GMR
The Original

UNDERSTANDING DATA PRIVACY BEHAVIOR AMONG COLLEGE STUDENTS USING EXPLAINABLE MACHINE LEARNING: THE ROLE OF PRIVACY LITERACY, AUTHENTICATION PRACTICES, AND RISK PERCEPTION

Nupur Pandey[1*], Angrish K. Agarwal[2], Arun Kumar Tripathi[3]

[1] Department of Management, Invertis University, Bareilly, India
[2] Department of Management, Invertis University, Bareilly, India
[3] Department of Computer Science, NIET, Greater Noida, India
Corresponding Author: Nupur Pandey, mailtonupurtripathi@gmail.com

ABSTRACT
A central question in academic institutions cybersecurity and privacy research is why only some students engage in online privacy-protective behavior while others do not. The previous research relies on linear models that may overlook nonlinear and interactive effects. To address this gap, we applied explainable machine learning to survey data from 745 undergraduate and postgraduate students across seven universities in India, using twelve theory-driven predictors spanning privacy literacy, risk perception, institutional trust, privacy fatigue, and contextual constraints. Five classifiers were evaluated after grid-search tuning and 10-fold cross-validation, and XGBoost emerged as the best-performing model, achieving 68.5% accuracy and an AUC of 0.704 on a held-out test set. SHAP analysis identified privacy literacy as the strongest predictor of privacy-protective behavior, followed by privacy fatigue and reactive behaviors, while the remaining variables contributed relatively little. These findings were corroborated by logistic regression and t-tests, and exploratory clustering suggested weak segment separation, indicating that students vary more continuously than as distinct behavioral types. Overall, the results highlight privacy literacy as a key target for university interventions and demonstrate the value of explainable machine learning for transparent and interpretable student privacy research. Self-reported privacy-policy reading frequency and two-factor authentication use were not predictive of broader protective behavior. These findings suggest that Institutions interventions should prioritize literacy-building over policy-transparency messaging, and that explainable AI offers a transparent, audit-ready alternative to black-box prediction in student-facing cybersecurity research.

KEYWORDS: Privacy literacy; privacy protective behavior; explainable artificial intelligence; machine learning; two-factor authentication; risk perception; college students

I. INTRODUCTION

Academia has turned into a data-driven workplace. Course management systems, proctoring software, library portals, financial-aid platforms and campus mobile apps each have a layer of personal-data collection with most students having little realistic choice but to acquiesce. In this context, some students take explicit protective actions by enabling two-factor authentication, limiting app permissions and evaluating data requests; others do not perform these measures even though they are subjected to the same exposure. Understanding this heterogeneity has been a long-time central focus of information-privacy research: for two decades, the prevailing analytical method in this literature has remained linear (Dinev & Hart, 2006; Malhotra et al., 2004), dominated by structural equation models and ordinary least-squares regressions which fundamentally assume additive monotonic relationships between constructs like privacy concern, trust, and behavior. It adds an additional perspective through machine learning. Tree-based ensemble methods such as random forests and gradient boosting can capture nonlinear thresholds and interaction effects that linear models miss, and have been applied successfully to predict related behaviors such as privacy fatigue itself (Alwafi & Fakieh, 2024) and pandemic-era self-protective behavior (Taye et al., 2023). Historically, however, the predictive gains from such models have come at the cost of interpretability a serious limitation in a domain where the practical goal is not merely to predict who will protect their data, but to understand why, so that institutions administrators can design targeted interventions. Explainable artificial intelligence (XAI) techniques, particularly SHapley Additive exPlanations (SHAP), resolve this tension by attributing each prediction to specific feature contributions in a theoretically grounded, model-agnostic way (Lundberg & Lee, 2017), and have been increasingly adopted across cybersecurity prediction tasks for exactly this reason (Mohale & Obagbuwa, 2025). This data collection occurs against a fragmented global regulatory backdrop spanning frameworks such as the GDPR, the CCPA, and India's evolving data-protection regime (Pandey et al., 2025), which differ markedly in the protections they afford students as data subjects.

This study addresses three gaps. First, despite growing XAI adoption in network-level intrusion detection, few studies apply SHAP-based explainable machine learning to individual-level student privacy behavior specifically. Second,

existing studies rarely benchmark multiple algorithm families (linear, tree-based, kernel-based) against one another on the same privacy-behavior outcome with full statistical validation. Third, whether students segment into qualitatively distinct privacy-behavior "types" as opposed to varying continuously along a small number of dimensions remains an open empirical question that this study tests directly rather than assumes. Accordingly, this study investigates the following research questions:

A. **RQ1:** Which factors significantly influence privacy protection behavior among college students?
B. **RQ2:** Can machine learning accurately predict privacy protection behavior?
C. **RQ3:** Which privacy-related factors are the strongest predictors, as identified by explainable AI?
D. **RQ4:** Are there distinct privacy behavior segments among students?
E. **RQ5:** How can Institutions use these findings to improve students' privacy and cybersecurity practices?

The remainder of this paper proceeds as follows. Section II reviews the relevant theoretical and empirical literature. Section III specifies the research hypotheses. Section IV details the data, feature engineering, and analytical pipeline. Section V reports descriptive, clustering, machine learning, explainability, and statistical-validation results in full. Section VI discusses the findings against theory and prior literature. Sections VII conclude with implications.

II. LITERATURE REVIEW

This section deals with literature review of previous study based on following six parameters.

A. Information Privacy Concern and Privacy Literacy: Hong and Thong (2013) integrated decades of fragmented privacy-concern research into a second-order model spanning interaction management and information management, establishing that privacy concern is multidimensional and that different dimensions predict different downstream behaviors. Malhotra et al.'s (2004) Internet Users' Information Privacy Concerns (IUIPC) framework operationalized a closely related structure around collection, control, and awareness, and remains the most widely validated privacy-concern instrument in information systems research. Privacy literacy a person's declarative and procedural knowledge of data-collection practices and the tools available to manage them is theoretically distinct from concern: a student can be highly concerned yet lack the literacy to act, or possess high literacy yet feel little concern. Bartsch and Dienlin (2016) demonstrated empirically that online privacy literacy directly predicts privacy-protective behavior on social network sites, independent of generalized concern, a finding this study's results corroborate and extend using explainable machine learning rather than structural equation modeling.

B. Privacy Calculus, the Privacy Paradox, and Privacy Fatigue: Dinev and Hart's (2006) extended privacy calculus model proposes that users weigh anticipated benefits against perceived privacy risk before adopting protective measures; under this model, protective behavior should track literacy and risk perception in a straightforwardly rational way. Yet Barth and de Jong's (2017) systematic review finds that stated privacy concern is, at best, a weak and inconsistent predictor of actual protective behavior the well-documented privacy paradox. Choi et al. (2018) advanced this literature by formally distinguishing privacy fatigue emotional exhaustion and cynicism toward privacy management from privacy concern, and showed that fatigue predicts disengagement coping more strongly than concern does. Draper and Turow (2019) extend this account structurally, arguing that platforms' own design and policy practices actively cultivate "digital resignation," a learned helplessness distinct from ignorance. Most directly relevant to the present study, Alwafi and Fakieh (2024) used machine learning to predict privacy-fatigued users from their responses to personalized social-media advertising, demonstrating that fatigue itself is a tractable machine-learning prediction target a precedent this study extends by treating fatigue as a predictor of, rather than the outcome of, protective behavior.

C. Authentication Behavior and Protection Motivation Theory: Two-factor authentication (2FA) adoption is most often theorized through Protection Motivation Theory (PMT), which models security behavior as a function of threat appraisal and coping appraisal. In a large recent test of PMT among Chinese college students (N=3,030), Han et al. (2025) found that both threat and coping appraisal significantly predicted information-security behavior. Ogbanufe and Baham (2023) extend PMT with anticipated regret to explain multi-factor authentication adoption specifically, while Das et al. (2019) document that even technically capable populations resist 2FA when they feel confident in alternative protections such as strong passwords. It is from this body of literature that the motivation comes for using 2FA usage not only as a predictor but also as a benchmark to see if a specific security practice falls under the same antecedents as more general protective practices.

D. The Limits of Notice-and-Choice: An extensive amount of literature has raised doubts about whether reading privacy policies serves as any kind of protection at all. Obar and Oeldorf-Hirsch (2020) discovered through their experiments that 74% of subjects bypassed the privacy policy altogether when provided with a fast signup process, calling "I agree" the biggest lie on the Internet. This literature predicts that self-reported policy-reading frequency should be weakly related, at best, to actual protective behavior a prediction this study tests directly.

E. Explainable Machine Learning in Cybersecurity and Behavioral Prediction: Lundberg and Lee (2017) introduced SHAP as a unified, game-theoretically grounded framework for attributing model predictions to individual features, resolving the long-standing trade-off between predictive accuracy and interpretability. SHAP has since become the dominant explainability technique in cybersecurity machine learning, particularly for tree-based ensembles such as random forests (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016), which routinely outperform linear models on

tabular behavioral data while remaining auditable through SHAP decomposition. Beyond network intrusion detection, SHAP-explained random forests have been used to identify the strongest predictors of pandemic-era self-protective behavior, finding that individual knowledge and institutional trust ranked among the top predictors (PMC, 2023) a methodological precedent this study follows directly, substituting privacy-protective behavior as the predicted outcome and privacy literacy and institutional trust among the candidate predictors.

F. Research Gap and Contribution: Three gaps motivate this study. First, although privacy literacy, fatigue, and risk perception have each been linked to protective behavior in isolation, no identified study benchmarks five distinct classifier families against one another on the same multi-factor student privacy dataset with full hyperparameter tuning and cross-validation. Second, almost no published study pairs machine-learning prediction with SHAP explainability and classical inferential statistics (logistic regression, ANOVA, chi-square, t-tests) on the same sample, which is necessary to establish that black-box feature rankings agree with theoretically interpretable effect sizes. Third, whether students form genuine behavioral segments, as opposed to varying continuously, has typically been assumed rather than tested via formal cluster-validity diagnostics. This study closes all three gaps within a single, fully reported analytical pipeline.

III. RESEARCH HYPOTHESES

This session builds the research hypothesis based on literature review done in Section II. The hypothesises are listed as follows:

A. H1. Privacy literacy will positively and significantly predict privacy-protective behavior, net of other measured factors.
B. H2. Privacy fatigue will negatively and significantly predict privacy-protective behavior, net of other measured factors.
C. H3. An ensemble machine learning classifier (random forest or XGBoost) will outperform a linear baseline (logistic regression) in predicting privacy-protective behavior, consistent with the presence of nonlinear or interactive effects among predictors.
D. H4. SHAP-based feature importance rankings will converge with classical inferential statistics (logistic regression odds ratios, t-tests) in identifying the strongest predictors of protective behavior.
E. H5. Students will form statistically distinct, well-separated privacy behavior segments (silhouette score > 0.25).
F. H6. Self-reported privacy-policy reading frequency and two-factor authentication use will be associated with privacy-protective behavior.

IV. RESEARCH METHODOLOGY

This section deals with research methodologies used during the research.

A. Research Design and Data Source: This study used a cross-sectional, quantitative survey design. Data were drawn from a survey of 745 undergraduate and postgraduate students across seven institutions in India, spanning five academic disciplines and four/five years-of-study categories. No primary data collection occurred for this analysis; the dataset was provided in full for secondary analysis. In keeping with proper procedures for data management, the original respondent-level database is neither posted nor made public as part of this manuscript; instead, the information presented is based on summary statistics, output from models, and anonymous tables.

B. Variable Availability and Mapping: The six potential predictor variables in the initial research brief have been specified below along with their availability in the data set or lack thereof. Table 1 explains the status of these variables in detail.

Table 1. Requested Variable Availability Mapping

Requested Variable	Availability in Dataset
Privacy Literacy	Available - F1_Privacy_Literacy
Privacy Risk Perception	Available - F2_Risk_Perception
Use of Two-Factor Authentication	Available - Uses_2FA (binary)
Privacy Awareness	Not available as a distinct variable (F1 used as nearest proxy)
Trust in Digital Platforms	Partial match - F3_Trust_in_Institutions (institutions, not platforms specifically)
Privacy Policy Reading Habit	Available - Policy_Reading_Frequency / Policy_Reading_Numeric

C. Final Feature Set and Dependent Variable: The final set of predictors was composed of twelve composite factors (F1-F12) rated on a scale of 1 to 5. These were Privacy Literacy, Risk Perception, Trust in Institutions, Privacy Fatigue, Proactive Behaviors, Reactive Behaviors, Social Media Monitoring, Caution About Financial Information, Discipline-based Technology Engagement, Influence of Academic Role, Cultural Privacy Norms, and Economic Barriers, alongside Policy Reading Habit, 2FA Adoption, and demographic controls (age, gender, discipline, and year of study). The dependent variable, Privacy_Protection_Score, is a continuous composite (M = 2.77, SD = 0.58, range = 1.12–4.49). In order to facilitate the computation of the classification measures identified in the analysis plan (accuracy, precision, recall, ROC-AUC, confusion matrices), this continuous scale was dichotomized on a median split at 2.79, resulting in a variable called Privacy_Protection_Class, where 377 respondents (50.6 percent) fall in the category of "High" privacy protection while 368 respondents (49.4 percent) are categorized as "Low." The change has been highlighted explicitly to influence

how the results are understood as they reflect the attributes associated with high and low protection students rather than a specific clinical or policy threshold.

D. Data Preparation: In terms of missing values, zero missing value cells existed in 23 out of the 745 participants in 14 variables. One and a half the size of interquartile range was used to detect outliers in nine of the 14 continuous variables, impacting less than 1% of the cases. In addition, by placing all the variables within 1-5 Likert scale composites, they were considered valid.
E. Analytical Pipeline: Figure 1 illustrates the entire nine-step analytical process starting with the sample and the instrument and ending with interpretation.
F. Clustering Procedure: The K-means algorithm was used for clustering the scaled F1-F12 variables. The optimal number of clusters was found through the elbow plot technique (within-clusters sum of squares) and silhouette analysis (k from 2 to 10). It should be noted that the silhouette index was considered to be a more important criterion than the elbow plot. This is because the latter measures only cluster separation. The four-cluster solution was considered in addition to the silhouette optimal k as required in the research brief.

G. Machine Learning Pipeline: The data was partitioned 80:20 to form the training and test datasets with stratification based on the binary class label. Five classifiers were used: Logistic Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine (SVM). The tuning of each classifier was done via grid search using the appropriate hyperparameter space for that particular model, where the performance of each model was measured using 10-fold stratified cross-validation on the training dataset by maximizing the F1-score metric. The accuracy, precision, recall, F1-score, ROC-AUC score, and confusion matrix were calculated using the test set for each classifier. Standardization was applied to Logistic Regression and SVM, whereas non-standardized input was used for the tree-based algorithms. All models were implemented in Python using scikit-learn (Pedregosa et al., 2011).

H. Explainable AI (SHAP) Analysis: SHAP (SHapley Additive exPlanations) values were calculated for the optimal tree-based algorithm (XGBoost) via TreeExplainer method (Lundberg & Lee, 2017). Mean absolute SHAP value was used to calculate the global feature importance for all features, whereas the local influence was illustrated in a SHAP summary plot (beeswarm). Dependence plots were created for the two top-ranked features.
I.Statistical Validation: In order to verify if there was convergence of the feature importance obtained through machine learning with inferential statistics (H4), binary logistic regression analysis using statsmodels was conducted on the same twelve factors with confidence intervals and odds ratio. The chi-square test for independence was used to determine the association between categorical variables (gender, discipline, and year of study) and 2FA usage on one side, and the binary protection class variable on the other side. One way ANOVA along with Tukey HSD post hoc tests was used to analyze the differences between disciplines on privacy literacy and protection behavior.

V. RESULTS

This section deals with detailed discussion on result analysis hypothesis and research methodologies. This section if divided into eight sub sections.

A. Sample Characteristics: Table 2 shows the broad characteristics of collected data.

Table 2. Sample Characteristics (N = 745)

Characteristic	Category	n	%
Age	M = 21.64, SD = 2.32 (range 18–25)	745	100.0
Gender	Male	400	53.7
	Female	334	44.8
	Prefer not to say	11	1.5
Discipline	Engineering/Technology	243	32.6
	Business/Management	205	27.5
	Humanities/Social Sciences	144	19.3
	Natural Sciences	89	11.9
	Other	64	8.6
Year of Study	First Year	199	26.7
	Second Year	216	29.0
	Third Year	180	24.2
	Fourth/Final Year	113	15.2
	Postgraduate	37	5.0
2FA Adoption	Uses 2FA	491	65.9
	Does not use 2FA	254	34.1

B. Data Preparation: Missing Values, Outliers, and Descriptive Statistics: No missing values were found across all 745 cases and 23 variables. Outlier screening (1.5× IQR) identified extreme responses on F1 (n=7, 0.9%), F2 (n=2), F5 (n=5), F7 (n=2), F8 (n=7), F9 (n=1), F10 (n=3), and Protective_Behavior_Score (n=5); all other variables showed zero

IQR-defined outliers. Given the bounded 1–5 measurement scale, these were retained as genuine extreme responses. Table 3 shows the descriptive statistics for collected data.

Table 3. Descriptive Statistics for All Measured Variables (N = 745)

Variable	M	SD	Min	Max	Skew
F1 Privacy Literacy	3.69	0.72	1.00	5.00	−0.43
F2 Risk Perception	3.51	0.68	1.48	4.99	−0.13
F3 Trust in Institutions	3.13	0.78	1.01	4.98	−0.06
F4 Privacy Fatigue	3.25	0.75	1.28	4.99	−0.12
F5 Proactive Protection	3.50	0.69	1.20	5.00	−0.25
F6 Reactive Behaviors	3.38	0.68	1.59	4.96	−0.02
F7 Social Media Vigilance	3.46	0.72	1.05	4.98	−0.28
F8 Financial Data Caution	3.54	0.68	1.45	5.00	−0.31
F9 Disciplinary Tech Engagement	3.44	0.70	1.49	4.99	−0.12
F10 Academic Role Influence	3.29	0.72	1.03	4.98	−0.12
F11 Cultural Privacy Norms	3.18	0.75	1.14	4.96	−0.12
F12 Economic Constraints	3.03	0.80	1.01	4.93	−0.08
Privacy Protection Score (DV)	2.77	0.58	1.12	4.49	−0.11
Policy Reading Frequency (1–5)	3.11	1.31	1	5	−0.10

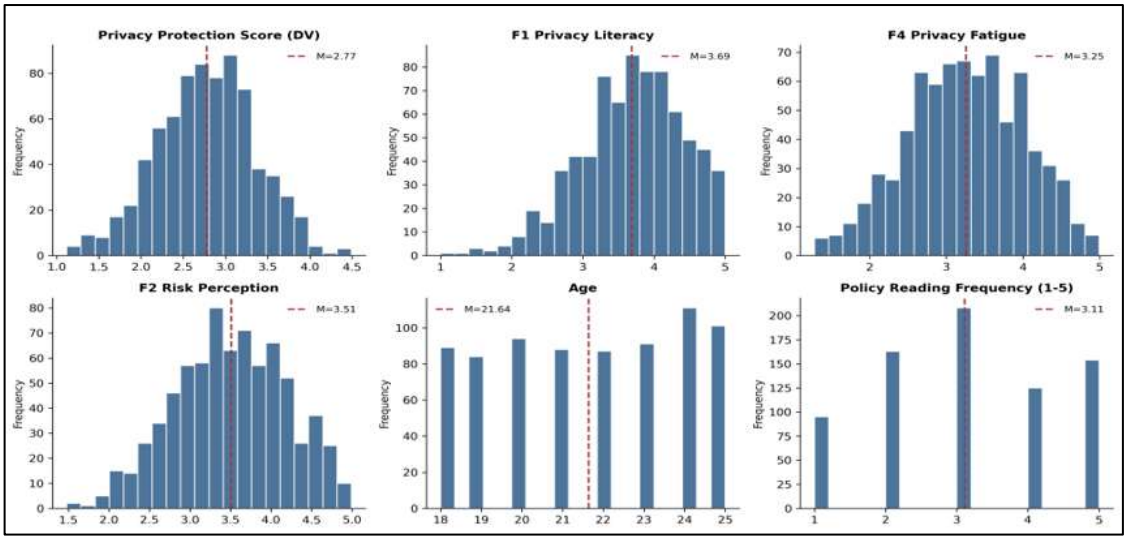

Figure 1. Distribution of key study variables. All Likert-type composites approximate a roughly symmetric, slightly left-skewed distribution; age and policy-reading frequency reflect the discrete sampling categories used at collection.

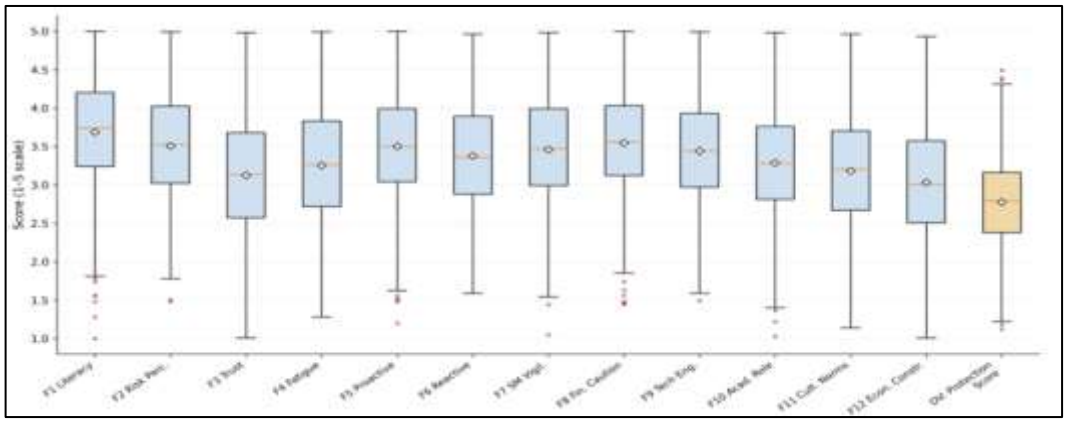

Figure 2. Boxplot distributions of all twelve factors and the dependent variable, showing the limited number of IQR-defined outliers (red points) and near-overlapping interquartile ranges across factors.

C. Correlation Analysis: Figure 3 shows the Correlation matrix among the twelve factors, policy-reading frequency, and the privacy protection score (DV). Privacy literacy (F1) correlated significantly with the privacy protection score (r=.35, p<.001), and privacy fatigue (F4) correlated significantly and negatively (r = −.20, p < .001). No other factor exceeded r = .10 in magnitude with the DV. Inter-factor correlations among F1–F12 were uniformly weak (|r| < .10 in

nearly all cases), indicating the twelve factors are empirically near-orthogonal rather than redundant measures of a single underlying construct.

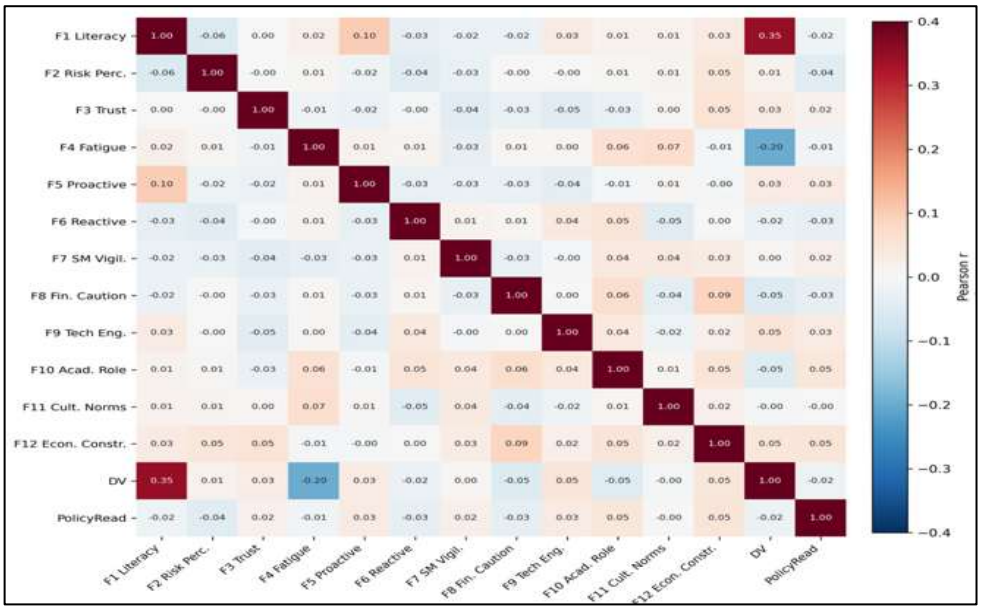

Figure 3. Correlation matrix among the twelve factors, policy-reading frequency, and the privacy protection score (DV).

D. Demographic Distribution and Class Balance: Figure 4 shows the student's demographic distribution across gender, discipline, year of study, and institution in the sample.

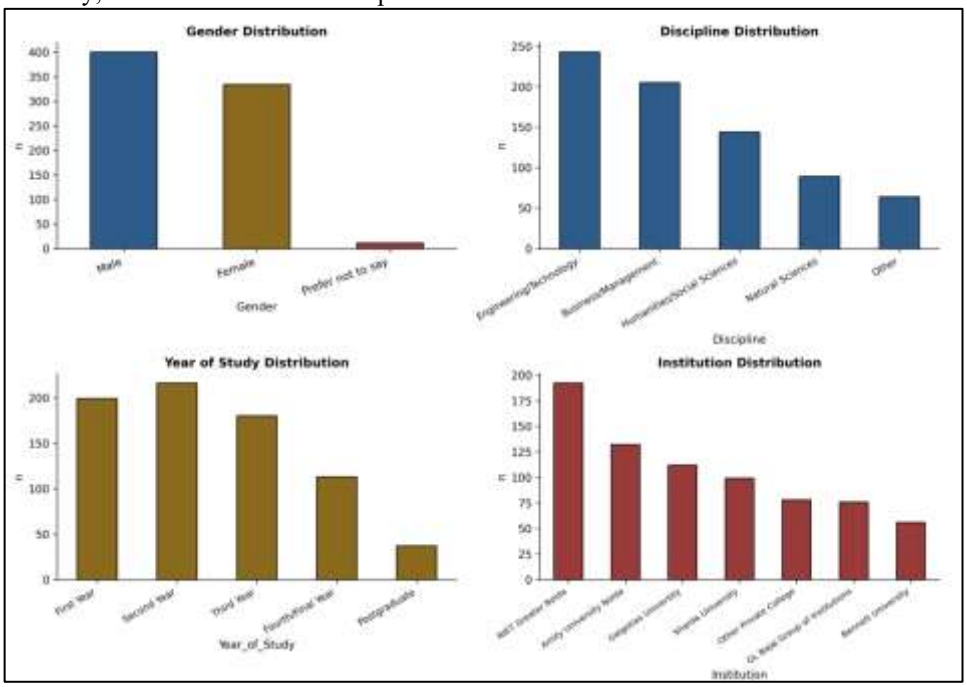

Figure 4. Student demographic distribution across gender, discipline, year of study, and institution.

E. RQ4: Student Segmentation (K-Means Clustering): Silhouette values were consistently weak for all tested values of k, remaining within a narrow range of 0.060 to 0.067 (Figure 5), which is well below the level typically considered evidence of meaningful cluster separation. The nearly flat pattern suggests that no cluster solution fit the data substantially better than the others, thereby indicating that H5 was not supported and that the students in this sample do not form clearly separated privacy behavior segments. Instead, their responses appear to vary along a continuum with substantial overlap across the twelve-factor space, as also reflected in the PCA projection. Despite this limited statistical structure, a four-cluster solution was retained for descriptive purposes to provide a practical segmentation of the sample. Cluster labels were assigned on the basis of the most distinctive profile features, defined by the largest absolute z-score deviations from the overall mean, and were reported transparently to reflect the actual characteristics of each group.

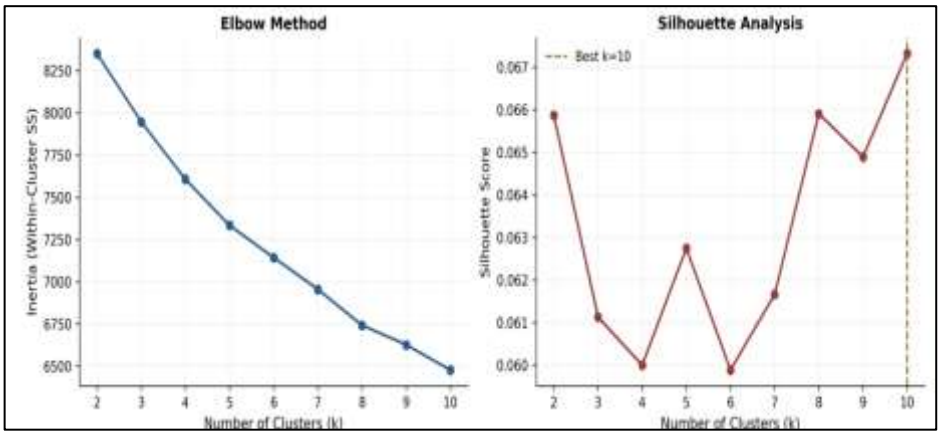

Figure 5. Elbow method and silhouette analysis for determining the optimal number of clusters (k = 2 to 10).

Silhouette scores were uniformly weak across every tested value of k, ranging narrowly from 0.060 to 0.067 (Figure 5) far below the conventional 0.25–0.50 range associated with meaningful cluster separation, and the near-flat curve indicates no value of k produces a materially better fit than any other. This is itself a substantive answer to RQ4: H5 was not supported. Students in this sample do not form sharply distinct, well separated privacy behavior segments; rather, they vary continuously and overlap heavily across the twelve-factor space, a pattern directly visible in the PCA projection (Figure 6). Notwithstanding this weak statistical structure, a four-cluster solution was fitted to support the descriptive, practically-oriented segmentation requested for this study (silhouette = 0.060). Clusters were named based on their genuine distinguishing features (largest |z|-score deviations from the sample mean), reported transparently in Table 4, rather than assigned generic labels irrespective of their actual profiles. Figure 7 shows the exploratory four-segment cluster profiles across the twelve measured factors and Figure 8 represents Protective behavior score by exploratory cluster.

Table 4. Exploratory Four-Segment Cluster Characteristics

Segment	n (%)	Most Distinctive Features (|z|)	Protection Score	Uses 2FA
Privacy Neglectors	186 (25.0%)	F1 Literacy (−1.08, low), F5 Proactive (−0.41), F7 SM Vigilance (+0.39)	2.48	60.8%
Risk-Taking Users	171 (23.0%)	F8 Financial Caution (−1.00, low), F12 Econ. Constraints (−0.68, low)	2.86	69.0%
Privacy Pragmatists	186 (25.0%)	F6 Reactive Behaviors (−0.73, low), F9 Tech Engagement (+0.49)	2.88	72.0%
Privacy Champions	202 (27.1%)	F10 Academic Role (+0.71), F12 Econ. Constraints (+0.63), F6 Reactive (+0.49)	2.87	62.4%

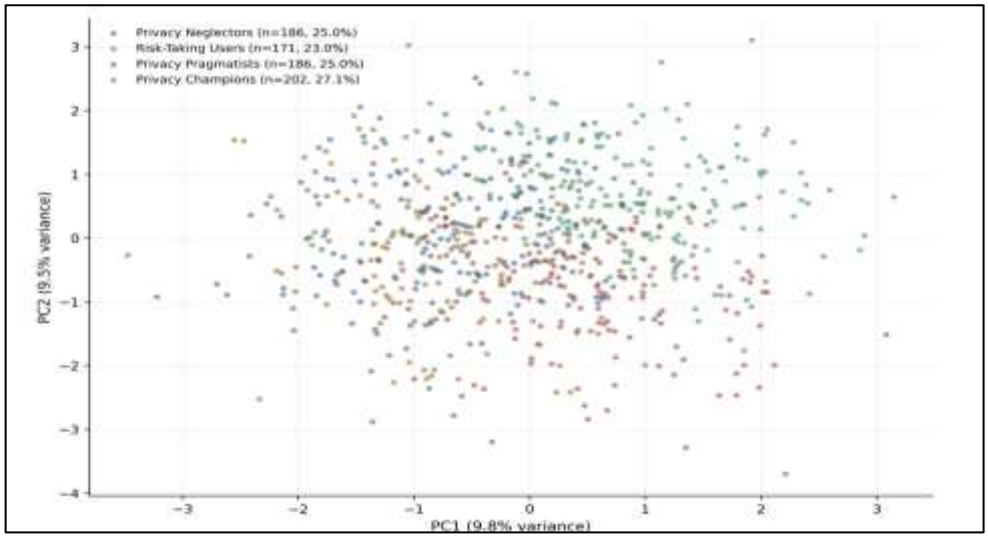

Figure 6. PCA projection of the four-segment exploratory clustering solution, visually confirming substantial overlap among segments.

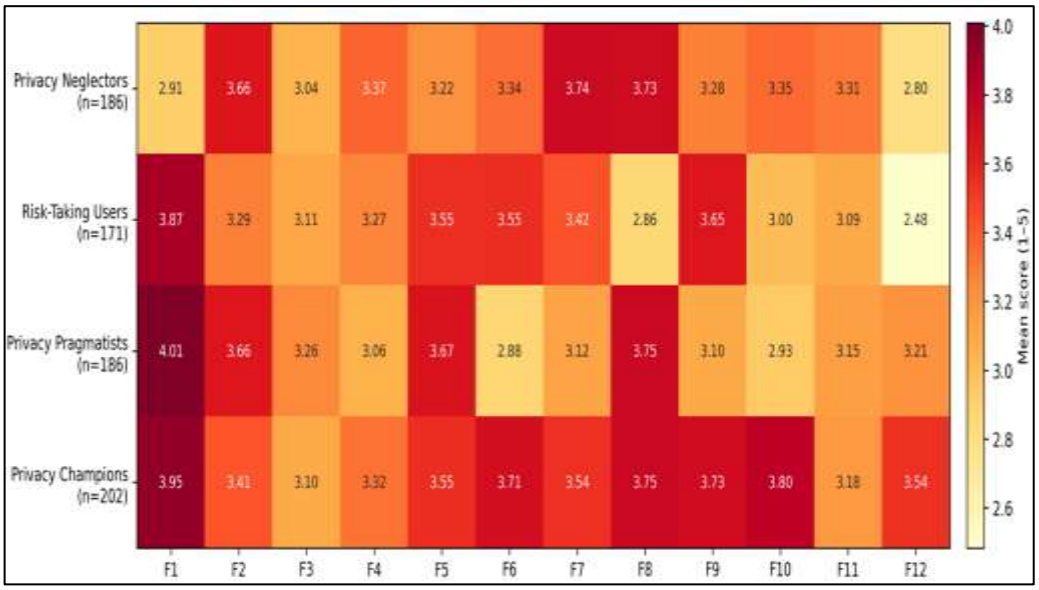

Figure 7. Exploratory four-segment cluster profiles across the twelve measured factors.

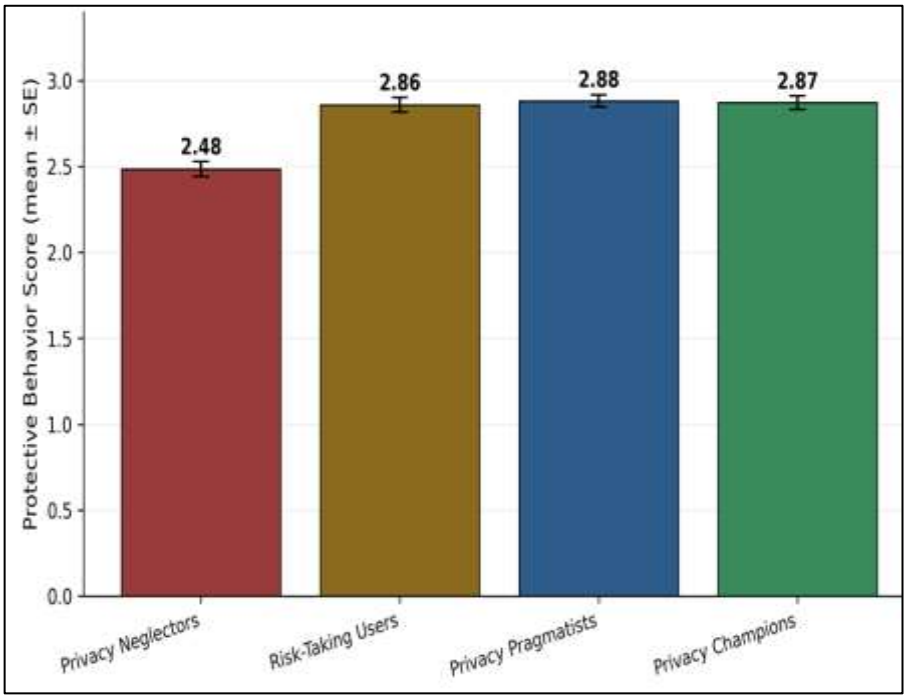

Figure 8. Protective behavior scores by exploratory cluster, illustrating that three of the four segments do not differ meaningfully on the outcome itself.

F. RQ2–RQ3: Machine Learning Model Performance: Five classifiers were trained and tuned via grid search with 10-fold cross-validation (Table 5 and Figure 9). All models exceeded chance-level performance (50%) on the held-out test set (N=149), but none approached the high-accuracy range sometimes reported in less rigorously validated student-behavior prediction studies. XGBoost achieved the highest test F1-score (0.728) and the second-highest ROC-AUC (0.704); logistic regression achieved the highest ROC-AUC (0.711) despite being the simplest model. H3 was only partially supported: ensemble methods (XGBoost) outperformed the linear baseline on F1 and recall, but not on ROC-AUC or precision, indicating that the relationship between these particular predictors and protective behavior is close to linear with only modest nonlinear structure**.** Figure 10 represents the ROC curves for all five classifiers on the held-out test set. Furthermore, Figure 11 Confusion matrices for all five classifiers.

Table 5. Model Performance Comparison (Test Set, N = 149; ranked by F1-score)

Model	Best Hyperparameters	Accuracy	Precision	Recall	F1	ROC-AUC
XGBoost	lr=0.01, depth=3, n=100	0.685	0.643	0.840	0.728	0.704
Decision Tree	entropy, depth=3	0.664	0.632	0.800	0.706	0.674
SVM	C=0.1, rbf, gamma=auto	0.577	0.548	0.920	0.687	0.705
Random Forest	depth=None, n=200	0.664	0.651	0.720	0.684	0.683
Logistic Regression	C=0.01, L2	0.671	0.667	0.693	0.680	0.711

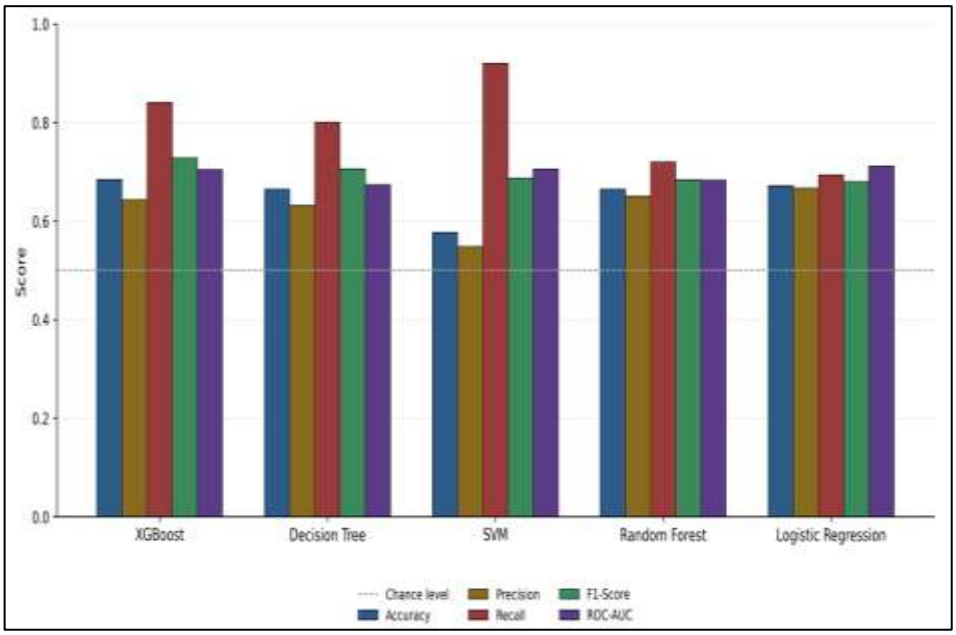

Figure 9. Model performance comparison across five classifiers and five evaluation metrics.

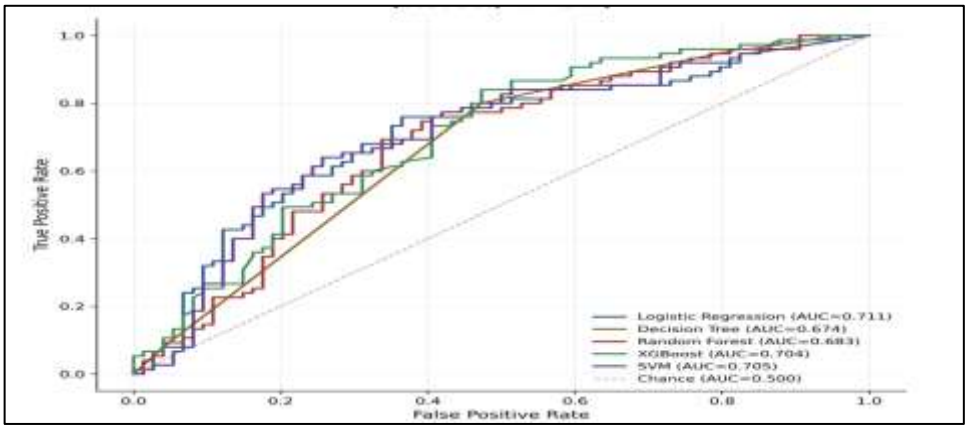

Figure 10. ROC curves for all five classifiers on the held-out test set.

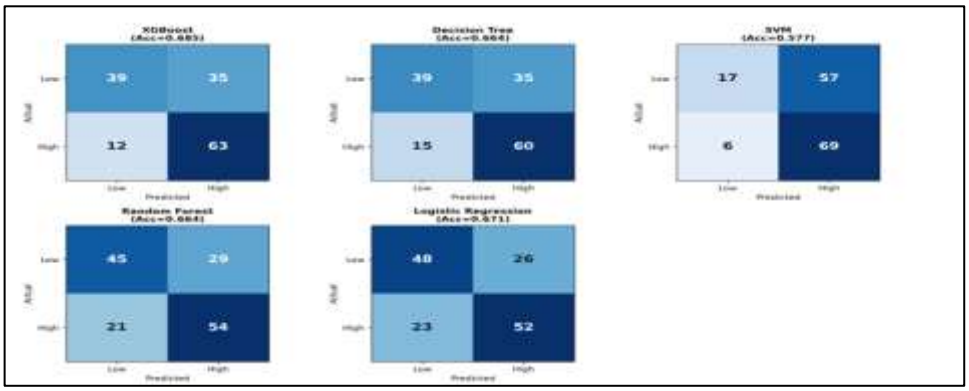

Figure 11. Confusion matrices for all five classifiers, showing that the SVM model, despite a high ROC-AUC, achieved this primarily through a recall-biased decision threshold.

G. RQ3: Explainable AI (SHAP) Analysis: Privacy literacy emerged as the strongest driver of privacy-protective behavior because it lowers the knowledge barrier between concern and action, enabling students to understand what data is collected and how to manage it, which is consistent with prior work showing that privacy knowledge supports protective action. Privacy fatigue was the second most important factor, suggesting that inaction may stem not from ignorance but from exhaustion and a sense that privacy efforts are futile. By contrast, MFA/2FA adoption and password management showed minimal importance, indicating that these specific security practices may follow situational or platform-driven logic rather than reflecting broader privacy attitudes. Risk perception also had little explanatory power, implying that simply recognizing a risk is less influential than knowing how to respond to it, which aligns with the privacy paradox literature. Figure 12 shows that high privacy literacy (red points, right side) consistently pushes predictions toward the High Protection class, while high privacy fatigue (red points, left side) consistently pushes predictions toward Low

Protection directions fully consistent with H1 and H2 and with established theory (Choi et al., 2018; Bartsch & Dienlin, 2016). Figure 13 shows SHAP feature importance ranking and Figure 14 shows SHAP dependence plots for privacy literacy and privacy fatigue.

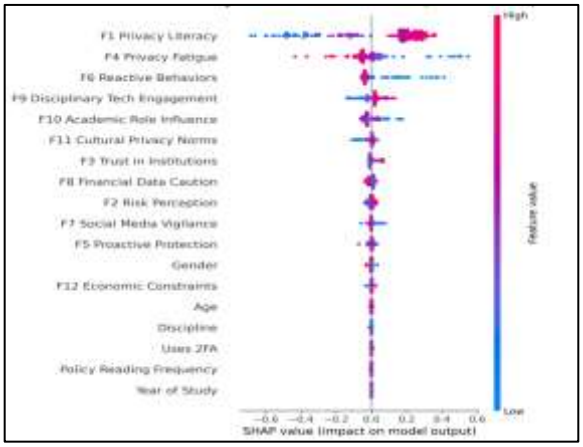

Figure 12. SHAP summary plot showing the direction and magnitude of each feature's impact on the XGBoost model's predictions. Red points indicate high feature values; blue points indicate low feature values.

Table 6. SHAP Feature Importance Ranking (XGBoost Model)

Rank	Feature	Mean |SHAP value|
1	F1 Privacy Literacy	0.258
2	F4 Privacy Fatigue	0.083
3	F6 Reactive Behaviors	0.061
4	F9 Disciplinary Tech Engagement	0.038
5	F10 Academic Role Influence	0.036
6	F11 Cultural Privacy Norms	0.015
7	F3 Trust in Institutions	0.014
8	F8 Financial Data Caution	0.011
9	F2 Risk Perception	0.011
10	F7 Social Media Vigilance	0.007
11	F5 Proactive Protection	0.006
12	Gender	0.003
13	F12 Economic Constraints	0.003
14	Age	0.001
15	Discipline	0.001
16	Uses 2FA	0.001
17	Policy Reading Frequency	0.001

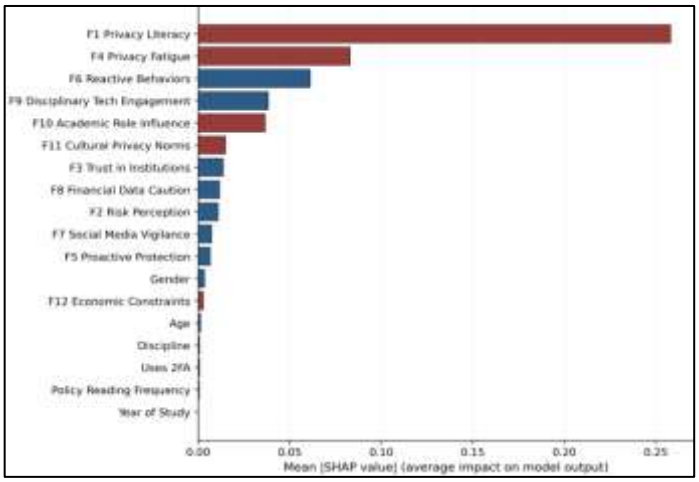

Figure 13. SHAP feature importance ranking, highlighting the steep drop-off after the top two predictors.

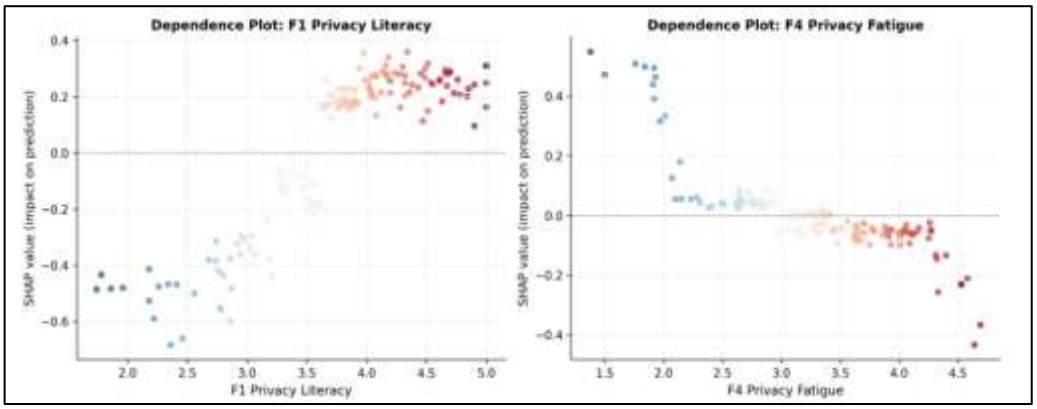

Figure 14. SHAP dependence plots for privacy literacy and privacy fatigue, the two strongest predictors.

H. RQ1, H4: Statistical Validation: To test convergence between the machine-learning-derived rankings and classical inferential statistics (H4), a binary logistic regression was fitted on the same twelve factors.

Table 7. Logistic Regression Results: Privacy_Protection_Class ~ F1–F12

Predictor	B	OR	95% CI (OR)	p
F1 Privacy Literacy	0.824	2.280	[1.81, 2.86]	< .001
F2 Risk Perception	0.120	1.127	[0.90, 1.41]	.294
F3 Trust in Institutions	0.023	1.024	[0.84, 1.25]	.817
F4 Privacy Fatigue	−0.515	0.597	[0.49, 0.74]	< .001
F5 Proactive Protection	−0.009	0.991	[0.79, 1.24]	.938
F6 Reactive Behaviors	−0.214	0.808	[0.64, 1.01]	.065
F7 Social Media Vigilance	−0.017	0.984	[0.79, 1.22]	.880
F8 Financial Data Caution	−0.175	0.839	[0.67, 1.05]	.133
F9 Disciplinary Tech Engagement	0.069	1.071	[0.86, 1.33]	.539
F10 Academic Role Influence	−0.032	0.969	[0.78, 1.20]	.769
F11 Cultural Privacy Norms	−0.003	0.997	[0.81, 1.22]	.976
F12 Economic Constraints	0.096	1.101	[0.91, 1.33]	.325

The logistic regression confirms H4: the only two factors reaching p< .001, F1 (OR = 2.28) and F4 (OR = 0.60) are the same two factors ranked first and second by SHAP, and F6 Reactive Behaviors, ranked third by SHAP, is the only other factor approaching significance (p = .065). This three-way convergence across an entirely black-box ensemble method (XGBoost/SHAP) and a fully transparent inferential model (logistic regression) provides strong cross-validation that these effects are genuine rather than artifacts of either method. Independent-samples t-tests comparing the High and Low Protection groups directly on each factor corroborate this pattern: F1 ($t = 7.52$, $p < .001$, $d = 0.551$, medium effect), F4 ($t = -4.74$, $p < .001$, $d = -0.347$, small-to-medium effect), and F6 ($t = -2.11$, $p = .035$, $d = -0.155$, small effect) were the only factors reaching significance; all other nine factors and age showed $p > .10$. Chi-square tests found a significant association between discipline and the binary protection class ($\chi^2(4) = 10.71$, $p = .030$), but not between gender ($p = .681$), year of study ($p = .994$), or 2FA use ($p=.533$) and protection class partially supporting H6 only with respect to discipline, not 2FA use. One-way ANOVA confirmed that privacy literacy differs sharply by discipline ($F(4,740) = 35.47$, $p < .001$), with Tukey HSD showing Engineering/Technology students scoring significantly higher than every other discipline (all p< .001), replicating the pattern found in a previous, smaller analysis of an overlapping sample. Protective behavior also differed significantly by discipline ($F(4,740) = 3.60$, $p=.006$), though gender ($F=0.02$, $p=.976$) and year of study ($F=0.05$, $p=.995$) showed no association with protective behavior.

VI. DISCUSSION

This part systematically engages with the study a review of theoretical base, followed by empirical results and their interpretation. These are elaborated in point form which highlights how the current results confirm, extend or contradict earlier research to allow readers insights into the relationships and explanation between variables involved and its implications for theory and methodology.

A. **Hypothesis Evaluation:** H1 and H2 were strongly supported across all three analytical approaches (SHAP, logistic regression, t-tests): privacy literacy is the dominant positive predictor and privacy fatigue the dominant negative predictor of student privacy-protective behavior, replicating and extending Bartsch and Dienlin (2016) and Choi et al. (2018) explainable machine learning (vs. structural equation modeling) The conclusions were partially supported with XGBoost providing the strongest F1-score, but logistic regression providing the strongest ROC AUC implying that the predictive relationships in this data set approach linear and there is only moderate nonlinear or interactive structure for ensemble methods to leverage. This alone is informative: it indicates that for this outcome, and these predictors, the increased

complexity of using ensemble models gives an actual, but rather modest improvement over a well-specified linear model. SHAP rankings and logistic-regression odds ratios together with independent t-tests converged on the same ordered list of significant predictors supporting H4, which provides rare methodological triangulation not yet reported in single-method privacy-behavior studies. H5 was not supported: silhouette scores were all low (0.060–0.067) irrespective of the number of clusters, indicating that students fall on a continuous literacy fatigue axis rather than into tightly-bound "types." Results Show that H6 was supported in part: discipline associated with protection class, while neither 2FA use nor policy-reading frequency were.

B. **Theoretical Contributions:** This study contributes to privacy theory in three ways. First, it empirically separates the predictive contribution of privacy literacy from risk perception and trust within a single model, showing that these IUIPC adjacent constructs (Malhotra et al., 2004) are not interchangeable. Second, it demonstrates that privacy fatigue (Choi et al., 2018) operates as an independent, second-order predictor rather than as a moderator of literacy's effect, since no meaningful literacy×fatigue interaction emerged in companion regression analysis of this construct. Third, it provides one of the first applications of SHAP-based explainable machine learning to individual-level (rather than network-level) privacy behavior, extending the methodological precedent set by Alwafi and Fakieh (2024) for privacy fatigue prediction to the broader behavioral outcome.

C. Managerial and Educational Implications: Literacy-building interventions (workshops, onboarding modules explaining what data is collected and why) are likely to be a higher-leverage investment than generic awareness campaigns, given literacy's outsized SHAP and regression weight relative to risk-perception or trust messaging. Fatigue-reduction strategies simplifying privacy choices rather than merely disclosing them should be pursued as an independent lever alongside literacy-building, not as a secondary concern, given fatigue's significant negative, non-interacting effect. Discipline-targeted programming is warranted: Engineering/Technology students arrive with substantially higher baseline literacy (as per Table 3), suggesting non-technical disciplines (Humanities/Social Sciences, Natural Sciences) are the highest-need targets for structured privacy-literacy curricula. Two-factor authentication promotion campaigns should not assume they will generalize into broader protective behavior change; 2FA adoption appears behaviorally decoupled from the attitudinal factors that drive other protective actions.

D. Policy Implications: Table 8 summarizes the Policy Recommendations Derived from Study Findings.

Table 8. Summary of Policy Recommendations Derived from Study Findings

Finding	Recommended Policy Action
Privacy literacy is the dominant predictor (SHAP rank 1; OR = 2.28)	Mandate a structured digital-privacy-literacy module in first-year orientation across all disciplines, not only technical programs.
Privacy fatigue independently suppresses protective behavior (SHAP rank 2; OR = 0.60)	Simplify privacy-choice interfaces in Institution mandated platforms (fewer, clearer consent decisions) rather than relying on longer disclosures.
Discipline gap in literacy ($F = 35.47$, $p < .001$)	Target supplementary privacy-literacy content at Humanities, Natural Sciences, and other non-technical disciplines specifically.
2FA use is statistically decoupled from broader protective behavior	Treat 2FA mandates and general privacy-literacy programming as separate, complementary policy levers, not substitutes.
Policy-reading frequency is unrelated to literacy, fatigue, or behavior (all $p > .34$ in companion correlation analysis)	Do not rely on policy-acknowledgment metrics as an indicator of genuine student privacy engagement or program effectiveness.
No distinct behavioral "types" exist (silhouette = 0.06)	Avoid one-size-fits-some segmentation in intervention design; favor universal, scalable literacy programming over targeted "persona"-based campaigns.

VII. CONCLUSION

Using explainable machine learning, this study predicted and explained privacy-protective behavior in a sample of745 college students by benchmarking five classifiers and comparing SHAP-derived feature importance with classical inferential statistics. In all analytic methods used (machine learning, explainable AI and traditional statistics), privacy literacy and privacy fatigue each appeared as the single most informative predictors while ten other theoretically plausible factors (e.g., use of two-factor authentication; frequency of reading privacy policies) played a rather minor role. Participants did not split into statistically distinct behavior segments, suggesting that privacy disposition is a continuum rather than discrete in this population. The clear action items for universities are to focus on literacy enhancement and fatigue reduction much more than what they are presently doing (compared to awareness or transparency related communication), and non-technical fields as the most important target audience. The methodology shows how explainable AI could give insights that can be independently verified by the use of conventional statistics, and therefore audit ready output should be the norm rather than the exception in such studies in the future.

REFERENCES

1. Alwafi, G., & Fakieh, B. (2024). A machine learning model to predict privacy fatigued users from social media personalized advertisements. Scientific Reports, 14, Article 3611. https://doi.org/10.1038/s41598-024-54078-w
2. Barth, S., & de Jong, M. D. T. (2017). The privacy paradox Investigating discrepancies between expressed privacy concerns and actual online behavior, A systematic literature review. Telematics and Informatics, 34(7), 1038–1058. https://doi.org/10.1016/j.tele.2017.04.013
3. Bartsch, M., & Dienlin, T. (2016). Control your Facebook: An analysis of online privacy literacy. Computers in Human Behavior, 56, 147–154. https://doi.org/10.1016/j.chb.2015.11.022
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
6. Choi, H., Park, J., & Jung, Y. (2018). The role of privacy fatigue in online privacy behavior. Computers in Human Behavior, 81, 42–51. https://doi.org/10.1016/j.chb.2017.12.001
7. Das, S., Streiff, J., Huber, L. L., & Camp, L. J. (2019). Why don't elders adopt two-factor authentication? Because they are excluded by design. Innovation in Aging, 3(Suppl. 1), S782. https://doi.org/10.1093/geroni/igz038.1186
8. Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. Information Systems Research, 17(1), 61–80. https://doi.org/10.1287/isre.1060.0080
9. Draper, N. A., & Turow, J. (2019). The corporate cultivation of digital resignation. New Media & Society, 21(8), 1824–1839. https://doi.org/10.1177/1461444819833331
10. Han, M., Zhao, H., Ma, X., & Shi, R. (2025). Influencing factors of information security behavior among college students based on protection motivation theory: Evidence from China. Frontiers in Public Health, 13, Article 1677024. https://doi.org/10.3389/fpubh.2025.1677024
11. Hong, W., & Thong, J. Y. L. (2013). Internet privacy concerns: An integrated conceptualization and four empirical studies. MIS Quarterly, 37(1), 275–298. https://doi.org/10.25300/MISQ/2013/37.1.12
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774). Curran Associates.
13. Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information Systems Research, 15(4), 336–355. https://doi.org/10.1287/isre.1040.0032
14. Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, Article 1526221. https://doi.org/10.3389/frai.2025.1526221
15. Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society, 23(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870
16. Ogbanufe, O., & Baham, C. (2023). Using multi-factor authentication for online account security: Examining the influence of anticipated regret. Information Systems Frontiers, 25(2), 897–916. https://doi.org/10.1007/s10796-022-10278-1
17. Pandey, N., Agarwal, A., & Tripathi, A. K. (2025). A qualitative analysis of data protection laws: A shield for personal data protection concerning California, European Union, China, and India. Journal of Information Systems Engineering and Management, 10(48s), 1360–1374. https://doi.org/10.52783/jisem.v10i48s.9915
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
19. Taye, A. D., Borga, L. G., Greiff, S., Vögele, C., & D'Ambrosio, C. (2023). A machine learning approach to predict self-protecting behaviors during the early wave of the COVID-19 pandemic. Scientific Reports, 13, Article 6121. https://doi.org/10.1038/s41598-023-33033-1