of movie clips, the knowledge is not a primary engagement factor. In addition, this suggests that prior Knowledge of multimedia content has a measurable but weak influence on emotional involvement and perceived experience quality. Nevertheless, the negative correlations with the variable EM, common to all variables, suggest that some specific emotions influence engagement in an inverse way. In addition to Spearman and Kendall correlations, the Somers' D test was computed to assess the strength and degree of similarity between ranking variables and to assess the significance of their relation. In practice, considering EN as the independent variable and the others (UX, KN, EX, and EM) as dependents, Somers' D measures whether increases in EN correspond to predictable changes in the other variables. The results revealed that EX is a stronger variable associated with EN, with a Somers' D of 0.7273, thus reinforcing its role as the primary predictor of user engagement. Likewise, UX has a moderate association (D = 0.5360), confirming its relevance as a secondary predictor. The KN variable has a weak association (D = 0.1789),

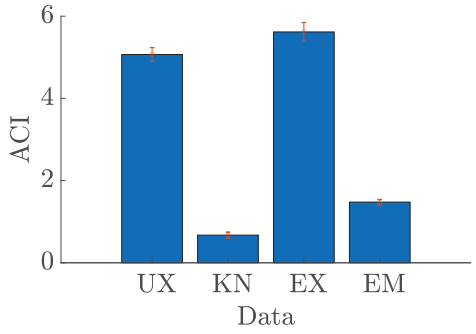


Fig. 6. Average cumulative importance (ACI) score between 1 and 5 with a confidence interval of 95% on the data.

indicating that prior knowledge of multimedia content may enhance engagement. EM also returns a negative Somers' D of –0.2787, suggesting that emotional valences may be inversely associated with the user's perceived engagement. These Somers' D results suggest a directional interpretation that reinforces the predictive hierarchy of user engagement: EX $>$ UX$>$ $\mid$ EM $\mid$ $>$ KN. However, this result deserves further investigation.

B. Performance Evaluation

The comparative performance evaluation of the three classifier designs of 5-class C_1, 3-class C_2, and 2-class C_3 was conducted by stratified 10-fold cross-validation 30 times using forty intelligent algorithms, similar to [7]. It should be noted that the 5-class, 3-class, and 2-class settings do not represent tasks of equivalent difficulty. As the number of classes decreases, the prediction problem becomes coarser and generally easier. Therefore, performance differences across these settings should be interpreted with caution and understood as reflecting a trade-off between classification granularity and predictive stability, rather than as a direct one-to-one comparison. Then, the mean accuracy and standard deviation (std), precision, recall, and weighted F1-score were measured. Only seven models are cited, which are the best ones for each system architecture design, as presented in Table IV. The other models are not presented for simplicity. As reported in the table, the Fine Tree C_1 classifier achieved an average accuracy of 64.6%, with $F1 \approx 0.65$, indicating that it is relatively difficult to distinguish five distinct levels of engagement. Whereas the 3-class classifier C_2 Wide ANN increased the accuracy to 75.1% and $F1 \approx 0.75$, the two binary configurations achieved

the highest scores. In particular, the design C_{3a} with Naive Bayes reached 86.5% accuracy, while C_{3b} with Wide ANN exceeded 88% accuracy with $F1 \approx 0.88$. However, these higher values should not be interpreted as a direct superiority over the 5-class setting, because the binary formulations define a coarser and inherently easier prediction task. Rather, the results highlight the expected trade-off between finer-grained engagement modeling and predictive robustness. Furthermore, by extending the design C_1 with the *optimization* layer, *i.e.* the C_{1_2}, C_{1a} and C_{1b} architectures, an increase in accuracy of up to 87.5% was observed. This result confirmed the importance of extending the model with an adaptive output layer depending on the application's context. Ultimately, these presented evaluations were validated with the Student t-test, which highlighted how the classification level represents a trade-off between the model's informativeness and predictive reliability.

C. Data Importance

The predictive power of the UX (User Experience), KN (Knowledge), EX (Emotional Experience), and EM (Dominant Emotion) data was assessed using the *Feature Permutation Importance* (FPI) technique [7]. This FPI technique relies on the out-of-bag estimates of a random forest model to assess feature importance. It measures the impact of randomly shuffling feature values and evaluating the resulting change in model performance across multiple decision trees. Variables whose permutation causes a larger reduction in performance are interpreted as having a stronger influence on model predictions. Each feature was assigned an *importance score* 30 times, and then their average and variance were measured. This permitted the computation of the *Average Cumulative Importance* (ACI) score for each feature as the sum of averages. So, the dataset presented in Section III-C was used to measure the ACI of the feature set. Fig. 6 shows the ACI score assigned to each feature with a 95% confidence interval. As shown in Fig. 6, EX has the highest ACI score, suggesting that perceived emotional experience while watching the content is the main variable driving engagement. This result is very important because it aligns with the correlation analysis presented in Section IV-A, in which EX has a strong positive relationship with the *level of engagement*. In contrast, the UX has a moderate ACI score, indicating a significant but less important contribution than EX. UX ranks second in importance, which indicates that the perceived quality of the interface and video presentation significantly influences user engagement. The EM contributes less, as shown by its lower ACI score and narrow confidence interval. In addition, EX's ACI is higher than EM and KN, indicating that the model learns more from subjective responses than evoked dominant emotion and past knowledge. This suggests that when the user perceives a strong emotional experience, the user tends to have behavioral changes (*e.g.*, attention, memory [23]) that the classifier associates with higher levels of engagement.

V. Discussion

The research results presented in this paper illustrate the effects on machine learning performance. The C_{3b} binary design may be more suitable for future adaptive systems, ensuring a distinction between *Disengaged* and *Engaged* users with reliability higher than 88%. Nevertheless, this advantage must be interpreted in light of the lower difficulty of the binary task, which is not directly comparable to the finer-grained 5-class formulation. This result supports the use and consideration of traditional models in operational scenarios, reducing the noise due to subjective engagement transitions. In parallel, the common results obtained between correlation (see Section IV-A) and data importance analysis (see Section IV-C) highlight the importance of EX as primary predictive variable. It is followed by UX, Dominant Emotion evoked by the population of participants, and prior knowledge as second, third, and fourth predictive position variables, respectively. Such a cross-result strengthens the proposed best model design's performance robustness and justifies the use of affective signals in the proposed system. However, the moderate accuracy of the 5-class model without the *optimization* layer may suggest integrating additional multimodal input variables, such as Physiological signals (e.g. heart rate variability, skin conductance) or temporal features (e.g. gaze patterns) could improve fine-grained discrimination. These multimodal input variables facilitate the design of architectures that are able to capture sequential patterns, such as recurrent networks or transformers. These findings give the scientific community a solid foundation to create intelligent systems based on subjective data, considering also human engagement. In the future, the proposed system could be integrated within frameworks, such as e-learning platforms or interactive video streaming platforms, where the system detects that a user/ student is disengaged and automatically stops the video or displays an interactive quiz to get attention back. These findings may be relevant for future personalized communication or recommendation systems, in which emotional experience could help platforms dynamically adapt content and context-aware algorithms to improve user retention under realistic usage conditions.

VI. Limitations

This study has several limitations that should be acknowledged. First, all predictors are based on subjective post-viewing self-reports, which limits the direct deployment of the current approach in real-time adaptive systems. Second, the constructs are measured through single questionnaire items, which may not capture the full complexity of user experience and affective response. Third, the analysis is based on one dataset and one experimental scenario involving movie clips, so the generalizability of the findings to other forms of human-computer interaction remains to be tested. Finally, the present work does not incorporate physiological, facial, vocal, or temporal signals; therefore, it should be interpreted as a benchmark for future multimodal extensions rather than as a full multimodal engagement recognition system.

VII. Conclusion

This study examined whether the *level of engagement* during multimedia viewing can be predicted from self-reported affective and experiential variables. The results show that useful predictive performance can be achieved, especially when engagement is modeled with reduced output granularity, and that emotional experience is the strongest predictor among the variables considered. These findings provide an interpretable empirical baseline for engagement prediction based on subjective post-viewing data. At the same time, the study does not yet demonstrate a real-time multimodal sensing framework, since no physiological, facial, vocal, or behavioral signals are used as input signals in the current model. These results should therefore be interpreted as a benchmark under constrained self-reported conditions rather than as evidence of a deployable multimodal sensing framework. Future work should therefore extend the current benchmark by integrating multimodal temporal signals, validating the approach on additional datasets and interaction settings, and assessing how context-aware models can be incorporated into adaptive multimedia, e-learning, or human-machine interaction systems in a transparent and ethically responsible way.

Data Availability Statement

The *Dataset DT22* analyzed in this study is publicly available through Zenodo at DOI: 10.5281/zenodo.10086788.

References

[1] Abeer Al-Nafjan, Manar Hosny, Yousef Al-Ohali, and Areej Al-Wabil. Review and Classification of Emotion Recognition based on EEG Brain-Computer Interface System Research: A Systematic Review. *Applied Sciences*, 7(12):1239, 2017.

[2] Mosab Alfaqeeh. TriMod Fusion for Multimodal Named Entity Recognition in Social Media. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, pages 1–9, 2024.

[3] Jesús A Ballesteros, Gabriel M Ramírez V, Fernando Moreira, Andrés Solano, and Carlos A Pelaez. Facial Emotion Recognition Through Artificial Intelligence. *Frontiers in Computer Science*, 6:1359471, 2024.

[4] 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.

[5] Gaochao Cui, Xueyuan Li, and Hideaki Touyama. Emotion Recognition Based on Group Phase Locking Value using Convolutional Neural Network. *Scientific Reports*, 13(1):3769, 2023.

[6] Alejandro de León Languré and Mahdi Zareei. Improving Text Emotion Detection Through Comprehensive Dataset Quality Analysis. *IEEE Access*, 12:166512–166536, 2024.

[7] Antonio Di Tecco, Pierfrancesco Foglia, and Cosimo Antonio Prete. Video Quality Prediction: An Exploratory Study With Valence and Arousal Signals. *IEEE Access*, 12:36558–36576, 2024.

[8] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human Emotion Recognition: Review of Sensors and Methods. *Sensors*, 20(3):592, 2020.

[9] Michael Eaves and Dale G Leathers. *Successful Nonverbal Communication: Principles and Applications*. Routledge, 2017.

[10] Anna Esposito, Antonietta M Esposito, and Carl Vogel. Needs and Challenges in Human Computer Interaction for Processing Social Emotional Information. *Pattern Recognition Letters*, 66:41–51, 2015.

[11] Vyvyan Evans. *The Emoji Code: How Smiley Faces, Love Hearts and Thumbs Up are Changing the Way We Communicate*. Michael O'Mara Books, 2017.

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

[13] Essam H Houssein, Asmaa Hammad, and Abdelmgeid A Ali. Human Emotion Recognition from EEG-based Brain–Computer Interface using Machine Learning: A Comprehensive Review. *Neural Computing and Applications*, 34(15):12527–12557, 2022.

[14] Zi-Yu Huang, Chia-Chin Chiang, Jian-Hao Chen, Yi-Chian Chen, Hsin-Lung Chung, Yu-Ping Cai, and Hsiu-Chuan Hsu. A Study on Computer Vision for Facial Emotion Recognition. *Scientific reports*, 13(1):8425, 2023.

[15] Wooksoo Kim, Isok Kim, Krisztina Baltimore, Ahmed Salman Imtiaz, Biplab Sudhin Bhattacharya, and Li Lin. Simple Contents and Good Readability: Improving Health Literacy for LEP Populations. *International Journal of Medical Informatics*, 141:104230, 2020.

[16] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

[17] Claudio Loconsole, Domenico Chiaradia, Vitoantonio Bevilacqua, and Antonio Frisoli. Real-Time Emotion Recognition: An Improved Hybrid Approach for Classification Performance. In *Intelligent Computing Theory: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10*, pages 320–331. Springer, 2014.

[18] Eva Pietroni, Alfonsina Pagano, Luigi Biocca, and Giacomo Frassineti. Accessibility, Natural User Interfaces and Interactions in Museums: The IntARSI Project. *Heritage*, 4(2):567–584, 2021.

[19] Shahzad Rizwan, Chee Ken Nee, and Salem Garfan. Identifying the Factors Affecting Student Academic Performance and Engagement Prediction in MOOC Using Deep Learning: A Systematic Literature Review. *IEEE Access*, 13:18952–18982, 2025.

[20] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, 2019.

[21] 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.

[22] Rukshani Somarathna, Tomasz Bednarz, and Gelareh Mohammadi. Virtual Reality for Emotion Elicitation–A Review. *IEEE Transactions on Affective Computing*, 14(4):2626–2645, 2022.

[23] Ekaterina R Stepanova, Denise Quesnel, and Bernhard E Riecke. Space—A Virtual Frontier: How to Design and Evaluate a Virtual Reality Experience of the Overview Effect. *Frontiers in Digital Humanities*, 6:7, 2019.

[24] Nazmi Sofian Suhaimi, James Mountstephens, and Jason Teo. EEG-based Emotion Recognition: A State-of-The-Art Review of Current Trends and Opportunities. *Computational intelligence and neuroscience*, 2020(1):8875426, 2020.

[25] James Z Wang, Sicheng Zhao, Chenyan Wu, Reginald B Adams, Michelle G Newman, Tal Shafir, and Rachelle Tsachor. Unlocking the Emotional World of Visual Media: An Overview of the Science, Research, and Impact of Understanding Emotion. *Proceedings of the IEEE*, 111(10):1236–1286, 2023.

[26] Mingqing Yang, Li Lin, and Slavko Milekic. Affective Image Classification based on User Eye Movement and EEG Experience Information. *Interacting with Computers*, 30(5):417–432, 2018.

[27] Yun Yi and Hanli Wang. Multi-Modal Learning for Affective Content Analysis in Movies. *Multimedia Tools and Applications*, 78:13331–13350, 2019.

[28] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion Recognition using Multi-Modal Data and Machine Learning Techniques: A Tutorial and Review. *Information fusion*, 59:103–126, 2020.