Journal of Cognitive Human-Computer Interaction (JCHCI)

♦ ISSN: 2771-1463 Vol: 11, No. 01, 2026 ♦ PP. 01–05

Article DOI: https://doi.org/10.54216/JCHCI.110101

FULL LENGTH ARTICLE

Task-Conditioned Early Prediction of Navigation Failure in Information Architecture Evaluation

Kharchenko Raisa[1,*] **Rahul Chauhan**[2] **Andino Maseleno**[3]

[1] North-West Institute of Management, RANEPA, Russia

[2] Unitedworld Institution of Management, Karnavati University, Gandhinagar, India

[3] Institut Bakti Nusantara, Lampung, Indonesia

Emails: kh9044947155r@gmail.com · rahulchauhan@karnavatiuniversity.edu.in · andino.maseleno@ibnus.ac.id

Received: October 06, 2025 Revised: November 17, 2025 Accepted: December 27, 2025 ⋆ Corresponding author

ABSTRACT

The interaction logs which researchers collected during their information-architecture evaluation process contain detailed proof which shows how users select between successful and unsuccessful navigation routes. The predictive signal displays its initial appearance during task execution yet users exhibit different navigation patterns depending on their current task and interface they are using. The researchers of this study developed an early navigation failure prediction system which uses public interaction data to create task-specific prefix classification models. The study analyzes data from an open dataset which includes 180 participants completing 1800 tasks across six testing conditions that evaluate tree testing and high-fidelity prototype navigation. A prefix-structural encoder works together with a regularized task-conditioned logistic model which predicts success based on the first k navigation actions. The researchers assessed model performance through participant-specific validation using three different machine learning techniques which included random forest, extra trees, and gradient boosting. The optimal configuration achieved 0.7833 accuracy, 0.7513 balanced accuracy, 0.8350 F1-score, and 0.7949 ROC–AUC performance at $k = 3$. The horizon analysis demonstration shows that predictive signals become accessible after users complete their first three actions. The ablation study proves that task conditioning functions as an essential component. The study results demonstrate that early trace analytics enable quick identification of navigation failures in information-architecture research while providing a useful method for customized assessment during usability testing.

Keywords: Information-architecture evaluation ▪ Navigation patterns ▪ Task-specific prefix classification models

1. INTRODUCTION

Human–computer interaction data analysis now depends on event-level traces instead of using post-hoc questionnaires and aggregated completion data as the only measurement method. The user action log in navigation-focused interfaces maintains the sequence of user choices which includes their initial path selection and their subsequent backtracking and repeating decisions that ultimately determine task completion results. Researchers can now conduct more precise studies of these signals through the availability of public HCI datasets which enable them to analyze website and menu assessment methods that use interaction traces to measure information architecture quality [1, 2, 3, 4].

Information architecture assessment relies on two methods which include tree testing and prototype-based findability studies. The methods produce results which organizations use to measure success rate and directness and completion time. The interaction process becomes reduced to final outcomes through these measures yet they fail to answer one critical

technical issue which concerns the prediction accuracy of navigation failure from initial user evidence. The evaluator can determine whether the user will follow the complete path by observing early path indicators which show confusion and shallow exploration and multiple branch switches. Tree-testing analytics and wider usability-data processing systems have developed new methods for tracking user behavior yet the field lacks advanced solutions for predicting outcomes in early stages [5, 6, 7].

This paper addresses that problem by modeling the first k navigation actions as a task-conditioned prefix. The generated representation keeps the initial route local structure intact while providing explicit pathway information about upcoming user movements.

2. RELATED WORK

The current state of interaction analysis research has developed methods which show user behavior through event-based analysis instead of measuring total success rates and using retrospective surveys. Public datasets now provide detailed user data which researchers can use to create navigational models that accurately reflect actual user behavior. Esposito et al. [2] released a website-interaction dataset which contains both user activity data and emotional impact assessments while Jeong et al. [3] demonstrated that visual analytics semantic interaction traces enable researchers to assess user behavior through latent goal tracking.

Research in this area aims to create abstract behavioral patterns which researchers can use to study actual user behavior through their interface interactions. Rebmann and van der Aa [8] developed an unsupervised method which enables users to identify task-based activities through their interaction sequences while Martínez-Rojas et al. [9] developed a task-mining framework which analyzes screenshots to identify which factors lead to different user behaviors. The research demonstrates that interaction logs maintain their original value even when the interface content and task definitions differ.

Information architecture evaluation continues to use tree testing and prototype testing because these methods provide direct measures of findability. Callejo and Macías [6] developed enhanced approaches for tree-testing analysis, and Kuric et al. [4] provided a cross-methodological comparison of tree testing variants and prototype user testing. These studies show that information architecture decisions should not be evaluated only through final success rates because path-level evidence provides a richer description of the interaction process.

Adaptive HCI and usability monitoring studies also motivate early prediction. Monitoring systems and adaptive interfaces can benefit from models that detect likely problems while a session is still unfolding [10, 7]. The present study builds on this line of work by treating early navigation prefixes as task-conditioned evidence for predicting navigation episode success.

3. PROBLEM FORMULATION AND METHOD

Consider a dataset of navigation episodes

$$D = \{(s_i, y_i, \tau_i, v_i, p_i)\}_{i=1}^{N}, \tag{1}$$

where $s_i = (a_{i1}, a_{i2}, \ldots, a_{iL_i})$ is the ordered interaction sequence for episode i, $y_i \in \{0,1\}$ is the final task outcome, τ_i is the task identifier, v_i is the interface condition, and p_i is the participant identifier. Each action a_{ij} is mapped to a structural state containing, when available, the selected node, resulting depth, branch transition, and event type.

For an early horizon k, only the prefix

$$s_i^{(k)} = (a_{i1}, a_{i2}, \ldots, a_{i\min(k,L_i)}) \tag{2}$$

is observed. The objective is to estimate the conditional probability of final success from the prefix:

$$\Pr(y_i = 1 \mid s_i^{(k)}, \tau_i, v_i). \tag{3}$$

A structural encoder $\phi(\cdot)$ maps the prefix to a feature vector

$$\phi(s_i^{(k)}) = \left[g_i^{(k)}, b_i^{(k)}, r_i^{(k)}, u_i^{(k)}, e_i^{(k)}, d_i^{(k)}\right]^{\top}, \tag{4}$$

where $g_i^{(k)}$ is cumulative depth gain, $b_i^{(k)}$ is the backtrack ratio, $r_i^{(k)}$ is the repeat-node ratio, $u_i^{(k)}$ is the number of unique visited nodes, $e_i^{(k)}$ is the number of expand/collapse events, and $d_i^{(k)}$ is the mean realized depth. Task and condition are encoded as sparse vectors t_i and v_i, giving the full design vector

$$x_i^{(k)} = \left[\phi(s_i^{(k)}), t_i, v_i, o_i\right]^{\top}, \tag{5}$$

where o_i denotes task order.

The proposed task-conditioned prefix logistic model (TC-PLM) estimates

$$\hat{y}_i = \sigma(w^{\top} x_i^{(k)} + b), \tag{6}$$

with $\sigma(z) = 1/(1+e^{-z})$. Parameters are learned by minimizing the regularized empirical risk

$$L(w,b) = -\sum_{i=1}^{N} [y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] + \lambda \|w\|_2^2. \tag{7}$$

The regularization term controls coefficient growth in the presence of sparse task coding and correlated path features. Since class proportions are imbalanced toward success, class-balanced fitting is used.

Grouped participant-wise validation is employed to avoid leakage between episodes generated by the same respondent. Let $P_1, \ldots, P_K$ denote disjoint participant partitions. For fold m, training and testing sets are

$$D_{\text{train}}^{(m)} = \{i : p_i \notin P_m\}, \quad D_{\text{test}}^{(m)} = \{i : p_i \in P_m\}. \tag{8}$$

Performance is measured by accuracy, balanced accuracy,

F1-score, and ROC–AUC. Balanced accuracy is

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \tag{9}$$

and the F1-score is

$$F1 = \frac{2TP}{2TP+FP+FN}. \tag{10}$$

Besides the proposed model, random forest, extra trees, and gradient boosting are evaluated on the same prefix representation.

Table 1. Task-conditioned early navigation failure prediction algorithm.

Step	Operation
1	For each episode $(s_i, y_i, \tau_i, v_i, p_i)$ in D, truncate the sequence to prefix $s_i^{(k)}$.
2	Compute structural features $\phi(s_i^{(k)})$ and build the task-conditioned vector $x_i^{(k)}$.
3	Split train and test folds by participant group P_m.
4	Standardize continuous attributes on the training fold.
5	Fit TCPLM with class-balanced optimization and predict success probabilities.
6	Store fold metrics, aggregate results, and repeat for baselines and horizons.

Figure 1 summarizes the analytic workflow in four phases. Phase 1 converts each navigation trace into an early prefix and derives structural descriptors such as depth gain, branch switching, repeated-node ratio, and backtracking. Phase 2 combines these descriptors with task identity, interface condition, and task order to construct the task-conditioned design matrix used by the classifier. Phase 3 performs participant-wise grouped cross-validation and compares TCPLM with tree-ensemble baselines across multiple prefix horizons. Phase 4 aggregates fold-wise predictions and supports error analysis through confusion matrices, ROC curves, ablation tests, and feature interpretation.

4. DATASET AND EXPERIMENTAL DESIGN

The analysis uses the public dataset released with the study of Kuric et al. [4]. The repository describes 180 participants and 1800 task completions collected across six conditions that combine tree-testing variants and high-fidelity menu navigation prototypes. Each condition contributes 300 task completions. The response variable in the present study is final task success. Additional outcomes such as direct success and directness were also examined, but the primary target was full success because it yields the most practically relevant failure-diagnosis setting.

The six conditions differ substantially in aggregate performance. As shown in Table 2, success rates range from 0.5433 for TC to 0.6800 for WTC and WTCI, already indicating that the interaction process is not homogeneous across evaluation modes. Task difficulty is similarly uneven, with Tasks 5, 6, and 10 representing the most failure-prone cases and Tasks 11, 7, and 4 representing the easiest cases.

The model is evaluated for early horizons $k \in \{1,2,3,4,5\}$, where k denotes the number of actions retained from the beginning of the trace. Continuous features are standardized within training folds only. Grouped cross-validation by participant is used throughout. The task-conditioned logistic model is compared with random forest, extra trees, and gradient boosting on the same feature space. An ablation study removes task identity, variant identity, or prefix-structural components in order to quantify the contribution of each modeling element.

Table 2. Observed success rates by experimental condition.

Condition	Success rate	Task completions
TC	0.5433	300
TP	0.5867	300
WM	0.5967	300
TO	0.6267	300
WTC	0.6800	300
WTCI	0.6800	300
Overall mean	0.6189	1800

5. RESULTS

5.1 Task Difficulty and Condition Variability

The empirical task profile is sharply non-uniform. The lowest success rates are 0.1556 for Task 5, 0.2667 for Task 6, and 0.2722 for Task 10. The highest success rates are 0.9000 for Task 11, 0.8167 for Task 7, and 0.8000 for Task 4. This spread confirms that navigation behavior must be interpreted relative to task semantics rather than treated as a single pooled process.

5.2 Early-Horizon Analysis

Table 3 reports the grouped cross-validation results for the proposed model across prefix horizons. The first action already provides useful discriminatory signal, but the strongest balance between parsimony and predictive quality is achieved around three to five actions. ROC–AUC improves from 0.7820 at $k = 1$ to 0.8008 at $k = 5$, while F1-score remains consistently above 0.827. The modest gain beyond $k = 3$ indicates that a substantial portion of failure information becomes visible very early.

Table 3. TCPLM performance across early horizons.

k	Accuracy	Balanced accuracy	F1-score	ROC–AUC
1	0.7711	0.7356	0.8272	0.7820
2	0.7750	0.7398	0.8300	0.7829
3	0.7739	0.7386	0.8292	0.7850
4	0.7717	0.7368	0.8272	0.7919
5	0.7750	0.7409	0.8295	0.8008

5.3 Model Comparison

The proposed task-conditioned formulation is most effective when implemented as a regularized logistic model. Table 4 shows that logistic regression outperforms the tested tree ensembles at $k = 3$, reaching an accuracy of 0.7833, balanced accuracy of 0.7513, F1-score of 0.8350, and ROC–AUC of 0.7949. This result suggests that, for the current prefix representation, the dominant predictive structure is captured well by a sparse linear decision surface conditioned on task identity.

Table 4. Model comparison at prefix horizon $k = 3$.

Model	Accuracy	BA	F1	AUC
Logistic regression (TCPLM)	0.7833	0.7513	0.8350	0.7949
Random forest	0.7628	0.7308	0.8187	0.7868
Extra trees	0.7417	0.7120	0.8003	0.7755
Gradient boosting	0.7739	0.7386	0.8292	0.7850

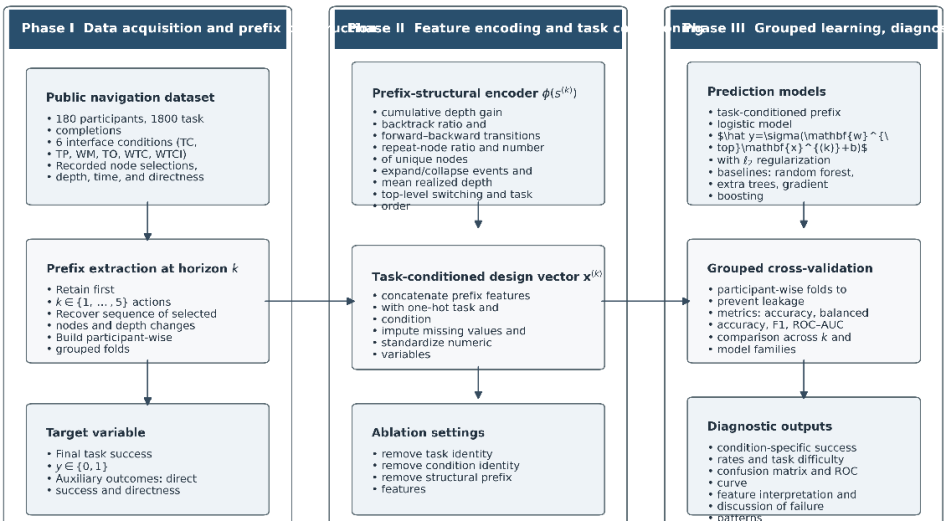


Figure 1. Detailed multi-phase framework of the proposed pipeline.

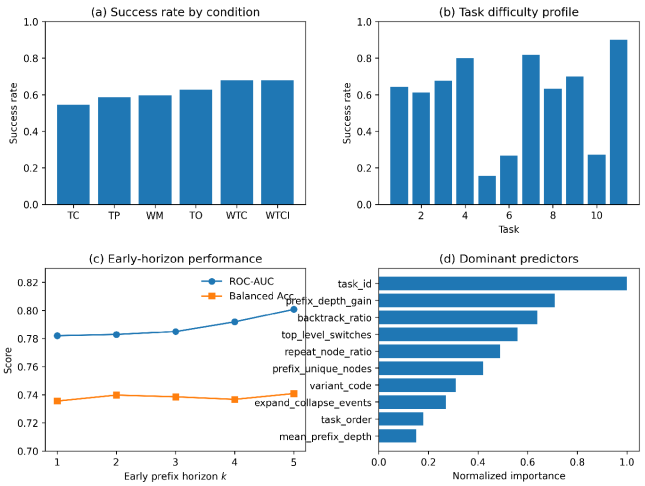
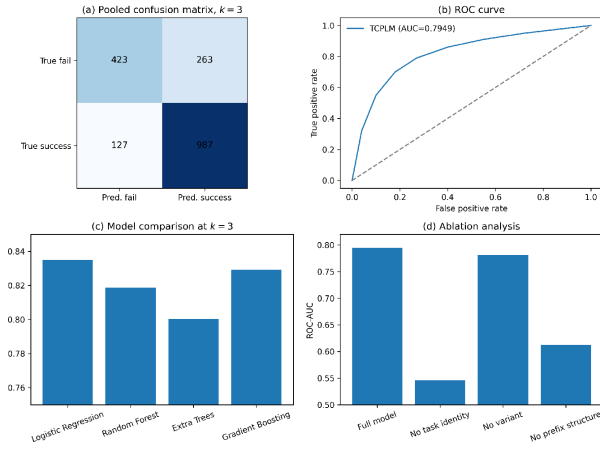


Figure 2. Dataset-level and model-level overview: (a) success rate by condition, (b) task difficulty profile, (c) early-horizon performance, and (d) dominant predictors.



gure 3: Diagnostic results for the best configuration: (a) pooled confusion matrix at $k = 3$, (b) ROC curve, (
odel comparison by F1-score, and (d) ablation analysis.

Figure 3. Diagnostic results for the best configuration: (a) pooled confusion matrix at $k = 3$, (b) ROC curve, (c) model comparison by F1-score, and (d) ablation analysis.

5.4 Ablation and Interpretation

The ablation study reveals that task conditioning is essential. Removing task identity reduces ROC–AUC from 0.7949 to approximately 0.5460, whereas removing variant identity produces only a minor drop. This result is technically important because it shows that early navigation traces are not self-explanatory: the same pattern of backtracking or branch switching may indicate confusion in one task but be completely reasonable in another. Prefix-structural features also matter. When those features are removed and only condition-level covariates are retained, discrimination collapses toward chance.

Figure 3 summarizes pooled diagnostic results, including the confusion matrix corresponding to the best logistic configuration. The pooled confusion pattern indicates a model that is especially strong at identifying eventual successes while maintaining a useful but less dominant failure-detection capability.

Figure 5 presents complementary analyses in a different visual style. Panel (a) reports class-wise precision, recall, and

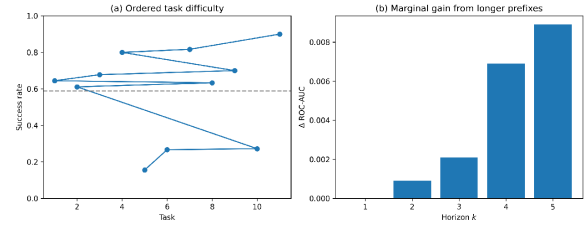


Figure 4. Ordered task difficulty and marginal gain from longer prefixes.

Figure 4. Ordered task difficulty and marginal gain from longer prefixes.

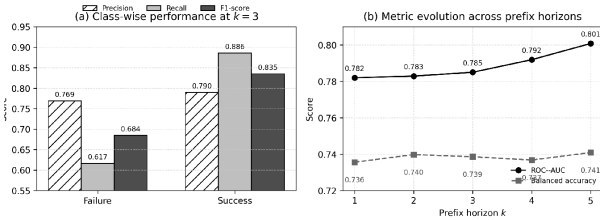


Figure 5. Additional results in an alternative visual style: (a) class-wise precision, recall, and F1-score for the best logistic model at $k = 3$ and (b) ROC–AUC and balanced accuracy across prefix horizons.

F1-score for success and failure, showing that failure cases are harder to capture than successes. Panel (b) compares balanced accuracy and ROC–AUC across prefix horizons and confirms that discrimination improves steadily as a longer portion of the early path becomes available.

6. DISCUSSION

There are three observations supported by the results. First, early prefixes provide sufficient information for prediction. Most information is encoded in the first 1–3 actions which indicates that interaction studies can detect problematic information architectures without complete paths. This is useful because it allows evaluators to identify problems before the interaction is complete, which alleviates end user burden, and allows evaluators to adaptively prompt users.

Second, conditioning on tasks is far more critical than model complexity. The dramatic decrease seen when we remove the task identity indicates that search paths should not be interpreted in isolation from the retrieval task. A short path with immediate switching could be a sign of confusion in one situation and effective disambiguation in another. The new formulation explicitly models this relationship, which is why the regularized logistic model performs better than more flexible ensembles for the same representation of data.

Our findings show how interaction-trace analytics is a key component of information-architecture research. The conventional tree-testing process relies on two sources of data for testing which are completion rates and feedback. The system's results continue to have value as evidence but evidence can come in other forms. Modelling at the prefix level uncovers the patterns that are present in branch-switching and backtracking and the depth gain, and this can inform the design of the system. The dataset included hard problems which had two characteristics: low gain in depth in early stage and inconsistent branch switching. Researchers can identify issues in the semantics of labels, categories which will cause path tracking errors and avoid full path tracking errors.

HCI data science studies impact other areas with their analysis. The research on public interaction logs and usability monitoring and adaptive interfaces demonstrate user-centered evaluation approaches which are more feasible for deployment in practice [1, 10, 7]. Our study follows this trend by demonstrating how a lightweight prediction model can be used to turn raw navigation traces into a diagnostic tool. The models are not replacements for existing usability measures because they assist in the decision-making about what tasks and conditions and path patterns need to be analysed.

7. CONCLUSION

This paper introduced a task-conditioned model for early prediction of navigation failure from public interaction logs. The approach models each episode via the first k actions, encodes structural aspects of the prefix, and employs a regularized logistic model dependent on the task and variant of the interface to make a prediction of the episode's success. The approach achieved a stable grouped validation result on a public dataset for experiments in information architecture, and showed that the information about success and failure is indeed present early on.

The experimental results suggest that predicting failure at navigation is task-dependent. The task being solved is not a nuisance variable, but part of the decision boundary. This approach can be extended in future work to incorporate path-graph encoders, more sophisticated uncertainty measures, and adaptive interventions to recommend alternative labels or structures in real time during navigation studies. More broadly, the early trace method gives interaction-focused HCI analytics a way forward, where event-level behavior is leveraged to inform design decisions at a faster pace and with simpler explanations.

REFERENCES

[1] L. Abb and J.-R. Rehse, "Process-related user interaction logs: State of the art, reference model, and object-centric implementation," *Information Systems*, vol. 124, p. 102386, 2024.

[2] A. Esposito, G. Desolda, and R. Lanzilotti, "A dataset of interactions and emotions for website user experience evaluation," *Scientific Data*, vol. 12, p. 1794, 2025.

[3] D. H. Jeong, B. K. Jeong, and S. Y. Ji, "Leveraging machine learning to analyze semantic user interactions in visual analytics," *Information*, vol. 15, no. 6, p. 351, 2024.

[4] E. Kuric, P. Demcak, and M. Krajcovic, "Validation of information architecture: Cross-methodological comparison of tree testing variants and prototype user testing," *Information and Software Technology*, vol. 183, p. 107740, 2025.

[5] I. Atoum, "Measurement of key performance indicators of user experience based on software requirements," *Science of Computer Programming*, vol. 226, p. 102929, 2023.

[6] A. Callejo and J. A. Macías, "Enhancing tree testing analysis to improve the usability evaluation of websites," *Behaviour & Information Technology*, vol. 45, no. 6, pp. 1–19, 2025.

[7] J. A. Macías and C. R. Borges, "Monitoring and forecasting usability indicators: A business intelligence approach for leveraging user-centered evaluation data," *Science of Computer Programming*, vol. 234, p. 103077, 2024.

[8] A. Rebmann and H. van der Aa, "Recognizing task-level events from user interaction data," *Information Systems*, vol. 124, p. 102404, 2024.

[9] A. Martínez-Rojas, A. Jiménez-Ramírez, J. G. Enríquez, and H. A. Reijers, "A screenshot-based task mining framework for disclosing the drivers behind variable human actions," *Information Systems*, vol. 121, p. 102340, 2024.

[10] R. Hamdani and I. Chihi, "Adaptive human-computer interaction for industry 5.0: A novel concept, with comprehensive review and empirical validation," *Computers in Industry*, vol. 168, p. 104268, 2025.