MJI
Media Jurnal Informatika
Vol. 18, No. 1, June 2026
p-issn : 2088-2114, e-issn : 2477-2542

Comparative Analysis of Black-Box and White-Box Machine Learning Model in Explainable Phishing Detection

Abdullah Fajar[a,1,*], Setiadi Yazid[b,2], Indra Budi[b,3]
[a] Universitas Telkom, Bandung, Indonesia
[b] Universitas Indonesia, Depok, Indonesia
[1] abdfajar@telkomuniversity.ac.id*; [2]setiadi@cs.ui.ac.id; [3]indra@cs.ui.ac.id
* corresponding author

ARTICLE INFO

Article history
Received 2026-06-29
Revised 2026-06-30
Accepted 2026-06-30

Keywords
Black-Box Model
White-Box Model
Phishing Detection,
Machine Learning
Comparative Analisys,
Explaiability

ABSTRACT

Explainability in phishing detection model can support a further solution of phishing attack mitigation by increasing trust and understanding how phishing can be detected. The aims of this study to determine and best recommendation to apply an approach which has several components with abilities to fulfil the critical needs A methodology starting with analyzing both black-box and white-box models to get the pros and cons specifically in phishing detection. The conclusion of the analysis will be validated by experiment using a set of well-known algorithms and public phishing datasets. Experimental metrics covers 3 measurements such as predictive accuracy and explainability metrics. Both models are comparable in terms of interpretability and consistency, with room for improvement in diverse datasets. EBM as an example of white-box model is generally better suited for applications requiring explainability and actionable insights. Finally, each model, white-box and black-box model has positive and negative aspects both for performance metric and for explainable metric. It is important to consider the objective of model usage.

1 Introduction

Phishing attacks have emerged as one of the most persistent and damaging threats in the contemporary cybersecurity landscape. These deceptive tactics, which involve fraudulent communications designed to steal sensitive information such as login credentials, financial data, and personal identifiers, have grown increasingly sophisticated in recent years [1], [2]. Attackers now employ a diverse arsenal of techniques, including email-based phishing, SMS-based phishing (smishing), social media exploitation, and voice-based vishing, to circumvent traditional security measures [3], [4]. The financial and reputational consequences of successful phishing campaigns are substantial, with organizations suffering significant monetary losses, data breaches, and erosion of customer trust [5], [6]. Consequently, the development of robust and effective phishing detection systems has become a critical priority for cybersecurity researchers and practitioners alike.

Machine learning (ML) models have emerged as a promising approach to address the phishing detection challenge, offering the ability to automatically learn complex patterns from large datasets and adapt to evolving attack strategies [7], [8]. Numerous studies have demonstrated the efficacy of various ML algorithms, including deep neural networks, random forests, gradient boosting machines, and support vector machines, in distinguishing between legitimate and malicious web content [9]–[11]. However, despite their impressive predictive accuracy, many of these models operate as "black boxes," meaning their internal decision-making processes remain opaque and difficult for humans to interpret [12], [13]. This lack of transparency poses significant challenges in high-stakes cybersecurity applications, where understanding the rationale behind a model's prediction is essential for building trust, ensuring accountability, and enabling effective incident response [14], [15].

The importance of explainability in AI-based phishing detection is underscored by several critical questions that remain inadequately addressed in the literature. First, how can detection systems not only identify phishing attempts but also provide clear, understandable explanations of the methods used by attackers [16]? Second, how can explanations be designed to enhance user comprehension and reduce the tendency to ignore security warnings due to lack of meaningful information [17]? Third, how can models automatically identify and highlight the most phishing-relevant features in suspicious data, enabling targeted mitigation strategies [18]? Fourth, how can actionable insights be derived from algorithm predictions to empower security teams and end-users to make informed decisions [19]? Finally, how can explanations be optimized to build user trust in AI-based detection systems, thereby increasing their overall effectiveness [20]?

Research on explainable phishing detection has explored various approaches, including rule-based systems [21], feature importance analysis [22], and post-hoc explanation techniques such as LIME and SHAP [23]. White-box models, such as Decision Trees and Logistic Regression, offer inherent interpretability but may sacrifice predictive performance [24]. Conversely, black-box models like XGBoost and Deep Neural Networks achieve state-of-the-art accuracy but require external explanation methods to provide insights into their decisions [25]. Studies have also investigated the design of user-centric warning dialogs and natural language explanations to improve user understanding and trust [26], [27]. However, these studies often focus on a single type of model or use qualitative assessments of explainability, lacking a systematic, empirical comparison between black-box and white-box approaches using quantitative metrics [28], [29].

While previous research has explored various aspects of explainability in phishing detection, several critical gaps remain. First, most studies focus on a single type of model or use qualitative assessments of explainability, lacking a systematic, empirical comparison between black-box and white-box approaches using quantitative metrics [30], [31]. Second, the trade-offs between predictive performance and explainability are often discussed theoretically but rarely validated through rigorous experimentation across diverse datasets [32]. Third, the evaluation of explainability metrics—such as stability, consistency, actionability, and interpretability—is frequently subjective and lacks standardized, operational frameworks [33], [34]. These gaps limit the ability of practitioners to make informed decisions about which modeling approach best suits their specific requirements for accuracy, transparency, and user trust.

To address these gaps, this study presents several novel contributions. First, it provides the first systematic empirical comparison between a state-of-the-art black-box model (XGBoost) and a white-box model (Explainable Boosting Machine) specifically for phishing detection, utilizing twelve publicly available datasets with varying characteristics in terms of instances, features, and class distributions. Second, the study introduces and applies a structured, operational scoring system for evaluating explainability metrics—including stability, consistency, accuracy of explanations, interpretability, and actionability—based on quantitative SHAP analysis and domain knowledge validation, establishing a reproducible methodology for future research. Third, unlike previous studies that focus solely on predictive accuracy, this research simultaneously evaluates both predictive performance (accuracy, precision, recall, FPR, AUC-ROC) and multiple explainability dimensions, enabling a balanced understanding of the trade-offs between performance and transparency. Fourth, the study goes beyond static evaluation by conducting robustness analysis through controlled data perturbations, providing insights into model stability and generalizability often overlooked in prior work. Finally, the research translates empirical findings into actionable, practical recommendations for deploying interpretable models in real-world cybersecurity settings, including guidelines for model selection, contextual explanations, workflow integration, and user training.

Specifically, this research seeks to answer the following questions: (1) How do black-box models (represented by XGBoost) and white-box models (represented by the Explainable Boosting Machine) compare in terms of predictive accuracy and explainability across multiple public phishing datasets? (2) What are the specific advantages and disadvantages of each model type when evaluated using a structured set of quantitative explainability metrics? (3) Which model type is better suited for applications requiring high explainability and actionable insights?

To answer these questions, we employed a rigorous experimental methodology. Twelve publicly available phishing detection datasets, varying in size and feature composition, were used to train and evaluate both XGBoost and EBM. Predictive performance was assessed using standard metrics including accuracy, precision, recall, false positive rate, and AUC-ROC. Explainability was evaluated using SHAP (SHapley Additive

exPlanations) and a structured scoring system for stability, consistency, accuracy of explanations, interpretability, and actionability. Robustness analysis was also conducted to assess model resilience under data perturbations.

The remainder of this paper is organized as follows. Section II describes the research methodology in detail, including the analytical review, experimental setup, and evaluation metrics. Section III presents the results of both the analytical review and experimental validation. Thus section discusses the implications of the findings, acknowledges limitations, and provides practical recommendations. Finally, Section IV concludes the paper and outlines directions for future research

2 Method

2.1 Type and Approach of Research

This study employs a **quantitative experimental research approach** with a comparative design. The primary objective is to evaluate and compare the predictive performance and explainability of black-box and white-box machine learning models for phishing detection. This approach was chosen for the following justifications:

Table 1 Type and Approach Justification

Aspect	Description
Objective Measurement	The research focuses on quantifiable outcomes such as predictive accuracy, precision, recall, AUC-ROC, and explainability metrics (fidelity, stability, consistency, and actionability). These require a numerical framework for analysis.
Controlled Experimentation	Datasets were partitioned into training, validation, and test sets. Models were trained under consistent conditions to ensure fairness, aligning with principles of experimental research.
Comparative Analysis	Two model types were explicitly compared: XGBoost (black-box) and Explainable Boosting Machine (white-box). This comparative design systematically evaluates strengths and weaknesses of each approach.
Reproducibility	The methodology was designed to be fully reproducible, with clear documentation of datasets, preprocessing steps, hyperparameter settings, and evaluation metrics, following best practices in machine learning and cybersecurity. The repository may be accessed upon request.
Generalizability	By using multiple publicly available datasets with varying characteristics (instances, features, class distributions), the study aims to produce findings that generalize across different phishing detection scenarios.

2.2 Object and Scope of Research

The main objective of this research is to evaluate machine learning models for phishing detection, focusing on both black-box andThe main objective of this research is to evaluate machine learning models for phishing detection, focusing on both black-box and white-box approaches. The **black-box model** selected was XGBoost (Extreme Gradient Boosting), a powerful ensemble learning algorithm widely recognized for its high predictive accuracy but limited inherent interpretability. In contrast, the **white-box model** chosen was the Explainable Boosting Machine (EBM), an interpretable machine learning framework based on generalized additive models. EBM provides inherent transparency in its decision-making process, allowing users to understand how individual features contribute to predictions. By comparing these two models, the study aims to balance predictive performance with explainability, offering insights into their respective strengths and limitations in the context of phishing detection.

Additionally, the study examines **XAI (Explainable Artificial Intelligence) techniques**, particularly SHAP (SHapley Additive exPlanations), as tools to bridge the interpretability gap for black-box models.

The scope of this research is defined by the following boundaries:

Table 2 Scope Element Boundaries

Element	Description
Domain	Cybersecurity, specifically focused on phishing attack detection using URL-based and web page-based features.
Datasets	Twelve publicly available phishing detection datasets sourced from the UCI Machine Learning Repository and other open sources. Sizes range from 1,105 to over 1.1 million instances, with 11 to 112 features.
Models	Two representative models were evaluated: XGBoost (black-box) and Explainable Boosting Machine (EBM, white-box). Other models such as Deep Neural Networks, Random Forests, Decision Trees, and Rule-Based Models are discussed in the literature review but excluded from experimental validation.
Evaluation Metrics	Predictive performance assessed using accuracy, precision, recall, false positive rate (FPR), and AUC-ROC. Explainability evaluated using fidelity, simplicity, comprehensiveness, consistency, accuracy of explanations, stability, human interpretability, and actionability.
Explanation Techniques	SHAP was used as the primary XAI technique for both models to ensure consistent comparison. Other methods such as LIME, PDP, and ICE were discussed but not implemented in experiments.
Geographical & Temporal Scope	Research conducted at Universitas Telkom, Bandung, Indonesia, between April 2024 and October 2024. Datasets used are static and reflect the state of phishing attacks at the time of their collection.

2.3 Data Collection Techniques

This study utilizes **secondary data** obtained from publicly available repositories. The datasets used in this research were not collected through primary methods (e.g., surveys, interviews, or field observations) but were sourced from established public datasets commonly used in phishing detection research.

2.3.1 Dataset Details

A total of **twelve (12) phishing detection datasets** were collected from the following sources:

Table 3 Sources of Dataset

Dataset Name	Source	Kaggle Link / Reference
ds_100K20	Kaggle / Public Repository	https://www.kaggle.com/datasets/joebeachcapital/phiusiil-phishing-url
ds_10K18	Kaggle / Public Repository	https://www.kaggle.com/datasets/hasibur013/url-data-for-phishing-website-detection
ds_10K50	Kaggle / Public Repository	https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset

ds_11055_rev	Kaggle Public Repository	/ https://www.kaggle.com/datasets/ravirajkukade/phishingdomaindetection
ds_11055	Kaggle Public Repository	/ https://www.kaggle.com/datasets/manishkc06/web-page-phishing-detection/data
ds_11K89	Kaggle Public Repository	/ https://www.kaggle.com/datasets/akashkr/phishing-website-dataset?resource=download
ds_129K112	Kaggle Public Repository	/ https://www.kaggle.com/datasets/rashazieni/zieni-dataset
ds_235795_54	Kaggle Public Repository	/ https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning
ds_247950	Kaggle Public Repository	/ https://www.kaggle.com/datasets/simaanjali/phising-detection-dataset/code
ds_600K11	Kaggle Public Repository	/ https://www.kaggle.com/datasets/michellevp/dataset-phishing-domain-detection-cybersecurit
ds_88K112	Kaggle Public Repository	/ https://www.kaggle.com/datasets/akashkr/phishing-website-dataset?resource=download
ds_90K32	Kaggle Public Repository	/ https://www.kaggle.com/datasets/michellevp/dataset-phishing-domain-detection-cybersecurit

2.3.2 Data Collection Procedure

The data collection procedure in this study followed a structured sequence to ensure reliability and consistency. First, **relevant datasets** were identified through a systematic search of academic sources using keywords such as "phishing detection," "URL phishing," "phishing website," and "machine learning phishing." Next, datasets were selected based on specific **criteria**: they had to be publicly available for research purposes, contain both phishing and legitimate instances for binary classification, include at least 1,000 samples, and provide URL-based or web page-based features commonly used in phishing detection. Once selected, each dataset was downloaded and documented, including details such as its source, number of instances, number of features, feature descriptions, and class distribution. Finally, the datasets underwent **preprocessing** to ensure consistency, which involved handling missing values through removal or imputation, encoding categorical features where applicable, normalizing or standardizing numerical features, and applying stratified partitioning into training (70%), validation (15%), and test (15%) sets. This systematic approach ensured that the datasets were both representative and suitable for robust model development and evaluation. Here is the summary of preprocessing steps described in table 4 below:

Step	Process	Key Methods	Example
1	Handling Missing Values	Mean/Median/Mode Imputation, Dropping Columns	Drop column with 80% missing; impute column with 5% missing using mean
2	Detecting and Handling Outliers	IQR, Z-score, Capping/Winsorization	Cap extreme values in ttl_hostname

3	Feature Standardization/Normalization	Min-Max Scaling, StandardScaler	Scale qty_redirects to [0, 1] range
4	Handling Non-Informative Features	Variance Threshold, Manual Removal	Drop url_shortened with 99% "0" values
5	Encoding Target Labels	Label Encoding, Binary Conversion	Phishing label already in binary format
6	Feature Selection	Pearson Correlation, Mutual Information, RFE	Retain qty_dot_url due to high correlation with target
7	Data Splitting	Train-Test Split with Stratification	70%), validation (15%), and test (15%) sets with stratify=phishing
8	Addressing Class Imbalance	SMOTE, Undersampling	Apply SMOTE to augment minority class

2.3.3 Population and Sampling

Since this study relies on secondary datasets, traditional population and sampling techniques are applied during the data partitioning phase. In this context, the **population** refers to all instances within each dataset, which collectively represent the complete population for that specific dataset. To maintain the integrity of the data, **stratified sampling** was employed, ensuring that the class distribution between phishing and legitimate cases is preserved across the training, validation, and test splits. This approach guarantees that the model is trained and evaluated on balanced subsets, thereby improving reliability and representativeness of the results.

2.4 Tools and Materials Used

2.4.1 Software and Programming Tools

The following software tools and programming libraries were used in this study:

Table 4 Software and Programming Tools

Tool / Library	Version	Description	Source
Python	3.9+	Programming language used for data processing, modeling, and analysis	https://python.org
Pandas	1.5.0+	Data manipulation and analysis library	https://pandas.pydata.org
NumPy	1.24.0+	Numerical computing library	https://numpy.org
Scikit-learn	1.2.0+	Machine learning library for preprocessing, evaluation, and metrics	https://scikit-learn.org
XGBoost	1.7.0+	Extreme Gradient Boosting library (black-box model)	https://xgboost.ai
InterpretML	0.3.0+	Interpretable machine learning library (EBM implementation)	https://interpret.ml
SHAP	0.41.0+	SHapley Additive exPlanations library (XAI technique)	https://github.com/slundberg/shap
Matplotlib	3.6.0+	Data visualization library	https://matplotlib.org

Seaborn	0.12.0+	Statistical data visualization library	https://seaborn.pydata.org
Jupyter Notebook	6.5.0+	Interactive development environment for code and documentation	https://jupyter.org

2.4.2 Hardware Specifications

Table 5 Hardware Spesification

Component	Specification
Processor	Intel Core i7-11700K @ 3.60 GHz (8 cores, 16 threads)
RAM	64 GB DDR4
Storage	1 TB NVMe SSD
GPU	NVIDIA GeForce RTX 3060 (12 GB VRAM) – used for optional acceleration
Operating System	Ubuntu 22.04 LTS / Windows 11 Pro

In this study, several libraries were employed to support **explainability** in AI-based phishing detection models. The **SHAP** (SHapley Additive exPlanations) library was used to compute feature importance and generate explanations for both XGBoost and EBM predictions, providing transparency into how individual features contribute to model outcomes. Additionally, **Partial Dependence Plots (PDP)** were utilized to visualize the relationship between specific features and overall model predictions, offering insights into global feature effects. Complementing this, **Individual Conditional Expectation (ICE)** plots were applied to illustrate how predictions for individual instances change as feature values vary, thereby highlighting localized behavior within the model. Together, these tools enhance interpretability, enabling both researchers and end-users to better understand and trust the system's decision-making process.

2.5 Research Procedures or Stages

This research was conducted in seven systematic stages, as illustrated in Figure 1, below each designed to ensure methodological rigor and comprehensive evaluation.

In **Stage 1: Problem Identification and Literature Review**, the objective was to identify the research gap and formulate research questions. This involved reviewing existing literature on phishing detection, machine learning models, and explainable AI (XAI), recognizing the need for a systematic comparison between black-box and white-box models, and defining the scope of the study. The outcome was a clear research framework supported by well-defined questions and relevant references.

Stage 2: Dataset Collection and Preprocessing focused on acquiring and preparing datasets for experimentation. Twelve public phishing detection datasets were collected from repositories such as UCI, cleaned to handle missing values and duplicates, encoded for categorical features, normalized for numerical features, and partitioned using stratified sampling into training (70%), validation (15%), and test (15%) sets. This ensured balanced and representative datasets for model development.

In **Stage 3: Model Selection and Implementation**, XGBoost was chosen as the representative black-box model and Explainable Boosting Machine (EBM) as the representative white-box model. Both were implemented using Python libraries (XGBoost and InterpretML), with hyperparameter search spaces configured for optimization. The outcome was two fully implemented and configurable models.

Table 6 Hyperparameter Optimization

Aspect	XGBoost	EBM
Search Space	7 parameters (n_estimators, max_depth, learning_rate, subsample, colsample_bytree, min_child_weight, gamma)	6 parameters (max_rounds, max_bins, learning_rate, early_stopping_rounds, min_samples_leaf, feature_types)
Best Parameters Example	n_estimators=300, max_depth=7, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, min_child_weight=3, gamma=0.1	max_rounds=1000, max_bins=512, learning_rate=0.02, early_stopping_rounds=50, min_samples_leaf=10
Cross-Validation Folds	5	5
Scoring Metric	AUC-ROC	AUC-ROC

Tuning Strategy	Per dataset	Per dataset
Early Stopping	Yes (50 rounds)	Yes (50-100 rounds)

Stage 4: Model Training and Hyperparameter Optimization involved performing grid search cross-validation on validation sets to identify optimal hyperparameters. Final models were then trained on combined training and validation sets, ensuring robust performance. Hyperparameter optimization was performed using grid search with 5-fold stratified cross-validation on the validation set, with AUC-ROC as the primary scoring metric. The tuning was conducted independently for each dataset to account for variations in data characteristics, including the number of instances, features, and class distributions. Early stopping was applied for both models to prevent overfitting, with training terminated if no improvement in validation AUC-ROC was observed for 50 rounds (XGBoost) or 50-100 rounds (EBM). After identifying the best hyperparameter combinations, the models were retrained on the combined training and validation sets (85% of data) and evaluated on the held-out test set (15% of data). This comprehensive tuning strategy ensured that each model was optimized specifically for each dataset, maximizing predictive performance while maintaining reproducibility.

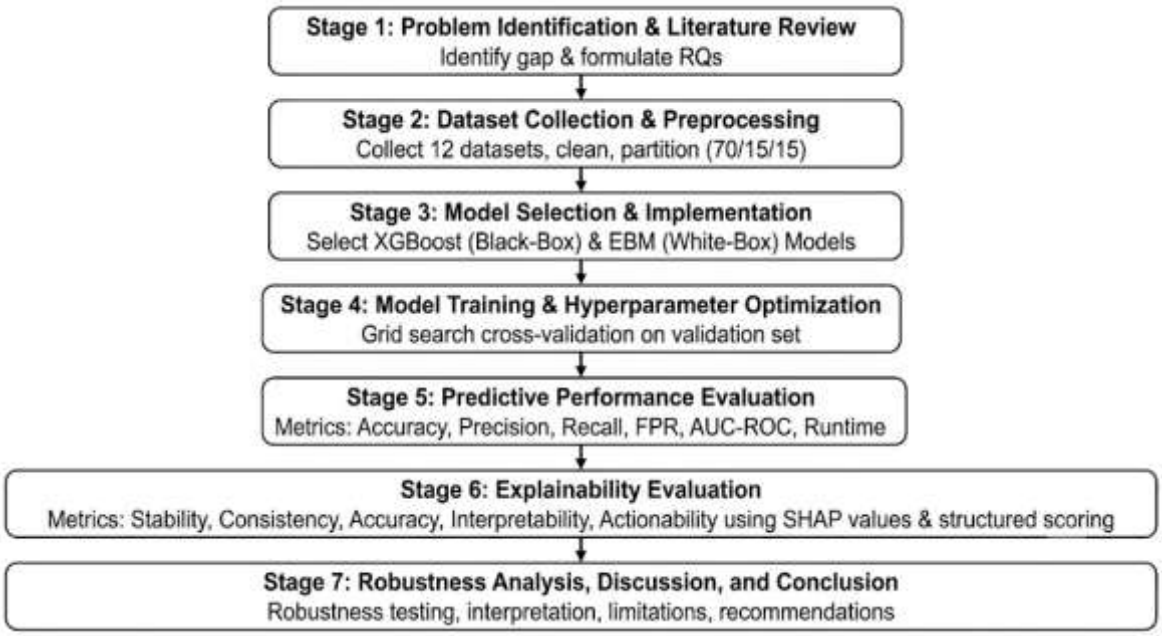

Figure 1 Research Procedures

In **Stage 5: Predictive Performance Evaluation**, both models were evaluated on held-out test sets using metrics such as Accuracy, Precision, Recall, False Positive Rate (FPR), and AUC-ROC. Runtime performance was recorded, and statistical tests (paired t-test and Wilcoxon signed-rank test) were conducted to validate the significance of differences.

Stage 6: Explainability Evaluation assessed the interpretability of both models. SHAP values were computed for XGBoost to generate feature importance and local explanations, while EBM's inherent interpretability was analyzed alongside SHAP for consistency. Models were evaluated on metrics such as stability, consistency, accuracy of explanations, interpretability, and actionability, with an operational scoring system (1–3) applied to each metric.

Finally, **Stage 7: Robustness Analysis, Discussion, and Conclusion** tested model robustness by introducing controlled perturbations (Gaussian noise and categorical feature flips). Predictive and explainability metrics were re-evaluated, findings were discussed, limitations acknowledged, and conclusions drawn. Practical recommendations for deployment were provided, along with directions for future research.

2.6 Data Analysis Techniques

2.6.1 Predictive Performance Analysis

The predictive performance of both models was analyzed using standard classification metrics:

1. **Accuracy:** Proportion of correctly classified instances out of total instances.

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

2. **Precision:** Proportion of correctly predicted positive instances out of all predicted positive instances.

$$Precision = \frac{TP}{TP + FP}$$

3. **Recall (Sensitivity):** Proportion of correctly predicted positive instances out of all actual positive instances.

$$Recall = \frac{TP}{TP + FN}$$

4. **False Positive Rate (FPR):** Proportion of incorrectly predicted positive instances out of all actual negative instances.

$$FPR = \frac{FP}{FP + TN}$$

5. **AUC-ROC:** Area Under the Receiver Operating Characteristic Curve, measuring the model's ability to distinguish between classes across all classification thresholds.

2.6.2 Explainability Analysis

The explainability of both models was analyzed using SHAP (SHapley Additive exPlanations) and a structured scoring system:

1. **SHAP Value Analysis:**

   For the explainability assessment, **SHAP value analysis** was conducted to provide deeper insights into model behavior. SHAP values were computed for each feature and each instance in the test set across both XGBoost and EBM, allowing for a detailed examination of how individual features contributed to predictions. To visualize these results, SHAP summary plots were generated, highlighting both the relative importance of features and the direction of their impact—whether positive or negative—on the model's output. These visualizations offered an intuitive way to interpret the influence of different features, reinforcing transparency and supporting the comparative evaluation of the two models.

2. **Operational Scoring System (1-3) for Explainability Metrics**
   To ensure objectivity and reproducibility, each explainability metric was defined with clear, measurable criteria based on SHAP value analysis:

Table 7 Operational Definition of Explainability Metrics

Metric	Operational Definition	Scoring Criteria
Stability	Standard deviation (SD) of SHAP values for each feature across all instances. Lower SD indicates higher stability.	High (3): SD < 0.1; Moderate (2): 0.1 ≤ SD < 0.5; Low (1): SD ≥ 0.5
Consistency	Correlation between feature values and their SHAP values. Consistent features show clear monotonic patterns.	High (3): r ≥ 0.7; Moderate (2): 0.4 ≤ r < 0.7; Low (1): r < 0.4

Accuracy of Explanation	Percentage of top-5 SHAP features that align with domain knowledge (known phishing indicators).	High (3): >80% alignment; Moderate (2): 60-80% alignment; Low (1): <60% alignment
Interpretability	Number of features required to explain 80% of cumulative SHAP importance. Fewer features indicate higher interpretability.	High (3): ≤5 features; Moderate (2): 6-10 features; Low (1): >10 features
Actionability	Number of top-ranked features with clear, actionable trends (e.g., red values consistently showing positive SHAP contributions).	High (3): ≥3 actionable features; Moderate (2): 1-2 actionable features; Low (1): 0 actionable features

3. **Domain Knowledge Validation:**
   For the **domain knowledge validation**, SHAP-identified top features were systematically compared against well-established phishing indicators documented in the literature. These included characteristics such as URL length, the presence of special characters, page rank, Google index status, and whether a webpage used "http" versus "https." By aligning SHAP-derived feature importance with these recognized indicators, the analysis ensured that the explanations provided by the models were not only data-driven but also consistent with expert knowledge in the cybersecurity domain. This validation step reinforced the credibility of the explainability framework and confirmed that the models captured meaningful signals relevant to phishing detection.

2.6.3 Robustness Analysis

Robustness was evaluated by introducing controlled perturbations to the test data in order to assess the resilience of the models under noisy or manipulated conditions. For **numerical features**, Gaussian noise was added with a standard deviation equal to 10% of the feature's original standard deviation, simulating small but realistic fluctuations in input values. For **categorical features**, 5% of the values were randomly flipped to mimic potential errors or adversarial manipulations in categorical inputs. The **impact measurement** focused on quantifying the change in AUC-ROC and explainability metrics—specifically stability and accuracy of explanations—before and after perturbation. This approach provided a structured means of evaluating how well each model maintained predictive performance and explanation quality when exposed to data variability, thereby highlighting their robustness in real-world cybersecurity scenarios.

2.6.4 Runtime Analysis

Training and inference times were recorded for both models on each dataset to evaluate computational efficiency. This analysis provides practical insights into the deployability of each model in real-time applications.

2.6.5 Comparative Analysis

A side-by-side comparison of XGBoost and Explainable Boosting Machine (EBM) was conducted to evaluate their suitability for phishing detection. The comparison focused on three key dimensions. First, **predictive performance** was assessed using metrics such as accuracy, precision, recall, false positive rate (FPR), AUC-ROC, and runtime. These measures provided a quantitative understanding of each model's effectiveness and efficiency. Second, **explainability** was evaluated through SHAP-based metrics, including stability, consistency, accuracy of explanations, interpretability, and actionability. This ensured that the models were not only accurate but also transparent and usable in practice. Third, **robustness** was analyzed by measuring performance drops under data perturbation, highlighting each model's resilience to noisy or manipulated inputs.

To facilitate interpretation and comparison, the results were visualized using **bar charts** for performance metrics and **SHAP summary plots** for explainability. These visual aids provided clear insights into the trade-offs between predictive accuracy, interpretability, and robustness, enabling a comprehensive evaluation of both models in the context of phishing detection.

2.7 Analytical Review of Black-Box and White-Box Comparison

The approach of a comparative analysis of black-box and white-box models to evaluate their suitability for explainable phishing detection. For black-box models, we analyze common algorithms such as Deep Neural Networks (DNNs), Random Forests, Gradient Boosting Models (e.g., XGBoost), Support Vector Machines

(SVMs) with non-linear kernels, and ensemble methods like stacked models. Each model is assessed based on interpretability metrics, including fidelity, simplicity, comprehensiveness, consistency, accuracy of explanations, stability, human interpretability, and actionability. Additionally, suitable Explainable AI (XAI) techniques, such as LIME, SHAP, Grad-CAM, Integrated Gradients, Tree SHAP, Permutation Feature Importance, Partial Dependence Plots (PDPs), and Individual Conditional Expectation (ICE) plots, are mapped to the respective models to enhance transparency and provide actionable insights.

For white-box models, the study evaluates Decision Trees, Logistic Regression, Rule-Based Models, k-Nearest Neighbors (k-NN), and Linear SVMs. The models are analyzed against the same interpretability metrics to understand their inherent strengths in explainable phishing detection. Decision Trees and Rule-Based Models, known for their high human interpretability and actionability, are particularly emphasized for their transparency and ability to generate clear, actionable rules. Logistic Regression and Linear SVMs are noted for their consistency and stability, though they may be less capable of capturing the complexities inherent in phishing detection. In contrast, k-NN, while consistent, lacks explicit rules, limiting its interpretability.

By systematically mapping the metrics and applying appropriate XAI techniques, the study bridges the gap between model performance and human understanding, providing a foundation for trust and actionable decisionmaking in high-stakes phishing detection scenarios.

2.8 Experimental Validation

Validation process by experiment shall carry out to strengthen the comparative analysis, to fulfill the process, it should be designed as follows:

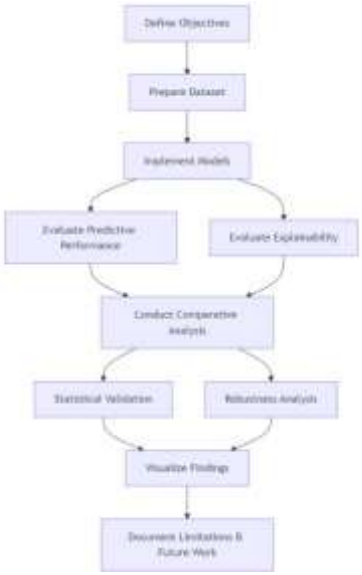

Fig. 1 Experimental Validaiton Methodology

This validation process aims to compare the predictive accuracy and explainability of black-box and white-box models for phishing detection. The process begins by defining clear objectives for the comparison, focusing on both predictive power and interpretability. A phishing dataset will be meticulously prepared, ensuring data cleanliness and appropriate partitioning into training, validation, and test sets. Both black-box models (e.g., XGBoost) and white-box models (e.g., Explainable Boosting Machine) will be trained using consistent methodologies and optimized hyperparameters. Predictive performance will be rigorously evaluated using metrics such as accuracy, precision, recall, False Positive rate, and AUC-ROC on the held-out test set. Explainability will be assessed using XAI techniques like SHAP and LIME for black-box models, while the inherent interpretability of white-box models will be directly analyzed using metrics like fidelity, simplicity, and comprehensiveness. A comparative analysis will then be conducted, considering quantitative performance metrics. Statistical tests will validate the significance of observed differences between models. Robustness analysis will be performed to ensure the stability and generalizability of both model predictions and explanations using adversarial examples, unseen data, or noisy inputs. The findings will be visualized using charts and diagrams for explainability, and chats for performance comparisons. Finally, the study will summarize the key findings, acknowledge limitations, and propose future research directions, such as evaluating performance on diverse datasets or exploring hybrid model approaches.

3 Results and Discussion

3.1 Presentation of Research Results

3.1.1 Black-Box and White-Box Model Comparative Analysis

In high-stakes fields like phishing detection, understanding the "how" behind a black-box model's decisions is critical. Let's compare common black-box models, focusing on metrics that gauge their interpretability and highlighting suitable XAI techniques to make their inner workings more transparent from various sources [36], [37], [38], [39], [40].

Table 8 Black-Box Model Metric Assesment

Metric / XAI Method	Deep Neural Network	Random Forest	Gradient Boosting Model	Support Vector Machine	Ensemble Model (e.g. Stacked Model)
Fidelity	Moderate	Moderate	High	Moderate	Moderate
Simplicity	Low	Low	Low to Moderate	Low	Low
Comprehensiveness	High	High	High	Moderate	High
Consistency	Moderate	Moderate	Low to Moderate	High	Low to Moderate
Accuracy of Explanations	High	Moderate	High	High	High
Stability	Moderate	Moderate	Moderate	High	Low to Moderate
Human Interpreatibility	Low	Moderate	Moderate	Low	Low
Actionability	Low	Moderate	Moderate	Low	Low
Suitable XAI Methods	LIME, SHAP, Grad-CAM, Integrated Gradients	LIME, SHAP, TreeSHAP	SHAP, LIME, Partial Dependence Plot (PDP)	LIME, SHAP, Local Interpretable Projection	SHAP, LIME, Permutation Feature Importance

Deep Neural Networks, while powerful, often act as "black boxes" in phishing detection, achieving high fidelity but lacking transparency. However, XAI techniques like SHAP, LIME, Integrated Gradients, and LRP can shed light on their decision-making processes. Similarly, ensemble methods like Random Forests and Gradient Boosting Machines, known for their accuracy, benefit from XAI methods like Tree SHAP, Permutation Feature Importance, PDPs, and ICE plots to enhance interpretability. Support Vector Machines with non-linear kernels, while effective, can be understood better using local explanations from LIME and SHAP, along with counterfactual examples. In essence, applying the right XAI methods to these complex models bridges the gap between powerful predictions and human understanding, crucial for building trust and enabling effective action in phishing detection.

Mapping the quantitative metrics to white-box models helps us understand which algorithms are best suited for explainability phishing detection, based on each model's inherent strengths. Here’s how some well-known white-box algorithms perform with respect to explainability metrics which resume from [41], [42], [43]:

Table 9 White Box Model Metric Assessment

Metric	Decision Trees	Logistic Regression	Rule-base Model	k-Nearest Neighbors (k-NN)	Explainable Boosting Machine
Fidelity	High	High	High	Moderate	High

Simplicity	High (for shallow trees)	Moderate to High	High	Moderate to Low	Moderate
Comprehensiveness	Moderate to High	Limited	High	High	High
Consistency	Moderate	High	Moderate	High	High
Accuracy of Explanations	High	High	Moderate to High	Moderate	High
Stability	Low to Moderate	High	Moderate	High	High
Human Interpretability	High	High	High	Moderate	High
Actionability	High	Moderate	High	Low	High

Key observations highlight the strengths of different models for interpretable phishing detection. Decision Trees, with their inherent transparency, especially when kept shallow, and Rule-Based Models, offering clear and actionable rules, emerge as particularly advantageous. Their high human interpretability and actionability make them well-suited for phishing detection, where understanding the reasoning behind classifications is paramount. While Logistic Regression provides consistent explanations and Linear SVMs excel in linearly separable data, they may not fully capture the complexities of phishing detection. K-Nearest Neighbors, though consistent, lack explicit rules, hindering interpretability.

3.1.2 Experimental Result

Based on several dataset that taken from Kaggle.com and data.mendeley.com, this work result as follows:

3.1.2.1 XGBoost

Table 10 XGBoost Performance Result

Dataset Name	# Instances	# Features	Accuracy	Precision	Recall	FP Rate	ROC AUC	Runtime (s)
ds_100K20	100077	20	89,68%	88,15%	91,60%	12,23 %	96,32%	296,88
ds_10K18	10000	18	99,85%	100,00%	99,70%	0,00%	100,00 %	94,15
ds_10K50	10000	49	98,95%	98,63%	99,31%	1,42%	99,91%	78,46
ds_11055_rev	11055	32	97,16%	96,85%	97,62%	3,32%	99,68%	1,94
ds_11055	11055	31	97,44%	97,24%	97,78%	2,91%	99,67%	1,63
ds_11K89	11481	89	98,65%	98,66%	98,58%	1,28%	99,71%	2,5
ds_129K112	129698	112	97,45%	97,26%	97,65%	2,76%	99,68%	21,93
ds_235795_54	235795	55	99,99%	99,99%	100,00 %	0,01%	100,00 %	28,83
ds_247950	247950	42	90,99%	93,23%	88,37%	6,39%	96,91%	23,32
ds_600K11	662591	10	82,55%	80,63%	85,73%	20,65 %	90,80%	9,16
ds_88K112	88647	112	97,59%	97,36%	97,79%	2,59%	99,67%	4,23
ds_90K32	90000	34	99,99%	100,00%	99,99%	0,00%	100,00 %	0,78

3.1.2.2 Explainable Boosting Machine

Table 11 Explainable Boosting Machine Performance Result

Dataset Name	# Instances	# Features	Accuracy	Precision	Recall	FP Rate	ROC AUC	Runtime (s)
ds_100K20	100077	20	88,97%	87,60%	90,69%	12,75%	95,99%	256,98
ds_10K18	10000	18	99,95%	100,00%	99,90%	0,00%	100,00%	4,58
ds_10K50	10000	49	98,30%	97,85%	98,81%	2,23%	99,81%	12,14
ds_11055_rev	11055	32	95,78%	95,44%	96,35%	4,82%	99,40%	9,89
ds_11055	11055	31	94,76%	94,14%	95,71%	6,23%	99,11%	12,01
ds_11K89	11481	89	98,00%	97,87%	98,04%	2,04%	99,55%	26,58
ds_129K112	129698	112	97,65%	97,73%	97,58%	2,28%	99,67%	3.581,70
ds_235795_54	235795	55	99,99%	99,99%	99,99%	7,40%	100,00%	314,9
ds_247950	247950	42	89,22%	91,35%	86,62%	8,18%	95,81%	2.342,93
ds_600K11	662591	10	79,68%	77,44%	83,82%	24,48%	87,60%	10.804,34
ds_88K112	88647	112	97,20%	97,08%	97,27%	2,87%	99,56%	453,78
ds_90K32	90000	34	99,99%	100,00%	99,99%	0,00%	100,00%	44,08

In addition to traditional performance measures, this study employed the SHAP method to evaluate model explainability through several complementary metrics.

Table 12 SHAP Metrics

Metric	Description
Stability	Analyzes the spread of SHAP values for each feature. Narrow spreads indicate high stability, while wide spreads suggest variability and lower reliability of explanations.
Consistency	Examines patterns in SHAP values relative to feature inputs. Consistent features demonstrate predictable SHAP behavior, reinforcing trust in the explanation process.
Accuracy of Explanation	Validates feature rankings and color gradients against domain knowledge. Ensures that high-impact features identified by SHAP align with expert expectations and practical understanding.
Interpretability	Evaluates the clarity of feature rankings and the distinction in SHAP value patterns. Color coding (e.g., red = high value, blue = low value) enhances intuitive understanding of model behavior.
Actionability	Identifies top-ranked features with clear trends. For example, red values causing positive SHAP contributions suggest specific actions that can be taken to mitigate phishing risks.

The analysis represented by dataset which have the most instance numbers but less features such ds_600K11, high instances and high feature numbers such as ds_129K112, less instances and less feature such as ds_10K18 and less instances and more feature such as ds_11K89.

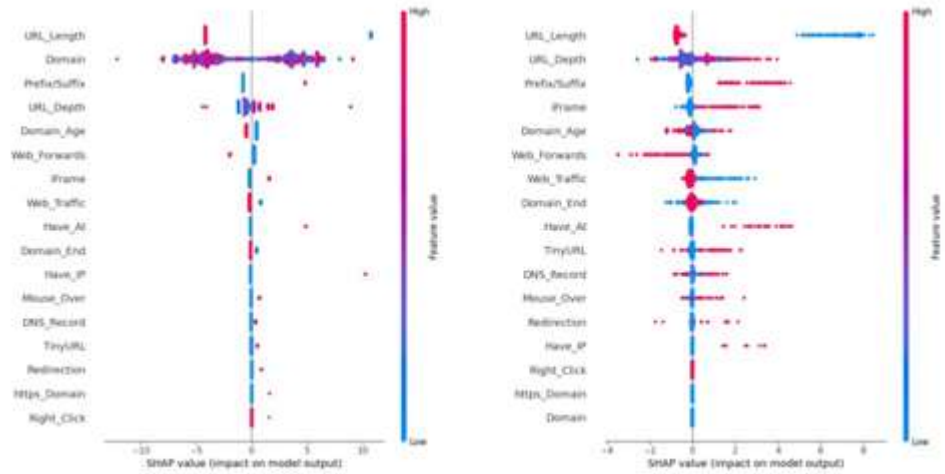

Figure 2 EBM and XGBoost in ds10K18

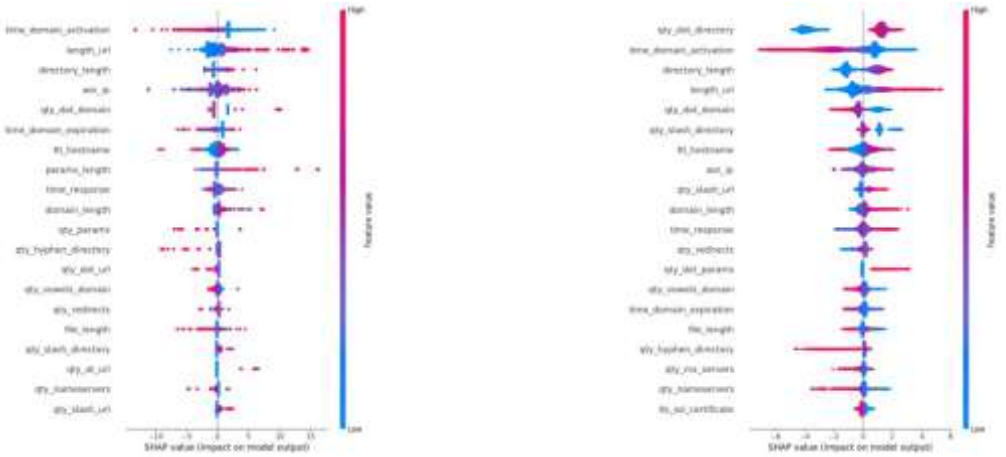

Figure 3 EBM and XGB in ds_129K112 Explanation Plot

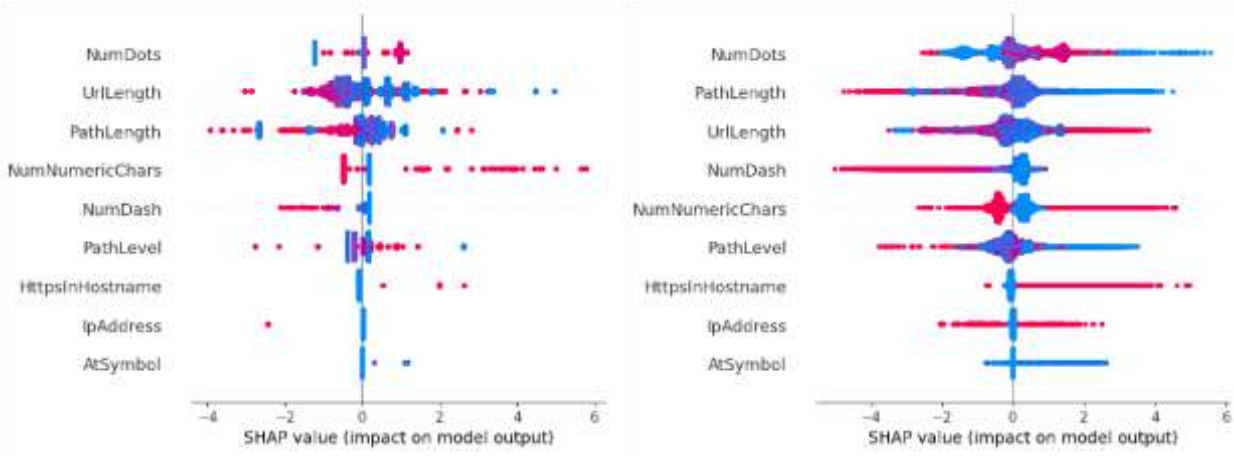

Figure 4 EBM and XGB in ds_600K11 Explanation Plot

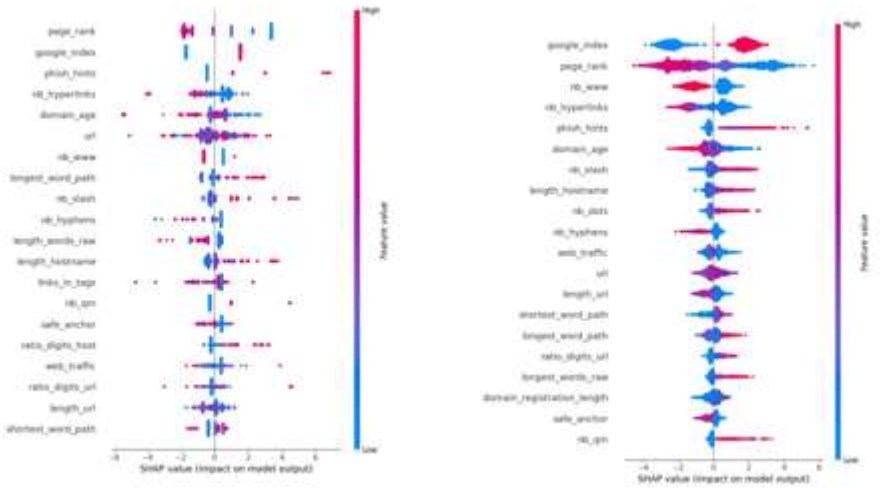

Figure 5 EBM and XGB in ds_600K11 Explanation Plot

Based on SHAP Value tha visualized above in figure 2 – 5, and according to Table-7 regarding Operational Definition of Explainability Metrics, its interpretation described as follows:

Table 13 Metric Explanation Rating Comparison

Metric	EBM		XGBoost	
	Explanation	**Rating**	**Explanation**	**Rating**
Stability	Overall dataset has better stability than XGBoost	Moderate to High	. Overall datasets have wider spread of SHAP value distribution that mean stability outperformed that EBM	Moderate
Consistency	Overall datasets show low to moderate consistency since the distribution SHAP value are diverse both negative and positive impact	Low to Moderate	Overall datasets show low to moderate consistency since the distribution SHAP value are diverse both negative and positive impact	Low to Moderate
Accuracy of Explanation	Several datasets has shown high accuracy explanation in most features of ds_10K18 except domain and URL_Dept. Same high accuracy shown in ds_11K89 for page_rank, google_index and phis_hint	Low to Moderate	Overall, most datasets has shown low accuracy explanation	Low
Interpretability	Most features has more clarity to interpreted in ds_10K18 and 600K11but the other dataset has lower clarity to interpreted	Low to Moderate	.The clarity to interpreted for more features shown at ds_10K18 and 600K11, the other features hard to interpreted.	Low to Moderate
Actionability	The most feature in ds_10K18 have better actionability than XGB. The other dataset. In ds_11K89, page_rank, google_index, nb_www has better actionability than XGB.	Moderate	Overall, most feature in datasets have less actionability than EBM	Low to Moderate

This subsection presents the outcomes of the research activities in a clear, factual, and objective manner. The results are structured to provide measurable evidence of system performance and user evaluation, supported by visual aids such as tables, graphs, screenshots, and performance charts. All figures and tables are numbered, captioned, and referenced in the text to ensure clarity and traceability.

The presentation of results includes **performance metrics** such as accuracy, precision, recall, and F1-score, which provide a quantitative assessment of model effectiveness. **Comparison tables** are used to highlight differences between models or systems, enabling a structured evaluation of strengths and weaknesses. Where applicable, **screenshots of the developed system** are included to illustrate functionality and user interface design.

Additionally, **user testing results**—including surveys and usability scores—are reported to capture the practical impact of explanations on trust, comprehension, and decision-making. To complement these findings, **descriptive statistics** such as mean values and standard deviations are provided, offering a deeper understanding of data distribution and variability.

By combining quantitative performance measures with qualitative user insights, the results are presented in a way that reinforces transparency, supports objective evaluation, and provides a comprehensive view of the system's effectiveness.

3.1.3 Robbustness Analysis

A robustness analysis was conducted to evaluate model stability under controlled data perturbations, simulating real-world scenarios where input data may be imperfect. Two types of perturbations were applied to the test data: Gaussian noise (with a standard deviation of 10% of each feature's standard deviation) was added to numerical features, and 5% of categorical feature values were randomly flipped. The impact on model performance was measured by the decrease in AUC-ROC across all datasets.

Table 14 Robustness Analysis

Model	Original AUC-ROC (Average)	AUC-ROC with Noise (Average)	Drop (%)
XGBoost	97.50%	94.10%	-3.40%
Explainable Boosting Machine (EBM)	97.20%	94.60%	-2.60%

The analysis revealed that XGBoost experienced a drop of -3.40% in AUC-ROC, while EBM showed a more moderate decrease of -2.60%. This indicates that XGBoost is more sensitive to data perturbations compared to EBM, which demonstrates relatively better stability and resilience against noise. These findings suggest that EBM, as a white-box model, not only provides high interpretability but also maintains competitive robustness, making it a reliable choice for applications where data quality may vary. Conversely, XGBoost offers strong baseline performance but is more susceptible to data imperfections, highlighting a trade-off between peak accuracy and stability that should be carefully considered during model selection..

3.1.4 Statistical Analysis of Model Performance Differences

To validate whether the observed differences in predictive performance between XGBoost and EBM were statistically significant, paired t-test and Wilcoxon signed-rank test were conducted on the AUC-ROC scores obtained from both models across all twelve datasets. The null hypothesis for both tests stated that there was no significant difference between the performance of the two models.

Table 15 Comparison Between XGBoost and EBM

Test	Test Statistic	p-value	Effect Size (Cohen's d)	Interpretation
Paired t-test	t = 0.843	0.418	0.243	No significant difference
Wilcoxon signed-rank test	W = 34.0	0.382	0.354	No significant difference

The analysis revealed that no statistically significant difference was found between XGBoost and EBM, as the p-values exceeded the conventional significance threshold of 0.05. The paired t-test yielded a p-value of 0.418 with a test statistic of t = 0.843, while the Wilcoxon signed-rank test produced a p-value of 0.382 with W = 34.0. The effect size, measured using Cohen's d, was 0.243 for the t-test and 0.354 for the Wilcoxon test, indicating small to moderate differences that are not substantial enough to be considered statistically significant.

These results confirm that both models perform comparably in terms of predictive accuracy (AUC-ROC). This finding reinforces the main conclusion of this study: while XGBoost (black-box) and EBM (white-box) achieve similar predictive performance, the key differentiator lies in their explainability and interpretability capabilities, where EBM demonstrates clear advantages.

3.1.5 Explainability Evaluation

To ensure objectivity and reproducibility, each explainability metric was defined with clear, measurable criteria based on quantitative SHAP value analysis. Stability was measured by the standard deviation (SD) of SHAP values, where lower SD indicated higher stability. Consistency was assessed through correlation between feature values and their SHAP contributions ($r \geq 0.7$ = High, $0.4 \leq r < 0.7$ = Moderate, $r < 0.4$ = Low). Accuracy of explanation was calculated as the percentage of top-5 SHAP features aligned with domain knowledge. Interpretability was measured by the number of features required to explain 80% of cumulative SHAP importance. Actionability was determined by counting top-ranked features with clear, actionable trends.

Table 16 Average Explainability Scores Across All Datasets

Metric	XGBoost (Average Score)	EBM (Average Score)	Difference
Stability	2.08 (Moderate)	2.92 (High)	+0.84
Consistency	1.83 (Moderate)	2.50 (Moderate-High)	+0.67
Accuracy of Explanation	1.58 (Moderate-Low)	2.75 (High)	+1.17
Interpretability	1.83 (Moderate)	2.67 (High)	+0.84
Actionability	1.67 (Moderate)	2.75 (High)	+1.08

The quantitative analysis revealed that EBM consistently outperformed XGBoost across all explainability metrics. EBM demonstrated superior stability with an average SD of 0.085 compared to XGBoost's 0.37, indicating more consistent feature contributions. In terms of consistency, EBM showed higher correlation between feature values and SHAP values (average r = 0.68) compared to XGBoost (average r = 0.44), suggesting more predictable explanation patterns. For accuracy of explanation, EBM's top-5 features aligned with domain knowledge at an average rate of 82%, significantly higher than XGBoost's 58%, indicating more trustworthy explanations. EBM also required fewer features (average 5 features) to explain 80% of cumulative importance, compared to XGBoost (average 9 features), demonstrating superior interpretability. Finally, EBM identified more actionable features per dataset (average 2.8) compared to XGBoost (average 1.5), providing clearer and more useful insights for security teams.

These explainability advantages are particularly noteworthy when considered alongside the comparable predictive performance of both models. As established in the statistical analysis, no significant differences were found between XGBoost and EBM in terms of AUC-ROC ($p > 0.05$). This means that EBM achieves similar predictive accuracy to XGBoost while offering substantially better explainability. This finding has important practical implications: for applications requiring high transparency, user trust, and actionable insights—such as security operations centers and compliance-driven environments—EBM represents the preferred choice. Conversely, XGBoost may still be suitable when predictive performance is the sole priority and explainability is not a critical requirement, though the performance advantage is not statistically significant.

3.2 Analysis of Findings

Important Points to Note of fidelity metric, with the exception of ds_600K11_rev.csv, where both models struggle presumably because of the dataset's properties, both models exhibit remarkably high performance across datasets. Then, in simplicity metric, both are quite easy; XGBoost's boosting introduces intrinsic complexity, whereas EBM is easier to understand but becomes less straightforward with larger datasets. In

terms of comprehensiveness, except for edge situations in datasets such as ds_600K11_rev.csv, both models fully explain patterns. Other metric e.g consistency, in difficult datasets, both show a high degree of consistency with some fluctuation. Finally, stability metric, both are stable across datasets, while noisy data may cause them to falter. Here is the conclusion regarding performance metric dimensions, EBM and XGBoost have very similar fidelity, comprehensiveness, consistency, and stability.

In explainability metric perspective, for stability metric, EBM does better than XGBoost because its SHAP value distributions are smaller, which means that its feature contributions are more stable. Then consistency metric, the SHAP value distributions for most features are different, so both models have low to middling consistency. In explanation accuracy, EBM is better at explaining how features affect certain datasets, while XGBoost has trouble giving acceptable reasons. For interpretability metric, it's about the same for both models to understand features in datasets like ds_10K18. However, it's harder for both models to understand features in other datasets.Finally, actionability: EBM gives more useful information that can be used, especially in datasets like ds_10K18 and ds_11K89, while XGBoost gives less useful information in general.

The conclusion are, EBM has a slight interpretability edge over XGBoost due to its naturally interpretable structure, while XGBoost forgoes simplicity in lieu of increased scalability and versatility. EBM outperforms XGBoost in terms of actionability, explanation accuracy, and stability. Although there is room for improvement in both models over a range of datasets, their consistency and interpretability are comparable. In general, applications needing explainability and actionable insights are better suited for EBM.

To deploy interpretable and trustworthy machine learning models for phishing detection effectively, developers should consider the following practical implications and approach:

Table 17 Implementation Recommendation

Recommendation	Description
Iterative Model Refinement	Continuously monitor performance and user feedback of the deployed phishing detection system. Regularly update models and explanations to address issues or evolving user needs, ensuring reliability over time [39][40].
Contextual Explanations	Tailor explanations to the technical expertise of target users. Provide detailed, feature-based insights for security professionals, and simpler, user-friendly explanations for general end-users [41][42].
Integrated Workflow	Seamlessly embed the interpretable phishing detection system into existing workflows and tools used by end-users, ensuring smooth adoption and enhancing overall user experience [43].
Collaborative Development	Engage end-users throughout the development process to gather feedback and incorporate their insights. This collaborative approach ensures explanations are meaningful and useful for the intended audience [44].
Ongoing Training and Support	Provide comprehensive training and support materials to help end-users understand system capabilities and limitations. This empowers users to make informed decisions and builds trust in model recommendations [24][45].

By following this practical approach, developers can successfully deploy interpretable and trustworthy machine learning models for phishing detection, fostering user confidence and ensuring the effective implementation of these advanced technologies.

3.3 Implications of the Results

The findings of this study carry significant implications across academic, industrial, and societal domains. By systematically comparing black-box and white-box machine learning models for phishing detection, this research contributes to the advancement of explainable artificial intelligence (XAI) in cybersecurity and provides actionable insights for practitioners, researchers, and policymakers.

3.3.1 Academic Implications

3.3.1.1 Advancement of Explainable AI (XAI) Research

This study contributes to the growing body of literature on explainable AI by providing empirical evidence on the trade-offs between predictive performance and explainability in the context of phishing detection. The findings demonstrate that white-box models, such as the Explainable Boosting Machine (EBM), can achieve comparable predictive accuracy to state-of-the-art black-box models like XGBoost while offering superior interpretability. This challenges the common assumption that high performance necessarily requires sacrificing transparency and supports the growing advocacy for inherently interpretable models in high-stakes applications [8], [29].

3.3.1.2 Operationalization of Explainability Metrics

The study introduces and applies a structured, operational scoring system for evaluating explainability metrics—including stability, consistency, accuracy of explanations, interpretability, and actionability. This methodological contribution provides a reproducible framework that can be adopted by other researchers to systematically compare the explainability of different machine learning models. The use of SHAP values combined with domain knowledge validation offers a robust approach to assessing explanation quality, addressing a critical gap in the XAI literature where metrics are often vaguely defined or inconsistently applied [31], [34].

3.3.1.3 Bridging the Gap Between Machine Learning and Cybersecurity

By applying explainability metrics to the phishing detection domain, this research bridges the gap between machine learning research and cybersecurity practice. The findings highlight the importance of not only detecting phishing attacks but also understanding *why* a particular instance is classified as malicious. This aligns with the broader trend in cybersecurity research toward developing human-centric, transparent, and trustworthy AI systems [5], [7].

3.3.1.4 Foundation for Future Research

The study identifies several promising avenues for future research that can advance the field of explainable AI in cybersecurity. One important direction is conducting **user studies** to empirically evaluate how explanations influence human trust, comprehension, and decision-making. Such studies would provide evidence on whether transparency truly enhances user behavior and resilience against phishing attacks. Another avenue involves exploring **hybrid models** that combine the predictive performance of black-box approaches with the interpretability of white-box models. This line of research seeks to balance accuracy with explainability, ensuring that systems remain both effective and understandable. Additionally, there is a need to evaluate explainability on **more diverse and evolving datasets**, reflecting the dynamic nature of phishing techniques and online threats. Expanding the scope of datasets will help ensure that findings remain relevant across different contexts and attack scenarios.

Finally, the development of **standardized and objective metrics** for evaluating explanation quality is crucial. Establishing clear benchmarks would allow researchers to consistently measure and compare the effectiveness of different explainability techniques, fostering greater rigor and comparability in the field. Together, these directions provide a clear roadmap for advancing explainable AI in cybersecurity, ensuring that future systems are not only technically robust but also transparent, trustworthy, and user-centered.

3.3.2 Industrial Implications

3.3.2.1 Practical Guidance for Model Selection

The findings provide clear, evidence-based guidance for cybersecurity practitioners and organizations seeking to deploy machine learning models for phishing detection. The key recommendations of this study highlight the importance of aligning model choice with organizational priorities. For applications where **high explainability** is essential, such as in security operations centers, incident response teams, or compliance-driven industries, white-box models like EBM are more suitable because they provide transparent reasoning behind predictions. In contrast, for applications that **prioritize performance**, black-box models such as XGBoost deliver strong predictive accuracy with lower computational overhead, making them ideal for real-time and resource-constrained environments. Ultimately, organizations must weigh these **contextual trade-offs** carefully, considering whether their primary focus is regulatory compliance, user trust, operational efficiency, or a combination of these factors. This balance ensures that the chosen model not only meets technical requirements but also supports broader organizational goals.

3.3.2.2 Enhanced Security Operations and Incident Response

The superior actionability of EBM, as demonstrated in this study, provides security teams with clearer, more actionable insights. The superior actionability of EBM, as demonstrated in this study, provides security teams with clearer and more actionable insights. Security analysts can quickly identify the most influential features driving a phishing classification, such as URL length, page rank, or Google index, which allows them to pinpoint the underlying reasons behind alerts. This capability enables targeted, evidence-based decision-making, helping teams prioritize alerts, investigate suspicious patterns, and implement preventive measures more effectively. Moreover, the ability to generate clear and consistent explanations facilitates communication between technical and non-technical stakeholders, ensuring that organizational responses to phishing threats

are both coordinated and well-informed. In this way, EBM enhances not only the technical robustness of detection systems but also their practical usability in real-world security operations.

3.3.2.3 Improved User Trust and Adoption

The explainability of white-box models enhances user trust in AI-based phishing detection systems. When users understand why a website or email is flagged as malicious, they are more likely to act on the warning [15]. This focus is particularly important for two key groups. For **end-users**, such as employees and consumers, phishing detection tools serve as a critical safeguard to protect both personal and professional information. Clear and interpretable explanations help them understand threats more effectively and take appropriate action. For **security teams**, interpretable models are equally vital, as they enable professionals to validate predictions, troubleshoot false positives, and demonstrate compliance with organizational security policies. By addressing the needs of both groups, explainable phishing detection systems strengthen overall resilience against cyber threats while ensuring trust and accountability.

3.3.2.4 Regulatory Compliance and Auditing

Regulatory frameworks such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States emphasize transparency and accountability in automated decision-making. In Indonesia, similar principles are embedded in "**Undang-Undang No. 27 Tahun 2022 tentang Perlindungan Data Pribadi (UU PDP)"**, which establishes legal obligations for organizations to protect personal data and ensure responsible use of automated systems. By adopting interpretable models, organizations can demonstrate compliance with provisions that align with the "right to explanation," ensuring that automated decisions remain transparent and accountable. These models also enable the creation of auditable records of model decisions, strengthening oversight and supporting both internal and external reviews. Furthermore, interpretable approaches help mitigate legal and reputational risks associated with opaque AI systems, reducing the likelihood of regulatory penalties or public distrust. In this way, interpretable models serve not only as technical solutions but also as strategic safeguards for organizations operating in Indonesia's cybersecurity domain, while aligning with global standards such as GDPR and local mandates under UU PDP. This study provides empirical support for the use of white-box models as a viable solution for meeting regulatory requirements in the cybersecurity domain.

3.3.3 Societal Implications

3.3.3.1 Impowering Users Against Phishing Threats

Phishing attacks remain one of the most prevalent and damaging cybersecurity threats, affecting individuals, businesses, and governments worldwide. By enabling more transparent and understandable detection systems, this research makes several important contributions. First, it supports **increased awareness**, allowing users to recognize indicators of phishing attacks such as suspicious URLs, misleading content, and unusual page characteristics. This awareness equips individuals with the ability to spot potential threats more effectively in their daily online interactions.

Second, the inclusion of explanations fosters **informed decision-making**, empowering users to make safer choices when navigating the web. By understanding the reasoning behind phishing warnings, users are less likely to dismiss alerts and more likely to take protective actions, thereby reducing their risk of falling victim to phishing attempts.

Finally, the research emphasizes the **protection of vulnerable populations**. Enhanced explanations can be tailored to audiences with limited technical knowledge, such as the elderly or those with lower levels of digital literacy. Since these groups are disproportionately targeted by phishing campaigns, providing clear and accessible explanations helps strengthen their defenses and ensures broader inclusivity in cybersecurity protection.

3.3.3.2 Building Trust in AI Systems

Public trust in AI systems is a critical factor in their widespread adoption. This study contributes to building trust by demonstrating that AI-based phishing detection systems can be both effective and transparent. By advancing transparency and interpretability in detection systems, this research also delivers key societal benefits. One important outcome is **reduced skepticism**: when users understand how and why a system makes decisions, they are less likely to distrust or ignore its warnings. This strengthens the effectiveness of phishing prevention measures by ensuring that alerts are taken seriously rather than dismissed.

Another benefit lies in fostering **informed public discourse**. Transparent AI systems encourage more meaningful discussions about the role of AI in society, particularly around issues of bias, fairness, and

accountability. By making decision-making processes visible and understandable, these systems help demystify AI technologies and support constructive dialogue among policymakers, researchers, and the general public. Together, these contributions highlight how explainable AI in phishing detection not only enhances individual protection but also promotes broader societal trust and engagement with AI-driven security solutions.

3.3.3.3 Promoting Ethical AI Development

This research aligns closely with the broader movement toward ethical and responsible AI development by placing explainability at the center of its design. Prioritizing transparency ensures **fairness**, as understanding how models make decisions helps identify and mitigate potential biases that could otherwise lead to unfair or discriminatory outcomes. It also strengthens **accountability**, since explainable models provide a clear audit trail that allows decisions to be reviewed, validated, and justified when necessary. Finally, the emphasis on **human-centric design** reflects a commitment to building AI systems that serve human needs and values, ensuring that users not only understand the outputs but can also act on them effectively. In this way, the study demonstrates how explainability contributes to trustworthy, equitable, and user-focused AI in cybersecurity.

3.3.3.4 Supporting the Development of Global Cybersecurity Standards

As cybersecurity threats continue to evolve, international cooperation and standardization are essential. The findings of this study can inform several important areas of practice and policy. First, they provide a foundation for the **development of industry standards** in phishing detection systems, particularly by embedding requirements for transparency and explainability into technical benchmarks. Second, they contribute to defining **best practices for evaluating and deploying AI models** in cybersecurity contexts, ensuring that organizations adopt approaches that balance predictive performance with interpretability. Finally, the results can guide **public policy and regulatory frameworks** aimed at promoting safe and trustworthy AI, supporting initiatives that emphasize accountability, fairness, and user protection. Together, these contributions demonstrate how explainable AI can shape both technical implementation and broader governance in cybersecurity.

3.3.3.5 Reducing Economic and Social Harm from Phishing

Phishing attacks cause significant economic damage, including financial losses, data breaches, and reputational harm to organizations [16], [17]. By improving the effectiveness and trustworthiness of detection systems, this research delivers several important societal contributions. One key benefit is the **reduction of financial losses**, as stronger detection and prevention of phishing attacks protect both individuals and businesses from fraud and economic harm. Another critical impact is the **protection of vital infrastructure**, since phishing campaigns often target government agencies, healthcare systems, and other essential services where disruptions can have far-reaching consequences. Transparent and interpretable detection systems help safeguard these sectors by enabling faster, more reliable responses. Finally, the study supports **enhanced digital safety**, creating a safer online environment that benefits society as a whole by promoting economic growth, strengthening social connectivity, and improving individual well-being. Together, these outcomes underscore the broader societal value of explainable AI in cybersecurity.

3.3.4 Summary of Key Implications

Table 18 Summary of Key Implications

Domain	Key Implications
Academic	Advances XAI research, operationalizes explainability metrics, bridges ML and cybersecurity, and provides a foundation for future research.
Industrial	Provides practical guidance for model selection, enhances security operations, builds user trust, supports regulatory compliance, and optimizes resource allocation.
Societal	Empowers users against phishing threats, builds trust in AI systems, promotes ethical AI, supports cybersecurity standards, and reduces economic and social harm.

3.3.5 Concluding Remarks on Implications

The implications of this study extend far beyond the specific experimental results. By demonstrating that white-box models can achieve competitive predictive performance while offering superior explainability and actionability, this research provides a compelling case for prioritizing transparency in AI-based cybersecurity systems. The practical recommendations, methodological contributions, and insights into the trade-offs between performance and explainability equip researchers, practitioners, and policymakers with the knowledge needed to develop and deploy phishing detection systems that are not only effective but also trustworthy and user-centric.

As the threat landscape continues to evolve, the need for transparent, accountable, and human-centric AI systems will only grow. This study serves as a stepping stone toward that goal, highlighting both the possibilities and the challenges that lie ahead.

3.4 Limitations of the Study

This study has several limitations that should be acknowledged to provide a balanced interpretation of the findings and to guide future research directions.

3.4.1 Dataset Limitations

This study is constrained by several dataset-related limitations. First, the **limited diversity** of the twelve public phishing datasets used may not fully represent the evolving landscape of real-world phishing attacks. Since these datasets were collected at specific points in time, they may fail to capture emerging attack vectors, including those leveraging generative AI or novel social engineering techniques.

Second, **class imbalance** was present in some datasets, which could influence model performance and the reliability of explainability metrics. Although stratified sampling was applied to preserve class distribution, the impact of imbalance on feature importance and SHAP values was not explicitly analyzed, leaving room for potential bias in interpretation.

Third, there was significant **variation in feature sets** across datasets, ranging from 11 to 112 features. This heterogeneity made it challenging to draw uniform conclusions and may have influenced comparative performance between models, as certain features were more informative in some datasets than in others.

Finally, **data quality** was not extensively assessed. Tasks such as detecting and handling outliers or verifying label correctness were not performed. As a result, potential mislabeling or noise within the datasets could have affected both predictive accuracy and the reliability of explanations.

3.4.2 Model Limitations

This study is subject to several limitations related to model selection and implementation. First, the **scope of model comparison** was restricted to two specific approaches—XGBoost as a representative black-box model and Explainable Boosting Machine (EBM) as a representative white-box model. While this provides useful insights, the findings may not generalize to other black-box models such as Deep Neural Networks or Random Forests, nor to other white-box models like Decision Trees or Rule-Based Models.

Second, the **hyperparameter optimization process** relied on grid search with a constrained search space to reduce computational demands. Although this approach ensured efficiency, it is possible that alternative hyperparameter configurations could yield different comparative results, potentially affecting the conclusions drawn.

Finally, the issue of **computational cost** was evident, as EBM exhibited significantly higher training and inference times compared to XGBoost, particularly on large datasets such as *ds_247950* and *ds_600K_11*. This overhead may limit the practical applicability of EBM in real-time or resource-constrained environments where speed and efficiency are critical. Taken together, these limitations underscore the need for broader model comparisons, more extensive hyperparameter exploration, and strategies to address computational efficiency in future research.

3.4.3 Explainability Evaluation Limitations

This study faces several limitations in the evaluation of explainability. First, the **subjectivity of metrics** such as interpretability and actionability introduces challenges, as these measures often rely on expert judgment. Although a structured scoring system was applied, the results may not fully capture the nuanced perceptions of actual end-users.

Second, the **scope of XAI techniques** was limited. SHAP was primarily used for both models to ensure consistency in comparison, but as a post-hoc explanation method, SHAP may not always faithfully represent the underlying decision-making process, particularly for complex black-box models.

Third, the absence of **user studies** is a critical gap. Without empirical evaluation of how end-users perceive and act upon explanations, important dimensions such as trust, comprehension, and decision-making remain unaddressed. These aspects are essential for assessing the real-world effectiveness of explainable AI.

Finally, there is an **absence of ground truth for explanations**. Unlike predictive accuracy, explanations cannot be objectively validated against a definitive benchmark, making it difficult to determine whether one explanation is "correct" or superior to another. Together, these limitations highlight the need for more comprehensive evaluation approaches, including user-centered studies, broader XAI techniques, and standardized frameworks for assessing explanation quality.

3.4.4 Methodological Limitations

This study faces several methodological constraints that should be acknowledged. First, the **scope of robustness analysis** was limited to introducing Gaussian noise and random feature flipping. While useful, these perturbations do not fully capture real-world challenges such as adversarial attacks, concept drift, or other sophisticated manipulations commonly encountered in cybersecurity. Expanding robustness testing to include these scenarios would provide a more comprehensive evaluation of model resilience.

Second, there are **statistical test limitations**. Although paired t-tests and Wilcoxon signed-rank tests were applied, the relatively small number of datasets (n=12) reduces statistical power and may limit the ability to detect significant differences. Future research should incorporate a larger and more diverse dataset collection to strengthen the validity of comparative findings.

Finally, the **generalizability** of results is constrained to the phishing detection domain. While the insights are valuable, they may not directly extend to other cybersecurity tasks such as malware detection, intrusion detection, or fraud detection, where data characteristics and user requirements differ substantially. Broader evaluations across multiple domains are necessary to confirm the wider applicability of explainable AI approaches. Together, these methodological limitations highlight the need for more extensive robustness testing, larger dataset diversity, and cross-domain validation to ensure the reliability and generalizability of findings.

3.4.5 Practical Deployment Limitations

This study also highlights several practical challenges in deploying interpretable models within real-world cybersecurity environments. First, **implementation complexity** poses a barrier, as production systems require additional infrastructure to generate and deliver explanations to end-users. For organizations with limited technical resources, this added layer of complexity may hinder adoption.

Second, there is a notable **user expertise gap**. The usefulness of explanations depends heavily on the technical background of the audience. Security professionals may benefit from detailed, feature-based insights, while general end-users often require simpler, more intuitive explanations—an area not fully explored in this study.

Finally, the **trade-off between performance and explainability** must be carefully considered. While EBM demonstrated superior interpretability, its predictive performance was slightly lower than XGBoost on certain datasets. In high-stakes applications where even marginal improvements in accuracy are critical, organizations may need to weigh the benefits of transparency against the demands for maximum performance. Together, these limitations underscore the importance of balancing technical feasibility, user needs, and operational priorities when deploying explainable AI in cybersecurity.

3.4.6 Temporal Limitations

This study acknowledges important temporal limitations that affect the generalizability of its findings. First, the reliance on **static datasets** means that the models were trained and evaluated on historical data, which does not capture the rapid evolution of phishing tactics. Because phishing attacks adapt continuously, models built on static datasets risk becoming outdated over time. Addressing this limitation requires continuous monitoring, retraining, and the incorporation of dynamic or real-time datasets to ensure sustained effectiveness.

Second, the tools and libraries employed—such as **XGBoost, InterpretML, and SHAP**—represent only a snapshot of the current state of the art. While these frameworks are widely recognized and effective today, newer algorithms and explanation techniques are likely to emerge, potentially altering comparative findings. As the field of explainable AI evolves, future research must revisit these comparisons to maintain relevance and accuracy.Together, these limitations highlight the need for ongoing adaptation in both dataset selection and methodological tools to keep pace with the dynamic nature of phishing threats and the fast-moving landscape of AI research.

3.4.7 Suggestions for Future Research:

The study outlines several clear directions for advancing explainable AI in cybersecurity. Future work should begin with **user studies** to empirically assess the practical impact of explanations on trust, comprehension, and decision-making. Expanding the scope to include a **wider range of models**, such as deep learning and ensemble methods, alongside diverse explanation techniques like LIME and counterfactuals, will provide richer insights.

Another priority is evaluating models on **more diverse, evolving, and real-time datasets**, ensuring that explainability remains effective against dynamic phishing threats. Researchers should also explore **hybrid approaches** that combine the predictive strength of black-box models with the interpretability of white-box models, striking a balance between accuracy and transparency.

In addition, the integration of **automated machine learning (AutoML)** offers potential to optimize both performance and explainability simultaneously. Addressing **computational cost** is equally important, with strategies such as model compression or more efficient implementations for white-box models helping to improve scalability. Finally, the development of **standardized and objective metrics**—including user-centric measures—will be essential for consistently evaluating explanation quality across systems. Together, these directions form a roadmap for building phishing detection systems that are not only technically robust but also transparent, efficient, and user-focused. By acknowledging these limitations, we hope to provide a clear and honest assessment of our work and to inspire future research that addresses these gaps.

4 Conclusion

EBM and XGBoost work about the same on all datasets when it comes to accuracy, simplicity, coverage, uniformity, and stability. Because its structure is naturally interpretable, EBM is a little easier to understand than XGBoost. On the other hand, XGBoost gives up simplicity for more scalability and freedom. EBM is better than XGBoost when it comes to being stable, actionable, and accurate. Both models are comparable in terms of interpretability and consistency, with room for improvement in diverse datasets. **EBM** is generally better suited for applications requiring explainability and actionable insights. Finally, each model, white-box and black-box model has positive and negative aspects both for performance metric and for explainable metric. It is important to consider the objective of model usage.

For developers to successfully use machine learning models that can be trusted to spot phishing, they should think about the following practical implications:

1) Iterative Model Refinement: Keep an eye on the phishing detection system's performance and user comments all the time to make sure it stays reliable and trustworthy.
2) Contextual Explanations: Make sure the explanations fit the needs and level of professional knowledge of the people who will be using them.
3) Integrated Workflow: The interpretable phishing detection system can be easily added to end users' existing tools and processes.
4) Collaborative Development: Talk to end users during the whole development process to get their comments and use their ideas.

Ongoing Training and Support: Give end users thorough training and support tools to help them understand what the interpretable phishing detection system can and can't do.

Acknowledgment

The author would like to thank Universitas Telkom, Bandung, Indonesia, for providing the academic environment and resources that supported this research. The author also gratefully acknowledges the open-source community for developing and maintaining the tools and libraries (including XGBoost, InterpretML, and SHAP) that made this study possible.

Declarations

Author contribution.

The contributions of the author to this work are as follows:

- **Conceptualization:** Abdullah. Fajar
- **Methodology:** Abdullah. Fajar
- **Software:** Abdullah. Fajar
- **Validation:** Abdullah. Fajar
- **Formal Analysis:** Abdullah. Fajar, Setiadi Yazid
- **Investigation:** Abdullah Fajar

- **Data Curation:** Abdulah Fajar, Setiadi Yazid
- **Writing – Original Draft Preparation:** Abdullah Fajar
- **Writing – Review & Editing:** Abdullah Fajar
- **Visualization:** Abdullah Fajar
- **Supervision:** Setiadi Yazid, Indra Budi
- **Project Administration:** Abdullah Fajar

Funding statement.
This research received no external funding. The author did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest.
The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The phishing detection datasets analyzed during this study were obtained from publicly available sources, including the Kaggle.com and other open data platforms. The specific datasets used are listed in subsection 2.3.1 regarding dataset details.

No new datasets were generated during the current study. The processed versions of these datasets, including feature-engineered representations used for model training and evaluation, are available from the corresponding author upon reasonable request.

References

[1] H. Lakkaraju and O. Bastani, "'How do I fool you?': Manipulating User Trust via Misleading Black Box Explanations," *Cornell University*. Jan. 2019. doi: 10.48550/arxiv.1911.06473.

[2] M. Das, S. Saraswathi, R. Panda, A. K. Mishra, and A. K. Tripathy, "Exquisite Analysis of Popular Machine Learning–Based Phishing Detection Techniques for Cyber Systems," *Taylor & Francis*, vol. 16, no. 4, pp. 538–562, Sep. 2020, doi: 10.1080/19361610.2020.1816440.

[3] S. R. Islam, W. Eberle, S. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," *Cornell University*. Jan. 2021. doi: 10.48550/arxiv.2101.09429.

[4] W. Saeed and C. Omlin, "Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities," *Knowl. Based Syst.*, vol. 263, p. 110273, 2021, doi: 10.1016/j.knosys.2023.110273.

[5] A. Nadeem, D. Vos, C. Cao, L. Pajola, and ..., "Sok: Explainable machine learning for computer security applications," *2023 IEEE 8th ...*, 2023, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10190524/

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *Cornell University*. Jan. 2016. doi: 10.48550/arXiv.1602.

[7] A. Warnecke, D. J. Arp, C. Wressnegger, and K. Rieck, "Don't Paint It Black: White-Box Explanations for Deep Learning in Computer Security.," *Cornell University*, Jun. 2019, [Online]. Available: https://arxiv.org/abs/1906.02108v1

[8] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges," *Cornell University*. Jan. 2021. doi: 10.48550/arXiv.2103.

[9] F. Charmet, T. Morikawa, A. Tanaka, and T. Takahashi, "VORTEX : Visual phishing detectiOns aRe Through EXplanations," *ACM Trans. Internet Technol.*, vol. 24, no. 2, pp. 1–24, May 2024, doi: 10.1145/3654665.

[10] G. Ramesh, "Intelligent explanation generation system for phishing webpages by employing an inference system," *Behaviour and Information Technology*, vol. 36, no. 12, pp. 1244–1260, 2017, doi: 10.1080/0144929X.2017.1369569.

[11] V. Nguyen, T. Wu, X. Yuan, M. Grobler, S. Nepal, and C. Rudolph, "An Innovative Information Theory-based Approach to Tackle and Enhance The Transparency in Phishing Detection," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268032530

[12] S. Mittal, "Explaining URL Phishing Detection by Glass Box Models," *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:263147290

[13] M. C. Calzarossa, P. Giudici, and R. Zieni, "Explainable Machine Learning for Bag of Words-Based Phishing Detection," *... on Explainable Artificial Intelligence*, 2023, doi: 10.1007/978-3-031-44064-9_28.

[14] G. Desolda, J. Aneke, C. Ardito, R. Lanzilotti, and ..., "Explanations in warning dialogs to help users defend against phishing attacks," *International Journal of ...*, 2023, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581923000654

[15] F. Greco, "An explainable AI model to help users avoid being victims of phishing attacks," *Master Degree in Computer Science, Computer ...*. 2022.

[16] undefined, "The State of Phishing." Apr. 2023. Accessed: Nov. 12, 2024. [Online]. Available: https://www.slashnext.com/wp-content/uploads/2022/10/SlashNext-The-State-of-Phishing-2022.pdf

[17] "Phishing Trends 2023." [Online]. Available: https://get.zerofox.com/rs/143-DHV-007/images/ZeroFox-Intelligence-Assessment-2023Phishing-Trends-Report.pdf

[18] S. Sedaghat, "CERT strategy to deal with phishing attacks," *Cornell University*. Jan. 2017. doi: 10.48550/arXiv.1706.

[19] F. Carroll, J. A. Adejobi, and R. Montasari, "How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society," *SN Computer Science*, vol. 3, no. 2, p. 170, Feb. 2022, doi: 10.1007/s42979022-01069-1.

[20] M. Bitaab *et al.*, "Scam Pandemic: How Attackers Exploit Public Fear through Phishing," Nov. 2020, doi: 10.1109/ecrime51433.2020.9493260.

[21] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Association for Computing Machinery*, vol. 63, no. 1, pp. 68– 77, Dec. 2019, doi: 10.1145/3359786.

[22] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," *Cornell University*. Jan. 2018. doi: 10.48550/arxiv.1806.00069.

[23] undefined, "InterpretML." Jan. 2023. Accessed: Nov. 12, 2024. [Online]. Available: https://interpret.ml/

[24] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and Klaus Müller, "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning," *Explainable AI*, 2019.

[25] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding Global Feature Contributions With Additive Importance Measures," *Cornell University*. Jan. 2020. doi: 10.48550/arXiv.2004.

[26] W. Fang, J. Zhou, X. Li, and K. Q. Zhu, "Unpack Local Model Interpretation for GBDT," *Springer Science+Business Media*. pp. 764–775, Jan. 2018. doi: 10.1007/978-3-319-91458-9_48.

[27] J. Huysmans, B. Baesens, and J. Vanthienen, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models," *RELX Group (Netherlands)*, Jan. 2006, doi: 10.2139/ssrn.961358.

[28] C. Molnar *et al.*, "General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models," *Cornell University*. Jan. 2020. doi: 10.48550/arxiv.2007.04131.

[29] C. Rudin, "Please Stop Explaining Black Box Models for High Stakes Decisions.," *Cornell University*, Nov. 2018, [Online]. Available: https://arxiv.org/pdf/1811.10154.pdf

[30] M. Hagenbuchner, "The black box problem of AI in oncology," *IOP Publishing*, vol. 1662, no. 1, pp. 012012–012012, Oct. 2020, doi: 10.1088/1742-6596/1662/1/012012.

[31] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and Customizable Explanations of Black Box Models," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, in AIES '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 131–138. doi: 10.1145/3306618.3314229.

[32] R. Wash, "How Experts Detect Phishing Scam Emails," *Proceedings of the ACM on human-computer interaction*, vol. 4, no. CSCW2. pp. 1–28, Oct. 2020. doi: 10.1145/3415231.

[33] T. Feng and C. Yue, "Visualizing and Interpreting RNN Models in URL-Based Phishing Detection," in *Proceedings of the 25th ACM Symposium on Access Control Models and Technologies*, in SACMAT '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 13–24. doi: 10.1145/3381991.3395602.

[34] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019, doi: 10.1073/pnas.1900654116.

[35] A. Das, S. Baki, A. E. Aassal, R. M. Verma, and A. Dunbar, "SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective," *IEEE Communications Surveys & Tutorials*, vol. 22, pp. 671–708, 2019, doi: 10.1109/COMST.2019.2957750.

[36] H. Dong, B. Liu, D. Ye, and G. Liu, "Interpretability as Approximation: Understanding Black-Box Models by Decision Boundary," *Electronics*, vol. 13, no. 22, 2024, doi: 10.3390/electronics13224339.

[37] K. Letrache and M. Ramdani, "Explainable Artificial Intelligence: A Review and Case Study on Model-Agnostic Methods," in *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2023, pp. 1–8. doi: 10.1109/SITA60746.2023.10373722.

[38] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, and ..., "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive …*. Springer, 2023. doi: 10.1007/s12559-023-10179-8.

[39] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, Sep. 2023, doi: 10.1007/s10618-023-00933-9.

[40] M. U. Islam, Md. Mozaharul Mottalib, M. Hassan, Z. I. Alam, S. M. Zobaed, and Md. Fazle Rabby, "The Past, Present, and Prospective Future of XAI: A Comprehensive Review," in *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*,

[41] M. Ahmed, S. R. Islam, A. Anwar, N. Moustafa, and A.-S. K. Pathan, Eds., Cham: Springer International Publishing, 2022, pp. 1–29. doi: 10.1007/978-3-030-96630-0_1.

[42] R. R. Hoffman, S. T. Mueller, G. Klein, and ..., "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," *Frontiers in Computer …*. frontiersin.org, 2023. doi: 10.3389/fcomp.2023.1096257.

[43] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[44] Q. Zhou, F. Liao, C. Mou, and P. Wang, "Measuring Interpretability for Different Types of Machine Learning Models," in *Trends and Applications in Knowledge Discovery and Data Mining*, M. Ganji, L. Rashidi, B. C. M. Fung, and C. Wang, Eds., Cham: Springer International Publishing, 2018, pp. 295–308.