Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 288–293
https://doi.org/10.54941/ahfe1007182

Can Machine Learning Replace Expert Evaluation? Towards an AI Platform for XR, Tangible, and Haptic User Interfaces Automated User Tests

Mohammad Mustafa and Ahmed Seffah

Symbio Living Lab (Human-Centric AI & Sustainability Software), College of Technological Innovation, Zayed University, Abu Dhabi Campus, UAE

Abstract

The rapid evolution of HCI (Human Computer Interaction) paradigms such Extended Reality (XR), tangible natural interfaces, and haptic systems has fundamentally challenged the adequacy of the existing user interface usability evaluation and user test methods. Did heuristic evaluations, psychometric tests, usability metrics, thinking aloud protocol, event logging systems for behavior research still valid? How can we test XR user experiences? These questions and similar ones are the motives of this research that aims to develop an innovative AI-powered, metric-based platform to test and benchmark contemporary human-computer interaction paradigms. The platform integrates enhanced event logging, biometric sensing, and a suite of machine learning algorithms — specifically Random Forest, Long Short-Term Memory (LSTM) networks, and Support Vector Machines (SVM) with RBF kernels — to predict core well established usability attributes efficiency, effectiveness, and satisfaction, as well as emerging dimensions including emotional engagement (happiness), cognitive load, and social presence. Automating usability evaluation and user testing of these post-GUIs interfaces through a portable, deployable architecture, the AI-powered platform will aim, at the long term, to achieve accuracy comparable to expert-conducted heuristic evaluations, thereby accelerating the iterative design of next-generation interactive AI systems such as autonomous vehicle, drones, and robots that are showcases justifying this research.

Keywords: AI as user testing platform, XR, Haptics, Wearable, Human-AI interaction, Usability evaluation and user tests

INTRODUCTION

The proliferation of novel user interface paradigms—including Extended Reality (XR), tangible natural interfaces, and haptic systems—combined with evolving user experience attributes such as emotional and social interactions, has fundamentally transformed human-computer interaction (Norman & Nielsen, 2020). These emerging interaction modalities, increasingly integrated with AI systems, challenge the adequacy of traditional UX/usability evaluation methods developed for conventional graphical user interfaces (Følstad et al., 2021). Human-AI interaction introduces unique design complexities that differ fundamentally from traditional user interface

Received January 31, 2026; Revised February 26, 2026; Accepted March 7, 2026; Available online April 1, 2026

design, requiring new considerations for transparency, explainability, and user trust (Yang et al., 2020). Amershi et al. (2019) propose that human-AI systems must balance automation with user control, provide clear feedback about system capabilities and limitations, and support efficient error correction—principles that necessitate novel evaluation approaches. As Xu (2019) argues, designing effective human-AI interaction requires shifting from technology-centered to genuinely human-centered approaches that prioritize user understanding, agency, and meaningful collaboration with AI systems. This research proposes a metric-based, AI-powered evaluation platform designed to test and benchmark contemporary user interfaces and interaction modalities systematically. The platform employs enhanced event logging and behavioral coding mechanisms to capture comprehensive user interaction data, including performance metrics, behavioral patterns, and temporal sequences (Hilbert & Redmiles, 2000). Machine learning algorithms analyze this multidimensional data to predict standard usability attributes—efficiency, effectiveness, and satisfaction—as defined by ISO 9241-11 (2018), alongside emerging UX dimensions such as emotional engagement, cognitive load, and social presence (Hassenzahl & Tractinsky, 2006).

Platform Architecture and Key Components

The platform integrates multiple technical components to enable comprehensive evaluation across diverse interaction modalities Figure 1 portrays the platform:

- **Hardware Configuration:** High-performance computing system with triple 27-inch monitor setup, professional-grade graphics processing units for XR rendering, eye-tracking sensors, haptic feedback devices, and motion capture cameras for gesture-based interaction analysis
- **Data Collection Infrastructure:** Enhanced event logging system capturing keystroke dynamics, mouse movements, gaze patterns, task completion times, error rates, and physiological responses through integrated biometric sensors
- **Training Datasets:** Curated repositories of annotated interaction data from previous usability studies, expert evaluation benchmarks, standardized task performance baselines, and cross-cultural UX assessment data
- **Machine Learning Models:** Ensemble algorithms combining supervised learning for usability prediction, unsupervised clustering for behavioral pattern identification, and natural language processing for analyzing user feedback and think-aloud protocols
- **Custom-Engineered Transportable Case** with shock-resistant mounting, integrated power management system, and modular connectivity supporting rapid deployment in diverse testing environments

Figure 1: A view of the platform.

The central hypothesis posits that AI-driven automated evaluation can achieve comparable accuracy and reliability to expert-conducted heuristic evaluations and cognitive walkthroughs. This hypothesis will be validated through rigorous A/B controlled experiments comparing automated predictions against expert assessments across diverse interface types and user populations (Kohavi et al., 2020).

The platform architecture's portability enables flexible deployment for conducting evaluations across varied contexts—laboratories, field studies, or organizational settings—addressing the ecological validity concerns inherent in traditional lab-based usability testing (Kjeldskov & Skov, 2014).

AI Algorithms for User Testing: An Exploratory Investigation

To identify the most suitable AI algorithms for automated usability evaluation, we conducted a systematic benchmark study examining a range of candidate approaches. Based on these preliminary findings, we propose that the platform would benefit from a strategically selected combination of three complementary AI algorithms — Random Forest, LSTM, and SVM — which together offer an innovative, multi-dimensional evaluation framework tailored to the complexities of XR user interfaces. It is important to note, however, that these propositions remain hypotheses at this stage. Rigorous validation through controlled experiments and A/B testing will be essential to confirm their predictive accuracy and reliability against established expert benchmarks. Such validation constitutes a central objective of our long-term research agenda, which extends well beyond the scope of the present paper. What is reported here represents an early-stage account of preliminary

investigations, intended to lay the conceptual and technical groundwork for that broader research program.

A. Random Forest Ensemble Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks or mean prediction for regression. Proposed by Breiman (2001), this algorithm employs bootstrap aggregating (bagging) to create diverse tree models by randomly sampling the training dataset with replacement. Each tree is trained on a different subset of data, and at each node split, only a random subset of features is considered, introducing controlled randomness that reduces correlation among trees and mitigates overfitting.

For usability evaluation, Random Forest excels at handling the heterogeneous nature of interaction data, including continuous metrics (task completion time, error rates), categorical variables (interaction modality, user expertise level), and high-dimensional feature spaces from multimodal sensors. The algorithm naturally provides feature importance rankings through mean decrease impurity or permutation importance measures, enabling researchers to identify which interaction patterns most strongly predict usability outcomes. Its robustness to noisy data and missing values makes it particularly suitable for real-world testing scenarios where sensor failures or incomplete logging may occur. The ensemble approach aggregates predictions across numerous trees, typically 100-500, providing stable probability estimates for usability classifications while maintaining computational efficiency through parallel tree construction. This interpretability and reliability make Random Forest an excellent baseline model for automated usability assessment.

B. Long Short-Term Memory (LSTM) Recurrent Neural Network

Long Short-Term Memory networks, introduced by Hochreiter and Schmidhuber (1997), represent a specialized recurrent neural network architecture designed to capture long-range temporal dependencies in sequential data. Unlike traditional feedforward networks, LSTMs maintain internal memory states that persist across time steps, enabling them to learn patterns spanning extended interaction sequences. The architecture comprises memory cells regulated by three gating mechanisms: input gates controlling information flow into memory, forget gates determining what historical information to discard, and output gates regulating what memory content influences predictions.

For usability evaluation, LSTMs process temporally-ordered interaction events—mouse trajectories, keystroke dynamics, gaze patterns, navigation sequences—as time-series data where the order and timing of actions convey critical information about user experience. The model learns to recognize behavioral signatures associated with confusion (hesitation patterns, repeated actions), flow states (smooth continuous interactions), and error recovery strategies. Each memory cell maintains a context vector encoding relevant historical interactions, allowing the network to distinguish between similar action sequences that differ in timing or context. During training, backpropagation through time with gradient clipping prevents vanishing and exploding gradient problems common in standard RNNs. The model

outputs continuous usability predictions at each timestep or provides session-level assessments by processing complete interaction sequences. This temporal modeling capability makes LSTMs uniquely suited for capturing the dynamic, sequential nature of human-computer interaction.

C. Support Vector Machine (SVM) With Radial Basis Function Kernel

Support Vector Machines, developed by Vapnik and colleagues (1995), implement the principle of structural risk minimization to find optimal decision boundaries separating different classes in feature space. The core algorithm identifies the maximal-margin hyperplane that maximizes the distance between the nearest training examples (support vectors) of different classes, providing robust generalization even with limited training data. When data is not linearly separable in the original feature space, kernel functions implicitly map observations into higher-dimensional spaces where linear separation becomes possible without explicitly computing the transformed coordinates. The Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, computes similarity between data points based on their Euclidean distance, with a bandwidth parameter (gamma) controlling the influence radius of individual training examples.

For usability evaluation, the RBF kernel excels at capturing complex, non-linear relationships between interaction metrics and usability outcomes. The model learns decision boundaries that accommodate the multidimensional, non-convex distribution of usability patterns in feature space. Regularization through the C parameter balances margin maximization against training error minimization, preventing overfitting to expert evaluation labels. SVMs provide probabilistic predictions through Platt scaling, enabling confidence estimates for automated assessments. Their effectiveness with high-dimensional data and relatively small training sets makes them particularly valuable when expert evaluation benchmarks are expensive to obtain, while their mathematical foundation in convex optimization ensures convergence to globally optimal solutions.

A Concluding Remark

This paper is one step of a long-term fundamental research project. This research presents an early exploratory study redefining how usability evaluation is conducted in an era of increasingly complex human-AI and multi-modal interfaces using wearable, XR/VR and haptic and tangible UI. By combining the interpretability of Random Forest, the temporal sensitivity of LSTM networks, and the high-dimensional classification power of SVM, the proposed platform can offer a comprehensive and scalable solution to traditional expert-driven evaluation methods and user testing methods such as A/B testing and controlled experiments and tools such event logging systems. The portable architecture further addresses long-standing concerns around validity, enabling evaluation across laboratory, field, and organizational contexts alike. Still that rigorous A/B controlled experiments remain necessary to fully train and validate the platform's predictive capacity and accuracy, the foundational work presented paves the road for AI-powered testing

platform to make evaluation and user testing more accessible, consistent, and efficient — ultimately supporting better-designed AI interactive systems including robots, drones and autonomous vehicle.

ACKNOWLEDGMENT

This platform is being developed, thanks to a Research-Intensive Grant from Zayed University and as part of a Collaboration between Symbio Living Lab as Nostra Technologies in joint funded venture with UAE Ministry of Culture. Project is aiming to explore AI-Enabled XR for visually exploring cultural heritage assets and for learning.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Lucking, S., & Hobert, S. (2021). Future directions for chatbot research: An interdisciplinary research agenda. *Computing*, *103*(12), 2915–2942. https://doi.org/10.1007/s00607-021-01016-7

Hassenzahl, M., & Tractinsky, N. (2006). User experience – A research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys*, *32*(4), 384–421. https://doi.org/10.1145/371578.371593

ISO 9241-11:2018. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*. International Organization for Standardization.

Kjeldskov, J., & Skov, M. B. (2014). Was it worth the hassle? Ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services* (pp. 43–52). ACM. https://doi.org/10.1145/2628363.2628398

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.

Norman, D., & Nielsen, J. (2020). The definition of user experience (UX). Nielsen Norman Group. https://www.nngroup.com/articles/definition-user-experience/

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3313831.3376301