A Data-Driven Usability Evaluation of E-Commerce Interfaces Using Machine Learning

***Abstract*—Usability measures how effectively a system enables users to complete tasks easily and satisfactorily. In e-commerce applications, usability is essential for designing intuitive systems that enhance user experience and trust. Traditional usability testing methods like surveys are often slow and subjective than a data-driven approach. To attain this objective in this research, at first, a Usability Score Algorithm has been developed. 50 has been considered as a neutral initial value to handle both negativity and equilibrium score throughout the process. Secondly, a Clickstream dataset is reshaped to count user events. And finally, machine learning models like Linear Regression, Random Forest, SVR, KNN, Decision Tree and XGBoost were trained to predict the Usability Score. As outcomes, this study shows that, XGBoost achieved the best performance with an R^2 of 0.98. Random Forest achieved the second-best performance with an R^2 of 0.95. Feature analysis identifies that purchase is the most influential event, while TotalSessions has the least impact. This approach offers a quick and unbiased usability evaluation. It supports designers to enhance e-commerce platforms.**

***Index Terms*—Human-Computer Interaction (HCI), Usability Evaluation, User Experience, E-commerce, Machine Learning**

I. Introduction

Usability evaluation is one of the important factor to evaluate the performance of web applications [1]. Usability evaluation basically explores how well a system works and ease-to-use from the user's perspective. Evaluating usability became a crucial issue for websites and web-applications that people rely on regularly [2]. The number of e-commerce platforms is growing day by day, as is their usability. The growing number of e-commerce platforms also raises the concern of usability. The more user-friendly a site is, the more profit it can generate. In the world of e-commerce, where competition is fierce and user attention spans are short, good usability can make or break a platform [3]. A smooth, user-friendly experience often means more time spent on the site, more purchases, and a higher chance that users will come back. Usability of e-commerce interface can determine the success or failure of the platform.

There are various usability evaluation methods available over the recent years. The traditional usability evaluation methods, such as manual survey and questionnaires, rely solely on user input data [4]. As a result, the evaluation becomes time-consuming and difficult. There are other approaches, like semiotic evolution [5], heuristic evolution [6], which involve too much uncertainty in their final products. The developer can only guess but cannot be sure whether the end users like it, feel comfortable using it, or find it attention-grabbing. Recent studies mention that the use of machine learning algorithms for usability evaluation of such platforms failed to provide any simple scoring techniques. Additionally, many studies does not even utilize user behavioral data or clickstream data to evaluate the platform. Some proposed eye-tracking [7] for usability evaluation, which is costly and sometimes not comfortable for end-users. Hence, there is a need for utilizing user behavioral or clickstream data to evaluate the usability of e-commerce platform accurately.

Therefore, the objective of this research is to utilize clickstream data. To achieve this objective, supervised machine learning models were used to predict the system usability of e-commerce interfaces. Additionally, Usability Score has been proposed and evaluated, rather than relying on traditional methods. The contributions of this research are as follows:

- Preprocess the clickstream data to extract user behavioral features such as purchase, add-to-cart, click, login, logout, etc.
- Proposed Usability Score algorithm for calculating usability scores.
- Train and evaluate machine learning models to predict usability scores and illustrate comparison between them.

The remainder of this research is organized as follows. Section II reviews related works on usability evaluation of e-commerce and machine learning implementations in usability evaluation. Section III describes the proposed methodology including data preprocessing and model training. Section IV discusses experimental results of the proposed methodology. Section V concludes this research and discusses future direction of usability evaluation of e-commerce platforms.

II. Related Works

This section briefly introduces the studies focused on the adoption of machine learning for usability evaluation.

Usability assessment in e-commerce has long relied on standardized instruments and psychometric models. Bangor et al. [8] provided one of the most comprehensive benchmarks for the System Usability Scale (SUS) by analyzing 2,324 surveys across 206 studies, confirming its reliability (Cronbach's α = 0.911) and introducing an adjective rating scale to aid interpretation (e.g., $>$ 70 "passable," $>$ 80 "good," $>$ 90 "excellent"). By enhancing these quantitative scales, Van Schaik and Ling [9] examined the interplay of pragmatic (usability) and hedonic (stimulation and identification) qualities within a Wikipedia-style site. It shows that perceived enjoyment significantly drives intention to use alongside usability and usefulness, although their findings were constrained by artificial tasks and a homogeneous student sample.

After psychometric evaluation, Su and Chen [10] combined an improved leader clustering algorithm with rough set theory to identify overlapping e-commerce interest clusters like women's dresses, diversified household goods, and electronics using nearly three million records (198,325 after preprocessing). Their rough leader clustering outperformed traditional leader and K-medoids in efficiency and compactness. Yet, it is sensitive to threshold settings and limited by single-day data. Similarly, Kanaan and Kheddouci's [11] Sequential Event Pattern Mining (SEPM) algorithm enables the discovery of time-aware navigation sequences without repeated dataset scans. Despite strong scalability, SEPM's performance depends on duration estimation and threshold stability .

Predictive modeling of user behavior has also benefited from machine learning approaches. Ahmed et al. [12] developed a hybrid recommender that fuses user-user and item-item collaborative filtering with a weighted scoring mechanism, achieving superior prediction accuracy on the XING dataset, though its offline evaluation on anonymized, noisy data and lack of online testing limit generalizability. Sahi and Geetanjali [13] illustrated the effectiveness of backpropagation neural networks (BPNN) in modeling perceived usefulness (PU) on Indian B2C sites, identifying system quality and trust as primary drivers. More recently, Tokuc and Dağ [14] compared tree-based (LightGBM, Random Forest) and linear models to predict purchase behavior from clickstream sessions. They achieved an AUC-ROC of 0.9865 using LightGBM with Tree-structured Parzen Estimator tuning. However, their framework lacks advanced sequential modeling and struggles with class imbalance and interpretability granularity.

Integrated multi-criteria decision-making (MCDM) methods have been proposed to jointly assess usability and security. Kumar et al. [15] applied AHP, VIKOR, and TOPSIS to datasets from three e-commerce sites (smplazza.com, opsmart.in, helloshoppee.com), revealing gaps in standalone models and offering a structured comparative framework; their static ML-based evaluation, however, considers only a limited set of attributes and may miss dynamic usability–security interactions.

Zhang et al. [16] proposed a Multi-dimensional Visual Performance Evaluation Model (AEU Model) to evaluate user evaluations of e-commerce promotion pages through quantitative experiments. It involve correlation and multiple regression analysis. They used a dataset of 67 promotional pages from Taobao, Tmall, and Amazon, and responses from 34 Chinese participants aged 22–34. The study revealed three key visual performance indicators. They are aesthetics, ease of use, and information usefulness. This influences overall satisfaction. Results illustrated contextual factors like standard vs non-standard goods significantly affecting user perception. They offered design insights for improving e-commerce visual marketing effectiveness. However, the study was limited to Chinese users and a specific age range. It focused on visual perception. Manually selected pages that may introduce bias.

Overall, all of these studies underscore significant advances in user behavior modeling, e-commerce usability evaluation, and decision support. However, there are a limited number of works mentioning the implementation of machine learning. There are still many opportunities that exist to improve and address the gaps of existing predictive modeling. Therefore, this research focuses on improving the predictive modeling of recent works by utilizing machine learning for usability evaluation of e-commerce.

III. Methodology

This research evaluates the usability of e-commerce interfaces by utilizing user behavioral patterns on the platform and applying machine learning algorithms to these patterns to predict the usability score. This section outlines all the details of the proposed methodology.

A. Dataset Description

The dataset consists of 74,817 user interaction records, which were collected from an e-commerce platform between January 2024 and July 2024. This dataset was downloaded from Kaggle [17]. It contains event data from 1,000 unique users. This dataset provides simulated data for user interactions on an e-commerce platform. Each record captures user activity within sessions, making it suitable for analyzing clickstream paths and transaction sequences. The dataset contains a total of seven columns.

- **UserID**: Unique identifier for each user
- **SessionID**: Identifier for each session
- **Timestamp**: The time of the interaction
- **EventType**: Type of user action (e.g., page_view, product_view, add_to_cart, purchase)
- **ProductID**: Product identifier (only present for product-related events)
- **Amount**: Monetary value, recorded only for purchase events
- **Outcome**: Result of the interaction (e.g., purchase for successful transactions)

EventType includes sequences of events. Each time product_view, add_to_cart and purchase event triggers for any product, ID of that product is added to the ProductID. For other events, it is none. If a user purchases a product, the amount of the product is added to the Amount column and the Outcome is labeled as "purchase". It means the user has purchased the product. There are seven event in total in the EventType column. Each event represent the user's actions across various stages of interaction within the e-commerce platform.

- **add_to_cart**: Indicates that a product has been added to the cart.
- **click**: Represents user clicks for navigating to the next page.
- **login**: Indicates a user login event.
- **logout**: Indicates a user logout event.
- **page_view**: Records each instance a user views a page.
- **product_view**: Records each instance a user views a product.

- **purchase**: Indicates that a product has been successfully purchased.

Table I illustrates the initial dataset structure with a few records from the dataset.

B. Data Preprocessing

The raw data is first loaded and analyzed to determine the types of user events recorded. Each row of the raw data corresponds to a single event for a particular user and an associated event type within the session. To conduct usability evaluation, the dataset is restructured in a way that each UserID is associated with distinct columns for each event type. These columns represent the user's actions across various stages of interaction within the e-commerce platform. The data is grouped by user ID, and the frequency of each event type is counted for that user. These counts are then pivoted into separate columns, transforming the dataset from an event-level format to a user-level format. Additionally, the total number of sessions is calculated, which indicates the total number of sessions per user. The number of sessions represents how many times the user has visited the site. As shown in Table II, each row represents a particular user, where the frequency of different events is illustrated in separate columns along with total sessions. For example, a user with UserID 2 has a total of 18 add_to_cart events within their timeline, which means that the user has added products to the cart 18 times. The Usability Score algorithm is applied over the events to calculate the usability of the platform for each user. Any missing values are then dropped to ensure data quality, and the dataset is split into training and testing sets for model training. Figure 1 shows the data preprocessing steps along with model training and feature importance.

C. Usability Score Calculation

To determine the usability score of e-commerce interfaces based on restructured dataset, the Usability Score algorithm was developed, which assigns a usability score to each user. The score reflects how efficiently a user is able to navigate the platform and complete their purchases.

In Table III, events are statistically calculated to determine the threshold. Definitions of each column are given below:

- **Event-Type**: This column represents the types of events from the data.
- **Minimum**: Minimum frequency of that event.
- **Maximum**: Maximum frequency of that event.
- **Difference**: Difference between the Maximum and Minimum frequency of the event.

$$Difference = Maximum - Minimum \quad (1)$$

- **Quarter**: Quarter is calculated by dividing Difference by four. Rounded value is taken.

$$Quarter = round(Difference/4) \quad (2)$$

- **Center**: Center is the mean value of Minimum and Maximum frequency. Center is calculated as follows,

$$Center = (Minimum + Maximum)/2 \quad (3)$$

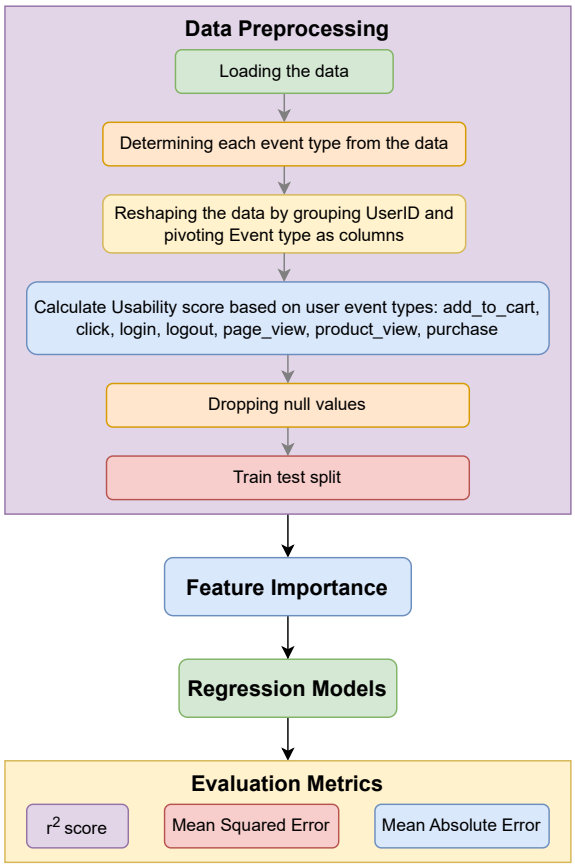

Fig. 1. Data preprocessing with model training and feature importance.

To determine the threshold δ, the sum of quarters is calculated from Quarter column in Table III. Then it is divided by the total number of events. The threshold value is determined as follows,

$$\delta = \frac{\sum Quarter}{\text{Total number of unique event types}} \quad (4)$$

$$\delta = 4.85 \approx 5$$

This threshold δ was used in order to set the interval of each event's frequency count.

The website is scored on a scale of 0 to 100. Half of 100, which is 50, has been taken as an initial score to balance negativity and max-score adjustments. It serves as a neutral baseline. A maximum of 50 points can be added, considering the highest possible score of 100, while a maximum of 50 points can be deducted, resulting in the lowest possible score of 0.

Based on Fig. 3, a discussion between six expert individuals has been organized. The concluded decisions suggested that, among the different event types, purchase bears the most importance. So, it has the maximum score ± 22.5. On the other hand, login has the lowest importance cause the platform logs users out automatically before the expected session end. This inflates the number of login events, but those logins are not strong indicators of engagement. They are often just forced

TABLE I
RAW ECOMMERCE CLICKSTREAM TRANSACTIONS DATASET

UserID	SessionID	Timestamp	EventType	ProductID	Amount	Outcome
1	1	2024-07-07 18:00:26.959902	page_view	NaN	NaN	NaN
1	1	2024-03-05 22:01:00.072000	page_view	NaN	NaN	NaN
1	1	2024-03-23 22:08:10.568453	product_view	prod_8199	NaN	NaN
1	1	2024-03-12 00:32:05.495638	add_to_cart	prod_4112	NaN	NaN
1	1	2024-02-25 22:43:01.318876	add_to_cart	prod_3354	NaN	NaN

TABLE II
RESHAPED ECOMMERCE CLICKSTREAM TRANSACTIONS DATASET

EventType	UserID	add_to_cart	click	login	logout	page_view	product_view	purchase	TotalSessions
0	1	23	5	11	6	15	14	8	10
1	2	18	10	9	9	10	4	13	10
2	3	11	10	5	13	9	10	6	10
3	4	14	9	13	12	17	9	9	10
4	5	13	15	8	13	14	14	7	10

TABLE III
STATISTICAL MEASUREMENT FOR EACH EVENT TYPE IN THE DATASET

Event-Type	Minimum	Maximum	Difference	Quarter	Center
add_to_cart	2	23	21	5	12.5
click	3	23	20	5	13.0
login	1	19	18	4	10.0
logout	2	22	20	5	12.0
page_view	3	24	21	5	13.5
product_view	2	22	20	5	12.0
purchase	3	22	19	5	12.5

re-logins. So, it has the lowest maximum score ± 2.5. Other 5 event has nearly equal importance. So they have average maximum score,

$$max - score = \pm \frac{(100-50)-(22.5+2.5)}{5} = \pm 5 \quad (5)$$

Purchase indicates the success of an e-commerce platform. Successful purchases indicates that the platform has higher usability. Here, the total session per user is 10. For the ratio,

$$value = Purchase\_count/TotalSessions \quad (6)$$

If the value exceeds 1, it indicates that the average number of purchases per visit is greater than one. So, the usability score is increased by 22.5. Hence, If the ratio is smaller than 1, it indicates fewer purchases per visit. Therefore, 11.25, Half of previous score has been added. If there is no purchase at all will be bad for an e-commerce site. So, for no purchase(purchase count = 0), the score is decreased by 22.5.

For the other six events, a mathematical equation was derived. The equation is shown below:

$$S(e) = \begin{cases} +2\alpha, & c_e \geq m_e - \delta, \\ +\alpha, & c_e \geq m_e - 2\delta, \\ -\alpha, & c_e \geq m_e - 3\delta, \\ -2\alpha, & \text{otherwise.} \end{cases} \quad (7)$$

Where,
$c_e = user\ event\ count\ for\ event\ e$
$m_e = maximum\ event\ count\ for\ event\ e$
$\delta = threshold\ unit$
$\alpha = score\ weight$

This equation 7 implies that for event-type: add_to_cart, logout, we are considering, $\alpha = 2.5$ and $\delta = 5$ (Equation 4).

$$S(e_{add\_to\_cart,logout}) = \begin{cases} +5, & c_e \geq m_e - 5, \\ +2.5, & c_e \geq m_e - 10, \\ -2.5, & c_e \geq m_e - 15, \\ -5, & \text{otherwise.} \end{cases} \quad (8)$$

For event-type: click, page_view, product_view, we are considering, $\alpha = -2.5$

$$S(e_{\substack{click,\ page\_view,\\ product\_view}}) = \begin{cases} -5, & c_e \geq m_e - 5, \\ -2.5, & c_e \geq m_e - 10, \\ +2.5, & c_e \geq m_e - 15, \\ +5, & \text{otherwise.} \end{cases} \quad (9)$$

For event-type: click, page_view, product_view, we are considering, $\alpha = -1.25$

$$S(e_{login}) = \begin{cases} -2.5, & c_e \geq m_e - 5, \\ -1.25, & c_e \geq m_e - 10, \\ +1.25, & c_e \geq m_e - 15, \\ +2.5, & \text{otherwise.} \end{cases} \quad (10)$$

Here, score weight, α value vary based on the importance of a feature.

To have a varition on user experience, user experience has been graded on a scale from 'A+' to 'F', where 'A+' represents excellent user experience and 'F' indicates very poor user experience. Figure 2 shows distribution of user experience based on usability score.

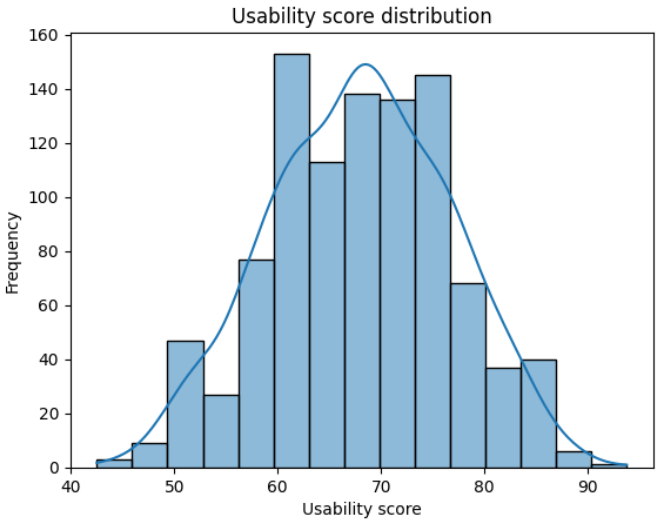

Fig. 2. User experience distribution based on usability score.

D. Model Training and Evaluation

Regression models such as Linear Regression (LR), Random Forest (RF), XGBoost, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree (DT) are trained on the reconstructed dataset in order to predict the usability scores based on user behavior data. The performance of the regression models was evaluated using the coefficient of determination (R^2 score), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

IV. Results and Discussion

This section outlines the performance of the proposed method. Although all regression models were able to capture the pattern from the refined dataset, some models performed noticeably better than other models.

A. Model Evaluation

The restructured dataset was trained on several regression models such as Linear Regression (LR), Random Forest (RF), XGBoost, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Decision Tree (DT). Among the models, RF and XGBoost performed significantly better than the other models. Table IV shows the comparison of the performance of models along with the best model.

TABLE IV
Performance comparison of models for proposed methodology

Model	R^2 Score	MSE	MAE
Linear Regression	0.77	19.99	3.67
Random Forest	0.95	4.11	1.61
XGBoost	**0.98**	**1.65**	**1.00**
Support Vector Regression	0.42	50.57	5.92
K-Nearest Neighbors (KNN)	0.72	24.47	3.81
Decision Tree	0.92	7.08	1.93

Table IV demonstrate that XGBoost achieved R^2 Score of 0.98, MSE of 1.65, MAE of 1.00. Performance of XGBoost is followed by Random Forest, which achieved R^2 Score of 0.95, MSE of 4.11, MAE of 1.61. XGBoost and RF are both tree based ensemble learning model. This indicates that tree based ensemble learning are better at capturing user behavioral patterns.

B. Feature Importance

Random Forest Feature importance analysis is performed to identify which user behavior features most influence the predicted usability scores. Higher importance indicates that a feature is more useful in predicting the target variable. Purchase achieved the highest importance score of 0.443 while model training. Login has the lowest importance score of 0.013. Figure 3 shows the sorted feature importance scores.

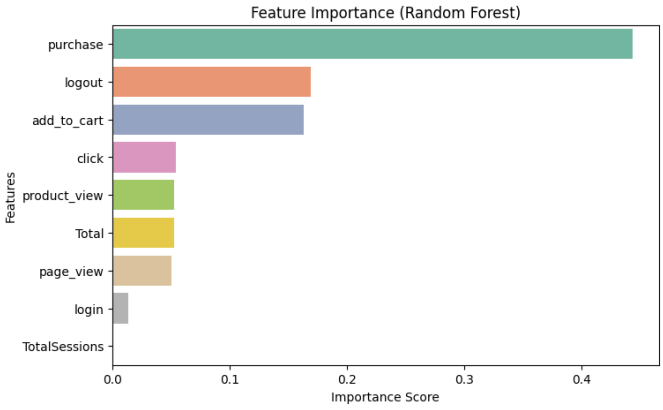

Fig. 3. Random Forest feature importance of event types.

C. Visualization of Prediction Performance

As shown in Table IV, XGBoost and RF are more capable of handling user behavioral patterns from the dataset. Further illustration of the models' predictive capability can be shown using an actual vs. predicted scatter plot.

Figure 4 shows that the calculated usability scores closely align with the predicted scores for the XGBoost model. The diagonal red line represents the perfect prediction line. Although there is very little variance, the plot indicates that the calculated usability scores and the predicted scores closely align with the perfect prediction line.

Figure 5 demonstrates that there is more spread of prediction points in the Random Forest model compared to XGBoost. The plot indicates that the majority of the prediction points are closely aligned with the perfect prediction line.

V. Conclusion

This research focused on machine learning approach to evaluate e-commerce website usability by analyzing user behavior data. A Usability Score algorithm quantified usability based on users event type. Among tested models, XGBoost have the

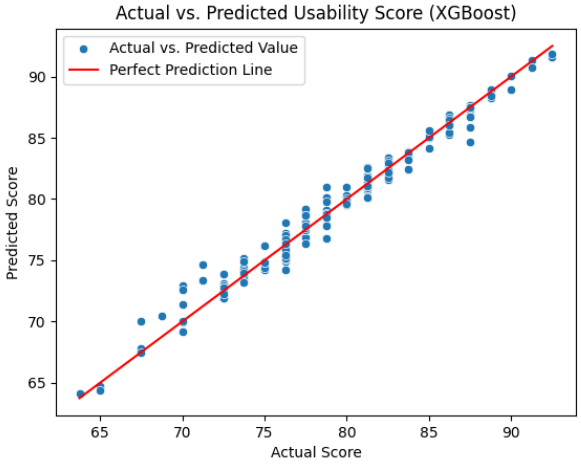

Fig. 4. Actual vs. predicted scatter plot of XGBoost.

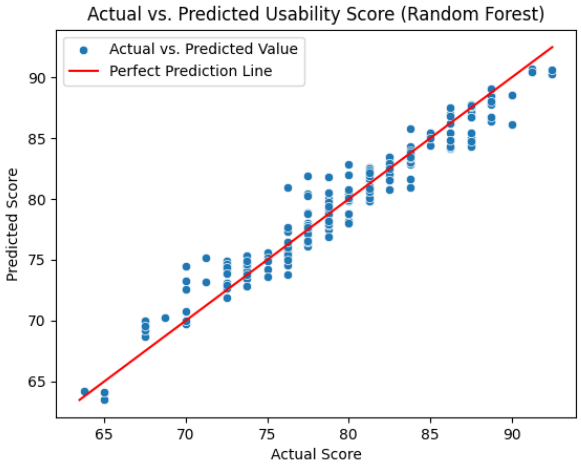

Fig. 5. Actual vs. predicted scatter plot of Random Forest.

highest accuracy ($R^2 = 0.98$), demonstrating the effectiveness of tree-based ensemble methods.

Feature analysis showed that "purchase" had the most impact, while "TotalSessions" had less impact on usability prediction as it's same for every user. This work highlights the potential for faster, data-driven usability assessment without manual surveys.

Future work includes incorporating richer user interaction data (e.g., time on page, scrolling), exploring unsupervised learning methods, and applying advanced deep learning to capture sequential behaviors. Enhancing the scoring system with user feedback and expert input can improve accuracy. Real-time usability monitoring and linking usability scores to business metrics are promising directions to optimize user experience and commercial outcomes.

References

[1] F. Liu, "Usability evaluation on websites," in *2008 9th international conference on computer-aided industrial design and conceptual design*. IEEE, 2008, pp. 141–144.

[2] M. N. Islam, S. A. Rahman, and M. S. Islam, "Assessing the usability of e-government websites of bangladesh," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2017, pp. 875–880.

[3] L. Hasan, A. Morris, and S. Probets, "A comparison of usability evaluation methods for evaluating e-commerce websites," *Behaviour & Information Technology*, vol. 31, no. 7, pp. 707–737, 2012.

[4] F. Paz and J. A. Pow-Sang, "Current trends in usability evaluation methods: a systematic review," in *2014 7th International Conference on Advanced Software Engineering and Its Applications*. IEEE, 2014, pp. 11–15.

[5] M. N. Islam, H. Bouwman, and A. N. Islam, "Evaluating web and mobile user interfaces with semiotics: An empirical study," *IEEE Access*, vol. 8, pp. 84 396–84 414, 2020.

[6] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1990, pp. 249–256.

[7] M. Hua and F. Qian, "An evaluation research on usability of taobao's homepage and main search engine based on eye tracking," in *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, vol. 2, 2010, pp. 23–26.

[8] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[9] P. van Schaik and J. Ling, "An integrated model of interaction experience for information retrieval in a web-based encyclopaedia," *Interacting with Computers*, vol. 23, no. 1, pp. 18–32, 2011.

[10] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electronic Commerce Research and Applications*, vol. 14, no. 1, pp. 1–13, 2015.

[11] M. Kanaan and H. Kheddouci, *Mining Patterns with Durations from E-Commerce Dataset: Volume 1 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018*, 01 2019, pp. 603–615.

[12] S. Ahmed, M. Hasan, M. N. Hoq, and M. A. Adnan, "User interaction analysis to recommend suitable jobs in career-oriented social networking sites," in *2016 International Conference on Data and Software Engineering (ICoDSE)*, 2016, pp. 1–6.

[13] G. Sahi, "Performance evaluation of artificial neural network for usability assessment of e-commerce websites," in *2018 3rd International Conference for Convergence in Technology (I2CT)*, 2018, pp. 1–6.

[14] A. Aylin Tokuç and T. Dag, "Predicting user purchases from clickstream data: A comparative analysis of clickstream data representations and machine learning models," *IEEE Access*, vol. 13, pp. 43 796–43 817, 2025.

[15] B. Kumar, S. Roy, K. U. Singh, S. K. Pandey, A. Kumar, A. Sinha, S. Shukla, M. A. Shah, and A. Rasool, "A static machine learning based evaluation method for usability and security analysis in e-commerce website," *IEEE Access*, vol. 11, pp. 40 488–40 510, 2023.

[16] F. Zhang, C. Lan, T. Wang, F. Gao, and E. Liu, "Research on visual performance evaluation model of e-commerce websites," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2020, pp. 1075–1080.

[17] W. Ali, "https://www.kaggle.com/datasets/waqi786/e-commerce-clickstream-and-transaction-dataset."