big data and cognitive computing

MDPI

Article

Cognition Orientation Risk Evaluation—A Personality Driven Integrated Model for Phishing Susceptibility

Chih-Hong Kao [1], Chia-Wei Tsai [2], Yu-Ting Kao [3] and Chao-Lung Chou [3,4,*]

[1] Department of Intelligent Vehicle and Energy, Minth University of Science and Technology, Hsinchu 307, Taiwan; oakhc888@o365.mitust.edu.tw
[2] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung 404, Taiwan; cwtsai@nutc.edu.tw
[3] Master's Program of Information and Communication Security, Feng Chia University, Taichung 407, Taiwan; m1416803@o365.fcu.edu.tw
[4] Department of Information Engineering and Computer Science, Feng Chia University, Taichung 407, Taiwan
* Correspondence: clchou@o365.fcu.edu.tw

Abstract

Phishing attacks exploit human vulnerabilities through social engineering techniques; therefore, a multidimensional framework integrating personality and cognition is crucial for tailoring personalized defenses against phishing threats. To prevent phishing attacks, focusing on psychological mechanisms has become the primary approach to address the human-centric nature of these threats. In response, we propose an integrated framework that synthesizes dimensions from the Five-Factor Model (FFM) and the Myers–Briggs Type Indicator (MBTI), grounded in Dual Process Theory to explore the cognitive drivers underlying decision-making under threat. To operationalize this framework, we developed a decision tree classifier to quantify the predictive significance of various personality traits and their hierarchical interactions. The results indicate that the personality types associated with the highest phishing risk profiles are ENFP, ESFP, ESTP, and ENFJ. These hierarchical classification results are projected onto the C.O.R.E. Quadrant (Cognition Orientation Risk Evaluation Quadrant) which enables the systematic representation of risk patterns across all 16 personality types. By providing a structured visualization of personality-driven risk patterns, the C.O.R.E. Quadrant offers a practical foundation for developing personalized defense mechanisms and tailored cybersecurity training strategies, moving beyond one-size-fits-all security protocols.

Keywords: phishing attacks; human vulnerability; social engineering; MBTI; Dual Process Theory; cybersecurity

Check for updates

Academic Editors: Ming Hour Yang, Vijayalakshmi Murugesan and Mercy Shalinie Selvaraj

Received: 10 March 2026
Revised: 18 July 2026
Accepted: 27 July 2026
Published: 3 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

According to Lastdrager's definition, phishing is an act of deception whereby impersonation is used to obtain information from a target [1]. Such attacks commonly involve a range of social engineering techniques, including persuasion, manipulation, deception, and inducement [2]. Unlike conventional cyberattacks, phishing attacks predominantly target human vulnerabilities in cognitive judgment, emotional responses, and social interaction. Empirical cybersecurity reports consistently indicate that phishing ranks as the fifth leading cause of security incidents globally and remains a leading contributor to data breaches worldwide [3]. Moreover, contemporary phishing techniques have grown increasingly sophisticated, encompassing email phishing, short message service (SMS) phishing

(smishing), social media phishing, voice phishing (vishing), and even cross-media social engineering attacks incorporating deepfake technologies [4]. Research has shifted from system-level defenses to psychological and behavioral mechanisms, as technical measures fail to counter evolving threats effectively.

For example, aligned with Dual Process Theory, in phishing contexts, attackers strategically deploy psychological cues—such as urgency, authority, and social validation—to elicit rapid responses that circumvent scrutiny. These tactics disproportionately trigger heuristic processing in susceptible individuals, thereby suppressing systematic reasoning and rigorous risk evaluation. This mechanism conceptualizes human cognition as operating through two parallel modes: heuristic processing, characterized by intuitive, automatic, and low-effort judgments, and systematic processing, which entails deliberate, analytical, and cognitively demanding evaluation [5]. Empirical studies further indicate that individuals with different personality traits exhibit variation in information-processing strategies, risk perception, and behavioral responses when confronted with phishing scenarios [6–8]. Collectively, these findings suggest that personality traits play a pivotal role in determining how cognitive pathways are activated under deceptive conditions, ultimately shaping phishing susceptibility.

Since personality traits influence judgment under stress conditions, time pressure, authority cues, and emotional stimulation, many studies have adopted the Five-Factor Model (FFM) as an analytical framework. Butavicius et al. [9] reported that individuals high in Extraversion and Openness were more capable of identifying phishing emails. In contrast, Albladi and Weir [10] and Lawson et al. [11] found that higher Extraversion increased susceptibility to phishing. These conflicting findings suggest that, within the FFM framework, empirical evidence remains contradictory [12,13]. Such inconsistencies indicate that trait-based models alone may be insufficient to capture the complex human-factor dynamics inherent in phishing scenarios. Understanding personality traits has become essential in cybersecurity research, as seen by prior research highlighting the profound impact of personality on security policy compliance and protective measures adoption [14–16].

While FFM is effective in describing who an individual is, it is considerably less capable of explaining how personality dimensions influence final decision-making behavior dynamically, as information-processing styles. Exclusively relying on trait-oriented analysis overlooks the critical interaction between personality dimensions and cognitive processes. To address this limitation, this study argues for the adoption of another personality framework that more directly reflects cognitive processes and behavioral tendencies involved in decision-making. Unlike FFM, the Myers–Briggs Type Indicator (MBTI) focuses on individuals' preferential patterns of cognition and behavior when processing information, coping with pressure, and making decisions.

There is an obvious convergence between MBTI and FFM. Based on the correlation coefficients established in three seminal studies [17–19]. In the field of information security, integrating MBTI with phishing strategies provides a robust framework for identifying the sensitivity of specific personality types to manipulative tactics. Personality traits play a dual role within an organization's security defense architecture. Cybersecurity errors do not randomly occur; rather, they are closely coupled with cognitive biases stemming from a specific personality. This study confirms that distinct MBTI dimensions exhibit significant variances in susceptibility when confronted with cyber threats. In information security, integrating MBTI with phishing tactics offers a framework to identify personality-specific sensitivities to manipulative strategies [20].

Accordingly, this study adopts MBTI as the primary personality framework and examines phishing message processing. This framework offers a comprehensive perspective

that captures how individuals allocate attention, favor heuristic or systematic processing, and respond to time pressure and persuasive cues when encountering phishing messages. Building on this theoretical shift, the study integrates MBTI dimensions and Dual Process Theory to propose an integrated risk assessment framework: the C.O.R.E. Quadrant (Cognition Orientation Risk Evaluation Quadrant). This quadrant systematically categorizes MBTI personality types into distinct risk groups based on cognitive processing tendencies, providing an interpretable and theory-driven representation of personality-based differences in phishing susceptibility.

The primary objective of this study is to model and predict the correlation between personality traits and phishing susceptibility by leveraging the C.O.R.E. Quadrant. The necessity of this integrated approach is driven by the critical need to bridge the gap between technical detection and human vulnerabilities. While AI and Large Language Models (LLMs) have demonstrated advanced semantic analysis capabilities in identifying subtle social engineering patterns [21,22], purely technical countermeasures often overlook the 'Human Factor.' Empirical evidence [23,24] underscores this limitation, proving that personality traits serve as pivotal indicators for predicting high-risk online behaviors.

Consequently, the paradigm of cybersecurity defense is shifting from static technical filtering toward behavioral risk prediction. Incorporating personality traits into risk models not only addresses the blind spots of conventional detection but also represents a definitive trend in future research.

2. Materials and Methods

This study seeks to quantify the relationship between MBTI personality types and phishing susceptibility by integrating machine learning and statistical analysis. The proposed approach is structured into four principal stages, as illustrated in Figure 1.

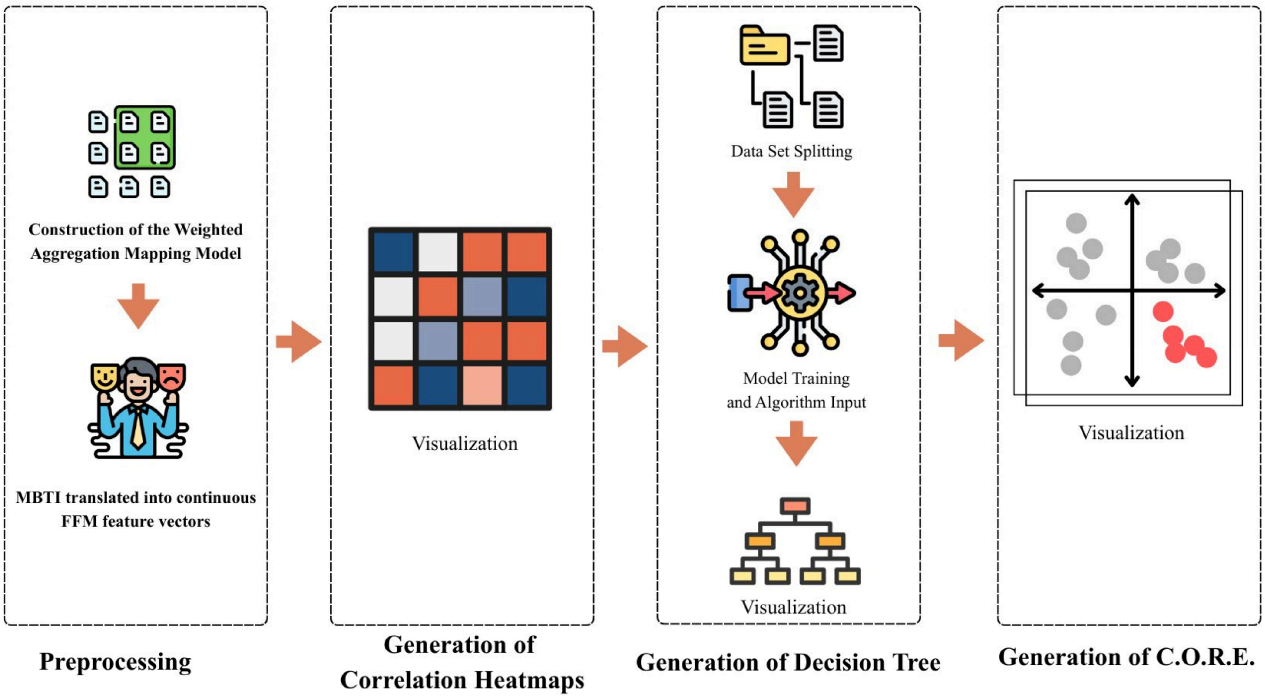


Figure 1. Flowchart of proposed method.

2.1. Preprocessing

The 'Personality Classification Data: 16 Personalities' dataset, a large-scale synthetic dataset specifically engineered for machine learning classification tasks, was adopted for

analysis in this study. This dataset is from Kaggle [25], providing a diverse sample of 59,999 individual profiles. The author utilized a Likert Scale distribution to map questionnaire responses onto a numerical interval between −3 and 3. Responses are encoded on a 7-point scale: 3 (Fully Agree), 2 (Partially Agree), 1 (Slightly Agree), 0 (Neutral), −1 (Slightly Disagree), −2 (Partially Disagree), and −3 (Fully Disagree). This process generates high-dimensional features that provide a clear classification of complex personality traits. The dataset samples contain questionnaire features corresponding to various MBTI dimensions. During the data preprocessing phase, this study established a theoretical correlation between MBTI dimensions and FFM traits through an extensive literature review [17–19], further deriving a cross-model weighted transformation mechanism. This research integrated mapping data from multiple datasets in the construction of the Weighted Aggregation Mapping Model as defined in Equation (1). By incorporating the sample size (n) from each respective source as a weighting coefficient ($w_{ij}^{combined}$), to ensure statistical stability and cross-study consistency, the formula calculates aggregated weights that facilitate a robust personality model transformation.

$$w_{ij}^{combined} = \frac{n_A \times w_{ij}^A + n_B \times w_{ij}^B + n_C \times w_{ij}^C}{n_A + n_B + n_C} \quad (1)$$

$w_{ij}^{combined}$ denotes the final aggregated mapping coefficient, representing the overall association between the *i* dimension of MBTI and the *j* trait in FFM. The terms w_{ij}^A, w_{ij}^B, and w_{ij}^C correspond to the correlation coefficients. The variables n_A, n_B, and n_C indicate the sample sizes of these studies. The denominator, $n_A + n_B + n_C$, represents the total sample size, ensuring that the aggregated coefficient is properly normalized. Based on the aggregated weights, the raw MBTI categorical data are translated into continuous FFM feature vectors via Equation (2). This linear combination model operates on the four binary dimensions and their corresponding weighting coefficients (w_{ij}), introducing an error term (ε) to account for residual variances inherent in the mapping process.

$$BF_j = \sum_{i=1}^{4} w_{ij} \times MBTI_i + \varepsilon \quad (2)$$

2.2. Generation of Correlation Heatmaps

This stage aims to validate and illustrate the cross-theoretical mapping relationship between MBTI dimensions and FFM traits through visualization. This approach visualizes correlation intensities and establishes statistical feature vectors for decision tree training. To address the limitations of combined personality dimensions found in existing research, this study deconstructs MBTI pairings into eight discrete binary variables, facilitating their inclusion in linear statistical modeling. Upon establishing this framework, we utilized Equation (1) to aggregate data, with weighting levels determined by the respective sample sizes of the reviewed studies. This approach gives more weight to studies with larger sample sizes, ensuring that they have a greater influence on the results. Finally, leveraging Python 3.13-based computational tools, the weights were projected onto the MBTI samples to generate continuous FFM feature vectors following Equation (2).

In the visualization of the model, the color gradient serves as an analytical indicator: intense red denotes a strong positive correlation, deep blue reflects a strong negative correlation, and neutral tones suggest negligible association. Specifically, a rigorous inspection of the heatmap revealed that the Neuroticism trait exhibited consistently near-zero correlation coefficients across all eight MBTI dimensions. Consequently, to refine the predictive power of the model, the Neuroticism feature was strategically removed during the feature pruning phase, ensuring that only the most relevant personality constructs were involved. The heatmap generated during this phase offers two primary contributions to this study. First, it

enhances the model's capacity to interpret individual cognitive processes by transforming categorical personality labels into mathematically meaningful continuous vectors. Second, it enables researchers to rapidly identify the various personality dimensions to cybersecurity risk, thereby establishing the empirical data foundation for the subsequent construction of the C.O.R.E. Quadrant.

2.3. Generation of Decision Tree

This model specifically aims to measure how different personality constructs influence the cognitive shift between heuristic intuition and systematic processing during threat assessment. This study employs a decision tree algorithm to perform machine learning analysis on the 59,999 samples, aiming to identify which personality dimensions possess the highest predictive contribution toward phishing susceptibility. During the modeling and feature engineering phase, this study utilized MBTI binary encoded features (E-I, S-N, T-F, J-P) extracted from raw samples as the primary model inputs.

To enhance predictive precision, this stage incorporates the interaction effects between personality dimensions. The generated decision tree utilizes Extraversion as the root node for initial branching, followed by hierarchical filtering based on the Sensing-Intuition dimension and decision-making functions (Thinking-Feeling). This structure illustrates how various personality traits form distinct risk pathways. The classification and construction process of the model is formalized in the following steps:

1. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the sample set, where each sample x_i consists of a feature vector $F_{feature} = \{E,I,S,N,T,F,J,P\}$ derived from MBTI binary encoding.
2. Let $S_{susceptibility}$ be the set of individual phishing susceptibility scores, with the calculation function defined as Equation (3):

$$S_{susceptibility} = \bigcup_{k=1}^{n} \left\{ \sum_{j \in F} w_j \times x_{kj} + \Delta S_{interaction} + \varepsilon \right\} \quad (3)$$

The set $S_{susceptibility}$ serves as the output of the algorithm, representing the set of phishing risk scores calculated for all individual samples in the dataset; $\bigcup_{k=1}^{n}$ (union symbol) represents the compiled set of individual scores computed for all 59,999 samples in the dataset; w_j represents each MBTI dimension (e.g., E, I, S, N). A positive weight signifies an increase in susceptibility, whereas a negative weight indicates a contribution to risk mitigation; x_{kj} represents the value of the kth sample on the jth personality dimension, utilizing binary encoding (1 or 0) to indicate the presence of a specific personality trait. This value is used to compute the weighted sum of the sample across the entire feature set $F_{feature}$; $\Delta S_{interaction}$ accounts for the additional impact generated by specific combinations of personality traits.

For instance, the simultaneous presence of Feeling (F) may trigger a multiplicative effect that further amplifies an individual's vulnerability beyond the sum of individual traits; ε denotes the error term for factors outside the model and the behavioral uncertainty found in practical decision-making.

3. G(pow) denotes the function used to evaluate and execute splits at each node within the decision tree structure, defined as Equation (4):

$$G(\text{pow}) \begin{cases} p, \text{ if Optimal_Split}(S, \text{ Param}) = p \\ p', \text{ if Grid_Search}(\text{depth}, \ldots) = p' \end{cases} \quad (4)$$

Gini Impurity serves as the evaluation criterion for node splitting, aiming to identify the personality dimensions that most effectively discriminate between 'high-risk' and 'low risk' samples; Optimal_Split (S, Param) serves as the searching for the specific attributes, such as Extraversion (E), that minimizes node impurity and maximizes

classification gain within the dataset $S_{susceptibility}$; p represents the core splitting dimension with the highest feature importance weight. In this model, Extraversion (E) is frequently identified as the primary splitting factor for the root node; Grid_Search (depth, . . .) represents an automated identification of optimal model hyperparameters, such as tree depth and minimum samples per leaf node; p' denotes the state of node distribution that possesses the optimal generalization capability.

4. L_{Risk} = {Low, Medium, High} defines the risk categories, which are mapped through the percentile threshold function Z (pow):

$$Z(\text{pow}) = \begin{cases} \text{Low} : S_{\text{susceptibility}}(k) < P_{33} \\ \text{Medium} : P_{33} \leq S_{\text{susceptibility}}(k) \leq P_{66} \\ \text{High} : S_{\text{susceptibility}}(k) > P_{66} \end{cases} \quad (5)$$

where P_{33} and P_{66} represent the 33rd and 66th percentiles of the phishing susceptibility distribution, respectively.

Visualized through the decision tree, the algorithmic output facilitates the accurate identification of high-risk trait combinations. Visualization establishes the quantitative foundation for mapping individual risks within the C.O.R.E. Quadrant.

2.4. Generation of C.O.R.E.

In this stage, the predictive outputs of the model are further transformed into a cybersecurity risk map with practical significance, named the Cybersecurity Orientated Risk Evaluation Quadrant (C.O.R.E. Quadrant). By mapping personality vectors and cognitive preferences onto a two-dimensional system of Risk and Resilience, this study precisely positions trainees within four risk quadrants. Specifically, the quadrant is primarily composed of individuals characterized by high social influence susceptibility and a preference for heuristic processing, typically associated with Extraverted (E) and Feeling (F) types. Conversely, the Resilient Defense quadrant corresponds to Sensing (S) individuals who tend toward systematic processing and maintain a high level of alertness toward anomalous information.

Analysis of predictive significance demonstrates that Extraversion is the most influential determinant of phishing vulnerability. The decision tree utilizes Extraversion as its primary splitting factor to account for trust-driven cognitive biases in threat assessment. Detection performance is subsequently governed by the Sensing-iNtuitive dimension, while its predictive efficacy is moderated by Feeling (F) characteristics, which frequently fluctuate for an individual's vigilance during threat evaluation. In this study, feature importance is defined as the weighted sum of the total reduction in Gini Impurity generated during the node-splitting process. Features with negligible correlations to MBTI dimensions, such as Neuroticism, are filtered via significance thresholds to optimize predictive performance. The decision tree flowchart generated by this research is shown in Figure 2, with detailed experimental parameters, which are discussed in Session 3.

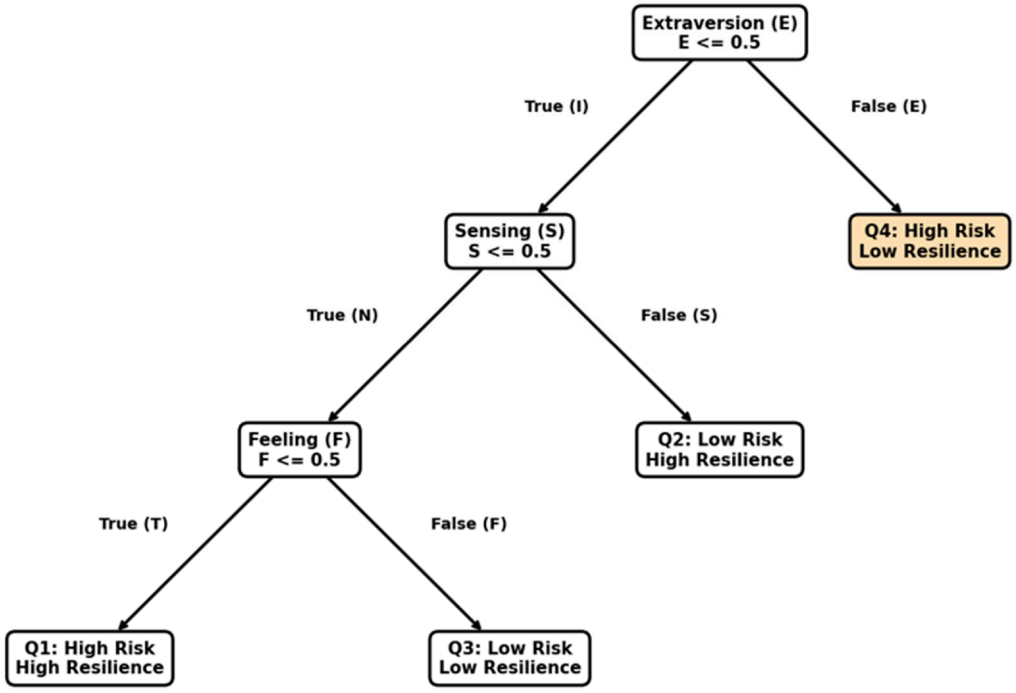


Figure 2. The decision tree flowchart.

3. Results

3.1. Experimental Environment

The experimental workflow was designed to handle large-scale psychometric data. The experimental procedures (encompassing data preprocessing, MBTI binary encoding, and the construction of the decision tree model) were implemented using the Python programming language. The primary execution environment was Google Colaboratory, a cloud-based platform that provides the necessary computational resources for handling extensive datasets without local hardware constraints.

Most existing empirical MBTI datasets exhibit inherent skewness or imbalances in personality type distributions. Such imbalances can introduce significant bias when validating theoretical mappings established in the literature. To meet the rigorous requirements of large-scale data analysis and to demonstrate the scalability of our model, we selected the "Personality Classification Data: 16 Personalities" synthetic dataset due to its substantial volume (59,999 samples) and relatively balanced distribution of personality types [25] as Dataset 1 (DS1). This dataset provides a comprehensive repository of 59,999 individual responses to an MBTI-based assessment, offering a diverse and representative sample for analyzing personality-driven behaviors. To ensure model evaluation, the dataset was split into a training set, validation set, and test set according to a ratio of 8:1:1. The large-scale nature of this dataset ensures that the resulting risk patterns and the structural differences in susceptibility identified in the C.O.R.E. Quadrant are statistically significant.

The dataset effectively translates original items from the '16 Personalities Test' into numerical features using a 7-point Likert scale. By encoding responses within a range from 'Fully Agree (+3)' to 'Fully Disagree (−3),' the data preserves the granular intensity of the original psychometric instrument. Although the data is synthetic, its core logic involves meticulous feature scaling and calibration tailored to specific personality dimensions, such as Intuition, Judging, and Thinking. To verify its empirical utility, the dataset underwent iterative training and hyperparameter optimization via an Automated Machine Learning (AutoML) framework. Consequently, the feature of engineering ensures that interdimensional statistical correlations remain intact, establishing this dataset

as a reliable statistical foundation for evaluating complex personality classification and cybersecurity susceptibility.

The primary role of this synthetic dataset in our study is to serve as a theoretical baseline. It was specifically employed to validate the projection of MBTI types onto the Five-Factor Model (FFM) based on foundational research, rather than to directly validate the final C.O.R.E. quadrants. This baseline provides a standardized reference that is essential for future researchers when assessing the model's generalizability across diverse real-world datasets that may exhibit varying degrees of environmental noise.

3.2. Heatmap of MBTI-FFM Mapping

To visualize the magnitude of each MBTI dimension's contribution to FFM traits, a heatmap was generated using Python based on Equation (2). In Figure 3, the color gradient depicts positive and negative correlations, illustrating the linear correspondences between MBTI and FFM traits. As indicated by the correlation heatmaps, notable differences were observed in the relationships between the primary MBTI dimensions (E-I, S-N, T-F, J-P) and the FFM traits. Extraversion, Openness, Agreeableness, and Conscientiousness showed moderate to strong correlations with at least one MBTI dimension, reflecting stable cross-theoretical mappings. By contrast, Neuroticism displayed consistently low correlations across all MBTI dimensions, suggesting a weak linear association with the MBTI framework.

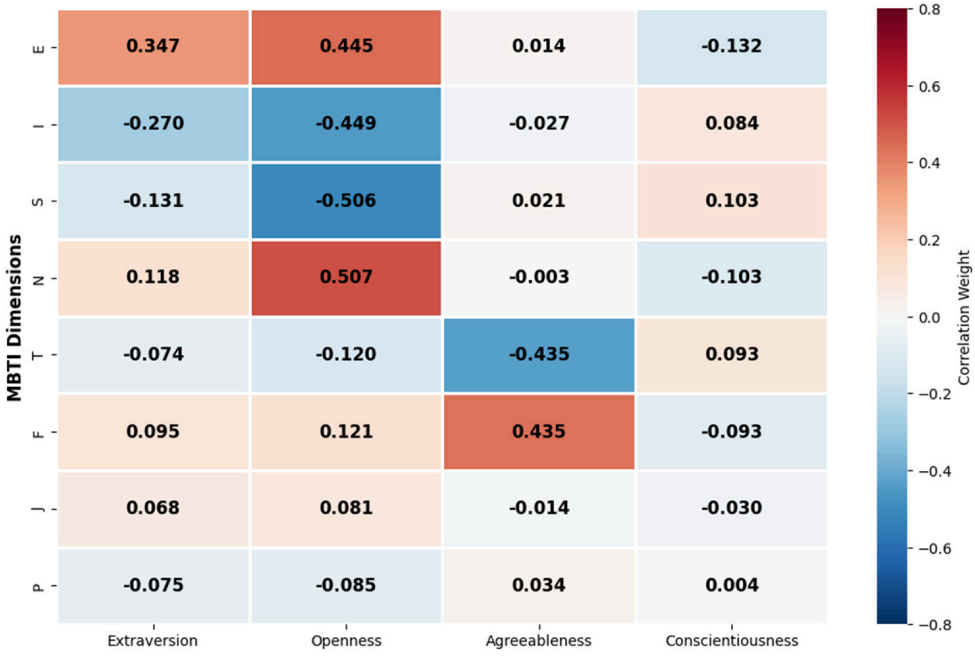


Figure 3. Heatmap of the mapping between MBTI dimensions and the four retained FFM traits.

Considering theoretical consistency and statistical stability, Neuroticism was excluded from feature engineering and subsequent machine learning analyses to enhance predictive performance. Accordingly, only Extraversion, Openness, Agreeableness, and Conscientiousness were retained as mapping to four MBTI dimensions. This selection ensures that the input variables exhibit sufficient psychological validity and statistical relevance. The results demonstrate strong theoretical coherence and meaningful correspondences between MBTI dimensions and the four retained FFM traits. For instance, Extraversion in MBTI shows a moderate positive correlation with FFM's Extraversion, Intuition exhibits a stable and significant positive association with Openness, and the Feeling dimension displays a moderate positive relationship with Agreeableness. In contrast, the Thinking dimension is negatively correlated with Agreeableness.

Notably, among the four FFM traits, Conscientiousness exhibits generally weak correlations with most MBTI dimensions, with only marginal positive associations for specific dimensions such as Judging. This pattern can be explained from two perspectives. First, The J-P dimension lacks direct conceptual equivalence to Conscientiousness, as it prioritizes lifestyle structure (flexibility vs. order) over behavioral discipline. Then, the MBTI's emphasis on cognitive styles, rather than the responsibility-driven attributes of the FFM, constrains its predictive reliability for this specific personality dimension.

3.3. C.O.R.E. Quadrant of MBTI Types to Phishing Susceptibility

In Figure 2, a decision tree was conducted to examine the effects of MBTI dimensions on phishing susceptibility. The tree is rooted at Extraversion, with the left branch further split by Sensing and subsequently by the Feeling dimension, leading to being classified as low susceptibility. This decision path indicates that individuals with lower Extraversion and specific perceptual tendency are consistently categorized into the low-risk groups. This pattern indicates that individuals with higher levels of Extraversion, when combined with specific perceptual and decision-making styles, are more likely to fall into moderate-to-high phishing susceptibility groups. Table 1 lists the hyperparameters configured for the generation of the decision tree.

Table 1. Decision tree hyperparameter table.

Parameter	Value	Description
criterion	gini	Function to measure split quality (gini = Gini impurity).
max_depth	3	Maximum depth of the tree (limits overfitting).
min_samples_split	500	Minimum samples required to split an internal node.
min_samples_leaf	1000	Minimum samples required to be at a leaf node.
max_features	0.3	Number of features to consider for best split.
random_state	42	Seed for random number generator (reproducibility).
max_leaf_nodes	8	Maximum number of leaf nodes (None = no limit).
min_impurity_decrease	0.001	Minimum impurity decrease required for split.
ccp_alpha	0.05	Cost-complexity pruning parameter (larger = more pruning).

Specifically, the interaction between Extraversion and Sensing, with a third interaction involving Feeling, can effectively differentiate susceptibility groups. This finding implies that susceptibility emerges from specific combinations of personality traits. Prior studies also indicate that personality traits modulate individuals' sensitivity to persuasive cues and risk detection capabilities [26,27]. This study first compared the 16 MBTI personality types in terms of heuristic and systematic information-processing tendencies.

To enhance interpretability, this study constructed a four-quadrant risk Quadrant, termed the Cognitive-Oriented Risk Evaluation (C.O.R.E.) Quadrant (Figure 4). The quadrant utilizes dual-processing theory to identify cognitive styles that predict phishing susceptibility. These styles govern how individuals process information and make decisions when encountering phishing messages, thereby serving as critical indicators of vulnerability. The horizontal and vertical axes represent phishing risk and defense resilience, respectively. Positions closer to the lower-right quadrant indicate greater reliance on rapid, intuitive decision-making, serving as potential phishing susceptibility. This visualization integrates personality traits and dual-processing theory, providing an empirical basis for personalized cybersecurity education and risk prediction models.

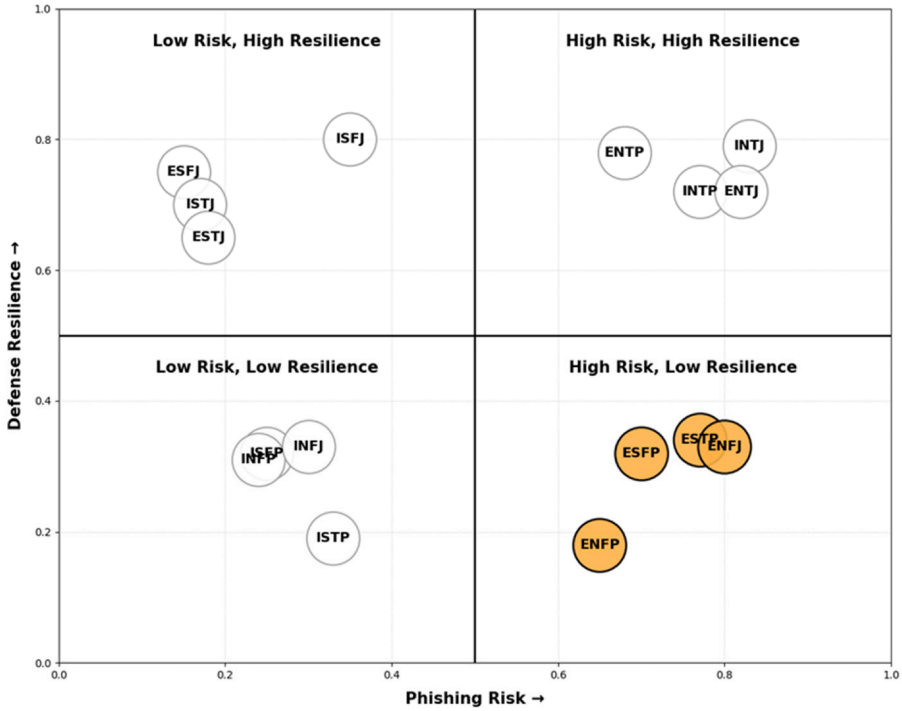


Figure 4. The Cognitive-Oriented Risk Evaluation (C.O.R.E.) Quadrant utilizes a dual-axis framework where the *x*-axis quantifies phishing risk, and the *y*-axis represents defense resilience. The arrows in the figure represent stronger phishing risk and defense resilience. Individuals categorized into the fourth quadrant exhibit the most significant phishing risk among all MBTI types. Notably, the colored data points in the fourth quadrant represent the personality types with the highest risk profiles (ENFP, ESFP, ESTP, ENFJ).

3.4. Evaluation and Statistical Validation

To validate the statistical reliability of the MBTI-FFM mapping model proposed by Furnham et al. [17] within the large-scale dataset used in this study, reliability analyses were conducted using Python. Detailed outcomes are summarized in Tables 2–5.

Table 2. Split-half reliability of personality mapping model.

Personality Traits	Split-Half Reliability	Interpretation
Extraversion	0.041	Low
Openness	0.011	Low
Agreeableness	−0.012	Negative
Conscientiousness	0.029	Low

Table 3. Construct validity of personality mapping odel.

Dimensions	Extraversion	Openness	Agreeableness	Conscientiousness
E	0.880	0.647	0.048	−0.611
I	−0.880	−0.647	−0.048	0.611
S	−0.355	−0.733	0.027	0.583
N	0.355	0.733	−0.027	−0.583
T	−0.242	−0.175	−0.997	0.527
F	0.242	0.175	0.997	−0.527
J	0.203	0.119	−0.055	−0.095
P	−0.203	−0.119	0.055	0.095

Table 4. Primary feature contributions of personality mapping model.

Personality Traits	Primary Feature Contributions	Interpretation
Extraversion	18.507	Important
Openness	18.875	Important
Agreeableness	17.585	Important
Conscientiousness	19.920	Important

Table 5. Factorial validity of personality mapping model.

Analysis Type Correlation Coefficient(r)	Interpretation
Theory-Empirical Consistency	0.911 Very High Consistency

The proposed research framework and the integrated mapping model exhibit meaningful alignment with theoretical expectations. Regarding the reliability indices, although the split-half reliability scores for traits such as Conscientiousness (0.029) and Openness (0.011) appear lower than conventional psychometric thresholds, these outcomes are not uncommon when dealing with the aggregation of diverse datasets and the conversion of typological variables into linear models.

Specifically, the construct validity analysis reveals an alignment between MBTI dimensions and their FFM counterparts, as evidenced by significant loading patterns in the factor analysis. The unstandardized primary feature contributions (or feature weights) range from 17.585 for Agreeableness to 19.920 for Conscientiousness, indicating a substantial representation of the underlying personality constructs within the large-scale dataset. Although these values in Table 4 exceed 1.0, this mathematical outcome stems from the fact that the scikit-learn module computes unstandardized loadings by default. Coupled with the highly structured multicollinearity resulting from the linear mapping of MBTI dimensions through a fixed weight matrix, the underlying eigenvalues are naturally scaled upward. Consequently, these results should be conceptualized as indicators of eigenvalue-based structural importance rather than traditional standardized correlations. Rather than a statistical anomaly, this expansion underscores that the framework possesses an exceptionally high degree of construct validity. A critical highlight of the model is its theoretical consistency, which yielded a correlation coefficient (r) of 0.911. This high level of consistency indicates that the transformation from MBTI typological data to FFM continuous vectors aligns well with the psychological constructs established in the previous literature.

Beyond statistical validation, model efficacy was measured using key performance indicators, confusion matrices, and ROC-AUC analysis, with corresponding results presented in Figures 5 and 6 and Table 6.

Table 6. Model evaluation metrics of DS1 [25].

Set	Accuracy	Precision	Recall	F1-Score
Training	0.9425	0.9416	0.9433	0.9425
Validation	0.9407	0.9388	0.9426	0.9407
Test	0.9395	0.9419	0.9366	0.9392

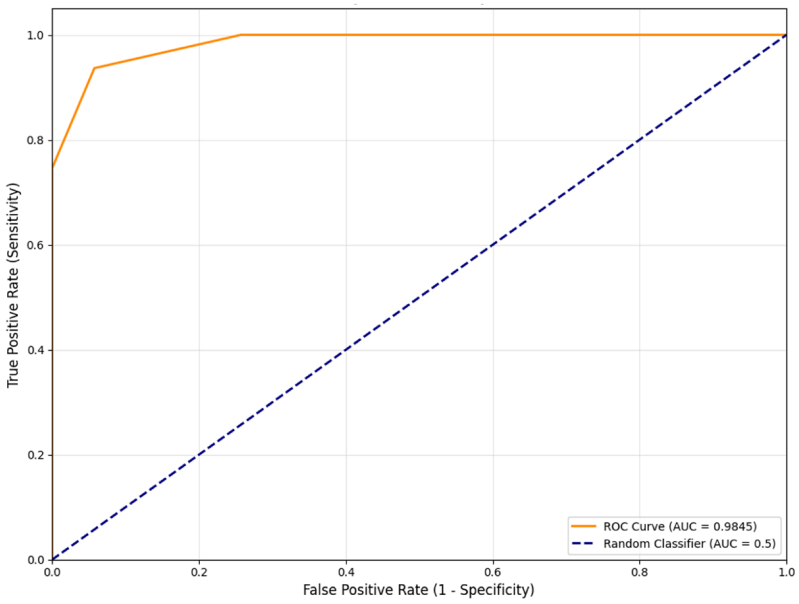


Figure 5. The ROC-AUC curve of DS1 [25].

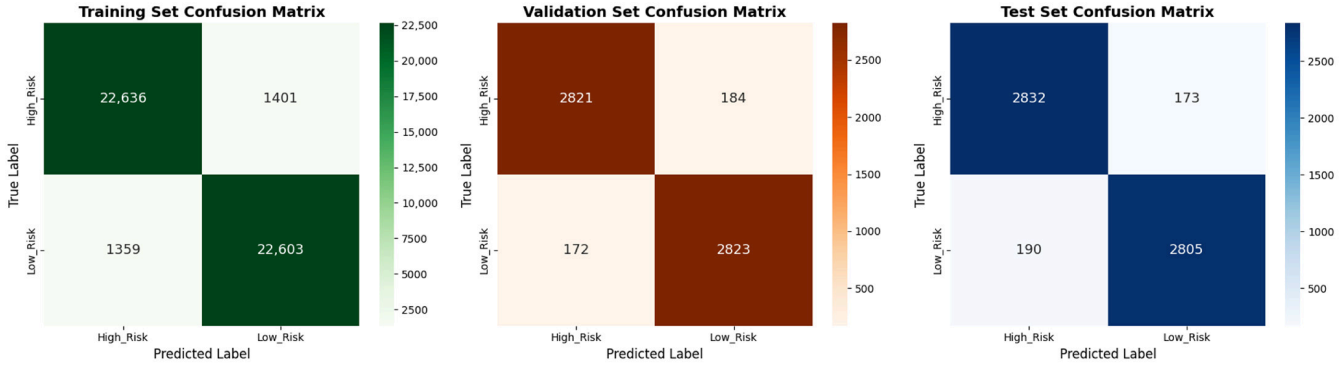


Figure 6. Confusion matrices of training, validation, and test sets of DS1.

The binary phishing risk classification model demonstrates exceptional predictive accuracy, as evidenced by the Receiver Operating Characteristic (ROC) curve reaching an Area Under the Curve (AUC) of 0.9845. As illustrated in Figure 5, the ROC curve (represented by the orange line) exhibits a steep ascent toward the upper-left coordinate, indicating a high true positive rate (sensitivity) maintained at a very low false positive rate (1-specificity).

As illustrated in Figure 6, the diagonal elements representing true positive and true negative predictions dominate each matrix, indicating a minimal occurrence of Type I (false positive) and Type II (false negative) errors. Specifically, the test set confusion matrix (right) reflects the model's robustness, achieving a balanced distribution of correct classifications with 2832 High-Risk and 2805 Low-Risk samples accurately identified. The marginal variance in error rates across the three sets suggests that the model is effectively regularized and avoids overfitting. These results confirm that the personality-driven features integrated into the decision tree provide stable and discriminative signals for binary phishing risk assessment.

As evidenced by consistent accuracy and F1-scores exceeding 0.93 across the training, validation, and test partitions summarized in Table 6, the model achieved a high Accuracy of 0.9425 and an F1-Score of 0.9425 on the training set. Critically, the performance remained stable upon transition to the validation and test sets, yielding accuracies of 0.9407 and 0.9395, respectively. The marginal discrepancy (less than 1%) between the training and test metrics

indicates that the model is effectively regularized and avoids overfitting. Furthermore, the balanced Precision (0.9419) and Recall (0.9366) observed in the test set confirm the model's reliability in accurately identifying phishing susceptibility while maintaining a low false positive rate. These results validate that the integrated personality features provide highly discriminative and stable signals for risk classification.

3.5. Robustness and Generalizability of the C.O.R.E. Framework

To further evaluate the discriminative power of the binary phishing risk classifier, Figure 7 illustrates the Receiver Operating Characteristic (ROC) curve alongside its corresponding Area Under the Curve (AUC) value. The model achieves an exceptional ROC-AUC score of 0.9486. It proves that transforming discrete typological data into integrated continuous feature dimensions provides stable, structured, and reproducible signals for behavioral risk classification.

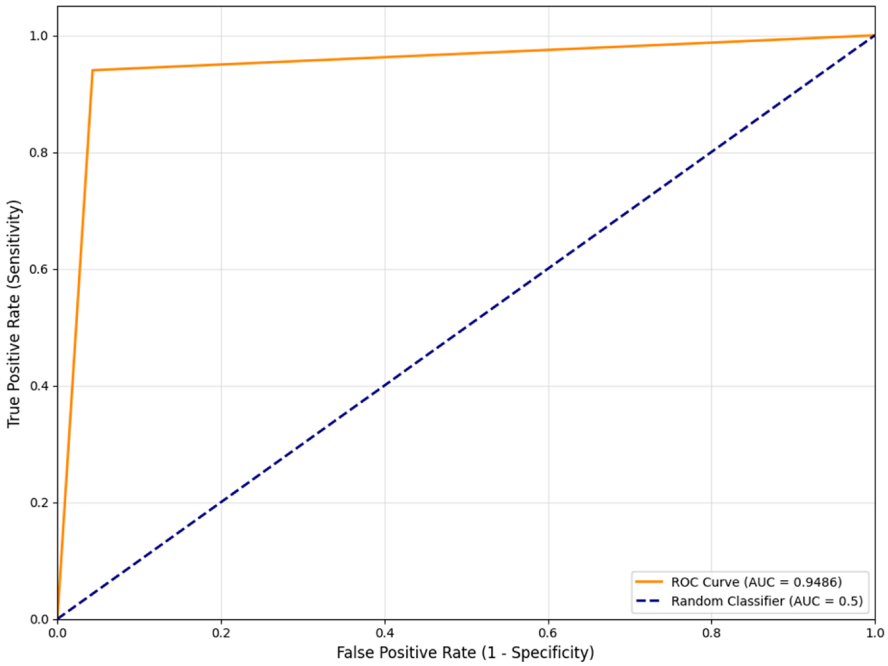


Figure 7. The ROC-AUC curve of DS2 [28].

To validate the proposed C.O.R.E. framework within empirical datasets, this study utilized a real-world dataset from Kaggle [28], which comprises over 7800 individual records sourced from Twitter via the Twitter API. To enhance labeling accuracy as Dataset 2 (DS2). A heuristic filtering strategy was applied, targeting explicit self-disclosures such as "I am an [MBTI Type]," "My MBTI is...," and "My personality type is...," thereby identifying users who explicitly reported their MBTI profiles. To ensure linguistic richness and data quality, the inclusion criteria were restricted to users who contributed a minimum of 200 words across their posts, facilitating the analysis of their communication patterns and behavioral tendencies. Figure 8 depicts the distribution of data within the C.O.R.E. Quadrant after applying the empirical dataset. While this empirical dataset provides a diverse representation of personality types within a naturalistic social media environment, we acknowledge the presence of inherent methodological constraints. Specifically, reliance on self-reported personality types and predefined search tags may introduce sampling biases, favoring individuals who are more actively engaged in personality-related discourse. Furthermore, all data collection and utilization procedures were conducted in strict adherence to established ethical standards governing social media research and public data privacy.

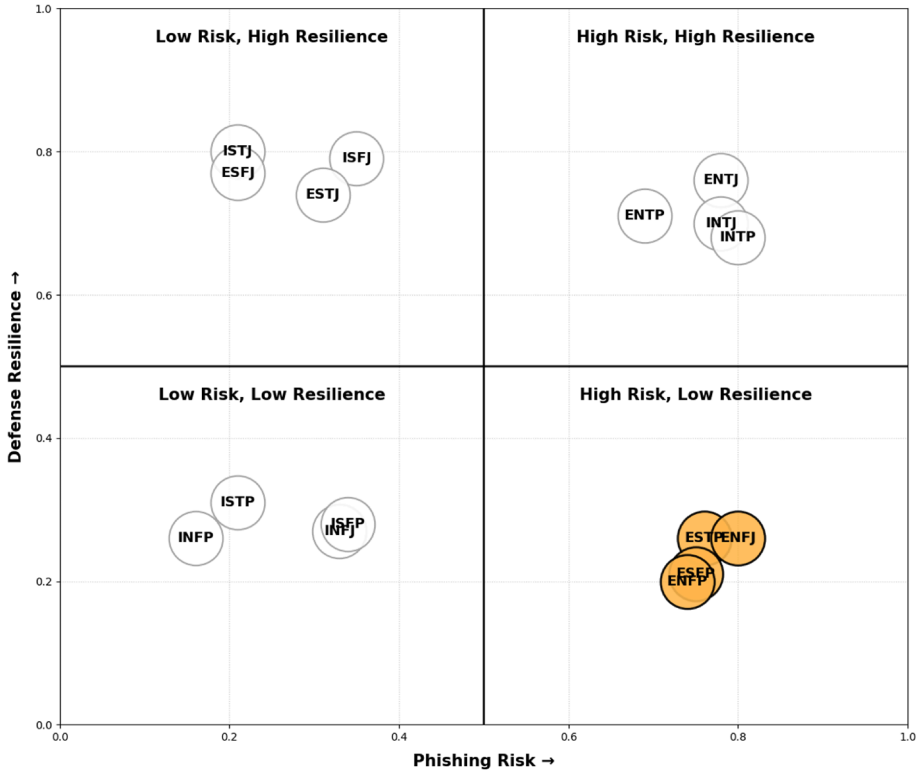


Figure 8. The Cognitive-Oriented Risk Evaluation (C.O.R.E.) Quadrant of MBTI Personality Type Twitter Dataset [28]. The arrows in the figure represent stronger phishing risk and defense resilience.

The reduced sample size results in higher sampling variance, increased sensitivity to outliers, and potential class imbalance across MBTI types. Collectively, these factors contribute to unstable estimations of both phishing risk and defense resilience, leading to observable local shifts in the positional distribution of personality types within the quadrant space. In addition, the presence of a non-normal (skewed) distribution constitutes a critical source of variation. When the data exhibits either positive or negative skewness, the meaning no longer serves as a reliable measure of central tendency, which may systematically bias the estimated scores. Consequently, certain MBTI types may be overestimated or underestimated along the risk and resilience dimensions.

This dataset was deliberately selected due to its distributional bias and small-scale characteristics. In sharp contrast to the original large-scale synthetic dataset, it serves as an ideal test case for validating the model's generalizability. The results demonstrate that, within the small empirical dataset, MBTI profiles exhibit pronounced positional shifts and increased dispersion across quadrants, whereas the synthetic dataset yields a more concentrated and structurally stable distribution. Nevertheless, despite these coordinate fluctuations, all personality types remain within the same risk quadrant, indicating that the quadrant-level classification remains invariant.

By integrating real-world behavioral data with large-scale synthetic benchmarks, this study aims to bridge the gap between theoretical personality mapping and the practical complexities of cybersecurity susceptibility in the digital age. Overall, although the specific coordinates of MBTI types vary across datasets due to sampling variance, distributional skewness, and real-world noise, the quadrant-level classification results remain substantially consistent.

The empirical validation and classification performance of the developed decision framework on the empirical dataset (DS2) are summarized in Table 7 and Figure 7, providing evidence of the model's predictive capability and generalization across real-world

behavioral samples. The model achieves an Accuracy of 0.9491 on the training set, which remains remarkably stable upon transition to the validation set (0.9424) and the independent test set (0.9501). The marginal variance across these partitions indicates that the classifier is effectively regularized and avoids overfitting artifacts.

Table 7. Model evaluation metrics of DS2 [28].

Set	Accuracy	Precision	Recall	F1-Score
Training	0.9491	0.9527	0.9205	0.9363
Validation	0.9424	0.9306	0.9277	0.9291
Test	0.9501	0.9373	0.9403	0.9388

4. Applications and Limitations

4.1. Applications

4.1.1. Handling Internal Threats

Organizational cybersecurity culture can be strengthened through risk-stratified educational interventions that enable HR and IT departments to deploy differentiated and resource-efficient training strategies. The C.O.R.E. Quadrant enables a strategic transition toward tailored cybersecurity education by mapping individual cognitive vulnerabilities to specific training workflows. This framework ensures that high-risk individuals receive prioritized vigilance training, thereby strengthening the reinforcement of secure behaviors against phishing attacks triggers. Similarly, individuals identified as low risk may be cultivated as "Security Defenders," receiving advanced instruction to enhance their ability to detect sophisticated anomalies and high-fidelity deceptive cues that often evade automated technical controls. By allocating intensive training resources to high-vulnerability groups while strategically empowering low-risk cohorts, organizations can establish a proactive and adaptive defense architecture grounded in data-driven risk differentiation.

4.1.2. Designing Cybersecurity Training

The results of this study indicate that personality traits and information-processing preferences are critical determinants of phishing susceptibility. Therefore, the C.O.R.E. Quadrant can serve as a foundational tool for establishing personalized educational intervention strategies, moving away from the limitations of traditional, uniform training materials. The C.O.R.E. Quadrant facilitates a proactive security culture by aligning cybersecurity training with individual cognitive profiles derived from an integrated personality framework. This enables organizations to optimize resource efficiency, prioritizing high-risk groups while maintaining a resilient defense architecture among low-vulnerability employees. These applications improve organizational cybersecurity management, personalized employee training, and design intelligent security platforms, aiming for the transition of research findings from theoretical validation to actionable operational workflows.

4.2. Limitations

While the C.O.R.E. framework demonstrates strong methodological consistency, its predictive outputs must be interpreted with caution, as they are inherently constrained by the characteristics of the evaluation datasets. The near-perfect performance observed in the synthetic Dataset 1 (DS1) reflects its controlled, noise-reduced design, which supports theoretical validation but does not represent real-world predictive capability. Instead, these results indicate structured behavioral tendencies rather than operational effectiveness in dynamic attack scenarios.

Dataset 2 (DS2), derived from Twitter, introduces more realistic yet challenging conditions. Its reliance on unverified self-reported MBTI profiles introduces self-presentation

bias and label noise, while natural data distributions further affect precision by causing positional variability within the C.O.R.E. Quadrant. Although quadrant-level classifications remain relatively stable, fine-grained spatial accuracy is sensitive to data quality and environmental noise.

To enhance validity, future work should prioritize primary data collection, including standardized psychometric assessments, behavioral experiments, and real-world security training data. The C.O.R.E. framework should be applied alongside externally validated behavioral measures to ensure robust and contextually grounded risk evaluation.

5. Conclusions

The proposed model serves as a core tool for employee cybersecurity training, enabling differentiated and personalized educational interventions. In our research, E-S and E-F types, like employees whose MBTI types are ENFP, ESFP, ESTP and ENFJ, should receive training beyond basic cybersecurity concepts, emphasizing advanced attack techniques and response strategies. Training design may incorporate scenario simulations of complex attacks and social engineering, such as spear phishing, deepfake-based deception, cross-platform social manipulation, and Business Email Compromise (BEC). Overall, the model overcomes the "one-size-fits-all" limitation of conventional corporate cybersecurity training by providing a framework for personalized interventions. Cybersecurity resilience is strengthened when organizations utilize the C.O.R.E. Quadrant to calibrate training content against employee risk profiles, ensuring that defenses are adaptive to individual cognitive vulnerabilities.

Although this study successfully constructed a highly effective predictive model using a large-scale synthetic dataset (N = 59,999), achieving strong classification performance (AUC = 0.9845), potential discrepancies between synthetic data and real-world conditions must be interpreted with caution. Despite the use of Likert-scale encoding and rigorous feature standardization to enhance realism, the inherently simulated nature of the dataset may not fully capture the psychological dynamics exhibited by individuals under high-pressure, real-world phishing attack scenarios. Accordingly, the risk assessment outcomes produced by the C.O.R.E. Quadrant should be interpreted as indicators of behavioral intent rather than definitive psychological diagnoses.

Future research will focus on empirical validation and dynamic deployment of the proposed framework to further strengthen human-centered cybersecurity defenses. First, through industry–academia collaboration, longitudinal data collected from real-world operational environments will be leveraged to evaluate the external validity of the cross-model transformation mechanisms, ensuring that the weighted integration of FFM and MBTI dimensions accurately reflects observable behavioral patterns. Second, in response to the growing threat of deepfakes and generative-AI-driven social engineering attacks, the C.O.R.E. Quadrant will be extended into an adaptive intelligent training system. By integrating real-time cognitive load indicators with personality traits, the framework aims to enable dynamic intervention, evolving from static risk assessment toward a proactive and responsive defensive architecture.

Author Contributions: Conceptualization, C.-H.K., C.-W.T. and C.-L.C.; methodology, Y.-T.K.; software, Y.-T.K.; validation, C.-H.K., C.-W.T. and C.-L.C.; formal analysis, C.-L.C.; writing—original draft preparation, Y.-T.K.; writing—review and editing, C.-H.K., C.-W.T. and C.-L.C.; visualization, Y.-T.K.; supervision, C.-H.K.; project administration, C.-L.C.; funding acquisition, C.-W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council, ROC, under Project No. NSTC-114-2221-E-025-006-MY2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Kaggle, reference number [25,28]. These data were derived from the following resources available in the public domain: [25] https://www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt (accessed on 20 March 2026) and [28] https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset (accessed on 20 March 2026). According to Kaggle's Terms of Service, publicly shared user submissions, including datasets, may be accessed and used by Kaggle users as permitted by the platform's functionality and applicable terms. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used Gemini 3.5 Flash for the purposes of generating the conceptual framework diagram (Figure 1) and grammar checking and polishing. The prompts used were 'Create a flowchart representing the C.O.R.E. framework and generate the relevant flowchart elements for the C.O.R.E. framework and connect them sequentially.' The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lastdrager, E.E. Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Sci.* **2014**, *3*, 9. [CrossRef]
2. Islam, A.; Rashid, M.; Othman, F.; Kaosar, M.G.; Islam, L. Identifying personality traits associated with phishing susceptibility. *Secur. J.* **2025**, *38*, 18. [CrossRef]
3. Verizon. 2019 Data Breach Investigations Report (DBIR). 2019. Available online: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf (accessed on 20 March 2026).
4. Modic, D.; Lea, S.E.G. How neurotic are scam victims, really? The Big Five and Internet scams. In Proceedings of the International Conference of the Confederation for the Advancement of Behavioral Economics and Economic Psychology, Exeter, UK, 13–16 July 2011.
5. Griffin, R.J.; Neuwirth, K.; Giese, J.; Dunwoody, S. Linking the heuristic-systematic model and depth of processing. *Commun. Res.* **2002**, *29*, 705–732. [CrossRef]
6. Urieta, P.; Aluja, A.; Garcia, L.F.; Balada, F.; Lacomba, E. Decision-making and the alternative five factor personality model: Exploring the role of personality traits, age, sex and social position. *Front. Psychol.* **2021**, *12*, 717705. [CrossRef] [PubMed]
7. Aluja, A.; Blanch, A. Neuropsychological behavioral inhibition system (BIS) and behavioral approach system (BAS) assessment: A shortened sensitivity to punishment and sensitivity to reward questionnaire version (SPSRQ–20). *J. Pers. Assess.* **2011**, *93*, 628–636. [CrossRef] [PubMed]
8. Toledo, F.; Carson, F. Neurocircuitry of personality traits and intent in decision-making. *Behav. Sci.* **2023**, *13*, 351. [CrossRef] [PubMed]
9. Butavicius, M.; Parsons, K.; Pattinson, M.; McCormac, A. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. In Proceedings of the Australasian Conference on Information Systems (ACIS), Adelaide, Australia, 30 November–4 December 2015.
10. Albladi, S.M.; Weir, G.R.S. User characteristics that influence judgment of social engineering attacks in social networks. *Hum. Centric Comput. Inf. Sci.* **2018**, *8*, 5. [CrossRef]
11. Lawson, P.; Zielinska, O.; Pearson, C.; Mayhorn, C.B. Interaction of personality and persuasion tactics in email phishing attacks. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications: Los Angeles, CA, USA, 2017; Volume 61, pp. 1331–1333. [CrossRef]
12. Oyibo, K.; Orji, R.; Vassileva, J. Investigation of the influence of personality traits on Cialdini's persuasive strategies. In Proceedings of the Personalization in Persuasive Technology Workshop (PPT 2017), Amsterdam, The Netherlands, 4 April 2017.
13. Ge, Y.; Lu, L.; Cui, X.; Chen, Z.; Qu, W. How personal characteristics impact phishing susceptibility: The mediating role of mail processing. *Appl. Ergon.* **2021**, *97*, 103526. [CrossRef] [PubMed]
14. Frauenstein, E.D.; Flowerday, S.; Mishi, S.; Warkentin, M. Unraveling the behavioral influence of social media on phishing susceptibility. *Inf. Manag.* **2023**, *60*, 103858. [CrossRef]

15. López-Aguilar, A.; Urruela, C.; Batista, E.; Machin, J.; Solanas, A. Phishing vulnerability and personality traits: Insights from a systematic review. *Comput. Hum. Behav. Rep.* **2025**, *20*, 100784. [CrossRef]
16. Halevi, T.; Memon, N.; Nov, O. Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. *SSRN* **2015**. [CrossRef]
17. Furnham, A. The Big Five versus the Big Four: The relationship between the MBTI and the NEO-PI five-factor model of personality. *Pers. Individ. Dif.* **1996**, *21*, 303–307. [CrossRef]
18. MacDonald, D.; Anderson, P.E.; Tsagarakis, C.I.; Holland, C.J. Examination of the relationships between the Myers Briggs Type Indicator and the NEO Personality Inventory. *Psychol. Rep.* **1994**, *74*, 339–344. [CrossRef]
19. McCrae, R.; Costa, P. Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *J. Pers.* **1989**, *57*, 17–40. [CrossRef] [PubMed]
20. Luo, X.; Zhang, W.; Burd, S.; Seazzu, A. Investigating phishing victimization with the heuristic-systematic model: A theoretical framework and an exploration. *Comput. Secur.* **2013**, *38*, 28–38. [CrossRef]
21. Pallapu, S.; Mounika, P.; Vishanvi, A.; Rishitha, P. PhishGuard—AI Phishing Email & URL Detector. *Int. J. Data Sci. IOT Manag. Syst.* **2024**, *4*, 46–51. [CrossRef]
22. Saranya, G.; Yamini Priya, M.R. AI-Powered Phishing Detection. In Proceedings of the 2025 International Conference on Data Science, Chennai, India, 28–29 March 2025; pp. 1–6.
23. Frauenstein, E.D.; Flowerday, S. Susceptibility to phishing on social network sites: A personality information processing model. *Comput. Secur.* **2020**, *94*, 101862. [CrossRef] [PubMed]
24. Eftimie, S.; Moinescu, R.; Răcuciu, C. Spear-Phishing Susceptibility Stemming From Personality Traits. *IEEE Access* **2022**, *10*, 73548–73561. [CrossRef]
25. Mehta, A. Personality Classification Data: 16 Personalities. Kaggle. 2022. Available online: https://www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt (accessed on 20 March 2026).
26. Aleroud, A.; Zhou, L. Phishing Environments, Techniques, and Countermeasures. *Comput. Secur.* **2017**, *68*, 160–196. [CrossRef]
27. Vishwanath, A.; Harrison, B.; Ng, Y.J. Suspicion, cognition, and automaticity model of phishing susceptibility. *Commun. Res.* **2018**, *45*, 1146–1166. [CrossRef]
28. Mazlum, I. MBTI Personality Type Twitter Dataset. Kaggle. 2023. Available online: https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset (accessed on 18 April 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.