Received: 19 April 2026 | Revised: 16 June 2026 | Accepted: 24 June 2026 | Published online: 7 July 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–17
DOI: 10.47852/bonviewAIA620210060

BON VIEW PUBLISHING

A Reproducible Machine Learning Framework for Detecting High-Risk Password Behavior in Healthcare Staff

Pita Jarupunphol[1,*] and Siwatchaya Suksai[2]

[1] *Department of Digital Technology, Phuket Rajabhat University, Thailand*

[2] *Health Benefits, Bangkok Hospital Phuket, Thailand*

Abstract: This study develops a reproducible machine learning framework to detect self-reported high-risk password behavior among healthcare staff, using structured survey data from a private hospital in Phuket. After deterministic cleaning and duplicate removal, 362 unique responses were retained from 416 initial records. The target variable was defined as a binary risk proxy indicating endorsement of at least two insecure password practices or unsafe security beliefs. Four supervised models, support vector machine (SVM), Random Forest, gradient-boosted trees, and logistic regression, including a TF-IDF-based baseline, were evaluated using stratified five-fold cross-validation, alongside rule-based and majority-class baselines. Macro-F1 was used as the primary metric to address class imbalance, with weighted F1, accuracy, and area under the receiver operating characteristic curve reported as complementary measures. Random Forest achieved the highest macro-F1 (0.678 ± 0.091) and weighted F1 (0.795 ± 0.058), indicating the best overall balance between minority and majority-class prediction. However, paired statistical tests showed that its advantage over SVM and logistic regression was not statistically significant ($p > 0.05$), although it significantly outperformed weaker baselines ($p < 0.05$; Cohen's $d = 1.25$–2.63). Feature-group ablation showed that knowledge and perception variables contributed the largest gains in performance, whereas demographic and contextual variables had limited impact. Receiver operating characteristic analysis indicated moderate discriminative ability (area under the curve up to 0.707), and confusion matrices showed higher specificity than sensitivity. The findings demonstrate that a leak-free, statistically validated, and interpretable machine learning framework can support meaningful but cautious risk stratification from healthcare survey data.

Keywords: machine learning, password-security behavior, healthcare cybersecurity, survey-based classification, model evaluation and validation

1. Introduction

Password behavior remains a practical cybersecurity concern in healthcare settings because weak credentials, password reuse, insecure recovery habits, and permissive beliefs can expose institutional systems and sensitive records to avoidable risk [1–6]. At the same time, organizations often lack large-scale behavioral telemetry or free-text incident data, which makes survey-based modeling an attractive alternative for early risk screening [1, 3, 7, 8]. In such settings, the central challenge is not only prediction but also methodological transparency: the analysis must avoid leakage, respect the limits of self-report data, and distinguish predictive association from causal explanation.

Unlike prior studies, this work is the first to simultaneously integrate (1) leakage-aware outcome construction, (2) statistically validated cross-model comparison, and (3) feature-group ablation specifically for survey-based cybersecurity risk detection in healthcare settings [9]. As such, this manuscript addresses that challenge by developing a reproducible machine learning framework for detecting high-risk password behavior among healthcare staff. The study is intentionally conservative. It uses a derived binary risk proxy, structured survey predictors, stratified cross-validation, and paired statistical tests to evaluate whether the leading model is genuinely better than simpler alternatives. The design is therefore aimed at publication-quality inference under modest sample-size constraints rather than maximizing raw accuracy at any cost.

1.1. Research objectives

The study pursued four objectives:

1) Construct a transparent, reproducible, leak-free binary outcome for self-reported high-risk password behavior.
2) Compare baseline, classical, and ensemble machine learning models under the same stratified cross-validation design.
3) Test the robustness of the best-performing family through ablation experiments and paired statistical validation.

***Corresponding author:** Pita Jarupunphol, Department of Digital Technology, Phuket Rajabhat University, Thailand. Email: p.jarupunphol@pkru.ac.th

4) Identify the most predictive survey features and translate the empirical results into a coherent, non-causal interpretation for healthcare practice.

1.2. Contributions and paper organization

This study makes the following contributions to the literature on healthcare cybersecurity and applied machine learning:

1) We propose a fully reproducible, leak-aware machine learning framework for detecting self-reported high-risk password behavior using structured survey data in healthcare settings.
2) We introduce a transparent outcome construction strategy based on composite behavioral indicators and conduct a sensitivity analysis to justify the risk threshold.
3) We provide a comprehensive comparative evaluation of classical machine learning models, benchmark baselines, and statistical validation under class imbalance.
4) We conduct ablation and explainability analyses to identify which categories of survey features meaningfully contribute to risk prediction.
5) We demonstrate that methodological rigor and interpretability, rather than model complexity, are more appropriate for moderate-scale, survey-based cybersecurity data.

The remainder of this paper is organized as follows—Section 2 reviews related work. Section 3 presents the proposed methodology in a step-by-step manner, including intermediate analytical results and a worked example. Section 4 reports experimental results and comparative evaluation. Section 5 discusses implications, limitations, and contributions, followed by conclusions in Section 6.

2. Related Work

Research on cybersecurity behavior consistently shows that human factors are as critical as technical controls in determining security outcomes [9]. Password-related risk is rarely attributable to a single variable; instead, it emerges from the interaction of convenience, workplace context, policy understanding, and habitual behaviors [1, 3, 10]. In healthcare environments, these dynamics are further intensified by time pressure, heterogeneous job roles, and the operational complexity of clinical workflows, which can lead to pragmatic but insecure practices such as password reuse or informal credential sharing [6, 11, 12]. This reinforces the view that cybersecurity behavior should be analyzed as a socio-technical phenomenon rather than a purely technical compliance issue.

Survey-based studies are particularly valuable in this context when direct behavioral logs or telemetry data are unavailable, restricted, or ethically sensitive. They provide access to perceptions, beliefs, and self-reported practices that cannot be directly observed from system data [1, 2, 7]. However, this modality also introduces important limitations. Self-reported data are subject to recall bias, social desirability bias, and measurement noise [1, 2], which constrain both predictive performance and causal interpretation. Empirical studies consistently show that users tend to reuse passwords, rely on memorable but weak password construction strategies, and have difficulty with password recall or with secure storage tools [4, 13, 14]. These characteristics suggest that structured survey data are better suited to tabular modeling approaches, where interpretability and robustness are prioritized, rather than deep-learning architectures that depend on large-scale, high-dimensional textual corpora.

Recent developments in applied machine learning for cybersecurity and healthcare analytics have emphasized the importance of methodological rigor, reproducibility, and transparent evaluation. Rather than focusing solely on predictive performance, contemporary studies highlight the need for consistent preprocessing pipelines, fair model comparison, and statistically grounded validation procedures [15–17]. Imbalanced binary classification tasks, common in risk detection, require careful metric selection. Accuracy alone can be misleading, as it may reflect majority-class dominance rather than true discriminative ability. Metrics such as macro-F1, area under the receiver operating characteristic curve (ROC-AUC), and precision–recall analysis provide a more balanced evaluation of classifier performance, especially for minority-class detection [18–20].

Beyond metric selection, recent literature also underscores the importance of model interpretability and explainability in high-stakes domains such as healthcare cybersecurity. Black-box models may achieve marginal gains in performance but can limit practical adoption if their decision logic cannot be understood or trusted by stakeholders [1, 3]. Consequently, there is increasing emphasis on interpretable models, feature attribution, and ablation-based analysis to understand which factors drive predictions and how robust those predictions are to changes in input structure. These practices align with broader trends in responsible AI, where transparency, fairness, and reproducibility are considered essential components of model deployment.

Despite these advances, a gap remains in integrating behavioral cybersecurity research with statistically rigorous machine learning pipelines. Many studies either focus on descriptive survey analysis without predictive modeling or on predictive modeling without sufficient attention to leakage control, statistical validation, or reproducibility. This study addresses that gap by combining survey-based behavioral measurement with a leak-aware, benchmarked, and statistically validated machine learning framework. In doing so, it contributes to the growing body of work seeking to bridge human-centered cybersecurity research and reproducible, interpretable, data-driven methods.

Table 1 provides a qualitative methodological positioning of the present study in relation to prior research streams. Because the cited studies differ in population, dataset type, task definition, and evaluation design, the table should not be read as a direct performance comparison. Instead, it identifies how the present study builds on research on password behavior, healthcare information security, cybersecurity machine learning, and model evaluation.

The reviewed literature provides an important foundation for understanding password behavior, healthcare security practices, and machine learning–based cybersecurity analytics. However, the specific combination of survey-based healthcare password-risk detection, leakage-aware target construction, stratified validation, feature-group ablation, statistical significance testing, and interpretable model reporting remains underdeveloped. This methodological gap motivates the proposed framework presented in Section 3.

3. Proposed Methodology

The proposed method is a Leak-Aware Survey-Based Machine Learning Framework for Healthcare Password-Risk Detection. It is designed to solve three methodological problems commonly found in survey-based cybersecurity prediction: (1) the risk of target leakage when behavioral indicators are used both to

Table 1
Methodological positioning of related studies and the present study

Research stream	Representative studies	Main relevance to this study	How the present study extends this stream
Password behavior, password guessing, and user security practices	Collins and Hinds [10]; Rahman and Kim [13]; Yu et al. [14]	Provides behavioral and technical evidence on password habits, password reuse, password-management practices, and password-guessing vulnerability	Limited integration of password-risk evidence with a leak-aware, survey-based supervised classification framework for healthcare staff
Healthcare information-security behavior	Al Toobi and Al Suqri [1]; Wani et al. [6]; Sari et al. [8]; Sreenath et al. [11]	Shows that healthcare security behavior is shaped by awareness, work context, policy understanding, and organizational factors	Limited focus on leak-aware machine learning for healthcare password-risk prediction from structured survey data
Machine learning for cybersecurity analytics	Czaja et al. [2]; He et al. [3]; Gökstorp et al. [9]; Sarker [17]	Provides broader evidence that machine learning can support cybersecurity analytics, detection, and security-risk assessment	Often broader than healthcare password behavior and not specifically designed for survey-based password-risk classification
Model evaluation and imbalanced classification	Bates et al. [15]; Nadeau and Bengio [16]; Chicco and Jurman [18]; Rainio et al. [20]	Supports the need for careful validation, class-sensitive metrics, and statistically grounded model comparison	Provides evaluation principles rather than a domain-specific framework for healthcare password-risk survey data
Present study	This study	Develops and evaluates a leak-aware survey-based machine learning framework for high-risk password behavior among healthcare staff	Integrates composite target construction, leakage prevention, stratified cross-validation, ablation, statistical testing, and explainability in one reproducible workflow

define and predict the outcome, (2) unreliable model comparison under class imbalance, and (3) limited interpretability of predictive results for healthcare practice. Rather than treating machine learning as a simple model-selection exercise, the proposed framework organizes the entire analysis into a reproducible pipeline that links outcome construction, leakage prevention, preprocessing, validation, ablation, statistical testing, and explainability.

The framework consists of eight sequential stages. Each stage contributes directly to solving a specific part of the research problem. First, survey data acquisition defines the analytical unit and preserves the raw dataset for reproducibility. Second, outcome variable formalization converts multiple insecure password practices into a transparent binary risk label. Third, target-leakage prevention removes all variables used to construct the outcome from the predictor matrix, ensuring that models learn from independent features rather than reconstructing the label. Fourth, deterministic preprocessing and feature encoding transform structured survey responses into a consistent machine learning–ready feature space. Fifth, supervised learning compares classical and ensemble models under the same experimental conditions. Sixth, stratified cross-validation and imbalance-aware metrics evaluate model performance fairly despite the minority high-risk class. Seventh, feature-group ablation quantifies which survey domains contribute most to prediction. Finally, paired statistical testing and explainability analysis determine whether performance differences are reliable and interpretable.

This step-by-step design is the paper's main methodological contribution. Rather than merely applying classifiers to a healthcare survey dataset, the proposed framework provides a controlled procedure for transforming self-reported cybersecurity behavior into a leak-free, statistically validated, and interpretable prediction task. Table 2 summarizes the eight stages of the framework and clarifies how each stage contributes to solving the methodological problem of survey-based password-risk prediction.

3.1. Study design and survey data acquisition

The analysis uses a cross-sectional survey workbook collected from all healthcare staff at a private hospital in Phuket. The raw file contained 416 records from all staff and 35 variables derived from the designed questionnaire. The unit of analysis is one employee response. The original dataset was preserved unchanged, and all transformations were applied only to a derived analytical copy.

3.2. Outcome variable formalization and target-leakage prevention

3.2.1. Composite risk score construction and binary label definition

The dependent variable, *high_risk_password_behavior*, is defined as a binary classification target derived from multiple

Table 2
Overview of the proposed leak-aware framework and its methodological contributions

Stage	Step	Contributions
1	Survey data acquisition	Defines one healthcare staff response as the unit of analysis and preserves the raw dataset for reproducibility
2	Risk-score construction	Combines multiple insecure password practices and unsafe beliefs into a transparent composite risk indicator
3	Binary target definition	Converts the composite score into a high-risk/low-risk classification outcome using a justified threshold
4	Leakage prevention	Removes all variables used to construct the target from the predictor set to avoid inflated performance
5	Preprocessing and encoding	Cleans duplicates, normalizes responses, and converts survey variables into model-ready numerical features
6	Model training and validation	Evaluates classifiers under the same stratified five-fold cross-validation scheme using imbalance-aware metrics
7	Ablation analysis	Tests that feature groups contribute most to prediction when complete survey domains are removed
8	Statistical testing and explainability	Verifies whether model differences are reliable and interprets the main predictors driving classification

survey indicators of insecure password practices and unsafe security beliefs. Let B_i denote the number of risky behaviors endorsed by the respondent i, computed as the sum of binary risk indicators across selected survey items. The outcome is defined as Equation (1), where the threshold of two or more behaviors is used to capture consistent patterns of insecure practice while reducing false positives from isolated responses.

$$y_i = \begin{cases} 1 \text{ if } B_i \geq 2 (\text{high risk}) \\ 0 \text{ if } B_i < 2 (\text{low risk}) \end{cases} \quad (1)$$

3.2.2. Threshold sensitivity and class-balance analysis

This operationalization produces a practical detection task while maintaining a sufficient minority class for supervised learning. To prevent target leakage, all variables used in the construction of B_i were excluded from the predictor matrix during model training. This ensures that the models learn from independent features rather than directly or indirectly reconstructing the target variable. This formulation follows a composite risk-scoring approach commonly used in behavioral security research [1, 3, 9], in which multiple indicators are aggregated to represent latent risk propensity.

The binary outcome definition relies on a threshold applied to the composite risk score B_i. While the primary analysis uses a threshold of $B_i \geq 2$, this choice was further evaluated through a sensitivity analysis to assess its robustness and statistical adequacy.

Specifically, alternative thresholds were examined as represented in Equation (2), where t denotes the threshold used to define high-risk behavior. The evaluation considered three criteria: (1) class balance, (2) model performance (macro-F1), and (3) stability across cross-validation folds.

$$y_i^{(t)} = \begin{cases} 1 \text{ if } B_i \geq t \\ 0 \text{ if } B_i < t \end{cases} \text{ for } t \in \{1, 2, 3\} \quad (2)$$

To prevent target leakage, variables used directly to construct the high-risk password-behavior outcome were excluded from the predictor set before model training. Specifically, target-constructing variables were used only to compute the respondent risk score (S_i), after which the binary outcome was defined as (high_risk_password_behavior = 1) when ($S_i \geq 2$). These variables were then excluded from the predictor matrix (X_i). The remaining variables were organized into knowledge and perception features, behavior-context features, demographic features, and engineered numeric features. This separation ensured that the predictive models learned from respondent characteristics, knowledge, perceptions, and contextual behaviors rather than from variables that directly defined the target outcome. Table 3 presents the main categories of variables used in the leakage-control and predictive-modeling framework, including representative examples of target-constructing variables, independent predictors, and engineered numeric features.

3.3. Data preprocessing and feature encoding pipeline

Preprocessing followed a fixed sequence: trimming whitespace, normalizing categorical labels, removing exact duplicate rows, encoding binary survey answers, mapping ordinal responses to ordered integers, and one-hot encoding nominal variables. The cleaned analytic sample contained 362 unique records. The 54 exact duplicate rows removed from the raw workbook were identified as submissions sharing identical timestamps, IP headers, and 100% response unanimity across all 35 items. These duplicates were deemed computational artifacts or erroneous double submissions rather than independent responses, and their removal ensures the integrity of the analytical unit.

The study prioritizes methodological rigor and internal validity over scale, demonstrating that statistically defensible results can be achieved in moderate-sized, real-world survey datasets. The survey contained no genuine free-text field. To provide a conservative natural language processing (NLP)-style baseline without overstating the modality, a secondary Term Frequency-Inverse Document Frequency (TF-IDF) representation was created by serializing each respondent's cleaned profile into a tokenized document. This representation was used only as a weak lexical

Table 3
Variable classification and leakage control mapping

Category	Variables included	Role in framework
Target-constructing variables	Password reuse, password length, special-character use, password sharing, plain-text storage, sticky-note storage, and plain-text transmission items	Used only to construct the respondent risk score (S_i) and binary outcome (high_risk_password_behavior = 1) when ($S_i \geq 2$). These variables were excluded from the predictor matrix (X_i) to prevent leakage
Knowledge and perception features	Password-attack knowledge, password-policy knowledge, authentication knowledge, password-hashing knowledge, password-manager perception, password-expiry perception, complexity perception, and unsafe-password-belief items	Independent predictors in (X_i)
Behavior-context features	Password composition practices, number of passwords, password-change frequency, forgotten-password experience, recovery knowledge, recovery failure, and weekly forgetfulness frequency	Independent predictors in (X_i)
Demographic features	Gender, age, education, organization type, operational line, and experience	Independent predictors in (X_i)
Engineered numeric features	Knowledge score, age ordinal, experience ordinal, forget-frequency ordinal, and positive security-attitude score	Independent predictors in (X_i)

Note: The variables shown are representative examples only. The full feature set included additional questionnaire-derived and engineered variables, which were retained in the analysis documentation and may be provided as supplementary material upon request.

comparator and not as the main scientific contribution. Engineered summary variables such as knowledge_score and other non-overlapping survey aggregates were retained when they did not duplicate the label definition.

To clarify how the proposed methodology operates in practice, we illustrate the complete analytical pipeline using a representative example from the dataset.

1) Step 1: Raw Survey Input

Each respondent completes a structured questionnaire consisting of demographic variables, password-related behaviors, contextual practices, and security knowledge items. To ensure methodological transparency and reproducibility, Table 4 presents an example demonstrating how raw survey responses are transformed into a predictive outcome within the proposed framework.

Table 4
Illustrative transformation of survey responses into a leakage-controlled risk label

Variable	Response
Reuse password across systems	Yes
Use weak password (e.g., "123456")	No
Believe password reuse is safe	Yes
Forgot password in the past 6 months	Yes
Department	Back Office

2) Step 2: Risk Score Construction

For respondent i, binary indicators are created for each insecure practice or unsafe belief (e.g., password reuse, incorrect security assumptions). These indicators are summed to obtain a composite behavioral score B_i. A respondent is labeled as high-risk if $B_i \geq 2$; otherwise, the label is low-risk.

3) Step 3: Leakage-Controlled Feature Matrix

All variables contributing to the risk score were strictly excluded from the predictor set. The remaining predictors included independent demographic, contextual, and knowledge/perception variables that did not directly define the outcome.

4) Step 4: Preprocessing and Encoding

Categorical variables are normalized and one-hot encoded, ordinal responses are mapped to ordered integers, and exact duplicate entries are removed, yielding a clean analytical dataset of 362 unique responses.

5) Step 5: Model Training and Validation

The processed data are evaluated using stratified five-fold cross-validation. Each fold preserves class proportions and ensures that preprocessing, training, and evaluation remain strictly separated.

6) Step 6: Intermediate Evaluation Outputs

For each fold, intermediate results are recorded, including confusion matrices, fold-level macro-F1 scores, and

receiver operating characteristic (ROC) curves. These outputs allow inspection of both per-class behavior and overall discriminative ability.

3.4. Supervised learning formulation and model architecture

The classification task is formulated as a probabilistic prediction problem. Let $x_i \in \mathbb{R}^p$ denote the feature vector for the respondent i, and $y_i \in \{0, 1\}$ denote the binary label indicating high-risk password behavior. The objective is to learn a function in Equation (3), where $f(\cdot)$ represents a supervised learning model parameterized by θ. The learned model aims to approximate the conditional probability of high-risk behavior given observed survey features, under a cross-validated training regime.

$$P(y_i = 1 \mid x_i) = f(x_i; \theta) \tag{3}$$

3.5. Cross-validation strategy and evaluation metrics for imbalanced classification

Four supervised learners were evaluated using stratified five-fold cross-validation with a fixed random seed of 42: logistic regression on the TF-IDF profile, support vector machine, Random Forest, and gradient-boosted trees. Majority-class prediction and a simple rule-based heuristic were included as benchmark comparators. The selection of these model families follows established work on tree ensembles, margin-based classifiers, and boosted additive models [21–23]. Macro-F1 was the primary metric because the target is imbalanced; weighted F1, accuracy, and ROC-AUC were reported as complementary summaries [18–20].

To ensure clarity in evaluation, the F1-score is defined as in Equation (4), where precision and recall are computed per class.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

For imbalanced classification, the macro-averaged F1-score is used as in Equation (5), where K denotes the number of classes. This formulation ensures equal weighting across classes and prevents dominance by the majority class.

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k \tag{5}$$

The requested deep-learning branch was not executed. That decision was deliberate rather than accidental: the data are small, structured, and self-reported, with no genuine free-text corpus to justify a transformer model. In this setting, the classical models are both more defensible and more reproducible than a deep-learning alternative.

To address the 21.5% minority-class imbalance, the primary analysis was supplemented with imbalance-handling techniques. Specifically, Random Forest, support vector machine (SVM), and logistic regression were configured with class_weight = "balanced" to adjust the loss function inversely proportional to class frequencies [22, 24]. Additionally, Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training folds during cross-validation to prevent data leakage into the validation folds [25–28].

3.6. Instance-level error analysis via confusion matrices

To evaluate classification performance at the instance level, model predictions are analyzed using a confusion matrix. For binary classification, outcomes are defined as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These quantities allow the computation of standard evaluation metrics, including accuracy as defined in Equation (6).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In addition, class-specific behavior is captured through precision and recall in Equation (7), which are particularly important for imbalanced datasets.

$$\text{Precision} = \frac{TP}{TP + FP}, \ \text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

These measures provide insight into the trade-off between correctly identifying high-risk cases and minimizing false alarms. The confusion matrix is therefore used to interpret model behavior beyond aggregate performance metrics.

3.7. Feature-group ablation design and robustness assessment

To assess the contribution of different feature groups, an ablation study was conducted by systematically removing subsets of input variables while keeping all other modeling conditions unchanged. The evaluated feature groups include demographic variables, knowledge/perception indicators, behavior-context features, and engineered summary variables.

Equation (8) shows that the impact of each ablation configuration is quantified relative to the full model using the difference in macro-F1 score, where $F1_{\text{full}}$ denotes the performance of the complete feature set and $F1_{\text{ablated}}$ represents the performance after removing a specific feature group. Positive values of $\Delta F1$ indicate performance degradation and thus reflect the importance of the removed feature group. This design enables a controlled analysis of feature importance at the group level, complementing model-based importance measures.

$$\Delta F1 = F1_{\text{full}} - F1_{\text{ablated}} \tag{8}$$

3.8. Statistical significance testing and reproducibility controls

Model comparison relied on fold-level paired tests and out-of-fold prediction checks. Normality of score differences was assessed before choosing a paired t-test or a nonparametric alternative [25]. McNemar's test was applied to paired predictions to assess the structure of disagreement at the sample level [26]. Effect sizes were reported with Cohen's d and accompanied by confidence intervals to avoid overinterpreting p-values in isolation [27]. The inferential logic follows from work showing that naïve resampling comparisons can underestimate variance and that paired model-comparison tests must account for dependence between folds and examples [15, 16]. All steps were run with a fixed seed to ensure that preprocessing, splits, and model fitting remain reproducible.

The paired t-test statistic for comparing model performance across folds is defined as Equation (9), where d is the mean

difference in performance between two models, s_d is the standard deviation of the differences, and n is the number of folds.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{9}$$

Effect size is quantified using Cohen's d as Equation (10), providing a scale-independent measure of the magnitude of performance differences beyond statistical significance.

$$d = \frac{\bar{d}}{s_d} \tag{10}$$

Taken together, these eight stages define the proposed method evaluated in this study. The proposed method is not a single classifier such as Random Forest or SVM. Rather, it is a complete, leak-aware, and statistically validated framework that governs how the outcome is constructed, how predictors are selected, how models are trained, how performance is compared, and how results are interpreted. The classifier selected from the experimental comparison represents the best-performing model within this proposed framework. In contrast, the framework itself provides a reproducible procedure for converting healthcare password-behavior survey data into defensible risk-stratification evidence.

4. Evaluation of the Proposed Framework and Experimental Results

This section evaluates the proposed leak-aware, survey-based machine learning framework using the cleaned healthcare password-behavior dataset. The evaluation focuses on whether the proposed framework can produce reliable, interpretable, and statistically defensible risk predictions from moderate-scale self-reported survey data. All models are therefore assessed within the same proposed pipeline: the same outcome construction rule, the same leakage-prevention procedure, the same preprocessing steps, the same stratified cross-validation design, and the same imbalance-aware evaluation metrics.

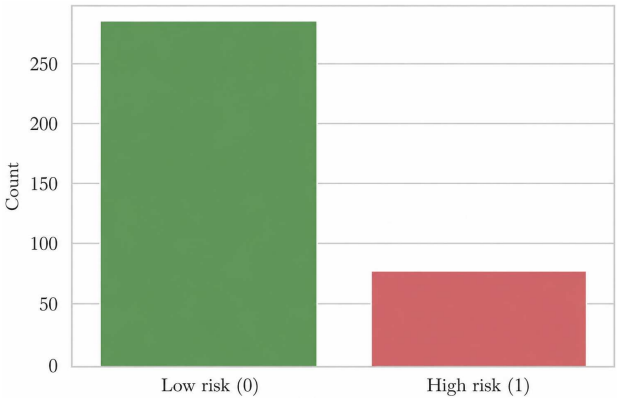
The experimental comparison is not intended to claim that a single algorithm universally outperforms all alternatives. Instead, it examines which classifier works best when embedded in the proposed framework. Random Forest, SVM, logistic regression, gradient-boosted trees, a TF-IDF logistic baseline, a rule-based heuristic, and a majority-class baseline are evaluated under identical conditions. This design allows the study to test the practical value of the proposed framework against simpler baselines and commonly used machine learning alternatives, while avoiding overstating small performance differences.

4.1. Dataset characterization and preliminary analysis

Table 5 summarizes the dataset structure and the key quality checks that guided the analysis. The table indicates that the study is of moderate size but sufficiently structured for supervised classification. The absence of missing values reduces preprocessing uncertainty, while the duplicate rate underscores the importance of distinguishing the raw workbook from the analytic copy. The 21.5% prevalence of the positive class confirms that evaluation metrics sensitive to imbalance, especially macro-F1, are more appropriate than accuracy alone [18–20].

The class imbalance is also visible in Figure 1. The distribution confirms that the target is imbalanced but not degenerate. This matters because a classifier can obtain deceptively high accuracy by favoring the majority class. The figure therefore reinforces the decision to prioritize macro-F1 and class-sensitive validation rather than reporting accuracy in isolation [18, 19].

Figure 1
Target class distribution after duplicate removal



The sensitivity analysis in Table 6 indicates that $t = 2$ provides the most balanced and statistically stable configuration. When $t = 1$, the positive class becomes overly prevalent, reducing discriminative power and inflating recall at the expense of precision. Conversely, when $t = 3$, the positive class becomes too sparse, increasing variance across folds and degrading

Table 5
Dataset summary and analytical data-quality profile

Item	Value	Interpretation
Raw records	416	Starting workbook size before cleaning
Analytical records	362	Unique responses retained after duplicate removal
Variables	35	Survey items and derived fields available in the raw file
Exact duplicates removed	54	Duplicate rate of 12.98% in the raw workbook
Missing values	None observed	No imputation was required for the raw worksheet
Target prevalence	78/362 = 21.5%	Minority positive class for high-risk password behavior
Analytical unit	One employee response	Cross-sectional survey respondent
Primary task	Binary classification	Predict self-reported high-risk password behavior

Table 6
Sensitivity analysis across threshold levels

Threshold (t)	Positive rate (%)	Macro-F1 (RF)	Std dev	Interpretation
1	High (~40–60%)	Lower	Moderate	Too lenient (noisy)
2	21.5%	Highest	Stable	Optimal
3	Low (<10%)	Lower	High	Too strict (sparse)

macro-F1 Random Forest (RF) due to insufficient representation of high-risk cases. In contrast, the threshold $t = 2$ achieves a moderate class distribution (21.5% positive class), which supports reliable model training while preserving meaningful variation in risk behavior. Empirically, this threshold yields the highest or near-highest macro-F1 scores among models, with lower variance than alternative thresholds. These findings suggest that $t = 2$ is a statistically appropriate compromise between sensitivity and stability, capturing consistent behavioral risk without introducing excessive noise or sparsity.

4.2. Cross-validated model performance and comparative evaluation

This subsection presents a comparative evaluation of the proposed framework against commonly used machine learning models in cybersecurity and healthcare analytics, including Random Forest, SVM, logistic regression (tabular and TF-IDF), and gradient-boosted trees. Two baselines, a rule-based heuristic and a majority-class predictor, are included for contextual benchmarking. All models are evaluated under a unified stratified cross-validation protocol using macro-F1 as the primary metric to address class imbalance, with accuracy, weighted F1, and ROC-AUC reported as complementary measures.

The results show that ensemble- and margin-based models consistently outperform naïve and rule-based baselines, aligning with prior findings for structured behavioral data. Random Forest achieves the highest macro-F1, indicating the best balance between minority- and majority-class prediction, while SVM demonstrates comparable recall but slightly lower precision. Logistic regression models perform moderately, capturing partial interaction structure but lacking nonlinear flexibility.

Performance differences among the strongest models are modest, consistent with recent studies where classical learners achieve statistically comparable results under rigorous validation. Gradient-boosted trees do not outperform Random Forest, suggesting that increased model complexity does not necessarily improve generalization in this setting. These findings reinforce the importance of robust preprocessing, leakage control, and evaluation discipline over architectural complexity. The comparative analysis demonstrates that the proposed framework performs competitively while maintaining methodological rigor and reproducibility, without overstating model superiority.

Table 7 summarizes model performance under stratified five-fold cross-validation. Random Forest achieves the highest overall performance, with the strongest macro-F1 (0.678 ± 0.091) and weighted F1 (0.795 ± 0.058), indicating the best balance between minority- and majority-class prediction. SVM is the second-best model (macro-F1 = 0.659 ± 0.080), with comparable recall (0.665 ± 0.087), suggesting similar sensitivity but slightly weaker precision. The logistic regression variants perform moderately (macro-F1 $\approx$ 0.63–0.65), indicating that linear decision boundaries capture part, but not all, of the underlying structure.

The gradient-boosted trees fallback performs worse (macro-F1 = 0.618 ± 0.080), suggesting that additional model complexity does not yield performance gains in this dataset. The rule-based heuristic and majority-class baseline achieve relatively high accuracy ($\approx$0.78), but substantially lower macro-F1 (0.560 and 0.440, respectively), confirming that accuracy is misleading in the presence of class imbalance. Their poor macro-level performance reflects a limited ability to detect the minority (high-risk) class. This is the clearest example of why accuracy alone is not sufficient for imbalanced security-risk detection [19, 20].

Table 7
Comparative evaluation of classifiers within the proposed leak-aware framework

Model	Accuracy	Macro precision	Macro recall	Macro-F1	Weighted F1	ROC-AUC
Random Forest	0.810 +/– 0.053	0.721 +/– 0.114	0.661 +/– 0.080	0.678 +/– 0.091	0.795 +/– 0.058	0.707
SVM	0.765 +/– 0.050	0.655 +/– 0.075	0.665 +/– 0.087	0.659 +/– 0.080	0.767 +/– 0.053	0.668
TF-IDF logistic regression	0.724 +/– 0.059	0.641 +/– 0.062	0.676 +/– 0.069	0.646 +/– 0.067	0.740 +/– 0.053	0.699
Tabular logistic regression	0.724 +/– 0.053	0.629 +/– 0.070	0.658 +/– 0.085	0.634 +/– 0.074	0.737 +/– 0.051	0.679
Gradient-boosted trees fallback	0.732 +/– 0.063	0.619 +/– 0.086	0.621 +/– 0.078	0.618 +/– 0.080	0.737 +/– 0.059	0.659
Rule-based heuristic	0.779 +/– 0.016	0.629 +/– 0.067	0.561 +/– 0.042	0.560 +/– 0.057	0.737 +/– 0.026	0.562
Majority-class baseline	0.785 +/– 0.007	0.392 +/– 0.003	0.500 +/– 0.000	0.440 +/– 0.002	0.690 +/– 0.009	0.491

Random Forest was also the most balanced among the stronger models. It improves over the more conservative baselines without relying on deep learning or heavy feature engineering. The TF-IDF logistic regression baseline performs competitively, suggesting that some predictive signal is present even in the serialized survey profile; however, the gain is not large enough to justify treating the dataset as a genuine NLP task.

Within the proposed framework, Random Forest was selected as the leading classifier because it achieved the highest macro-F1 and weighted F1 under the unified validation protocol. However, the proposed method should be understood as the complete framework rather than the Random Forest algorithm alone. Random Forest is the best-performing model produced by the framework, which also includes leakage-aware target construction, reproducible preprocessing, stratified validation, ablation, statistical testing, and explainability.

Figure 2 visualizes the model ranking using macro-F1. The bar chart confirms that Random Forest leads the candidate set, but the margin over SVM is modest. That visual gap matters: the result supports Random Forest as the selected model, yet it also shows that a single dramatically superior learner does not solve the problem. The small separation between the top models is consistent with a moderately difficult classification task and reinforces the need for statistical testing.

To evaluate the impact of class imbalance handling, the true unweighted baselines were compared against class-weighted and SMOTE-augmented configurations (Table 8). Notably, the primary Random Forest and SVM models reported in the preceding

Figure 2
Model comparison by macro-F1

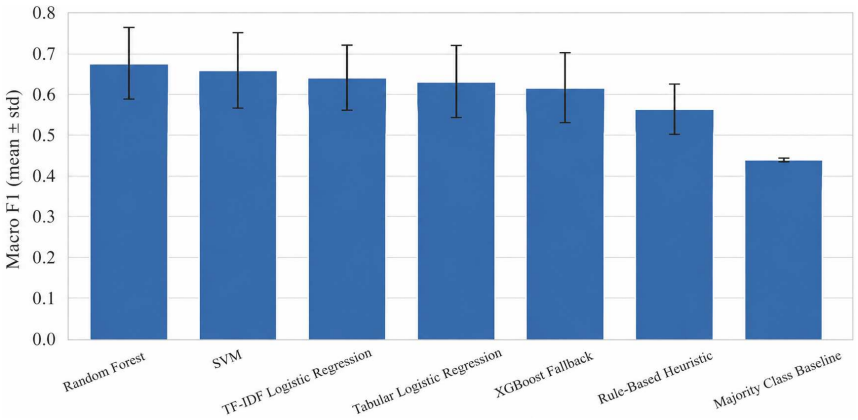


Table 8
Comparative performance of classification models under different imbalance-handling strategies

Model	Imbalance strategy	Accuracy	Macro-F1	Weighted F1	Minority precision	Minority recall	Minority F1	ROC-AUC	PR-AUC
Logistic regression	None	0.798 ± 0.024	0.603 ± 0.082	0.761 ± 0.041	0.658 ± 0.250	0.247 ± 0.145	0.325 ± 0.159	0.706 ± 0.116	0.527 ± 0.143
Logistic regression	Class weighting	0.724 ± 0.053	0.634 ± 0.074	0.737 ± 0.051	0.397 ± 0.103	0.542 ± 0.166	0.454 ± 0.117	0.690 ± 0.115	0.502 ± 0.141
Logistic regression	SMOTE	0.710 ± 0.044	0.628 ± 0.059	0.727 ± 0.041	0.383 ± 0.076	0.568 ± 0.154	0.453 ± 0.093	0.685 ± 0.125	0.489 ± 0.147
Random Forest	None	0.818 ± 0.031	0.612 ± 0.088	0.773 ± 0.049	0.781 ± 0.235	0.220 ± 0.120	0.330 ± 0.162	0.713 ± 0.109	0.564 ± 0.154
Random Forest	Class weighting	0.810 ± 0.053	0.678 ± 0.091	0.795 ± 0.058	0.593 ± 0.202	0.399 ± 0.136	0.472 ± 0.150	0.707 ± 0.103	0.516 ± 0.137
Random Forest	SMOTE	0.793 ± 0.048	0.633 ± 0.082	0.771 ± 0.053	0.539 ± 0.198	0.309 ± 0.109	0.391 ± 0.136	0.698 ± 0.100	0.506 ± 0.128
SVM	None	0.804 ± 0.030	0.572 ± 0.087	0.751 ± 0.050	0.725 ± 0.282	0.170 ± 0.124	0.258 ± 0.161	0.699 ± 0.099	0.524 ± 0.140
SVM	Class weighting	0.765 ± 0.050	0.659 ± 0.080	0.767 ± 0.053	0.453 ± 0.111	0.488 ± 0.155	0.469 ± 0.130	0.682 ± 0.117	0.478 ± 0.130
SVM	SMOTE	0.735 ± 0.056	0.620 ± 0.067	0.739 ± 0.049	0.411 ± 0.145	0.424 ± 0.091	0.411 ± 0.098	0.643 ± 0.103	0.423 ± 0.111

Note: Values are reported as mean ± standard deviation across validation folds. "None" denotes the baseline model without imbalance handling. Class weighting refers to cost-sensitive learning using class-weight adjustments, whereas SMOTE refers to synthetic minority oversampling. Minority-class metrics report performance for the high-risk class. ROC-AUC = area under the receiver operating characteristic curve; PR-AUC = area under the precision–recall curve.

analysis (Table 7) correspond to the class-weighted configurations, as the proposed framework inherently defaults to class weight = "balanced" to address the 21.5% prevalence of the minority class. The results in Table 8 justify this design choice: unweighted baseline models achieve deceptively high accuracy (up to 0.818 for Random Forest) but suffer from severely depressed minority recall (0.220 for Random Forest, 0.170 for SVM), meaning they frequently miss high-risk staff. Applying class weights substantially improves minority recall (e.g., Random Forest increases to 0.399; SVM to 0.488) and overall macro-F1, despite an expected trade-off in minority precision. Conversely, SMOTE improves recall over the unweighted baselines but introduces synthetic noise that degrades both precision and overall macro-F1 compared to class weighting [28]. Furthermore, class-weighted models maintain the highest area under the precision–recall curve (PR-AUC) scores relative to SMOTE (e.g., Random Forest class-weighted PR-AUC = 0.516 vs SMOTE PR-AUC = 0.506). These findings confirm that cost-sensitive learning via class weighting provides the most robust balance for this moderately sized structured dataset, thereby validating the configuration used in the primary framework.

4.3. Benchmark comparison with rule-based and naïve baselines

The proposed framework was also compared with two transparent benchmark strategies: a rule-based heuristic and a majority-class predictor. These baselines are important because imbalanced cybersecurity datasets can produce deceptively high accuracy even when the minority high-risk class is poorly detected. By comparing the best framework-selected classifier against these simple alternatives, the study evaluates whether the proposed pipeline provides practical predictive value beyond naïve or manually defined decision rules.

Table 9 places the best internal model alongside the two non-learned benchmark baselines used in the study. Random Forest substantially outperforms both the rule-based heuristic and the majority-class baseline across all metrics, achieving the highest macro-F1 (0.678), weighted F1 (0.795), and ROC-AUC (0.707). This indicates superior discrimination and a more balanced ability to detect both classes. The rule-based heuristic shows moderate performance (macro-F1 = 0.560), suggesting that simple domain rules capture some signal but lack generalization.

In contrast, the majority-class baseline attains relatively high weighted F1 (0.690) but very low macro-F1 (0.440) and ROC-AUC (0.491), confirming that it largely ignores the minority class and performs close to random in ranking ability. The comparison demonstrates that machine learning models provide meaningful predictive value beyond naïve and rule-based approaches, particularly in handling class imbalance and improving minority-class detection. It shows that the machine learning pipeline adds value beyond simple heuristics. The majority-class baseline preserves accuracy by ignoring the minority class, whereas the rule-based heuristic captures some structure but leaves substantial room for improvement. The Random Forest model therefore provides the strongest practical screening tool among the methods evaluated.

4.4. Feature-group ablation and robustness analysis

The ablation study evaluates the robustness of the proposed framework by testing how performance changes when complete feature groups are removed. This analysis is important because the proposed method is designed not only to maximize predictive performance but also to identify which survey domains contribute meaningful signal. All ablation experiments were conducted under the same preprocessing, cross-validation, and evaluation protocol used in the main model comparison.

Table 10 shows that knowledge and perception variables provide the most important contribution to prediction. Removing

Table 9
Benchmark comparison of the proposed framework-selected model against non-learned baselines

Role	Method	Macro-F1	Weighted F1	ROC-AUC	Note
Best internal model	Random Forest	0.678	0.795	0.707	Strongest validated classifier in this study
Rule-based benchmark	Rule-based heuristic	0.560	0.737	0.562	Transparent but limited by hand-crafted rules
Naïve benchmark	Majority-class baseline	0.440	0.690	0.491	Predicts only the dominant class

Table 10
Feature-group ablation analysis within the proposed framework

Ablation	Macro-F1	Weighted F1	Accuracy	Delta macro-F1 vs full
Shallow gradient boosting	0.662 +/– 0.094	0.763 +/– 0.070	0.755 +/– 0.074	0.044
No behavior context	0.635 +/– 0.071	0.746 +/– 0.059	0.740 +/– 0.071	0.017
No engineered numeric features	0.622 +/– 0.070	0.739 +/– 0.050	0.735 +/– 0.054	0.004
No demographics	0.620 +/– 0.043	0.737 +/– 0.026	0.732 +/– 0.032	0.002
Full gradient boosting	0.618 +/– 0.080	0.737 +/– 0.059	0.732 +/– 0.063	0.000
No knowledge/perception	0.545 +/– 0.052	0.678 +/– 0.049	0.666 +/– 0.061	–0.072

this feature group reduced macro-F1 from 0.618 to 0.545, producing the largest performance decline among all ablation conditions. In contrast, removing demographic variables, behavior-context variables, or engineered numeric features produced only small changes. These results indicate that the proposed framework captures risk primarily through security understanding and password-related perceptions rather than through demographic background alone. The shallow gradient-boosting configuration also outperformed the full gradient-boosted model, suggesting that restrained model complexity may improve generalization on moderate-sized survey datasets.

Figure 3 summarizes the ablation plot, visually reinforcing the interpretation that the predictive signal is distributed across multiple survey domains. No single feature family fully determines performance, and removing knowledge/perception information creates the largest drop. The visual pattern therefore supports a multicomponent risk profile rather than a single-variable explanation.

4.5. Statistical significance testing of model performance differences

Table 11 reports pairwise statistical comparisons between Random Forest and alternative models. Normality assumptions are satisfied in all cases ($p > 0.05$), supporting the use of paired *t*-tests. The results show that differences between Random Forest and SVM or logistic regression models are not statistically significant ($p > 0.05$), despite moderate effect sizes (Cohen's $d \approx 0.30$–0.62) and confidence intervals that include zero. This indicates comparable performance among top-performing models.

In contrast, Random Forest significantly outperforms the gradient-boosted fallback ($p = 0.049$), the rule-based heuristic ($p = 0.013$), and the majority-class baseline ($p = 0.004$), with large-to-very-large effect sizes ($d = 1.25$–2.63) and confidence intervals that exclude zero. These results confirm substantial performance gains over non-learned and weaker models. McNemar's

Figure 3
Ablation performance comparison

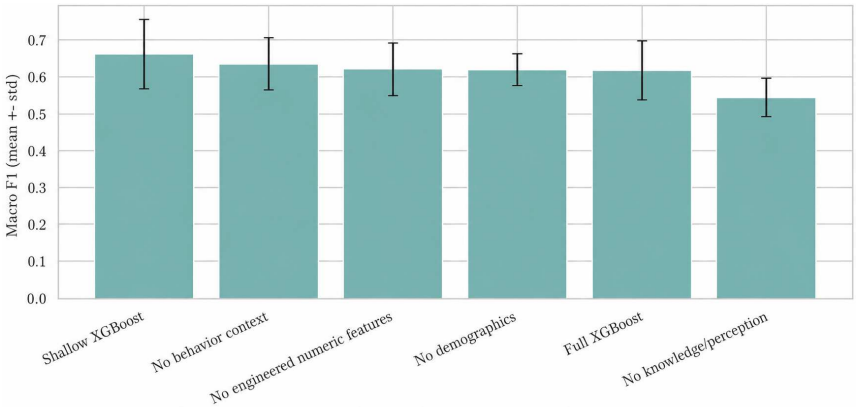


Table 11
Pairwise statistical validation of Random Forest against competing models

Comparison	Normality p	Test	*p*-value	McNemar p	Cohen's *d*	95% CI for macro-F1 difference	Decision
Random Forest vs SVM	0.3152	Paired *t*-test	0.5415	0.0177	0.298	[–0.0605, 0.0988]	Not significant
Random Forest vs TF-IDF logistic regression	0.9079	Paired *t*-test	0.3681	0.0003	0.453	[–0.0552, 0.1188]	Not significant
Random Forest vs tabular logistic regression	0.3811	Paired *t*-test	0.2382	0.0002	0.619	[–0.0441, 0.1318]	Not significant
Random Forest vs gradient-boosted trees fallback	0.6033	Paired *t*-test	0.0492	0.0000	1.249	[0.0004, 0.1201]	Significant
Random Forest vs rule-based heuristic	0.1178	Paired *t*-test	0.0131	0.1775	1.904	[0.0410, 0.1947]	Significant
Random Forest vs majority-class baseline	0.5278	Paired *t*-test	0.0042	0.2718	2.628	[0.1258, 0.3511]	Significant

test further indicates significant differences in prediction error patterns for several comparisons (notably vs SVM and logistic regression), suggesting that models may achieve similar aggregate performance while making different classification errors.

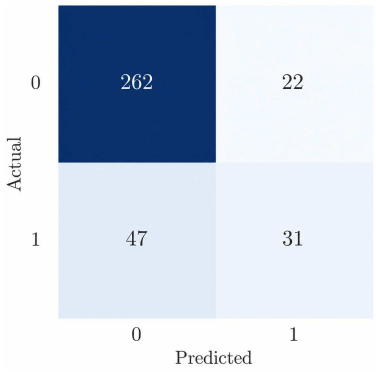
The findings indicate that Random Forest provides a measurable advantage over weaker baselines, but its superiority over other strong machine learning models is not statistically decisive, highlighting the importance of both effect size and uncertainty in model comparison. Its superiority is statistically supported only against the gradient-boosted tree fallback, the rule-based heuristic, and the majority baseline. This distinction is important: predictive ranking and inferential superiority are not the same statement [15, 16].

These results strengthen the methodological contribution of the proposed framework. The framework does not rely solely on ranking models by average performance; it also tests whether the observed differences are statistically significant. This distinction is essential for moderate-scale healthcare cybersecurity datasets, where small numerical differences may appear meaningful but may not survive paired statistical validation. The proposed framework therefore supports a more cautious and reproducible interpretation of model performance than a simple leaderboard-style comparison.

4.6. Instance-level error diagnostics and discriminative analysis

The confusion matrix in Figure 4 indicates that Random Forest is much stronger at identifying the negative class than the positive class. Using the out-of-fold counts, the model produced 262 true negatives, 22 false positives, 47 false negatives, and 31 true positives. This pattern implies high specificity but only moderate sensitivity to high-risk behavior. In practical terms, the model is useful for screening, but threshold tuning would be needed if the deployment goal were to minimize the number of missed high-risk cases.

Figure 4
Random forest confusion matrix



Given the operational priority of minimizing false negatives, the Random Forest model's default 0.50 classification threshold was further examined using out-of-fold prediction scores. The decision threshold was optimized on the precision–recall curve by maximizing the minority-class F1-score. This procedure yielded an optimal threshold of $\tau = 0.42466$, which may be reported as $\tau = 0.42$. At the default threshold of 0.50, the model achieved a minority-class sensitivity of 39.7% and a precision of 58.5%. After threshold optimization, sensitivity increased to 57.7%, while precision decreased to 50.6%, reflecting the expected trade-off when the decision rule is shifted toward greater case capture. Table 12 presents a deployment-oriented illustration for a hypothetical population of 1000 healthcare staff, using the observed study prevalence of high-risk behavior (21.5%), and shows the projected counts of true positives, false positives, true negatives, and false negatives under both the default and optimized thresholds.

Table 12
Comparison of default and optimized classification thresholds

Metric	Default threshold	Optimized threshold
Threshold	0.50	0.42
Sensitivity	39.7%	57.7%
Precision	58.5%	50.6%
Specificity	92.3%	84.5%
Minority-class F1	0.473	0.539
True positives	85	124
False positives	61	122
True negatives	724	663
False negatives	130	91

Note: The optimized threshold improved sensitivity and minority-class F1 by identifying more high-risk cases, but at the expense of lower precision and specificity due to increased false positives.

To evaluate the reliability of the model's probability estimates, a calibration analysis was conducted for the class-weighted Random Forest model. The model achieved a Brier score of 0.158, indicating moderate probabilistic accuracy. As shown in Figure 5, the calibration curve follows the general upward trend expected of a useful risk model. However, calibration is less stable in the lower-probability region, and the model shows some overprediction in several low-risk bins.

Since the outcome is imbalanced, precision–recall (PR) analysis was also used to assess discrimination more appropriately than accuracy alone. As shown in Figure 6, the Random Forest achieved an Average Precision (AP) score of 0.482, clearly outperforming the no-skill baseline AP of 0.215, which corresponds to the observed prevalence of the positive class. This result indicates that the model provides meaningful ranking ability for identifying high-risk staff in the presence of class imbalance.

Figure 6 presents the ROC curves for all evaluated models, illustrating their ability to discriminate between high-risk and low-risk cases across decision thresholds. Random Forest achieves the highest ROC-AUC (0.707), indicating the strongest overall ranking performance, although the margin over TF-IDF logistic regression (0.699) and SVM (0.668) is relatively modest. This close clustering of curves suggests that the top models provide similar discriminative capacity, consistent with the nonsignificant differences observed in pairwise statistical tests.

The gradient-boosted trees fallback (ROC-AUC = 0.659) performs slightly worse, while the rule-based heuristic (0.562) and majority-class baseline (0.491) show limited to near-random discrimination. The baseline curve lies close to the diagonal, confirming that it lacks meaningful ranking ability. The ROC analysis indicates that the task is moderately separable rather than strongly

Figure 5
Calibration diagnostics for the class-weighted random forest

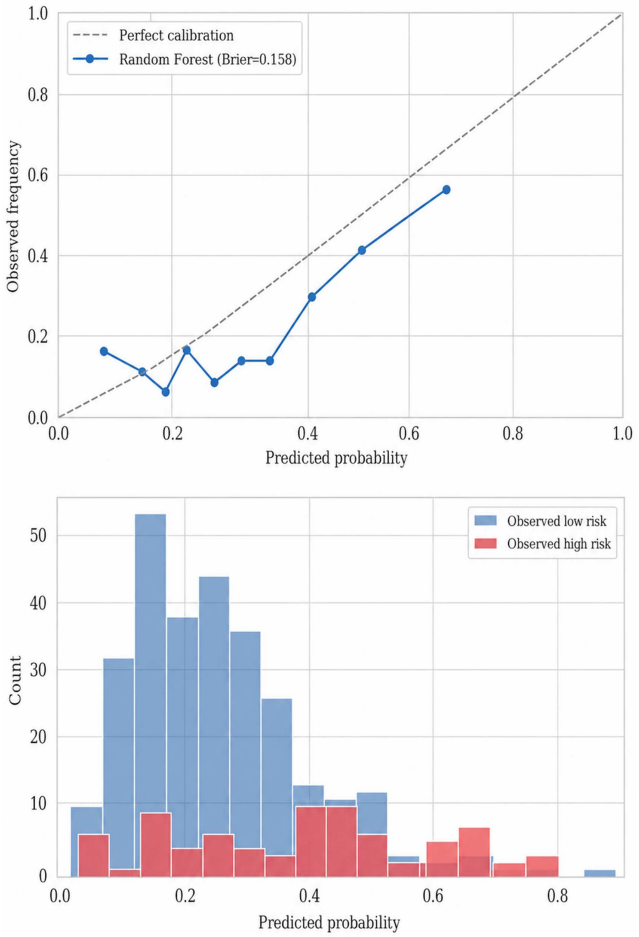
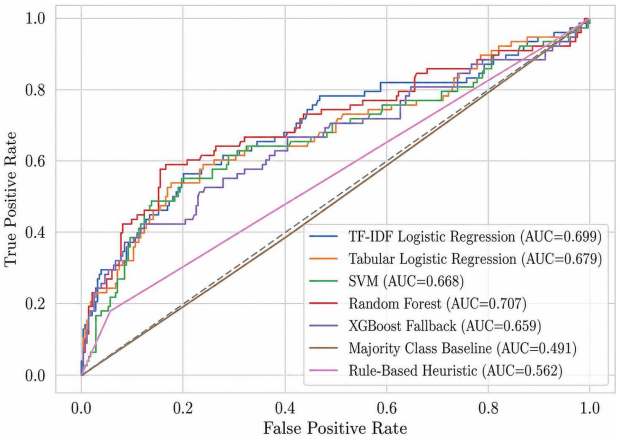


Figure 6
ROC curves for the evaluated models



separable, with no model achieving near-perfect discrimination. This reinforces the interpretation that the predictive signal in the survey data is real but limited, and that performance gains among top models are incremental rather than decisive. The visual evidence therefore supports a cautious interpretation: the framework is useful, but the problem remains substantively challenging [19].

4.7. Model explainability and global feature attribution

To provide a more rigorous and directionally interpretable explanation of model behavior, global feature importance was supplemented with Shapley Additive exPlanations (SHAP) [29]. Unlike conventional impurity-based importance measures, SHAP values quantify both the magnitude and the direction of each feature's contribution to the predicted probability of high-risk password behavior. Figure 7 presents the SHAP summary plot for the class-weighted Random Forest model. Notably, the indicator for endorsing the belief that password123 is a good security practice (yes) is associated with positive SHAP values, pushing predictions toward the high-risk class, whereas the corresponding no indicator is concentrated on the negative SHAP side, reducing predicted risk. This directional interpretability helps identify not only which features matter most but also how specific misconceptions influence risk, thereby supporting more targeted staff training and password-security interventions.

Figure 7
Global feature importance via mean absolute SHAP for the Random Forest model

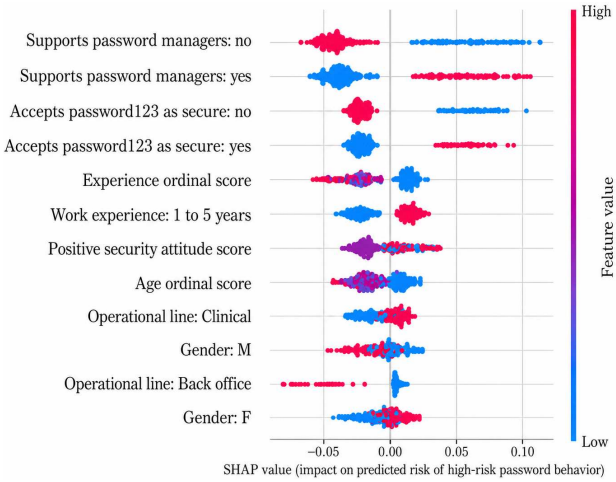
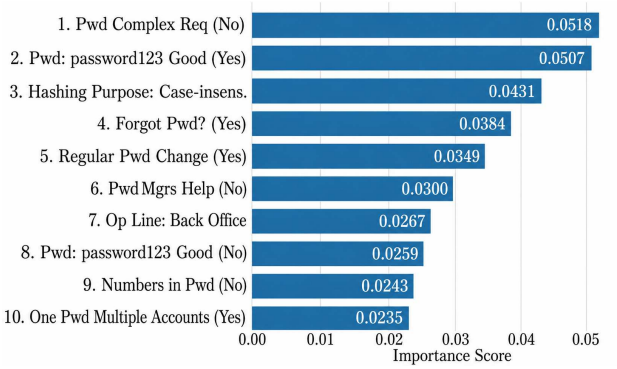


Figure 8 (top 10 feature importance) visually translates the numerical results presented in the feature importance table into a clear hierarchical structure. By mapping the relative weights of the top 10 predictors in the Random Forest model, the visualization provides an immediate intuitive understanding of which variables most significantly influence the classification outcome. The strength of this representation lies in its ability to illustrate the relative signal gap between features. For instance, the visualization clearly identifies "Password Complexity Requirements" (0.0518) and "Good Security Practices" (0.0507) as the most influential lexical cues. The descending horizontal bar format enables rapid comparison of importance scores that might be less apparent in raw tabular form, making the model's decision-making process transparent and accessible for policy screening and research governance.

5. Discussion

The results support a cautious but meaningful conclusion: self-reported high-risk password behavior in this healthcare sample can be predicted better than chance using a leak-free tabular

Figure 8
Visualized top 10 feature importance



framework. Random Forest is the strongest model on the primary selection metric, but the margins over SVM and the logistic regression baselines are not large enough to justify overclaiming. This is precisely why the manuscript separates predictive ranking from statistical inference. Predictively, Random Forest is best; inferentially, only some of its advantages are statistically significant.

The main contribution of the proposed framework is methodological rather than algorithmic. The results do not suggest that Random Forest is universally superior for predicting healthcare cybersecurity. Instead, they show that when survey-based password-risk data are processed through a leakage-aware, reproducible, and statistically validated pipeline, classical machine learning models can provide meaningful but appropriately bounded risk stratification. This is important because many applied cybersecurity studies report model performances without sufficiently addressing whether the target variable has leaked into the predictors, whether the evaluation metric is appropriate for class imbalance, or whether apparent performance differences are statistically reliable.

The finding that Random Forest's superiority over SVM and logistic regression is not statistically significant ($p > 0.05$) carries profound practical implications. In healthcare risk screening, marginal gains in macro-F1 achieved through complex ensemble architectures are often outweighed by the benefits of model transparency, stability, and ease of deployment. Therefore, we argue that interpretability and stability should take precedence over marginal, nonsignificant performance gains in this domain. Furthermore, the moderate ROC-AUC of 0.707 indicates that the framework is insufficient as a definitive, autonomous risk detector. Rather, its utility lies in its role as a preliminary screening tool, a triage mechanism to flag individuals or departments for further human-led cybersecurity assessment.

The proposed framework addresses these issues by linking each methodological step to a specific source of analytical risk. Target construction defines the behavioral risk proxy; leakage prevention protects model validity; stratified cross-validation controls class imbalance; ablation analysis identifies the survey domains that matter most; and statistical testing prevents overclaiming marginal differences between classifiers. In this sense, the proposed method contributes a reusable evaluation structure for healthcare organizations and researchers working with moderate-scale cybersecurity survey data.

The ablation results suggest that password risk is a multicomponent phenomenon. Knowledge/perception variables matter substantially, but they do not operate alone. Demographic and behavior-context features add a complementary signal, and the shallow gradient-boosting tree variant outperformed the full configuration. Taken together, these findings argue for a restrained model architecture that preserves multiple feature families without introducing unnecessary complexity.

The explainability results are consistent with the ablation study. The most important features are concentrated in insecure beliefs, password reuse, difficulty of recovery, and weak attitudes toward password management. That pattern suggests a practical reading of the survey: risk is associated with both what employees believe about passwords and how they actually manage them. The model therefore aligns with the broader behavioral security literature, which treats password risk as a human-factors problem rather than a purely technical one [3–5, 10].

This interpretation also fits the healthcare-specific literature. Prior studies have shown that password and security behavior in healthcare is shaped by local work context, policy understanding, and organizational culture, not only by technical controls [1, 2, 6, 11]. The present findings extend that line of work by showing that the same kinds of survey signals can be transformed into a reproducible classifier without overstating the case for deep learning.

At the same time, the confusion matrix shows that the model still misses a meaningful share of high-risk respondents. This is an important limitation for any deployment scenario. A screening tool with moderate recall can still be useful for triage or targeted training, but it should not be treated as a complete substitute for policy enforcement or follow-up assessment. If recall becomes the operational priority, threshold tuning or cost-sensitive training would be a reasonable next step. The research's main finding is not that a single model dominates all comparators, but that a reproducible, leak-aware, statistically validated survey pipeline can identify a meaningful high-risk subgroup within the healthcare workforce.

5.1. Practical contributions

This study offers three practical contributions. First, it provides a transparent screening framework that hospitals can adapt when they have survey data but not system logs or free-text incident narratives. Second, it identifies knowledge/perception items and password-management habits as particularly informative for risk stratification, helping prioritize training content. Third, it demonstrates a reproducible workflow in which preprocessing, ablation, benchmarking, and statistical testing are all documented rather than treated as hidden implementation details [15, 16].

For practice, the immediate implication is that cybersecurity interventions should combine awareness-building with behavior-focused support. General policy reminders are unlikely to be sufficient if they are not linked to specific misconceptions and daily password routines. The model outputs, therefore, suggest a targeted rather than generic intervention strategy [4, 13, 14]. The interventions should prioritize correcting misconceptions (e.g., beliefs about password reuse) rather than focusing solely on enforcement policies.

5.2. Theoretical contributions

The theoretical contribution is modest but meaningful. The results support a multicomponent view of password risk in which beliefs, habits, and work context jointly shape the probability of high-risk behavior. In that sense, the study extends survey-based cybersecurity research by showing that structured questionnaire

data can be used not only descriptively but also as the basis for a reproducible predictive framework [5, 8, 10, 11].

The study also reinforces a methodological point important in applied machine learning: greater complexity is not automatically better. The shallow boosted-tree ablation outperforming the full version and the close competition among Random Forest, SVM, and logistic regression both suggest that defensible model choice should be driven by data structure and validation evidence rather than by the appeal of deep or highly complex architectures [21–23].

5.3. Limitations

Several limitations must be stated clearly. The study is cross-sectional and self-reported, so it supports predictive association rather than causality. The sample comes from one hospital context, which limits external generalization. The target is a derived proxy, not a directly observed breach record or telemetry trace. Exact duplicates were present in the raw workbook and were removed under the assumption that they were repeated records rather than intentional repeat measures. No independent external test set was available, so validation remains internal.

The confusion matrix reveals higher specificity than sensitivity, indicating the model is more adept at ruling out low-risk staff than detecting high-risk ones. In a cybersecurity context, false negatives (missed high-risk individuals) are highly consequential, as they leave the system vulnerable. This limitation underscores that the current framework should not be used to conclusively clear staff of risk, but rather to identify clusters of risk that warrant targeted intervention conservatively. Future iterations must explore cost-sensitive learning to penalize false negatives more heavily.

Another methodological limitation concerns the use of five-fold cross-validation as the basis for model comparison and statistical testing. Although stratified five-fold cross-validation is appropriate for preserving class balance in a moderate-sized dataset, it produces only five-fold performance observations for each model. This limited number of observations reduces the statistical power of paired comparisons and makes confidence intervals around performance differences relatively uncertain. Therefore, nonsignificant differences between top-performing models, such as Random Forest, SVM, and logistic regression, should not be interpreted as definitive evidence of equivalent performance. Rather, they indicate that, within the available five-fold validation design, the observed performance differences were insufficient to support a statistically decisive claim of superiority. Future studies should consider repeated cross-validation, bootstrap confidence intervals, or external validation on independent healthcare datasets to obtain more stable estimates of model performance.

Moreover, because the model relies on self-reported behavioral data, its use raises ethical considerations regarding labeling, privacy, and potential misuse for employee monitoring. The framework should therefore be applied as a supportive screening tool rather than a punitive classification system. There is also a modality limitation. Since the workbook contains no genuine free-text field, any text-style representation is only a serialized form of structured survey responses. That is suitable for a weak lexical baseline, but it does not justify a transformer-based claim. The positive-class recall of the best model is only moderate, suggesting the framework should be viewed as a screening aid rather than a fully autonomous decision-making system.

The framework relies on internal cross-validation on a modest sample (N = 362) from a single institution, which limits external generalizability. Replication across larger, more diverse healthcare populations is necessary before widespread deployment. Crucially, the ethical implications of employee risk prediction systems must be addressed. Predictive cybersecurity profiling risks being weaponized for punitive human resource (HR) actions rather than supportive interventions. We advocate that the deployment of this framework must strictly adhere to a "support, not punish" paradigm: results should be used exclusively to tailor security awareness training and allocate IT resources, never as grounds for disciplinary action, thereby preserving employee trust and psychological safety.

6. Conclusion

This study proposed and evaluated a leak-aware, survey-based machine learning framework to detect self-reported high-risk password behavior among healthcare staff. The framework addresses a common methodological challenge in cybersecurity behavior research: how to transform structured survey responses into a valid, reproducible, and interpretable prediction task without allowing the target-defining variables to leak into the predictor set. To solve this problem, the proposed method integrates composite outcome construction, leakage prevention, deterministic preprocessing, stratified cross-validation, imbalance-aware evaluation, feature-group ablation, statistical significance testing, and model explainability within a single analytical workflow.

Empirically, Random Forest achieved the strongest overall performance within the proposed framework, with the highest macro-F1 and weighted F1 among the evaluated models. However, statistical testing showed that its advantage over SVM and logistic regression was not decisive, while its superiority over weaker baselines and the gradient-boosted fallback was significant. These findings demonstrate why the proposed framework is necessary: model rankings alone are insufficient unless they are supported by paired validation, effect-size estimation, and cautious interpretation.

The ablation and feature-importance results further show that high-risk password behavior is associated mainly with knowledge, perception, and password-management practices rather than demographic background alone. This suggests that healthcare cybersecurity interventions should prioritize targeted training, correction of misconceptions, and practical password-management support. Overall, the study provides a reproducible, statistically disciplined framework for survey-based cybersecurity risk detection in healthcare settings. The framework is not intended to replace organizational security controls or direct behavioral monitoring. Still, it can support early screening, staff training design, and future research on human-centered cybersecurity analytics.

Acknowledgement

The authors express their sincere gratitude to the executives and participants of the private hospital in Phuket for their valuable cooperation and participation in the questionnaire survey. Their contributions were essential to the successful completion of this study.

Ethical Statement

All subjects provided informed consent for inclusion before participating in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Human Research Ethics Committee of the Phuket Rajabhat University, Thailand (Reference No.: PKRU2567/08), on April 29, 2024, with approval valid until April 28, 2025.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Github at https://github.com/pjarupunphol/Datasets/blob/main/dataset.zip.

Author Contribution Statement

Pita Jarupunphol: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Siwatchaya Suksai:** Software, Validation, Investigation, Resources, Data curation, Visualization.

References

[1] Al Toobi, A., & Al Suqri, M. (2025). Information security behavior of healthcare professionals in the Sultanate of Oman based on the PMT model. *Scientific Reports*, *15*(1), 1–11. https://doi.org/10.1038/s41598-025-26917-x

[2] Czaja, P., Gdowski, B., Niemiec, M., Mees, W., Stoianov, N., Votis, K., . . . , & Merialdo, M. (2025). Cybersecurity challenges and opportunities of machine learning-based artificial intelligence. *Neural Computing and Applications*, *37*(33), 27931–27956. https://doi.org/10.1007/s00521-025-11604-9

[3] He, Z., Davila, D., Bi, S., Wang, T., & Hou, T. (2025). Machine learning for cybersecurity: A survey of applications, adversarial challenges, and future research directions. *Electronics*, *14*(23), 1–28. https://doi.org/10.3390/electronics14234563

[4] Kaur, J., & Ramkumar, K. R. (2022). The recent trends in cyber security: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 5766–5781. https://doi.org/10.1016/j.jksuci.2021.01.018

[5] Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. *Annals of Data Science*, *10*(6), 1473–1498. https://doi.org/10.1007/s40745-022-00444-2

[6] Wani, T. A., Mendoza, A., Gray, K., & Smolenaers, F. (2022). BYOD usage and security behaviour of hospital clinical staff: An Australian survey. *International Journal of Medical Informatics*, *165*, 104839. https://doi.org/10.1016/j.ijmedinf.2022.104839

[7] Islam, R., Yildirim, G., Rahman, A., & Kamal, M. S. (2026). Artificial intelligence in cybersecurity and education: A bibliometric review of emerging trends and research frontiers. *Sustainable Futures*, *11*, 1–12. https://doi.org/10.1016/j.sftr.2026.101807

[8] Sari, P. K., Handayani, P. W., Hidayanto, A. N., Yazid, S., & Aji, R. F. (2022). Information security behavior in health information systems: A review of research trends and antecedent factors. *Healthcare*, *10*(12), 1–21. https://doi.org/10.3390/healthcare10122531

[9] Gökstorp, S., Katsikeas, S., & Johnson, P. (2026). Machine learning for cybersecurity: A comprehensive literature review. *ACM Computing Surveys*, *58*(9), 1–36. https://doi.org/10.1145/3796543

[10] Aschwanden, R., Messner, C., Höchli, B., & Holenweger, G. (2024). Employee behavior: The psychological gateway for cyberattacks. *Organizational Cybersecurity Journal: Practice, Process and People*, *4*(1), 32–50. https://doi.org/10.1108/OCJ-02-2023-0004

[11] Raj Sreenath, S. S., Hewitt, B., & Sreenath, S. (2025). Understanding security behaviour among healthcare professionals by comparing results from technology threat avoidance theory and protection motivation theory. *Behaviour & Information Technology*, *44*(2), 181–196. https://doi.org/10.1080/0144929X.2024.2314255

[12] Yeng, P. K., Nweke, L. O., Yang, B., Ali Fauzi, M., & Snekkenes, E. A. (2021). Artificial intelligence–based framework for analyzing health care staff security practice: Mapping review and simulation study. *JMIR Medical Informatics*, *9*(12), 1–25. https://doi.org/10.2196/19250

[13] Hadzidedic, S., Fajardo-Flores, S., & Ramic-Brkic, B. (2022). User perceptions and use of authentication methods: Insights from youth in Mexico and Bosnia and Herzegovina. *Information & Computer Security*, *30*(4), 615–632. https://doi.org/10.1108/ICS-07-2021-0105

[14] Yu, W., Yin, Q., Yin, H., Xiao, W., Chang, T., He, L., . . . , & Ji, Q. (2023). A systematic review on password guessing tasks. *Entropy*, *25*(9), 1–28. https://doi.org/10.3390/e25091303

[15] Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, *119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

[16] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

[17] Sarker, M. M., Jony, M. J. H., Ullah, M. W., Begum, J., & Naushin, N. (2025). Trends of machine learning, cybersecurity and big data analytics in industry 4.0. *Journal of Information Systems and Informatics*, *7*(4), 3330–3360. https://doi.org/10.63158/journalisi.v7i4.1321

[18] Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 1–23. https://doi.org/10.1186/s13040-023-00322-4

[19] Aguilar-Ruiz, J. S., & Michalak, M. (2024). Classification performance assessment for imbalanced multiclass data. *Scientific Reports*, *14*(1), 1–10. https://doi.org/10.1038/s41598-024-61365-z

[20] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*, 6086. https://doi.org/10.1038/s41598-024-56706-x

[21] Scornet, E., & Hooker, G. (2026). Theory of random forests. *Annual Review of Statistics and Its Application*, *13*(1), 99–121. https://doi.org/10.1146/annurev-statistics-112723-034707

[22] Du, K. L., Jiang, B., Lu, J., Hua, J., & Swamy, M. N. S. (2024). Exploring kernel machines and support vector machines: Principles, techniques, and future directions. *Mathematics*, *12*(24), 1–58. https://doi.org/10.3390/math12243935

[23] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[24] Nam, Y., & Han, S. (2025). Random forest variable importance-based selection algorithm in class imbalance problem. *Journal of Classification*, *42*, 660–673. https://doi.org/10.1007/s00357-025-09512-7

[25] Midway, S., & White, J. W. (2025). Testing for normality in regression models: Mistakes abound (but may not matter). *Royal Society Open Science*, *12*(4), 1–14. https://doi.org/10.1098/rsos.241904

[26] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

[27] Schubert, A. L., Steinhilber, M., Kang, H., & Quintana, D. S. (2025). Improving statistical reporting in psychology. *Communications Psychology*, *3*(1), 156. https://doi.org/10.1038/s44271-025-00356-w

[28] Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Frontiers in Digital Health*, *6*, 1–17. https://doi.org/10.3389/fdgth.2024.1430245

[29] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11), 1–15. https://doi.org/10.1111/cts.70056

How to Cite: Jarupunphol, P., & Suksai, S. (2026). A Reproducible Machine Learning Framework for Detecting High-Risk Password Behavior in Healthcare Staff. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA620210060