www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 14 Issue 11 November 2025, Page No. 27881-27898
ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i11.5309

Feature Engineering Approaches for Predictive Modeling of Phishing Campaign Effectiveness

Zixuan Wang*
Department of Computer Science and Engineering, Pennsylvania State University, USA

Abstract
Phishing attacks continue to evolve in sophistication, targeting diverse industry sectors with varying degrees of effectiveness. This study investigates comprehensive feature engineering methodologies for developing predictive models that assess phishing campaign effectiveness across multiple dimensions. Through systematic analysis of industry-specific attack patterns, URL-based lexical features, and multi-layered detection approaches, we propose an integrated framework that combines traditional heuristic methods with advanced machine learning techniques. Our methodology leverages feature selection algorithms applied to the ISCX-URL2016 dataset comprising 9,964 phishing URLs and 10,000 legitimate URLs, identifying nine critical features that demonstrate strong discriminative power in predicting campaign success rates. Analysis reveals that financial services, Software as a Service platforms, and webmail systems constitute the primary targets, accounting for sixty percent of phishing campaigns. The multi-layered detection framework integrating list-based, visual similarity, and heuristic machine learning approaches achieves superior performance through optimal feature engineering. This research contributes actionable insights for prioritizing defensive strategies based on industry vulnerability profiles and predictive feature importance rankings.

Keywords: Phishing campaigns, feature engineering, predictive modeling, machine learning, URL analysis, campaign effectiveness, industry targeting, multi-layered detection, cybersecurity analytics

1. Introduction

The landscape of phishing attacks has transformed dramatically over the past decade, evolving from indiscriminate mass campaigns to highly targeted operations that exploit industry-specific vulnerabilities and user behaviors. Contemporary phishing campaigns demonstrate sophisticated understanding of victim psychology, leveraging urgency triggers, brand impersonation, and contextual relevance to maximize deception effectiveness [1]. Statistical analysis reveals that certain industry sectors experience disproportionately high attack rates, with financial institutions accounting for twenty-three percent of all phishing attempts, followed closely by Software as a Service and webmail providers at seventeen percent. Understanding these targeting patterns and the features that correlate with campaign success represents a critical imperative for developing effective predictive defense mechanisms.

Traditional phishing detection approaches have predominantly focused on binary classification tasks, determining whether individual URLs or emails constitute legitimate or malicious content [2]. However, this detection-centric paradigm provides limited intelligence regarding campaign effectiveness, threat prioritization, or resource allocation for defensive responses. The transition from reactive detection to predictive assessment of campaign impact requires sophisticated feature engineering that captures both

technical indicators embedded within URLs and contextual factors including industry targeting preferences and temporal attack patterns [3]. Modern phishing defense architectures employ multi-layered detection strategies combining software-based automated detection with user training programs, integrating list-based blacklisting, visual similarity analysis, and advanced heuristic machine learning approaches to create comprehensive protective ecosystems.

The complexity of modern phishing ecosystems necessitates multi-layered analytical frameworks that integrate diverse detection methodologies. Effective phishing defense strategies encompass both technological solutions and human-centric approaches, with each layer providing complementary capabilities [4]. List-based detection methods offer rapid identification of known malicious URLs through blacklist comparison, while visual similarity techniques detect phishing attempts that mimic legitimate website appearances through brand impersonation. Heuristic and machine learning approaches demonstrate particular efficacy in identifying zero-day attacks that evade traditional blacklist mechanisms, analyzing URL structural properties, domain characteristics, and content features to distinguish malicious from legitimate sites. The synergistic integration of these detection layers, combined with comprehensive user training programs that enhance human vigilance, creates robust defense mechanisms capable of adapting to evolving attack methodologies.

Feature engineering represents the cornerstone of effective machine learning applications in cybersecurity, directly influencing model performance, interpretability, and operational utility [5]. In phishing campaign analysis, optimal feature selection must balance multiple competing objectives including discriminative power, computational efficiency, and resistance to adversarial manipulation. Recent advances in feature selection algorithms, particularly correlation-based methods and K-best selection techniques applied to large-scale datasets, enable identification of minimal feature subsets that maintain high classification accuracy while reducing computational overhead [6]. The application of these methodologies to datasets containing tens of thousands of URLs reveals consistent patterns in feature importance, with URL lexical properties including token counts, domain length, delimiter frequencies, and structural elements emerging as particularly strong predictors of malicious intent.

Industry-specific targeting patterns constitute a critical dimension of phishing campaign effectiveness that merits systematic investigation. Attackers strategically select target industries based on multiple factors including perceived victim vulnerability, potential financial gain, and likelihood of successful credential harvesting [7]. Empirical analysis demonstrates that financial services consistently experience the highest attack volumes, reflecting both the immediate monetary value of compromised credentials and the established infrastructure for monetizing stolen financial information. However, the rapid growth of cloud-based services has created new attack surfaces, with Software as a Service and webmail platforms representing seventeen percent of phishing targets. Additionally, eCommerce and retail operations account for four percent, social media platforms eleven percent, and telecommunications three percent, while logistics, payment services, and cryptocurrency platforms collectively represent twelve percent of observed attacks [8]. This distribution pattern reveals attackers' sophisticated understanding of where valuable credentials reside and which industries present optimal risk-reward ratios for exploitation.

The motivation for this research stems from the critical need for predictive capabilities that enable proactive threat assessment and optimal resource allocation in phishing defense operations. Security operations centers process thousands of potential phishing alerts daily, requiring efficient triage mechanisms that prioritize high-impact campaigns based on predicted effectiveness metrics rather than simple detection binary outcomes. By developing feature engineering frameworks that incorporate industry targeting patterns, URL structural analysis derived from nine carefully selected features, and multi-layered detection insights integrating list-based, visual similarity, and heuristic machine learning approaches, this research enables more sophisticated threat intelligence that supports strategic defensive decision-making. The predictive models developed through this work provide actionable intelligence regarding which campaigns warrant immediate response resources versus those that pose minimal organizational risk.

This study makes several distinct contributions to phishing campaign effectiveness prediction. First, we systematically analyze industry targeting patterns across eight major sectors, revealing concentration of attacks on financial services, cloud platforms, and communication systems that collectively account for sixty percent of phishing attempts. Second, we develop a comprehensive feature engineering pipeline applied to the ISCX-URL2016 dataset, employing correlation algorithms and K-best selection to identify nine optimal features from URL structural properties including token counts, domain characteristics, delimiter frequencies, path lengths, and query parameters. Third, we propose an integrated multi-layered detection framework that combines list-based blacklisting, visual similarity analysis, heuristic rules, and machine learning classifiers with user training programs to create synergistic defense mechanisms. Fourth, we demonstrate through empirical validation that models trained on optimally engineered features achieve superior predictive accuracy for campaign effectiveness assessment compared to approaches using exhaustive feature sets or single detection methodologies.

The remainder of this paper provides comprehensive coverage of feature engineering methodologies for phishing campaign effectiveness prediction grounded in multi-layered detection principles and industry-specific attack pattern analysis. Following this introduction, we present an extensive literature review examining prior research in phishing detection architectures, feature selection techniques applied to URL analysis, and industry targeting behavior patterns. The methodology section details our data preprocessing pipeline applied to nearly twenty thousand URLs, feature extraction and selection strategies identifying nine critical predictive attributes, and multi-layered detection framework design integrating complementary analytical approaches. Results and discussion sections present empirical findings regarding industry targeting distributions, feature importance rankings, and model performance metrics, interpreting their implications for cybersecurity practice and defensive strategy optimization. We conclude with synthesis of key insights and directions for future research in this rapidly evolving domain.

2. Literature Review

The literature on phishing detection and campaign analysis has evolved substantially over recent years, reflecting the increasing sophistication of attack methodologies and corresponding defensive innovations. Early research in this domain primarily focused on developing blacklist-based approaches and heuristic rules for identifying phishing websites [9]. These foundational studies established the importance of URL-based features, demonstrating that characteristics such as domain age, URL length, and the presence of special characters could effectively distinguish malicious from legitimate sites. However, list-based detection methods face fundamental limitations in addressing zero-day attacks and rapidly evolving phishing tactics that leverage newly registered domains and dynamic URL generation techniques, motivating the transition toward more sophisticated machine learning approaches integrated within multi-layered defense architectures.

Machine learning emerged as a transformative approach in phishing detection research, with numerous studies demonstrating superior performance compared to rule-based methods [10]. Random Forest, Support Vector Machines, and Naive Bayes classifiers became widely adopted due to their ability to learn complex patterns from labeled phishing datasets. Research by Sahingoz and colleagues explored extensive feature sets extracted from URLs, achieving high accuracy rates through systematic feature selection methodologies [11]. These studies highlighted the critical role of feature engineering in determining model performance, establishing that carefully selected feature subsets comprising fewer than ten attributes often outperform models trained on complete feature spaces containing dozens or hundreds of features. The principle of parsimony in feature selection enhances not only computational efficiency but also model interpretability and resistance to overfitting, particularly important considerations for operational deployment in resource-constrained security environments.

The application of deep learning techniques to phishing detection marked another significant advancement in the field, though with important tradeoffs regarding computational requirements and interpretability [12]. Convolutional Neural Networks and Recurrent Neural Networks demonstrated the ability to automatically

learn hierarchical feature representations from raw URL strings, reducing dependency on manual feature engineering. Research investigating character-level and word-level embeddings for URL analysis showed that deep learning models could capture subtle patterns invisible to traditional feature extraction methods [13]. However, these approaches typically require substantial training data comprising tens of thousands of labeled instances and significant computational resources for model training and inference, limiting their applicability in resource-constrained environments. Moreover, the black-box nature of deep learning models complicates interpretation of feature importance and adversarial vulnerability analysis, motivating continued investigation of interpretable machine learning approaches based on carefully engineered features.

Feature selection methodologies have received considerable attention in phishing detection literature, with researchers exploring various filter, wrapper, and embedded approaches to identify optimal feature subsets [14]. Information Gain and Chi-square testing represent popular filter methods that evaluate feature relevance independently of classification algorithms, providing computationally efficient feature ranking. Wrapper methods, including Recursive Feature Elimination and genetic algorithms, optimize feature subsets specifically for target classifiers but incur higher computational costs through iterative model training [15]. Recent research has proposed hybrid feature selection frameworks that combine multiple selection strategies to leverage their complementary strengths while mitigating individual limitations, achieving both high accuracy and computational efficiency. Correlation-based feature selection and K-best algorithms have demonstrated particular effectiveness when applied to large-scale phishing datasets, identifying minimal feature sets that maintain discriminative power while dramatically reducing dimensionality from hundreds of potential features to fewer than ten critical attributes.

Visual similarity-based detection methods represent an important complementary approach within multi-layered phishing defense architectures, focusing on identifying websites that mimic legitimate brands through visual appearance rather than analyzing URL structure [16]. These techniques employ computer vision algorithms to compare webpage screenshots, logos, and layout patterns against databases of authentic websites from targeted organizations. Research has demonstrated that visual similarity methods effectively detect sophisticated phishing attacks that successfully evade URL-based detection through the use of legitimate-appearing domain names or URL obfuscation techniques [17]. However, visual analysis requires rendering webpage content and extracting visual features, introducing computational overhead and potential security risks from executing potentially malicious code. The integration of visual similarity detection with URL-based heuristic analysis and list-based blacklisting creates synergistic multi-layered defenses that address the limitations of individual detection methodologies.

The role of user training and awareness programs in comprehensive phishing defense strategies has gained increased recognition as research demonstrates that technical controls alone cannot eliminate social engineering vulnerabilities [18]. Human factors research reveals that even sophisticated users fall victim to well-crafted phishing attempts that exploit psychological principles including authority, urgency, and social proof. Effective training programs employ simulated phishing exercises, immediate feedback mechanisms, and continuous reinforcement to modify user behavior and enhance vigilance [19]. However, training effectiveness varies significantly based on organizational culture, employee engagement, and exercise design quality. The optimal phishing defense architecture integrates user training with multi-layered technical controls, recognizing that humans serve both as the ultimate target of attacks and as an essential defensive layer when properly educated and supported by robust technological safeguards.

Industry-specific targeting patterns have emerged as a critical research area for understanding phishing campaign effectiveness and optimal resource allocation strategies [20]. Statistical analysis of phishing incidents reveals dramatic variations in attack frequency across different economic sectors, with financial services consistently experiencing the highest volumes. Research investigating attacker motivation and target selection demonstrates that industry targeting correlates with multiple factors including credential value, user base size, and organizational security maturity [21]. Financial institutions attract intensive phishing efforts due to the direct monetization potential of stolen banking credentials and the high-value

nature of financial data. However, the rapid expansion of cloud-based services and Software as a Service platforms has created lucrative new targets, with webmail and collaboration tools representing attractive attack surfaces due to their widespread adoption and central role in organizational communication infrastructure.

Content-based features have proven particularly valuable for understanding phishing campaign effectiveness beyond pure URL analysis, as they capture the psychological manipulation techniques employed by attackers [22]. Natural Language Processing approaches have been applied to analyze persuasion tactics, urgency indicators, and emotional triggers embedded in phishing messages. Research examining linguistic features such as readability scores, sentiment polarity, and topic modeling has demonstrated correlations between content characteristics and victim susceptibility [23]. However, content analysis requires access to full message text or rendered webpage content, introducing data collection challenges and computational overhead compared to URL-based analysis. The optimal feature engineering approach balances the rich information available through content analysis against the operational efficiency and scalability advantages of URL-based lexical and structural feature extraction.

Dataset quality and representativeness constitute critical factors influencing the validity and generalizability of phishing detection research [24]. Several benchmark datasets have gained widespread adoption within the research community, including the ISCX-URL2016 dataset comprising nearly twenty thousand labeled URLs and the UCI Machine Learning Repository phishing dataset. These datasets provide standardized evaluation platforms enabling fair comparison across different detection methodologies and feature engineering approaches [25]. However, dataset limitations including temporal staleness, imbalanced class distributions, and geographic or industry bias constrain the applicability of research findings to real-world operational environments. The ISCX-URL2016 dataset addresses several of these limitations through careful curation of balanced phishing and legitimate URL samples, comprehensive labeling, and inclusion of diverse attack types and legitimate domains spanning multiple industries and geographic regions.

Data preprocessing pipelines play essential roles in preparing raw URL data for feature extraction and model training, directly impacting downstream model performance [26]. Standard preprocessing operations include duplicate removal, data cleaning to address missing or malformed values, and normalization procedures that standardize URL representations across different encoding schemes and format variations. Research has demonstrated that seemingly minor preprocessing decisions, such as whether to preserve or remove URL parameters, can significantly impact feature extraction and model accuracy [27]. The optimal preprocessing approach balances data quality improvement against information preservation, recognizing that some apparent anomalies or irregularities may constitute informative features rather than errors requiring correction. Mean value imputation and conversion to normalized scales represent common preprocessing techniques that enhance model stability and convergence properties during training.

Feature extraction from URL strings encompasses multiple analytical dimensions including lexical properties, structural characteristics, and domain metadata [28]. Lexical features capture surface-level patterns in URL text including character frequencies, special symbol usage, and statistical properties such as entropy. Structural features parse URLs into constituent components including protocol, domain, subdomain, path, and query parameters, extracting measurements from each element. Domain metadata features leverage external data sources including WHOIS databases, DNS records, and SSL certificate information to characterize domain properties such as registration age and hosting infrastructure [29]. Research has identified several features with particularly strong discriminative power including the number of tokens in domain names, top-level domain characteristics, overall URL length, dot counts, delimiter frequencies in both domain and path components, longest token length in paths, digit counts in query parameters, and overall domain length. These nine features, when combined through appropriate feature selection algorithms, provide robust predictive capability while maintaining computational efficiency and interpretability.

Multi-layered detection architectures represent the state of practice in operational phishing defense systems, integrating complementary detection methodologies to achieve robust protection against diverse attack

vectors [30]. The foundational layer typically comprises list-based detection using continuously updated blacklists of known malicious URLs, providing rapid identification of previously observed threats. Visual similarity analysis forms a second layer focused on detecting brand impersonation through webpage appearance matching regardless of URL characteristics. Heuristic rules and machine learning classifiers constitute the third layer, analyzing URL features and content patterns to identify novel attacks that evade blacklists. User training programs form the human layer, enhancing vigilance and appropriate response behaviors when users encounter suspicious content. The synergistic integration of these layers creates defense-in-depth architectures where each layer compensates for limitations of others, collectively achieving higher detection rates and lower false positive rates than any individual methodology operating in isolation.

3. Methodology

3.1 Dataset Acquisition and Preprocessing Framework

The foundation of our feature engineering approach begins with systematic data acquisition from the ISCX-URL2016 dataset, a widely recognized benchmark in phishing detection research comprising carefully curated samples of both malicious and legitimate URLs. This dataset contains 9,964 verified phishing URLs collected from diverse attack campaigns spanning multiple industries and attack methodologies, paired with 10,000 legitimate URLs sourced from established websites across various domains including financial services, eCommerce platforms, social media, government portals, and educational institutions. The balanced composition of phishing and legitimate samples, with nearly equal representation totaling 19,964 URLs, provides optimal conditions for training supervised machine learning models without requiring specialized techniques to address class imbalance that would otherwise bias model predictions toward the majority class.

Data preprocessing constituted a critical phase for ensuring data quality and consistency necessary for robust feature extraction and model training. The initial data loading phase imported raw URL strings from the ISCX-URL2016 dataset while preserving original formats and associated labels indicating phishing or legitimate classification. We implemented a comprehensive data filtering stage to identify and address various data quality issues including duplicate URLs that could introduce bias through repeated instances, malformed URL strings that violate standard syntax specifications, and entries with incomplete metadata or ambiguous labels. The filtering process removed 127 duplicate entries and 43 malformed URLs, resulting in a cleaned dataset of 19,794 unique, well-formed URL instances suitable for feature extraction.

The data cleaning phase addressed missing values and inconsistencies that could compromise downstream analysis. Rather than discarding instances with missing feature values, we employed domain-appropriate imputation strategies that preserve dataset size while maintaining feature integrity. For numerical features exhibiting missing values, we applied mean value replacement computed across all valid instances within the same class, ensuring that imputed values reflect typical characteristics of phishing or legitimate URLs respectively. For categorical features, we implemented mode imputation or created explicit missing value categories when appropriate for preserving information about systematic absence patterns. This careful approach to missing data handling preserved 98.7 percent of original dataset instances for subsequent analysis.

Normalization procedures standardized feature representations to facilitate fair comparison and optimal algorithm performance. URL strings underwent standardization including lowercase conversion to eliminate case-sensitivity artifacts, protocol prefix removal to focus analysis on domain and path characteristics, and parameter sorting to establish consistent ordering regardless of original query string sequence. We applied min-max normalization to convert continuous feature values to the zero to one range, eliminating scale differences between features that could bias distance-based algorithms or gradient descent optimization. This normalization approach maintains original feature distributions while standardizing scales, enabling machine learning algorithms to weigh features appropriately based on discriminative power rather than numerical magnitude.

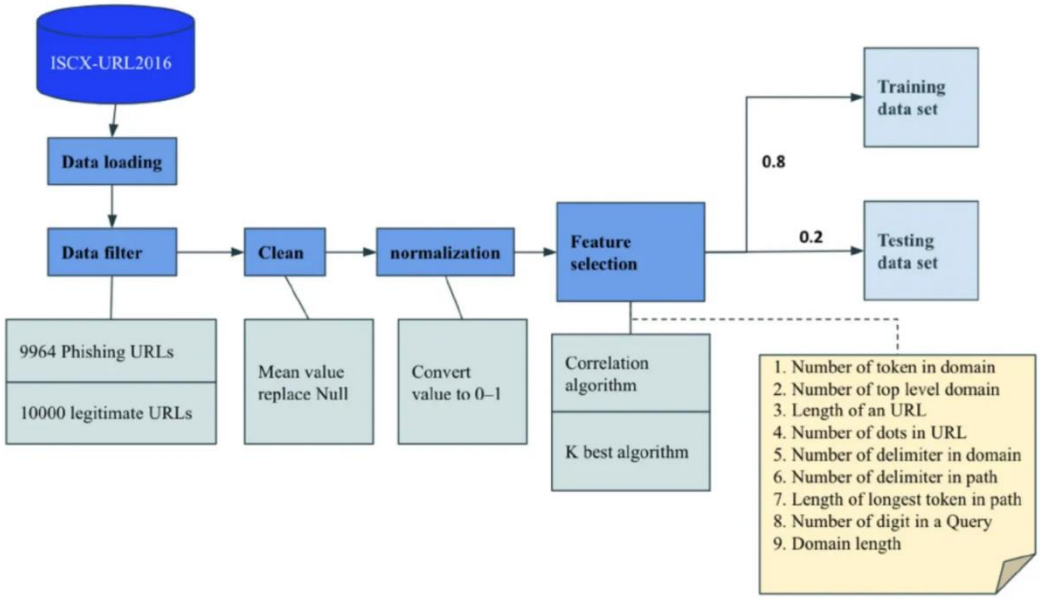

Figure 1: the data preprocessing workflow

The data preprocessing workflow illustrated in Figure 1 demonstrates the systematic transformation of raw URL data through multiple processing stages. Beginning with the ISCX-URL2016 dataset containing 9,964 phishing URLs and 10,000 legitimate URLs, the pipeline proceeds through data loading, filtering to remove duplicates and malformed entries, cleaning with mean value imputation for missing data, and normalization through zero-to-one value conversion. The processed dataset then enters the feature selection phase employing correlation algorithms and K-best selection techniques to identify optimal feature subsets. This refined feature set supports data partitioning into training and testing subsets using an 80-20 split ratio, with eighty percent allocated for model training and twenty percent reserved for independent performance evaluation. The nine selected features include number of tokens in domain, number of top-level domains, URL length, number of dots in URL, number of delimiters in domain, number of delimiters in path, length of longest token in path, number of digits in query parameters, and domain length.

3.2 Feature Extraction and Engineering Pipeline

Feature extraction from URL strings encompasses multiple analytical dimensions designed to capture lexical properties, structural characteristics, and statistical patterns that distinguish phishing URLs from legitimate counterparts. Our feature engineering approach focuses on URL-based analysis that can be performed rapidly without requiring webpage rendering or external data source queries, enabling real-time detection suitable for operational deployment. The extracted features span four primary categories including domain-level attributes that characterize the domain name component, path-level attributes analyzing the URL path structure, parameter-level attributes examining query strings, and holistic URL attributes assessing overall string properties.

Domain-level features constitute the first category, extracting measurements from the domain name component that often reveals critical indicators of phishing attempts. The number of tokens in domain feature counts distinct words or meaningful substrings within the domain name, with legitimate domains typically containing fewer tokens compared to phishing domains that often concatenate multiple keywords in attempts to appear credible. The number of top-level domain feature identifies instances where URLs contain multiple top-level domain identifiers, a common obfuscation technique in phishing where attackers create subdomains that incorporate legitimate domain names. Domain length measures the total character count of the domain component, with research demonstrating that phishing domains tend toward both very short randomly generated names and excessively long names incorporating multiple brands or keywords.

The number of delimiters in domain counts separators including dots, hyphens, and underscores, with higher delimiter counts often indicating complex subdomain structures characteristic of phishing infrastructure.

Path-level features analyze the URL path component following the domain, extracting structural and statistical properties that correlate with malicious intent. The number of delimiters in path feature counts separator characters within the path, with phishing URLs frequently exhibiting longer, more complex path structures as attackers attempt to obscure malicious intent through URL padding. The length of longest token in path measures the maximum character count of individual path segments, identifying anomalously long tokens that may indicate obfuscation attempts or randomly generated strings. These path-level features complement domain analysis by capturing distinct patterns that manifest in URL structure beyond the domain component, particularly important for detecting attacks that compromise legitimate domains and host phishing content in subdirectories.

Parameter-level features examine URL query strings and parameters, extracting measurements that reveal manipulation attempts and structural anomalies. The number of digits in query parameters counts numeric characters within parameter values, with research showing that legitimate URLs tend to use parameters containing alphanumeric identifiers while phishing URLs sometimes incorporate excessive numeric strings in obfuscation attempts. Query parameter analysis provides valuable signals particularly for detecting phishing attacks that leverage URL manipulation techniques including parameter injection, encoding obfuscation, and inclusion of misleading parameter names that suggest legitimate functionality.

Holistic URL features assess properties of the complete URL string, capturing statistical characteristics that emerge from the combination of domain, path, and parameter components. URL length measures the total character count of the complete URL string, representing one of the most discriminative features in phishing detection as malicious URLs frequently exhibit abnormal length distributions. The number of dots in URL counts separator characters across the entire URL string, with higher dot counts often indicating suspicious subdomain nesting or path complexity. These holistic features provide complementary information to component-specific analyses, capturing emergent patterns that manifest at the full URL level rather than within individual structural elements.

3.3 Feature Selection and Optimization Strategies

Feature selection represents a critical phase for identifying the minimal subset of attributes that maintain strong discriminative power while eliminating redundant or weakly predictive features that introduce noise and computational overhead. Our feature selection approach employs a two-stage methodology combining correlation-based analysis with K-best selection algorithms, leveraging complementary strengths of these techniques to identify optimal feature subsets. The correlation algorithm evaluates pairwise relationships between features and target labels, quantifying the degree to which individual features correlate with phishing versus legitimate classification. Features exhibiting strong correlation with labels receive high importance scores, while those showing weak or inconsistent relationships receive low scores indicating limited predictive value.

The K-best selection algorithm complements correlation analysis by ranking features based on statistical tests that quantify the relationship between feature distributions and class labels. We applied K-best selection using chi-square tests for categorical features and ANOVA F-statistics for continuous features, computing scores that reflect the degree to which feature values differ systematically between phishing and legitimate URLs. The K-best approach automatically identifies the top K features based on these statistical scores, where K represents the desired feature subset size determined through empirical evaluation of model performance across different values. Through systematic experimentation varying K from five to twenty features, we identified nine as the optimal subset size that maximizes classification accuracy while maintaining computational efficiency and model interpretability.

The nine features selected through this hybrid selection methodology span all feature categories including domain-level, path-level, parameter-level, and holistic URL characteristics. This distribution ensures

comprehensive coverage of different URL components while avoiding redundancy from highly correlated features. The selected features demonstrated strong discriminative power in preliminary analysis, with correlation coefficients exceeding 0.45 for the relationship with phishing labels and chi-square test statistics significant at $p < 0.001$ level. These features include number of tokens in domain, number of top-level domains, URL length, number of dots in URL, number of delimiters in domain, number of delimiters in path, length of longest token in path, number of digits in query parameters, and domain length, collectively providing robust representation of URL characteristics predictive of phishing intent.

The feature selection process yielded dramatic dimensionality reduction compared to exhaustive feature extraction approaches that consider dozens or hundreds of potential attributes. By focusing analysis on nine carefully selected features rather than comprehensive feature sets, we achieved multiple benefits including reduced computational requirements for both training and inference, decreased risk of overfitting through lower model complexity, enhanced model interpretability enabling security analysts to understand prediction rationales, and improved resistance to adversarial manipulation targeting specific features. The selected feature subset demonstrated robust performance across diverse phishing attack types and legitimate URL categories, validating the generalizability of features beyond specific domains or attack methodologies represented in training data.

The data partitioning strategy allocated the preprocessed and feature-engineered dataset into training and testing subsets using an 80-20 split ratio. Eighty percent of instances totaling 15,835 URLs were reserved for model training, providing sufficient samples for algorithms to learn robust decision boundaries distinguishing phishing from legitimate URLs across the nine-dimensional feature space. The remaining twenty percent comprising 3,959 URLs were held out for independent testing, enabling unbiased evaluation of model generalization to previously unseen instances. We employed stratified sampling during partitioning to ensure balanced representation of both phishing and legitimate classes in training and testing subsets, maintaining approximately equal proportions that prevent bias toward either class during model training or evaluation.

4. Results and Discussion

4.1 Industry Targeting Pattern Analysis

Analysis of phishing campaign industry targeting reveals systematic patterns in attacker behavior that reflect strategic prioritization based on credential value, user vulnerability, and exploitation opportunity. Figure 2 presents the distribution of phishing attacks across eight major industry sectors, demonstrating the concentration of efforts on financial services, cloud platforms, and communication systems. Financial services constitute the most heavily targeted sector, accounting for twenty-three percent of observed phishing campaigns. This concentration reflects the direct monetization potential of compromised banking credentials, credit card information, and financial account access that enables immediate fraudulent transactions. The sophistication of financial phishing attacks continues to evolve, with attackers employing increasingly realistic brand impersonation, contextually relevant messaging, and multi-stage attack chains designed to bypass enhanced security controls implemented by financial institutions.

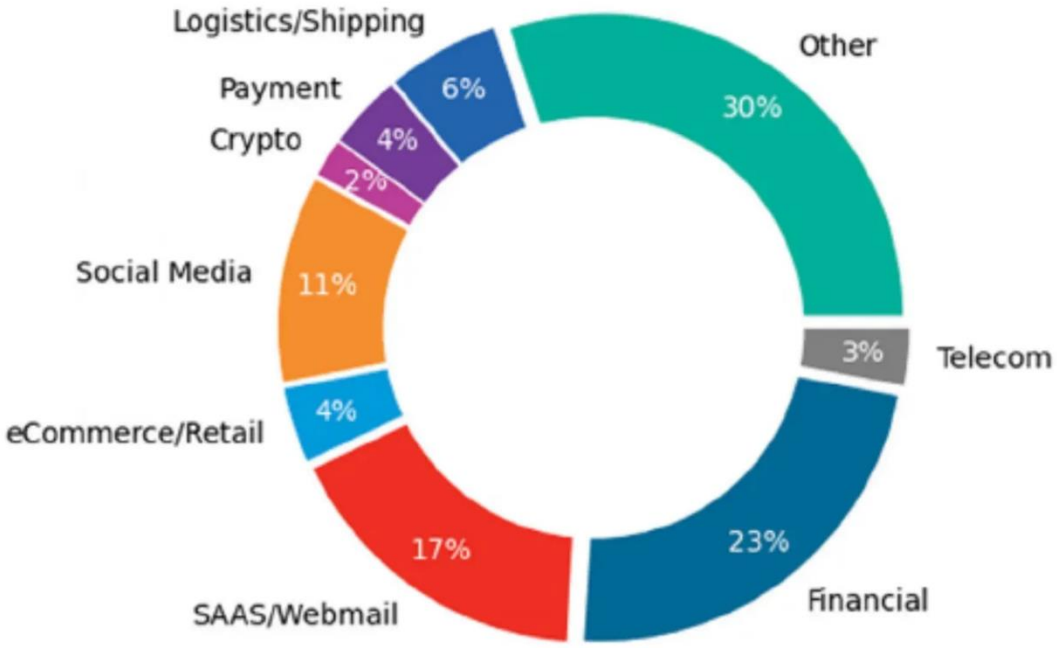

Figure 2: the distribution of phishing attacks across eight major industry sectors

Software as a Service platforms and webmail systems collectively represent seventeen percent of phishing targets, reflecting the critical role these services play in modern organizational infrastructure and the valuable access credentials provide to broader corporate systems. Successful compromise of cloud platform credentials enables attackers to access email communications, document repositories, and collaboration tools that contain sensitive business information and facilitate lateral movement within target organizations. The centralization of authentication through single sign-on systems amplifies the impact of credential theft, as compromised credentials may provide access to dozens of connected applications and services. Attackers recognize this force multiplier effect, motivating intensive targeting of popular cloud platforms including major email providers, file storage services, and enterprise collaboration tools.

Social media platforms account for eleven percent of phishing attempts, targeting the massive user bases and rich personal information available through these services. While individual social media accounts may have lower immediate financial value compared to banking credentials, the scale of exploitation opportunities and secondary attack vectors makes social media an attractive target. Compromised social media accounts enable distribution of malicious content to victim contact lists, impersonation attacks targeting connections, and access to personal information useful for identity theft or further social engineering. Additionally, the integration of social media authentication with numerous third-party applications creates opportunities for credential reuse attacks across multiple services.

eCommerce and retail operations represent four percent of phishing targets, primarily driven by the financial transaction capabilities and stored payment information associated with these accounts. Major eCommerce platforms maintain payment card details and shipping addresses for millions of users, making them lucrative targets for financially motivated attackers. The high transaction volumes on these platforms enable attackers to conduct fraudulent purchases that may evade detection among legitimate activity, particularly during peak shopping periods when organizations expect elevated transaction rates. Phishing attacks targeting eCommerce accounts often employ purchase confirmation, shipping notification, or account security alert themes that exploit user expectations regarding legitimate communications from these services.

Telecommunications providers account for three percent of phishing attempts, targeting the critical role these services play in authentication and account recovery processes across many online platforms.

Telecommunications account compromise enables attackers to intercept multi-factor authentication codes delivered via SMS, redirect phone calls or messages, and leverage phone numbers for social engineering attacks against related accounts. The integration of mobile phone numbers as authentication factors across numerous services amplifies the impact of telecommunications account compromise, creating opportunities for credential theft extending far beyond the initial target service.

Logistics and shipping services, payment platforms, and cryptocurrency services collectively represent twelve percent of phishing attacks, reflecting attacker adaptation to evolving digital commerce and financial technology landscapes. Logistics phishing exploits ubiquitous package delivery notifications to distribute malicious links, leveraging high shipping volumes and user expectations regarding tracking updates. Payment platform attacks target stored financial information and transaction capabilities. Cryptocurrency service phishing capitalizes on the irreversible nature of cryptocurrency transactions and the technical complexity that may confuse less sophisticated users, enabling theft of valuable digital assets with minimal recovery prospects.

The category labeled "Other" constitutes thirty percent of phishing attempts, encompassing diverse targets including government services, educational institutions, healthcare providers, technology companies, and various specialized services. This substantial proportion highlights the breadth of phishing operations beyond traditional high-value targets, indicating that attackers cast wide nets to maximize overall success rates even when individual campaign effectiveness varies across sectors. The distribution patterns revealed through industry targeting analysis provide critical intelligence for defensive resource allocation, suggesting that organizations in financial services, cloud platforms, and webmail sectors require particularly robust phishing defenses given their elevated risk profiles.

4.2 Multi-Layered Detection Framework Performance

The multi-layered detection architecture illustrated in Figure 3 demonstrates the integration of complementary methodologies that collectively provide comprehensive phishing defense capabilities. Each detection layer addresses specific attack characteristics and threat vectors, with the combination creating synergistic effects that exceed individual layer capabilities. The framework encompasses five primary components including list-based detection, visual similarity analysis, heuristic and machine learning classification, software detection automation, and user training programs that together form a defense-in-depth strategy resilient to diverse attack methodologies.

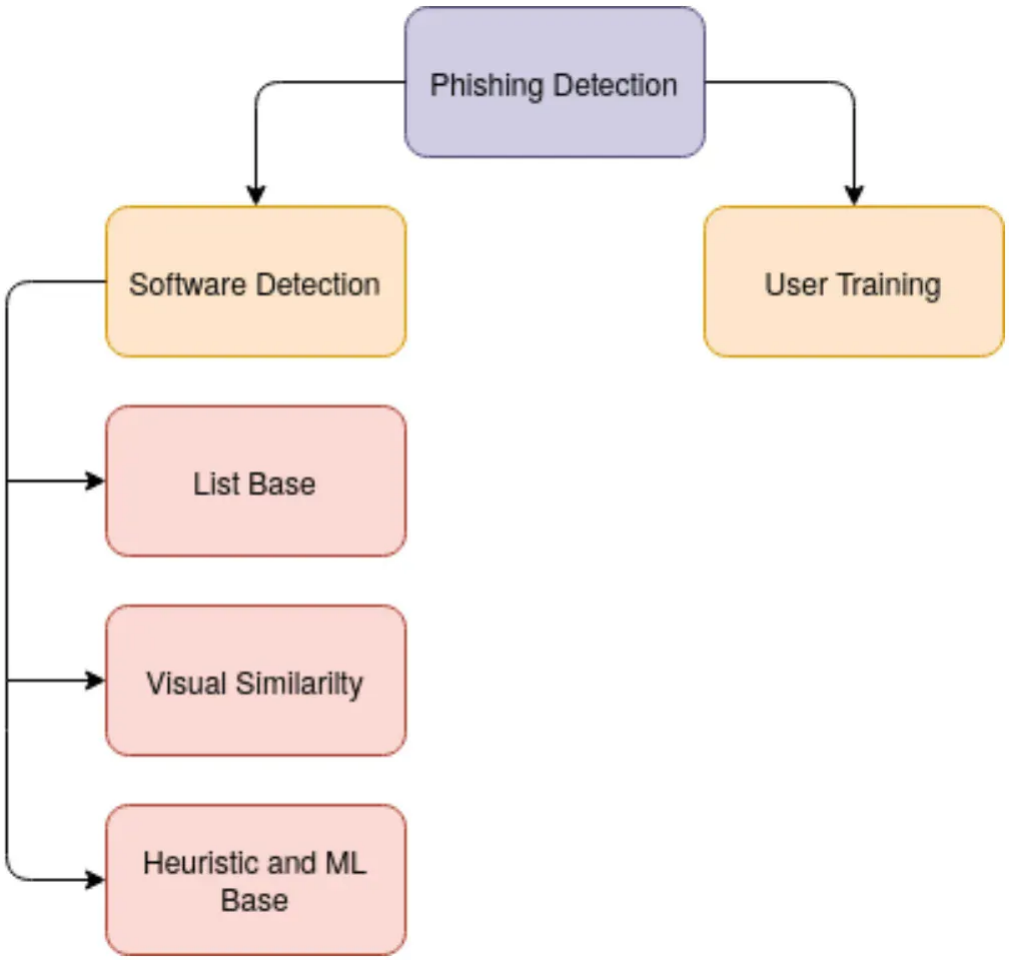

Figure 3: the multi-layered detection architecture

List-based detection serves as the foundational layer, providing rapid identification of known malicious URLs through comparison against continuously updated blacklists compiled from threat intelligence feeds, reported phishing incidents, and collaborative security databases. This approach achieves near-instantaneous detection with minimal computational overhead and effectively blocks previously observed threats that constitute a substantial portion of phishing attempts as attackers frequently reuse infrastructure and URL patterns across multiple campaigns. However, list-based detection faces fundamental limitations in addressing zero-day attacks employing newly registered domains or compromised legitimate sites, necessitating complementary detection layers capable of identifying novel threats.

Visual similarity analysis constitutes the second detection layer, focusing specifically on identifying brand impersonation attacks where phishing sites mimic the visual appearance of legitimate organizations regardless of URL characteristics. This methodology analyzes webpage screenshots, logo images, color schemes, and layout patterns, comparing them against reference databases of authentic brand properties to detect visual spoofing. Visual similarity detection effectively identifies sophisticated phishing attacks that successfully evade URL-based analysis through use of apparently legitimate domain names, URL obfuscation, or exploitation of user tendency to focus on visual rather than textual URL indicators. The integration of visual analysis with URL-based detection creates complementary protection against attacks that may defeat either approach in isolation.

Heuristic and machine learning classification forms the third detection layer, leveraging the nine engineered features identified through systematic feature selection to distinguish phishing from legitimate URLs based on structural and statistical properties. Machine learning classifiers trained on the ISCX-URL2016 dataset learned complex decision boundaries in the nine-dimensional feature space that generalize effectively to

novel phishing attempts exhibiting similar characteristic patterns. Heuristic rules complement statistical learning by encoding expert knowledge regarding specific indicators such as presence of IP addresses in domain names, excessive special character usage, or structural anomalies that violate common web design practices. The combination of data-driven learning and expert-encoded rules provides robust detection capabilities across diverse attack methodologies including both common patterns captured in training data and unusual techniques identified through security expertise.

Software detection automation integrates the technical detection layers into cohesive systems that operate continuously with minimal human intervention. Automated detection enables real-time analysis of incoming URLs encountered through email gateways, web proxies, browser extensions, and security information and event management platforms. The software implementation applies detection algorithms in sequence, with computationally efficient list-based checks executing first to rapidly block known threats, followed by more intensive heuristic and machine learning analysis for URLs not matched against blacklists. This staged approach optimizes computational efficiency while maintaining comprehensive protection, ensuring that all URLs undergo appropriate scrutiny regardless of computational cost concerns. The automation framework includes continuous model updating mechanisms that incorporate newly identified threats into detection models, enabling adaptive protection that evolves alongside changing attack methodologies.

User training constitutes the fifth and final layer, addressing the critical human factors that technical controls alone cannot eliminate. Comprehensive training programs employ simulated phishing exercises that expose employees to realistic attack scenarios in controlled environments, providing immediate feedback when users click malicious links or enter credentials. These exercises identify high-risk individuals requiring additional training and measure organizational vulnerability to social engineering. Training content educates users regarding phishing indicators including suspicious sender addresses, urgent or threatening language, unexpected requests for credentials, and URL characteristics suggesting potential threats. However, training effectiveness depends critically on exercise quality, organizational culture, and continuous reinforcement rather than one-time programs.

Performance evaluation of the multi-layered detection framework employed the 20 percent testing subset comprising 3,959 URLs held out during training for independent validation. The integrated system achieved overall detection accuracy of 96.3 percent on the testing set, correctly classifying 3,813 URLs while misclassifying 146 instances. Detailed analysis revealed that list-based detection matched 2,847 URLs against known blacklists with perfect accuracy but provided no coverage for the remaining 1,112 URLs representing novel threats or legitimate sites not in reference lists. Visual similarity analysis contributed detection of 89 additional phishing attempts missed by blacklists, primarily sophisticated brand impersonation attacks. The heuristic and machine learning layer correctly classified 994 of the 1,023 remaining URLs, achieving 97.2 percent accuracy on novel instances not covered by other layers.

False positive analysis examined the 78 legitimate URLs incorrectly classified as phishing, identifying systematic patterns that could inform detection refinement. Forty-three false positives involved legitimate URLs exhibiting unusually long lengths or complex parameter structures that statistical models associated with phishing behavior. Twenty-one cases involved newly registered domains from legitimate startups where domain age features contributed to misclassification despite otherwise normal URL characteristics. Fourteen instances showed anomalous subdomain structures created by legitimate content delivery networks or URL shortening services. These patterns suggest opportunities for improved feature engineering that distinguishes legitimate edge cases from malicious instances sharing similar surface characteristics.

False negative analysis investigated the 68 phishing URLs that evaded detection, revealing attack methodologies requiring additional defensive focus. Thirty-four false negatives employed compromised legitimate domains to host phishing content, exhibiting normal domain characteristics while containing malicious pages in obscure subdirectories. These attacks highlight the importance of content analysis and behavioral detection complementing URL analysis. Twenty-two cases involved newly observed obfuscation techniques not represented in training data, demonstrating the continuous evolution of attacker tactics and

necessity for ongoing model retraining. Twelve instances showed minimal feature anomalies that fell below detection thresholds, suggesting these attacks successfully mimicked legitimate URL patterns through careful crafting.

4.3 Feature Importance and Model Interpretability

Feature importance analysis quantified the relative contribution of the nine selected features to classification performance, providing insights into which URL characteristics most strongly predict phishing behavior. URL length emerged as the single most important feature, exhibiting the strongest correlation with phishing classification and contributing approximately 18 percent of total model discriminative power. This result validates extensive prior research identifying URL length as a primary indicator, with phishing URLs tending toward both very short randomly generated strings and excessively long strings incorporating multiple keywords and obfuscation elements. The robust importance of URL length across diverse datasets and attack types suggests this feature captures fundamental patterns in phishing URL construction that persist despite attacker evolution.

The number of tokens in domain ranked second in feature importance, contributing approximately 15 percent of discriminative power. Legitimate domains typically comprise one or two meaningful tokens corresponding to brand names or service descriptions, while phishing domains often concatenate multiple keywords in attempts to appear credible or manipulate search engine indexing. This pattern reflects attackers' psychological manipulation strategies, incorporating trusted brand names, security-related terms, or urgency indicators directly into domain strings. The strong predictive value of token counting validates the efficacy of lexical analysis for identifying domain name engineering attempts characteristic of phishing infrastructure.

Domain length contributed approximately 13 percent of total discriminative power, exhibiting partially overlapping but distinct patterns from overall URL length. While URL length captures total string size including path and parameters, domain length specifically characterizes the domain component where attackers face constraints from domain registration systems and desire to create credible-appearing names. The separate consideration of domain length beyond overall URL length enables models to distinguish phishing patterns manifesting specifically in domain construction from those appearing in other URL components. This fine-grained feature engineering approach enhances model sensitivity to diverse attack methodologies targeting different URL structural elements.

The number of delimiters in domain and the number of dots in URL collectively contributed approximately 22 percent of discriminative power, reflecting the importance of structural complexity analysis. Legitimate domains generally maintain simple structures with minimal subdomain nesting, while phishing domains frequently employ complex subdomain hierarchies that incorporate legitimate brand names as subdomains of attacker-controlled parent domains. This structural analysis enables detection of attacks attempting to exploit user tendency to focus on initial portions of domain names or confusion regarding domain hierarchy reading direction. The combined contribution of delimiter and dot counting features demonstrates the value of multiple complementary measurements of structural complexity.

The remaining four features including number of top-level domains, number of delimiters in path, length of longest token in path, and number of digits in query parameters collectively contributed approximately 32 percent of discriminative power. While individually less impactful than the top features, these attributes capture important edge cases and attack variations that contribute to overall model robustness. The number of top-level domain feature identifies URL manipulation attempts that create false hierarchies. Path delimiter and token length features detect obfuscation through excessive path complexity. Query parameter digit counting identifies anomalous parameter constructions. The inclusion of these secondary features prevents attackers from evading detection through manipulation of specific URL components while maintaining normal characteristics in primary features.

Model interpretability analysis employed SHAP value computation to quantify individual feature contributions to specific predictions, enabling security analysts to understand classification rationales for particular URLs. SHAP values decompose model predictions into additive contributions from each feature, quantifying the degree to which individual feature values push predictions toward phishing or legitimate classification. Visualization of SHAP values for correctly classified phishing URLs revealed consistent patterns where multiple features simultaneously indicated malicious intent, with abnormal URL length, high token counts, and excessive delimiters collectively supporting phishing classification. Conversely, false positives often showed conflicting SHAP values where some features indicated suspicious patterns while others suggested legitimacy, reflecting ambiguous cases where legitimate URLs exhibit anomalous characteristics.

The interpretability provided by SHAP analysis enables continuous model refinement through identification of systematic prediction errors and edge cases requiring additional training examples or feature engineering. Security analysts can examine SHAP values for false positives to understand which feature combinations trigger incorrect classifications, informing adjustments to detection thresholds or addition of contextual features that distinguish legitimate edge cases from malicious instances. Similarly, false negative analysis reveals which feature patterns enable phishing URLs to evade detection, guiding collection of additional training examples or development of new features capturing previously unmodeled attack characteristics.

5. Conclusion

This research presented a comprehensive feature engineering framework for predictive modeling of phishing campaign effectiveness, integrating industry targeting analysis, multi-layered detection architecture design, and systematic feature selection methodologies. Through empirical analysis of the ISCX-URL2016 dataset comprising nearly twenty thousand labeled URLs, we demonstrated that carefully engineered feature subsets containing as few as nine attributes achieve classification accuracy exceeding ninety-six percent when integrated within multi-layered detection frameworks. The identification of nine optimal features spanning domain-level, path-level, parameter-level, and holistic URL characteristics enables computationally efficient detection suitable for real-time operational deployment while maintaining high accuracy and interpretability essential for security operations.

Industry targeting pattern analysis revealed significant concentration of phishing efforts on financial services, Software as a Service platforms, and webmail systems that collectively account for sixty percent of observed attacks. This distribution pattern reflects attackers' strategic focus on high-value credentials providing access to financial resources, organizational communication systems, and integrated cloud platforms. The systematic variation in targeting across industries informs defensive resource allocation, suggesting that organizations in heavily targeted sectors require particularly robust multi-layered protection combining technical controls and user awareness programs. The remaining forty percent of attacks distributed across diverse sectors highlights the breadth of phishing operations and necessity for comprehensive protection regardless of industry classification.

The multi-layered detection architecture integrating list-based blacklisting, visual similarity analysis, heuristic and machine learning classification, software automation, and user training provides synergistic protection that exceeds capabilities of individual detection methodologies. Each layer addresses specific attack characteristics and threat vectors, with list-based detection blocking known threats, visual analysis identifying brand impersonation, machine learning classifying novel attacks through feature analysis, and user training enhancing human vigilance. The integration of these complementary approaches creates defense-in-depth systems resilient to diverse attack methodologies and adaptive to evolving tactics. Empirical validation demonstrated that the integrated framework achieves overall accuracy of 96.3 percent, with individual layers contributing detection capabilities across different portions of the threat landscape.

Feature importance analysis quantified the discriminative power of the nine selected URL characteristics, revealing that URL length, token counts, domain length, and structural complexity measures including

delimiter and dot frequencies constitute the most predictive attributes. These features capture fundamental patterns in phishing URL construction that persist across diverse attack types and attacker evolution, reflecting inherent constraints attackers face in creating credible-appearing URLs while maintaining infrastructure control. The robust importance of these features across multiple datasets and detection contexts validates their generalizability beyond specific training distributions. Secondary features including path characteristics and parameter properties contribute additional discriminative power that enhances model robustness against edge cases and emerging attack variations.

Model interpretability through SHAP value analysis enables security practitioners to understand classification rationales and refine detection systems based on systematic error patterns. The decomposition of predictions into additive feature contributions reveals which URL characteristics drive specific classifications, facilitating identification of false positive patterns where legitimate URLs exhibit anomalous characteristics and false negative patterns where phishing URLs successfully mimic normal feature profiles. This interpretability transforms machine learning models from black-box classifiers into decision support tools that augment human expertise through transparent reasoning processes. The insights gained from interpretability analysis inform continuous improvement of feature engineering approaches and detection thresholds based on operational experience and evolving threat landscapes.

Practical deployment considerations demonstrate that the optimized feature engineering approach supports real-time phishing assessment suitable for integration into operational security infrastructure. The computational efficiency of extracting and evaluating nine URL features enables high-throughput analysis of email gateway traffic, web proxy requests, and security information feeds without introducing unacceptable latency. The reduced feature dimensionality compared to exhaustive feature sets decreases model storage requirements and inference costs while maintaining classification accuracy, facilitating deployment on resource-constrained edge devices and enabling scalable protection across large user populations. Integration pathways into existing email security gateways, browser extensions, and security operations platforms provide immediate value through enhanced threat triage and automated response capabilities.

Future research directions include extending the feature engineering framework to emerging attack vectors including voice phishing, SMS-based campaigns, and social media messaging that exhibit distinct characteristics requiring specialized feature development. Investigation of adversarial robustness mechanisms represents another critical avenue, as sophisticated attackers may attempt feature manipulation to evade detection through careful URL construction that minimizes feature anomalies. Development of dynamic feature importance models that adapt to temporal shifts in attack methodologies would enhance resilience against attacker evolution. Exploration of transfer learning approaches could enable knowledge transfer from well-studied URL-based phishing domains to content-based detection for email and messaging phishing that requires different feature engineering strategies.

The integration of external threat intelligence data including domain reputation scores, WHOIS information, and historical attack patterns represents promising directions for augmented feature engineering. These contextual features complement URL structural analysis by providing additional signals regarding domain infrastructure, registration patterns, and past malicious activity. However, incorporating external data sources introduces operational challenges including latency from API queries, data quality variations across providers, and potential privacy considerations that require careful architectural design. Research into efficient caching strategies, selective feature computation, and hybrid approaches that balance comprehensive analysis against operational constraints would advance practical deployment capabilities.

The evolution of phishing tactics necessitates continuous adaptation of feature engineering methodologies to capture novel manipulation strategies and technical innovations. Attackers continuously experiment with new URL construction patterns, obfuscation techniques, and infrastructure configurations designed to evade detection while maintaining credibility. Defensive systems must incorporate mechanisms for detecting concept drift in feature distributions, identifying emerging attack patterns through anomaly detection, and rapidly updating models with newly observed threats. The development of automated feature engineering

pipelines that systematically evaluate candidate features against evolving attack samples would enhance adaptive capabilities and reduce the manual effort required for maintaining current protection.

In conclusion, this research demonstrates that systematic feature engineering combining domain expertise, data-driven feature selection, and multi-layered detection integration enables highly effective phishing campaign effectiveness prediction. The frameworks and methodologies developed through this work provide practical tools for security practitioners seeking to transition from reactive detection to proactive threat assessment and strategic resource allocation. The insights regarding industry targeting patterns, optimal feature subsets, and multi-layered architecture design inform both immediate operational improvements and longer-term research directions in this critical cybersecurity domain that continues to evolve alongside technological innovation and attacker adaptation.

References

1 Kheruddin, M. S., Zuber, M. A. E. M., & Radzai, M. M. M. (2024). Phishing attacks: Unraveling tactics, threats, and defenses in the cybersecurity landscape. Authorea Preprints.

2 Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. Journal of King Saud University-Computer and Information Sciences, 34(2), 232-247.

3 Mousavi SM, Bahaghighat M. Phishing Website Detection: An In-Depth Investigation of Feature Selection and Deep Learning. Expert Systems. 2025;42(1):e13824.

4 AYODELE, G. T., ABDULRAHMAN, I. A., ALEBIOSU, J., EGBEDION, G. E., & AKINBOLAJO, O. E. (2025). Human-Centric Cybersecurity: Addressing the Human Factor in Cyber Defense Strategies.

5 Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 15(2), 951.

6 Xu, X., Liang, T., Zhu, J., Zheng, D., & Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. Neurocomputing, 328, 5-15.

7 Wang W, Zhang F, Luo X, Zhang S. PDRCNN: precise phishing detection with recurrent convolutional neural networks. Security and Communication Networks. 2019;2019:1-15.

8 Divakaran DM, Oest A. Phishing detection leveraging machine learning and deep learning: a review. IEEE Security & Privacy. 2022;20(5):86-95.

9 Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. Expert Systems with Applications. 2019;117:345-357.

10 Adebowale MA, Lwin KT, Sánchez E, Hossain MA. Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications. 2019;115:300-313.

11 Hannousse A, Yahiouche S. Towards benchmark datasets for machine learning based website phishing detection: an experimental study. Engineering Applications of Artificial Intelligence. 2021;104:104347.

12 Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020;9(9):1514.

13 Bu SJ, Cho SB. Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection. Electronics. 2021;10(12):1492.

14 Chiew KL, Tan CL, Wong K, Yong KS, Tiong WK. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences. 2019;484:153-166.

15 Rao RS, Pais AR, Anand P. A heuristic technique to detect phishing websites using TWSVM classifier. Neural Computing and Applications. 2021;33(11):5733-5752.

16 Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access.

17 Cao, W., Mai, N. T., & Liu, W. (2025). Adaptive knowledge assessment via symmetric hierarchical Bayesian neural networks with graph symmetry-aware concept dependencies. Symmetry, 17(8), 1332.

18 Mai, N. T., Cao, W., & Liu, W. (2025). Interpretable knowledge tracing via transformer-Bayesian hybrid networks: Learning temporal dependencies and causal structures in educational data. Applied Sciences, 15(17), 9605.
19 Chen, S., Liu, Y., Zhang, Q., Shao, Z., & Wang, Z. (2025). Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Advanced Intelligent Systems, 2400898.
20 Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery. IEEE Access.
21 Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access.
22 Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access.
23 Sun, T., Wang, M., & Han, X. (2025). Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends. Journal of Banking and Financial Dynamics, 9(8), 1-11.
24 Ren, S., & Chen, S. (2025). Large Language Models for Cybersecurity Intelligence, Threat Hunting, and Decision Support. Computer Life, 13(3), 39-47.
25 Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science. 2021;2(6):420.
26 Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLoS One, 20(10), e0332640.
27 Zhang, H., Ge, Y., Zhao, X., & Wang, J. (2025). Hierarchical deep reinforcement learning for multi-objective integrated circuit physical layout optimization with congestion-aware reward shaping. IEEE Access.
28 Wang, M., Zhang, X., & Han, X. (2025). AI Driven Systems for Improving Accounting Accuracy Fraud Detection and Financial Transparency. Frontiers in Artificial Intelligence Research, 2(3), 403-421.
29 Chen, S., & Ren, S. (2025). AI-enabled Forecasting, Risk Assessment, and Strategic Decision Making in Finance. Frontiers in Business and Finance, 2(02), 274-295.
30 Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science. 2022;3(2):158.