Can Artificial Intelligence Predict Human Vulnerability to Cyber Threats?

A Machine Learning Approach to Cybersecurity Behaviour

Mahek Chanda[1], Prof. Nita Patil[2]
[1]*Marathwada Mitra Mandal's College of Commerce, Pune*
[2]*Research Guide, Marathwada Mitra Mandal's College of Commerce, Pune*
doi.org/10.64643/ijirt.208434-459

***Abstract*—Individual users, rather than the technical systems around them, are frequently a critical point of vulnerability in cybersecurity: a reused password, an unexamined link, a postponed software update, or a banking session conducted over an unsecured public network can expose individuals to significant cyber risks. Existing approaches to behavioural cybersecurity often reduce vulnerability to a single overall score, which may indicate that an individual is at risk without identifying the specific behaviours contributing to that risk. At the same time, the growing application of supervised machine learning in cybersecurity has focused predominantly on organisational networks, systems, and software rather than on individual behavioural profiles. This paper addresses this gap by developing a multidimensional framework for analysing human vulnerability to cyber threats. The framework incorporates eight behavioural domains—password management, authentication, phishing response, device security, privacy practices, network behaviour, digital payment security, and incident response—along with a transparent and leakage-aware procedure for behavioural risk scoring. It further proposes a supervised machine-learning pipeline using Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbours, with model assessment based on accuracy, precision, recall, F1-score, and ROC-AUC rather than accuracy alone.**

A SHAP-based explainability layer is incorporated to connect model classifications with the behavioural factors underlying them, making the resulting assessment more transparent and actionable. By integrating behavioural cybersecurity measurement, machine learning, and explainable artificial intelligence, the study presents a structured and non-stigmatising approach to understanding individual cybersecurity vulnerability and provides a foundation for future empirical investigation and targeted cybersecurity awareness strategies.

***Index Terms*—Cybersecurity behaviour; behavioural risk classification framework; explainable machine learning; SHAP; human factors in security; risk categorisation; digital payment security**

I. INTRODUCTION

Most cybersecurity incidents that affect an individual user begin with an ordinary decision rather than a sophisticated attack: a password reused across accounts, a link opened without checking its sender, a security update postponed, or a banking session started over an unsecured public connection. The technical safeguards surrounding these actions are usually intact; what fails is the behaviour of the person operating them. Research into the individual differences behind such behaviour indicates that this failure is not random — traits such as impulsivity and habitual, extended internet use are measurably associated with a broader pattern of risky security choices [1] — which suggests that behavioural risk varies systematically between people rather than being reducible to a single, uniform trait.

Tools built to assess this risk have not generally kept pace with that understanding. The most rigorously validated behavioural instrument available, the Security Behavior Intentions Scale, treats device securement, password generation, software-update habits, and proactive awareness as four distinct, independently correlated sub-scales rather than as one figure [2]. Outside formal psychometric research, however, awareness tools aimed at general users typically do exactly what SeBIS avoids: behaviour is compressed into a single composite rating that can indicate a person is at risk without ever identifying which behaviour is actually responsible.

A separate line of research has applied supervised machine-learning classification to cybersecurity problems with considerable success, but almost entirely at the level of software vulnerabilities, network traffic, or organisational infrastructure. Work that applies the same classification techniques directly to an individual's behavioural profile, and then explains in accessible terms which behaviours are driving that classification, remains comparatively uncommon [3]. This is the space the present paper occupies — not on the premise that no individual-level work exists, since closely comparable work has already applied several classifiers to self-reported student behaviour [3], but on the argument that multi-domain behavioural measurement, supervised classification, and explainable-AI interpretation have not yet been brought together into one transparent framework for a general, non-organisational population. The remainder of this paper develops that framework, situates it against the literature that motivates it, and sets out the objectives, design, and analytical structure through which it is intended to operate.

II. LITERATURE REVIEW

Four bodies of research inform the framework developed here, and each supplies a different part of the case for why it is needed.

The first concerns why behavioural risk varies between individuals at all. Hadlington's work linking impulsivity and habitual internet use to risky security behaviour indicates that this risk has multiple, partly independent sources rather than one underlying cause [1]. If that is correct, a measurement approach that produces only a single number for a person's risk discards information about which source is actually operating in their case — which is the direct justification for treating the eight behavioural domains used in this framework (see Methodology) as separately measured constructs rather than collapsing them from the outset.

The second concerns whether self-reported behaviour can be trusted as data at all. SeBIS provides the strongest available answer: across two studies from the same research group, its four sub-scales have been shown to predict independently observed actions, including recognising a fraudulent website, generating a stronger password, and keeping a device locked [2, 5]. This matters because the present framework, like SeBIS, depends on self-report; the documented correlation between stated intention and observed behaviour is what makes that dependence defensible, and it is reinforced by the Theory of Planned Behaviour's account of intention as the immediate precursor to action [4]. SeBIS, however, was never extended into a composite risk classification, and its four sub-domains stop short of privacy management, shared-network use, and digital payments domains that matter considerably to a student population that accesses the internet mainly through a personal smartphone.

The third body of work concerns the machine-learning side of the problem, where the literature is comparatively thin at the individual level. Most classification work in cybersecurity targets malware signatures, intrusion detection, or vulnerable code rather than a person's behaviour. Ting Tin Tin and colleagues are among the few exceptions: using self-reported behavioural data collected from 207 undergraduates, they trained five classifiers Logistic Regression, K-Nearest Neighbours, Decision Tree, Support Vector Machine, and Naïve Bayes to estimate risk across malware, social-engineering, and password-attack categories [3]. Their results establish that individual-level classification is technically achievable with a modest student sample, which is the precedent this paper's design relies on most directly, while their own framing of the work as filling a gap left by organisation-focused research confirms how uncommon this application still is.

The fourth strand concerns what should happen after a model produces a classification. A label such as "high risk" has limited practical value, and may be actively unhelpful, if it cannot be traced back to the behaviours that produced it, particularly where the intended use is supportive intervention rather than judgement of the person classified. SHAP addresses this by distributing a model's prediction across its contributing features using a method grounded in cooperative game theory, and has become close to a default choice for explaining both tree-based and general classifiers [6]; LIME offers a related but methodologically distinct alternative built on local approximation [7]. Neither method has yet been reported in combination with individual-level behavioural cybersecurity classification specifically, though both have precedent

in adjacent problems such as phishing-website detection.

Read together, these four strands establish that behavioural risk is multi-domain and can be captured through validated self-report [1, 2, 4, 5], that supervised classifiers can be trained on such data even with a modest sample [3], and that established tools exist for making the resulting classifications interpretable [6, 7]. No single study reviewed here, however, connects all three elements for a general population outside an organisational setting. The framework developed in this paper is built specifically to occupy that connecting position.

III. OBJECTIVES

1. To develop a multi-domain instrument capable of assessing self-reported cybersecurity behaviour across password, authentication, phishing, device, privacy, network, and digital-payment domains.
2. To synthesise the verified literature in order to identify which behavioural indicators are theoretically expected to be most strongly associated with elevated behavioural risk, stated as explicit hypotheses rather than findings.
3. To design a transparent, study-specific behavioural risk-scoring method that is clearly distinguished from any validated clinical or psychometric scale.
4. To specify a supervised machine-learning classification pipeline, with explicit safeguards against target leakage, capable of assigning individuals to behavioural risk categories from a behavioural feature profile.
5. To define a model-evaluation framework that reports multiple performance metrics rather than accuracy alone.
6. To establish a SHAP-based explainability approach for translating classifier outputs into practical, non-judgemental guidance for cybersecurity awareness interventions.

IV. PROBLEM STATEMENT

Security incidents affecting individual users are frequently the product of specific, identifiable behavioural choices rather than of failures in the technical controls around them. Existing tools for assessing this behaviour largely reduce it to a single composite score, which limits their diagnostic use: a low score signals elevated risk without indicating which behavioural domain is responsible, and so cannot by itself guide a targeted response. Machine-learning classification, meanwhile, has been applied extensively to cybersecurity data, but overwhelmingly at the level of software systems or organisational networks rather than to the behavioural profile of an individual. The result is a shortage of frameworks that operationalise several distinct behavioural domains as structured features, apply supervised classification to estimate an individual's behavioural risk category from those features, and explain which domains are driving that classification in terms a non-technical user could act on. This is the problem the present framework is designed to address, without extending its claims to the considerably stronger and empirically unsupportable question of whether a specific individual will actually be attacked.

V. NEED / RELEVANCE OF THE STUDY

- Academic relevance: the framework fills a specific integration gap between validated but score-level behavioural instruments and organisation-level machine-learning classification, drawing the two together into one individual-level, multi-domain design.
- Practical relevance: domain-specific classification, rather than a single score, would allow an institution to target awareness activities at the behaviours actually driving risk in its student population, rather than issuing generic advice.
- Social relevance: college students are frequently targeted by phishing and digital-payment fraud, so a framework capable of identifying which behavioural domains contribute most to their risk has direct protective value.
- Technological relevance: the design treats explainability as a requirement rather than an optional extra, aligning it with current expectations that applied machine-learning systems be interpretable rather than accurate-but-opaque.

VI. METHODOLOGY

This study presents a quantitative framework that combines behavioural cybersecurity indicators with a supervised machine-learning classification approach. The framework is designed to examine patterns

associated with individual vulnerability to cyber threats by converting cybersecurity-related behaviours into measurable variables and applying classification techniques. The approach is exploratory and explanatory rather than experimental, focusing on identifying behavioural associations rather than establishing causal relationships.

VII. BEHAVIOURAL DIMENSIONS AND MEASUREMENT

The framework covers eight behavioural domains: password behaviour, authentication, phishing response, device security, privacy practices, network behaviour, digital payment security, and incident response. These domains represent key areas in which individual practices may influence exposure to cyber threats. The framework draws conceptually on the Security Behavior Intentions Scale (SeBIS) while extending its coverage to privacy, network usage, digital payments, and incident response [2].

The behavioural indicators are measured using Likert-type scales, with coding structured so that higher scores consistently represent safer practices. Risk-oriented items are reverse-coded before domain scores are calculated. The framework does not involve the collection of passwords, OTPs, card numbers, or other confidential credentials.

Table 1. Behavioural Domains and Risk Interpretation

Behavioural Domain	Items	Measurement / Coding	Expected Relationship
Password behaviour	5	Likert composite; risk items reverse-coded	Weaker practices → higher vulnerability
Authentication	2	Composite measure	Lower MFA use → higher vulnerability
Phishing response	5	Likert composite; risk items reverse-coded	Greater susceptibility → higher vulnerability
Device security	4	Composite measure; risk items reverse-coded	Poor device protection → higher vulnerability
Privacy practices	3	Composite measure; risk items reverse-coded	Weaker privacy management → higher vulnerability
Network behaviour	3	Composite measure; risk items reverse-coded	Risky network use → higher vulnerability
Digital payment security	3	Composite measure; risk items reverse-coded	Unsafe payment practices → higher vulnerability
Incident response	2	Composite measure	Lower response awareness → higher vulnerability

Source: Author's framework, conceptually informed by Egelman and Peer [2] and extended to include privacy practices, network behaviour, digital payment security, and incident response.

Behavioural Risk Scoring

Since no universally validated composite instrument was identified for combining all eight behavioural domains, the framework uses a study-specific composite scoring method, rather than presenting the score as an established psychometric measure [2], [5].

Protective and risk-oriented indicators are identified in advance, with risk-oriented items reverse-coded so that higher values consistently represent safer behaviour. Scores within each domain are averaged and subsequently standardised to prevent domains with different numbers of indicators from disproportionately influencing the overall measure. The standardised domain scores are then combined through an unweighted average, as there is no established empirical basis for assigning different weights before analysis.

For interpretation, the resulting scores may be classified into Low, Moderate, and High vulnerability categories using relative thresholds such as terciles or quartiles. These categories are analytical classifications rather than universally applicable cybersecurity risk thresholds.

VIII. MACHINE-LEARNING CLASSIFICATION

The proposed machine-learning pipeline considers Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbours as primary classifiers because of their interpretability and use in comparable behavioural classification research [3]. Support Vector Machine and Gradient Boosting may be considered as secondary comparisons where appropriate.

The pipeline incorporates data cleaning, missing-value handling, categorical-variable encoding, exploratory analysis, feature construction, and feature selection. Both individual behavioural indicators and domain-level composite variables may be examined as potential predictors.

Model selection is structured around a stratified 80:20 train-test split and stratified five-fold cross-validation. Hyperparameter tuning is limited to settings where meaningful differences in cross-validated performance are observed. Preprocessing parameters such as scaling and encoding are fitted only within the relevant training data to reduce information leakage.

A major methodological concern is targeting leakage. If the same behavioural indicators are used to construct the vulnerability category and predict that category, model performance may be artificially inflated. The framework therefore recommends either withholding selected domains from the predictor set or, preferably, using an independently meaningful outcome—such as previous cybersecurity incident experience—as the classification target while retaining the composite vulnerability score as a descriptive measure.

Model Evaluation and Explainability

Model performance is assessed using accuracy, precision, recall, F1-score, and confusion matrices. Where appropriate, ROC-AUC may also be reported. Using multiple metrics is important because accuracy alone can be misleading when vulnerability categories are imbalanced. Precision and recall provide greater insight into how effectively different risk categories are identified.

Interpretability is incorporated through SHAP (SHapley Additive exPlanations) as the primary explanation technique [6]. For Logistic Regression, standardised coefficients provide an additional interpretable measure, while LIME may be used as a secondary robustness check [7]. Feature importance and SHAP results are interpreted as associations within the predictive model and are not treated as evidence of causal behavioural mechanisms.

IX. ETHICAL CONSIDERATIONS AND LIMITATIONS

The framework follows principles of informed participation, privacy, data minimisation, and responsible interpretation. It excludes sensitive credentials and recommends avoiding personally identifying information unless essential. Any individual-level data should be securely stored and used only for the stated research purpose. Vulnerability categories should be presented in neutral, improvement-oriented language rather than as labels of personal fault.

The framework has several limitations. Self-reported behavioural measures may not perfectly reflect actual cybersecurity practices, although previous research has identified relationships between behavioural-security measures and independently observed actions [5]. Non-probability sampling, where applied, may limit generalisability, while a cross-sectional design captures behaviour at a particular point in time. Finally, the composite vulnerability score is a study-specific, non-validated analytical construct and should not be interpreted as a definitive prediction of whether an individual will become a victim of a cyberattack. Behavioural vulnerability and actual cyberattack victimisation are related but distinct concepts.

Observation

The examination of the reviewed literature and the proposed behavioural framework reveals several important patterns relevant to the prediction of cybersecurity vulnerability. The first observation is that cybersecurity behaviour is multidimensional rather than limited to a single practice. The SeBIS literature identifies meaningful behavioural dimensions related to password management, device security, software updating, and proactive awareness [2], [5]. This provides a conceptual basis for examining cybersecurity behaviour through multiple measurable domains rather than treating vulnerability as a single behavioural characteristic.

A second observation is that individual behavioural differences can extend across multiple aspects of digital activity. Research examining behavioural and

personality-related factors indicates that characteristics such as impulsivity and patterns of internet use may be associated with broader risky online behaviours [1]. This supports the use of a multidimensional framework in which vulnerability is examined across password practices, authentication, phishing response, device security, privacy, network behaviour, digital payments, and incident response.

A third observation emerges from existing machine-learning research on cybersecurity behaviour. Comparable classification research has demonstrated the feasibility of applying supervised learning techniques to individual-level behavioural data, including models such as Logistic Regression, K-Nearest Neighbours, Decision Trees, Support Vector Machines, and Naïve Bayes [3]. This provides methodological support for considering interpretable classification algorithms within the proposed framework.

The structure of the proposed framework further indicates that different behavioural domains may contribute differently to overall cybersecurity vulnerability. Table 1 assigns a specific risk direction to each domain, allowing behaviours associated with weaker security practices to be distinguished from protective behaviours. This approach provides a systematic basis for converting qualitative aspects of cybersecurity behaviour into measurable variables suitable for computational analysis.

Another observation concerns the scope of existing behavioural-security measurement. While SeBIS provides an established foundation for several important security behaviours [2], the framework developed in this study extends the measurement perspective to include privacy practices, network behaviour, digital payment security, and incident response. These additional domains reflect the wider range of digital activities through which individuals may encounter cybersecurity risks.

Taken together, these observations indicate that human vulnerability to cyber threats can be approached as a multidimensional behavioural classification problem. The literature provides support for measuring individual security behaviours and for applying machine-learning methods to behavioural data, while the proposed framework integrates these strands into a broader structure. This creates a basis for examining whether combinations of behavioural indicators can be used to distinguish different levels or patterns of cybersecurity vulnerability.

Interpretation

The observations indicate that the proposed framework has a clear theoretical basis while also containing areas that require further empirical validation. Evidence from the SeBIS literature shows relationships between reported security behaviours and certain independently observed actions [2], [5]. This supports the use of behavioural indicators as meaningful measures of cybersecurity practices, particularly in areas such as password management, device security, and security awareness. However, this evidence should not be extended automatically to every domain included in the present framework. Privacy practices, network behaviour, and digital-payment security require separate validation, and are therefore treated as additional behavioural dimensions rather than established validated measures.

The literature linking impulsivity and habitual internet use with broader patterns of risky online behaviour [1] further supports a multidimensional understanding of cybersecurity vulnerability. This suggests that vulnerability may reflect the interaction of several behavioural tendencies rather than a single isolated practice. Consequently, examining individual domain scores alongside an overall vulnerability measure can provide a more informative representation of behavioural patterns. The eight-domain structure therefore allows different aspects of cybersecurity behaviour to be examined separately while still permitting them to be considered collectively.

Existing machine-learning research also provides methodological support for applying supervised classification to behavioural cybersecurity data. The study by Ting Tin Tin and colleagues demonstrates the use of several classification algorithms with self-reported behavioural information [3]. This supports the selection of multiple candidate models within the present framework. However, the findings of that study cannot be directly transferred to the proposed framework because the behavioural domains, feature structure, and research context differ. Comparing several algorithms and evaluation measures is therefore more appropriate than assuming that one particular classifier will perform best.

The proposed scoring framework also has an important interpretive implication. Since no

established composite measure covering all eight behavioural domains was identified in the reviewed literature [2], [5], the composite vulnerability score should be understood as a study-specific analytical construct rather than a universally validated measure. Its primary value lies in providing a consistent method for organising and comparing behavioural indicators. The use of relative categories such as Low, Moderate, and High vulnerability should similarly be interpreted as an analytical classification rather than as an externally established standard of cybersecurity risk.

The framework's extension beyond the behavioural dimensions represented in SeBIS provides an opportunity to examine cybersecurity behaviour more broadly. Privacy management, network practices, digital-payment behaviour, and incident response have become increasingly relevant to everyday digital activity, yet their integration into a single behavioural classification framework requires further empirical examination. Their inclusion therefore represents a conceptual extension of the existing literature rather than a claim that these dimensions have already been validated within the proposed model.

Overall, the interpretation suggests that machine learning can provide a structured approach to examining patterns of cybersecurity behaviour, provided that predictive outputs are interpreted cautiously. The framework is theoretically informed, incorporates established behavioural-security research, and includes safeguards against issues such as target leakage and over-reliance on accuracy. At the same time, its predictive performance and the validity of the proposed composite measure remain questions for empirical evaluation. The framework should therefore be viewed as a structured foundation for analysing human vulnerability to cyber threats rather than as a completed predictive model whose effectiveness has already been demonstrated.

X. CONCLUSION

This study develops a structured framework for examining human vulnerability to cyber threats through the integration of behavioural cybersecurity measurement, supervised machine-learning classification, and explainable artificial intelligence. The framework draws on established behavioural-security research, including the SeBIS literature [2], [5], research on individual differences in risky online behaviour [1], and previous applications of machine-learning techniques to behavioural cybersecurity data [3]. These strands of research are brought together into a unified approach for analysing cybersecurity behaviour at the individual level.

A central contribution of the framework is its multidimensional treatment of cybersecurity behaviour. Rather than representing vulnerability through a single behavioural characteristic, it considers password practices, authentication, phishing response, device security, privacy, network behaviour, digital-payment security, and incident response. This broader structure recognises that an individual's cybersecurity posture can be influenced by several interconnected aspects of everyday digital behaviour.

The framework also introduces a structured approach to behavioural risk scoring and classification. Domain-level measures can be combined into an overall analytical vulnerability score while retaining the individual domains for more detailed interpretation. The proposed machine-learning pipeline further provides a basis for comparing multiple classification algorithms using measures such as precision, recall, F1-score, and accuracy. Incorporating SHAP and other interpretability techniques adds an additional layer of transparency by allowing model outputs to be examined in relation to specific behavioural indicators [6], [7].

At the same time, the framework distinguishes between theoretically supported behavioural dimensions and those requiring further validation. The established literature provides stronger support for some security behaviours, while the additional domains incorporated into this framework represent an extension that requires empirical assessment. Similarly, the proposed composite vulnerability score is a study-specific analytical construct rather than a universally validated measure. These distinctions are important for ensuring that predictive cybersecurity research does not present methodological assumptions as established facts.

Overall, the study demonstrates how behavioural cybersecurity research and machine-learning techniques can be integrated into a transparent framework for examining individual vulnerability. Its value lies not in assigning predetermined risk labels, but in providing a systematic structure through which behavioural indicators can be measured, analysed, classified, and interpreted. The framework therefore

offers a foundation for future empirical investigation into whether combinations of everyday cybersecurity behaviours can meaningfully contribute to the prediction and understanding of human vulnerability to cyber threats.

REFERENCES

[1] L. Hadlington, "Human factors in cybersecurity: Examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours," *Heliyon*, vol. 3, no. 7, Art. no. e00346, 2017.

[2] S. Egelman and E. Peer, "Scaling the security wall: Developing a Security Behavior Intentions Scale (SeBIS)," in Proc. 33rd Annu. ACM Conf. Human Factors in Computing Systems (CHI '15), 2015, pp. 2873–2882, doi: 10.1145/2702123.2702249.

[3] T. T. Tin, J. X. Khiew, A. Aitizaz, K. T. Lee, C. K. Teoh, and H. Sarwar, "Machine learning based predictive modelling of cybersecurity threats utilising behavioural data," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 9, pp. 832–840, 2023.

[4] P. S. Bhosale, S. Shevkari, and G. A. Mule, "Machine learning in cybersecurity: Cutting-edge approaches to threat detection and protection," Int. J. Res. Publ. Rev., vol. 6, no. 11, pp. 7090–7092, 2025.

[5] S. Egelman, M. Harbach, and E. Peer, "Behavior ever follows intention? A validation of the Security Behavior Intentions Scale (SeBIS)," in Proc. 2016 CHI Conf. Human Factors in Computing Systems, 2016, pp. 5257–5261, doi: 10.1145/2858036.2858265.

[6] P. Shigvan, S. Shevkari, V. Auti, and G. Kumar, "Detecting and mitigating insider threats with artificial intelligence," Int. J. Innovative Inventions Social Sci. Humanities, 2025.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.