The Erosion of Password-Based Authentication Security: A Data-Driven Evaluation of Phishing-Based Session Hijacking and MFA-Bypass Techniques

Joe Basanta., Bernice Jennifer Bontalilid., Jose Arno Gamboa., Stephen John Gavaran., Tommi Michael Pallarco., Ace Jasper Peñaranda., Kristine Soberano

State University of Northern Negros, Sagay City, Philippines

DOI: https://doi.org/10.51244/IJRSI.2026.1307000331

Received: 01 August 2026; Accepted: 06 August 2026; Published: 18 August 2026

ABSTRACT

Password authentication remains widely used, but modern phishing campaigns increasingly target the authenticated session rather than the password alone. This study examines the gradual erosion of password-based security through a reproducible secondary-data analysis of phishing websites and network intrusions, while also assessing whether commonly used public datasets can support claims about phishing-based session hijacking and multifactor authentication (MFA) bypass. Three datasets formed the basis of the analysis. The UCI Phishing Websites dataset contained 11,055 records and 30 predictors, while the Vrbančič phishing dataset included 88,647 records and 111 predictors. The third dataset was a documented random sample from CICIDS2017, made up of 56,661 network flows and 77 predictors. Both logistic regression and random forest were tested using a hold-out method with an 80/20 split. The F1 scores for random forest algorithm were 0.972 and 0.957 for the two phishing detection datasets, and 0.996 for the binary classification of the CICIDS2017. Overall, random forest showed excellent results, with 0.994 accuracy and 0.980 macro-F1 in the case of a multi-class CICIDS2017 task. The results for the Infiltration class should be further verified, as the test set appears to have only seven instances of this class. Feature importance was largely driven by webpage, URL, domain, and network-flow characteristics. However, none of the analyzed datasets included MFA challenge events, session-cookie issuance, token capture, token replay, device binding, or post-authentication identity telemetry. A structured observability audit therefore found direct coverage for only three of the seven stages in the proposed compromise chain. These results show that strong phishing and intrusion classifiers can identify conditions that enable an attack, but they cannot, on their own, demonstrate the detection of session hijacking or MFA bypass. A more complete evaluation requires the integration of phishing, identity-provider, endpoint, and session telemetry.

Keywords: authentication, phishing, hijacking, MFA bypass, machine learning

INTRODUCTION

Passwords remain the standard way people access digital services because they are familiar, relatively inexpensive to manage, and already integrated into most account-recovery systems. Even so, they have long been a weakness in account security. When users reuse the same password across several services, a single compromise can put multiple accounts at risk. Attackers may also obtain passwords by guessing them, stealing them through malware, finding them in leaked data, or tricking users into entering them on a convincing phishing page. These risks have led many organizations to adopt multifactor authentication (MFA), but MFA does not prevent every form of attack. An attacker may still relay the second factor in real time or capture the authenticated session created after the user successfully signs in.For this reason, authentication security now involves more than protecting the password alone. The sign-in process itself, together with the session that follows, also needs to be protected.

Adversary-in-the-middle (AiTM) phishing demonstrates how this type of compromise can occur. Using a reverse proxy or a similar intermediary, an attacker can place themselves between the user and the legitimate

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

service, present what appears to be the normal sign-in experience, forward the user's credentials and MFA response to the actual platform, and capture the session cookie or token issued once authentication is completed. The stolen token may then be replayed, allowing the attacker to access the account as the user without going through the original sign-in process again (Microsoft, 2022, 2023). This is often described operationally as MFA bypass, although the second factor may have functioned exactly as designed. The bypass occurs because possession of the resulting bearer session can substitute for another authentication ceremony.

This development creates a measurement problem. Public phishing datasets generally describe URLs, domains, HTML elements, certificate states, or reputation indicators. Network-intrusion datasets describe flows, ports, packet lengths, timings, and attack labels. These resources are useful for reproducible machine-learning research, but they do not necessarily include identity-provider events, MFA prompts and outcomes, session-token issuance, browser or device binding, token replay, or post-authentication actions. Consequently, high phishing-detection or intrusion-detection accuracy can be misinterpreted as evidence that an end-to-end identity compromise has been measured.

The present study addresses that problem through actual execution of three public datasets and a structured audit of their observable attack stages. It asks: (1) how accurately can standard machine-learning models classify phishing websites and malicious network flows in the selected datasets; (2) which features contribute most to those classifications; (3) which stages of a phishing-to-session-hijacking chain are directly represented by the available labels and variables; and (4) what do the findings imply for the continued use of password-based and phishable MFA systems?

The study contributes in two ways. First, it reports reproducible held-out results rather than relying only on dataset descriptions or previously published accuracy values. Second, it separates predictive performance from attack-chain observability. This distinction is essential because a model can classify the available labels with high accuracy while the dataset remains incapable of evaluating the security outcome named in the research question. The analysis therefore treats performance metrics as evidence about the executed datasets, not as proof that session hijacking or MFA bypass has been directly detected.

Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are at the core of today's intelligent technologies. ML enables systems to learn from data, while DL, a subset of ML, utilizes neural networks to learn complex patterns. The increasing availability of data and computational power has accelerated the adoption of ML and DL across industries. The objective of this paper is to explain key ML and DL concepts, illustrate their applications, and highlight the challenges that remain in adopting these technologies at scale.

REVIEW OF RELATED LITERATURE

Password Authentication and Phishable MFA

Password-based authentication relies upon the security of an individual's account being placed into a reusable secret, which a user must both identify and recall so they may provide it to a system at time of authentication. The authors of Bonneau et al. (2012), have shown that new approaches to passwords will need to be compared and contrasted relative to their security advantages over current password systems; however, they also will need to be evaluated according to how easily they may be adopted by users, how easily they may be integrated into systems currently in use, and how well they will function in different environments. This could help explain why passwords continue to be one of the most common forms of authentication, despite the fact that many of the vulnerabilities surrounding passwords are now well known. Newer authentication techniques may offer greater protection than passwords against guessing and phishing attempts, but may also present unique challenges regarding issues of system migration, account recovery, compatibility, and customer service.

Multifactor authentication provides stronger security for an individual's account(s) because each user must authenticate themselves using multiple factors. Nevertheless, how strong a given multifactor authentication technique protects accounts, is contingent upon how securely it has been designed. Phishing, relay attacks, social engineering and repeated login requests may all be used to target a user's multifactor authentication

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

technique if either one-time passwords, SMS codes, or approval prompts are used as an example. According to NIST guidelines, since authenticator codes entered manually into a device are not cryptographically tied to the particular session of authentication (Temoshok et al., 2025), these types of codes are therefore not resistant to phishing. Thus, while there certainly exists an important difference between single-factor and multifactor authentication techniques; there is just as importantly a difference in terms of whether or not a particular technique is subject to phishing or not.

Public Key Credentials used by WebAuthn and FIDO2 resolve this problem by providing credentials that are bound to a specific Relying Party. Rather than producing a reusable secret to be presented to any webpage requesting access to an account, the authenticator produces a cryptographic response that is bound to the original website and its origin (World Wide Web Consortium, 2021). As Temoshok et al. (2025) point out in their review of NIST guidance on phishing-resistance techniques, Verifier-Name Binding (including the mechanism developed for WebAuthn) is an established method for creating phishing-resistance. The technologies mentioned reduce the effectiveness of credential attacks. But, other areas including account recovery, session management, legacy authentication routine, and device enrollment, must still be protected by a separate security control.

AiTM Phishing, Session Hijacking, and MFA Bypass

Traditional phishing attacks are often successful in obtaining an attackers' future use of a victim's password. Adversary-in-the-middle phishing, however, allows an attacker to enter into an ongoing authentication sequence. Microsoft (2022) reported several campaigns where attackers stole users' passwords, and/or their authentication session cookies, and subsequently accessed those victims' accounts using the stolen cookies - even though multi-factor authentication (MFA) had previously been enabled. Microsoft (2023), further described an additional campaign in which attackers utilized an indirect proxy approach; they sent an attacker's stolen credentials and MFA data to the legitimate identity provider, who provided them with the resultant session token. Both campaigns clearly illustrate that while authentication session data may not always be viewed by organizations as a primary security asset, but it should certainly be considered so.

Session Hijacking is the unauthorized reuse of an authenticating user's session token outside its intended environment. There are many measures organizations can take to mitigate this type of risk including shortening the lifetime of tokens, implementing token rotation, revoking unused tokens, configuring secure cookie options on devices, protecting session tokens via encryption/decryption techniques, continuously assessing each user's risk level during login sessions, checking whether the device attempting to authenticate is compliant with organizational requirements prior to allowing the device to complete its initial authentication attempt, requiring all sensitive transactions or actions require renewed authentication attempts after the original login session has expired. While organizations can implement various types of countermeasures to minimize the possibility of Session Hijacking occurring, none of these countermeasures will provide sufficient assurance about their efficacy simply based upon what appears in the URL request or normal network flow attributes. To truly assess whether the implemented countermeasures are effective, organizations need a comprehensive view of the phishing attack attempt, the user authentication activity that followed the phishing attack attempt, how and when the session token was created, how the session token was utilized after its creation, details related to the device being utilized to create/login to the account(s), and what activities occurred once the user successfully logged in.

Therefore, caution needs to be taken when using the term "MFA Bypass". An attacker could utilize an insecure recovery mechanism, exploit an older version of an authentication protocol, trick a member of the organization's Help Desk to perform certain functions on behalf of the attacker, have the user approve a malicious message or notification, forward a one-time code received by the user through another channel, or utilize a valid session token after the user has properly completed Multi-Factor Authentication. As noted above, each method will produce differing types of artifacts/evidence and represent a distinct weakness in the overall system. Therefore, if an organization utilizes a dataset consisting solely of generic web-based attack/brute force labels to evaluate MFA Bypass, it may very well not accurately reflect the true nature of MFA Bypass occurrences.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Public Cybersecurity Datasets and Evaluation Limits

The UCI Phishing Website's dataset includes 30 engineered attributes detailing aspects of URLs and domains which provide an indication of URL's composition and domain's characteristics, certificate condition, web page behavior and external reputation indicators (Mohammad & McCluskey, 2012). This dataset is often selected as it is small, well-structured and supports binary classification. While this dataset includes attributes to identify potential phishing sites; however, Mohammad & McCluskey (2012), created the original dataset prior to the appearance of current AI/ TM campaigns and did not consider either the authentication session or MFA event.

Vrbančič et al. (2020), presented two larger phishing detection datasets including 58,645 and 88,647 website instances, where each instance contained 111 attributes. In general most of the features were descriptive of the URL and each component of the URL. The other features described domain resolution and external website metrics. With a larger sample size for model training and testing will improve the reliability of the models. As previously mentioned Vrbančić et al. (2020), stated that the target label indicates if the site was legitimate or a phishing site but does not state if an account was compromised, MFA was relayed, or if an authorized access token was stolen.

CICIDS2017 was constructed as a labeled intrusion-detection benchmark representing benign network traffic and multiple types of attacks through network-flow features (Sharafaldin et al., 2018). CICIDS2017 has been widely accepted by researchers examining the reliability of its construction. Specifically, Engelen et al. (2021), have demonstrated concerns regarding attack generation, flow construction, feature extraction and labeling within the process of constructing CICIDS2017. They demonstrate that factors in creating a dataset can impact what researchers report as successful in their results. Additionally, since researchers can randomly divide their training and testing sets so that closely related traffic records or similar characteristics from scenarios appear in both groups; a researcher's model could be learning the characteristics specific to the particular dataset and not patterns they can apply to real world attacks.

Taken together, these datasets provide useful but incomplete perspectives on the problem. Phishing datasets primarily represent the malicious webpage, URL, or other characteristics of the deceptive material, while CICIDS2017 represents selected network attacks and patterns of network-flow behavior. Neither type of dataset directly captures the identity and session information required to determine whether MFA was relayed or an authenticated session token was reused. A defensible data-driven investigation must therefore clearly state both what its models are capable of classifying and what cannot be observed from the selected datasets.

RESEARCH METHODOLOGY

Research Design

The study used a quantitative, comparative secondary-data design with two linked components. The first component executed supervised classification models on phishing and network-intrusion datasets. The second component audited dataset schemas and labels against a seven-stage authentication-compromise chain. No live phishing infrastructure, credential collection, MFA interception, token theft, or attack against an operational account was performed.

The unit of analysis differed by dataset: a website record for the phishing datasets and a network flow for CICIDS2017. Results were therefore evaluated within each dataset rather than pooled into a single model. This avoided treating heterogeneous features as though they described the same event.

Datasets

Three datasets were executed. The first was the UCI Phishing Websites dataset obtained in ARFF format. It contained 11,055 observations, 30 predictors, and a binary target. The original target coding was converted so that phishing was the positive class. The second was the full 88,647-record dataset published by Vrbančič et al. (2020), containing 111 numeric predictors and a phishing target. The third was a documented random sample

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

of CICIDS2017 containing 56,661 flows, 77 predictors, and seven labels: BENIGN, Bot, BruteForce, DoS, Infiltration, PortScan, and WebAttack.

The user-specified Kaggle CICIDS2017 mirrors were evaluated as potential sources. Direct execution of the complete collection was not feasible in the available environment because the Kaggle downloads required additional access and the full collection is substantially larger. The analysis therefore used a publicly documented random CICIDS2017 sample. This is explicitly reported to prevent the sample results from being represented as full-corpus benchmarks. The additional Vrbančič dataset was selected in place of relying exclusively on the smaller Kaggle phishing mirror because it has a peer-reviewed data descriptor, substantially more records, and a broader engineered feature set.

Table 1. The Arrangement of Channels

Dataset	Records	Features	Positive n	Negative n	Missing/non-finite
UCI Phishing	11,055	30	4,898	6,157	0
Vrbančič Phishing	88,647	111	30,647	58,000	0
CICIDS2017 Sample	56,661	77	33,930	22,731	162

Preprocessing and Model Execution

Each dataset was divided once into 80% training and 20% test partitions using stratified random sampling and a fixed random seed of 42. The UCI and Vrbančič datasets had no missing predictor values. Positive and negative infinity in CICIDS2017 were converted to missing values, producing 162 cells requiring imputation. Median imputation was fitted only on the training partition within a scikit-learn pipeline.

Two binary classifiers were executed for each dataset. Logistic regression served as an interpretable linear baseline and used standardized predictors, balanced class weights, and a maximum of 2,500 iterations. Random forest used 300 trees, square-root feature sampling, balanced subsample weights, and parallel processing. For CICIDS2017, a separate seven-class random forest used 120 trees because of execution-time constraints. The models were intentionally conventional: the objective was to establish a transparent benchmark and evaluate observability, not to maximize accuracy through extensive hyperparameter tuning.

Binary performance was assessed using accuracy, precision, recall, F1 score, receiver operating characteristic area under the curve (ROC-AUC), and confusion-matrix counts. Multiclass performance was summarized using accuracy, macro-averaged precision, recall, and F1, weighted F1, and per-class measures. Random-forest feature importances were extracted to determine whether the most influential predictors represented phishing artifacts, network flows, identity events, or session behavior.

The execution used Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, and Matplotlib 3.10.8. A machine-readable result record and scripts were retained as supplementary materials.

Attack-Chain Observability Audit

A seven-stage chain was used to assess whether the datasets could directly support the paper's central security claims: (1) phishing artifact or delivery, (2) password or credential attack, (3) MFA challenge and response, (4) session-token issuance or capture, (5) token replay or session hijacking, (6) post-authentication abuse, and (7) network telemetry. A stage was marked directly observed only when dataset variables or labels represented the event itself. A stage was not counted merely because it could occur before or after a labeled event. The coverage rate was computed as the number of directly represented stages divided by seven.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Ethical and Validity Considerations

All data were publicly released for research. The analysis did not collect personal credentials, reproduce malicious kits, or interact with real users. Internal validity was supported through fixed preprocessing pipelines and a held-out test partition. External validity remains limited by dataset age, engineered features, random sampling, benchmark artifacts, and the use of a CICIDS2017 subset. Feature importance indicates predictive contribution within the fitted forest; it does not establish causality.

RESULTS AND DISCUSSION

Binary Classification Performance

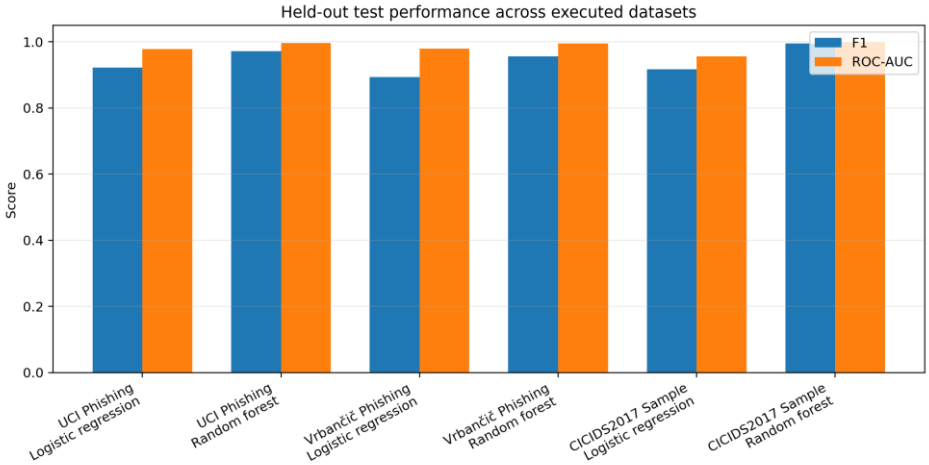
Table 2 reports held-out performance. Random forest outperformed logistic regression on all three datasets. On UCI Phishing, random forest achieved 0.975 accuracy, 0.972 F1, and 0.997 ROC-AUC. On the larger Vrbančič dataset, it achieved 0.970 accuracy, 0.957 F1, and 0.995 ROC-AUC. On binary CICIDS2017 classification, it achieved 0.995 accuracy, 0.996 F1, and 0.999 ROC-AUC. The corresponding confusion matrices contained 55 phishing errors on UCI, 529 phishing errors on Vrbančič, and 60 binary errors on the CICIDS2017 test partition.

Table 2. Held-Out Binary Classification Results

Dataset	Model	Acc.	Prec.	Recall	F1	AUC	FP/FN
UCI Phishing	Logistic regression	0.932	0.930	0.915	0.922	0.979	68/83
UCI Phishing	Random forest	0.975	0.979	0.964	0.972	0.997	20/35
Vrbančič Phishing	Logistic regression	0.922	0.847	0.945	0.893	0.979	1046/338
Vrbančič Phishing	Random forest	0.970	0.955	0.959	0.957	0.995	278/251
CICIDS2017 Sample	Logistic regression	0.900	0.910	0.924	0.917	0.956	620/516
CICIDS2017 Sample	Random forest	0.995	0.996	0.995	0.996	0.999	28/32

Note. Positive class = phishing for the website datasets and malicious flow for CICIDS2017. Acc. = accuracy; AUC = ROC-AUC; FP = false positives; FN = false negatives. All values are from a single stratified 20% test partition.

Figure 1. F1 and ROC-AUC on the held-out test partitions.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The linear baseline remained competitive, especially in ROC-AUC, but showed larger trade-offs between false positives and false negatives. On the Vrbančič data, logistic regression recalled 94.5% of phishing records but produced 1,046 false positives, reducing precision to 0.847. Random forest reduced false positives to 278 while retaining 95.9% recall. This suggests useful nonlinear interactions among URL, directory, file, and domain variables.

These values demonstrate that the selected feature sets are highly separable under the chosen random split. They do not demonstrate equivalent performance on newly emerging phishing kits, unseen organizations, temporally separated traffic, or production systems. The CICIDS2017 result is particularly susceptible to benchmark-specific regularities. Engelen et al. (2021) showed that collection and labeling issues can strongly influence machine-learning evaluation on this corpus. The high scores should therefore be interpreted as reproducible within-sample benchmarks, not as proof of operational intrusion-detection effectiveness.

Feature Importance and What the Models Actually Learned

The strongest UCI random-forest features were SSLfinal_State (importance = 0.315), URL_of_Anchor (0.250), web_traffic (0.071), having_Sub_Domain (0.063), and Prefix_Suffix (0.045). These variables describe certificate or transport indicators, hyperlink behavior, reputation, and URL structure. In the Vrbančič model, the leading features were directory_length (0.059), time_domain_activation (0.058), qty_equal_file (0.029), file_length (0.029), and length_url (0.029). These are likewise artifact-level characteristics.

The leading CICIDS2017 binary features were Init_Win_bytes_forward (0.089), Packet Length Std (0.042), Bwd Packet Length Mean (0.041), Init_Win_bytes_backward (0.040), and Average Packet Size (0.038). These variables describe transport and flow behavior. No highly ranked feature in any model represented an MFA prompt, authentication method, token identifier, token binding, session creation, cookie replay, device identity, impossible travel, mailbox rule creation, or other post-authentication event. Thus, the feature analysis supports a narrow interpretation of the performance results: the models detected phishing-like artifacts and dataset-specific network patterns.

CICIDS2017 Multiclass Results

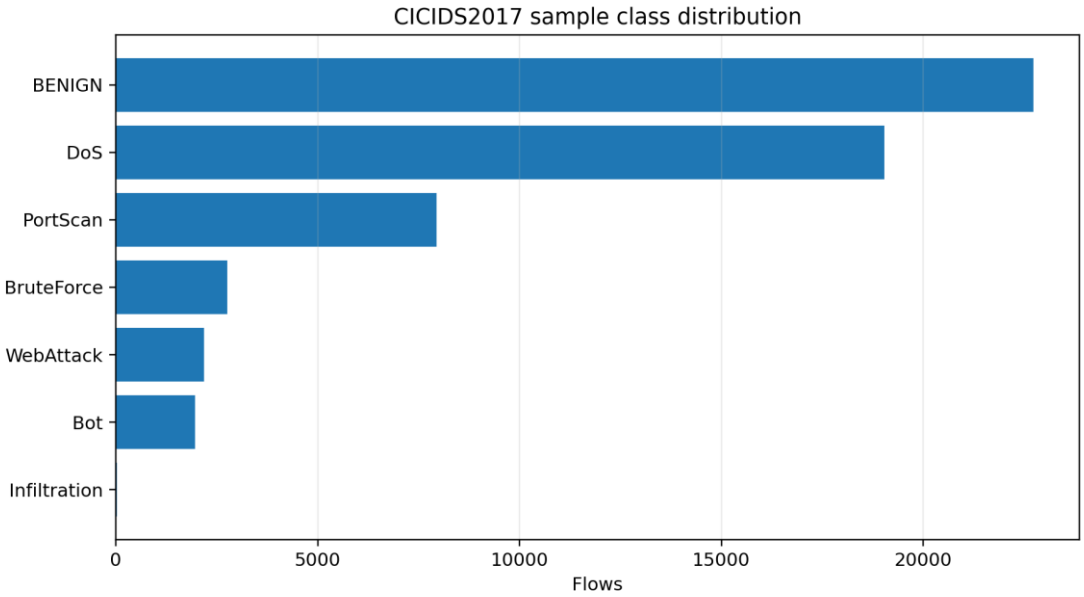
The seven-class CICIDS2017 random forest achieved 0.994 accuracy, 0.980 macro-F1, and 0.994 weighted F1. Per-class F1 exceeded 0.965 for every class except Infiltration, which obtained 0.923. That value is unstable because the held-out partition contained only seven Infiltration observations. The apparent strength of a class score with such small support should not be generalized.

Table 3. CICIDS2017 Multiclass Random-Forest Performance

Class	Precision	Recall	F1	Test support
BENIGN	0.993	0.993	0.993	4,547
Bot	0.946	0.985	0.965	393
BruteForce	0.998	0.996	0.997	554
DoS	0.997	0.998	0.997	3,807
Infiltration	1.000	0.857	0.923	7
PortScan	0.999	0.998	0.998	1,589
WebAttack	0.998	0.970	0.984	436
Overall / macro	0.990	0.971	0.980	11,333

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Note. *The overall row reports macro-averaged precision, recall, and F1. Overall accuracy was 0.994 and weighted F1 was 0.994.*



Class imbalance is visible in Figure 2. BENIGN and DoS together account for most flows, while Infiltration represents only 36 records in the complete sample. Balanced subsample weights reduce the direct influence of class frequency during tree construction, but they cannot create new behavioral diversity for a rare class. Future work should use the corrected full corpus where feasible, preserve temporal or scenario boundaries, and report confidence intervals or repeated validation.

Attack-Chain Observability

Table 4 separates direct observation from inference. The phishing datasets directly represent the malicious artifact. CICIDS2017 directly represents selected credential-related attacks through its BruteForce label and supplies network-flow telemetry. Together, the sources directly cover three of seven stages, or 42.9%. None of the datasets directly observes MFA interaction, token issuance or capture, token replay, or post-authentication identity abuse.

Table 4. Direct Observability of the Phishing-to-Session-Hijacking Chain

Attack-chain stage	**Directly observed?**	**Evidence in executed data**
1. Phishing artifact or delivery	Yes	UCI and Vrbančič phishing labels and webpage/URL features
2. Password or credential attack	Yes, limited	CICIDS2017 BruteForce; no verified credential submission outcome
3. MFA challenge and response	No	No prompt, factor, approval, denial, or relay variables
4. Session-token issuance or capture	No	No cookie/token issuance, extraction, or theft label
5. Token replay or session hijacking	No	No replay identifier, session reuse, or browser-context evidence
6. Post-authentication abuse	No	No mailbox, application, privilege, or account-action

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

		telemetry
7. Network telemetry	Yes	CICIDS2017 flow-level timing, packet, window, and volume features

Note. *Coverage = 3 directly represented stages / 7 stages = 42.9%. "Limited" is counted as direct only because a credential-attack label is present; it does not establish successful account compromise.*

This result resolves the apparent contradiction between near-perfect classifier scores and limited evidence about MFA bypass. A phishing classifier can accurately identify a fraudulent webpage while remaining entirely unaware of whether a user visited it, entered a password, completed MFA, or received a session. An intrusion classifier can accurately separate labeled flows while remaining unaware of the identity provider's authentication state. The layers are related in an attack chain but are not interchangeable measurements.

Microsoft's AiTM analyses show that the decisive event is often capture and replay of the authenticated session (Microsoft, 2022, 2023). The executed datasets stop before this identity-session evidence appears. Consequently, the study cannot estimate the prevalence of session hijacking, compare specific MFA-bypass techniques, or train a direct token-replay detector. It can demonstrate that common public benchmarks are structurally insufficient for those tasks.

The security implication is that password-based systems remain vulnerable when authentication outputs can be relayed and resulting sessions can be used outside their intended context. Conventional MFA improves security against password-only attacks but is not equivalent to phishing resistance. NIST defines phishing resistance through cryptographic binding to the channel or verifier and explicitly excludes manually entered OTP outputs from that category (Temoshok et al., 2025). WebAuthn implements relying-party-scoped public-key credentials, reducing the attacker's ability to proxy a valid assertion for a different origin (World Wide Web Consortium, 2021).

Even phishing-resistant sign-in does not remove the need for session security. Stolen endpoint cookies, malware, compromised recovery, or misconfigured federation may still expose authenticated state. A complete evaluation should correlate four data planes: phishing artifact telemetry; identity-provider authentication and risk events; endpoint or browser session evidence; and application-level post-authentication actions. Labels should identify whether a token was issued, captured, replayed, revoked, bound to a device, or used for a sensitive action.

The main limitation is that the CICIDS2017 analysis used a public random sample rather than the complete user-specified Kaggle collection. The models were tested with one stratified random split and were not tuned or validated through repeated cross-validation, temporal holdout, or external datasets. The phishing datasets rely on engineered features and may contain patterns tied to their collection periods. The results are therefore best understood as transparent baseline executions and a measurement-gap analysis, not a deployable detector or a prevalence study.

CONCLUSION

The executed experiments show that conventional machine-learning models can classify the selected phishing and intrusion datasets with high held-out performance. Random forest achieved F1 scores of 0.972 on UCI Phishing, 0.957 on the larger Vrbančič dataset, and 0.996 on binary CICIDS2017 classification. The multiclass CICIDS2017 model achieved 0.980 macro-F1, although rare-class support was inadequate for strong conclusions about Infiltration.

These metrics do not establish detection of phishing-based session hijacking or MFA bypass. Feature importance was confined to URL, webpage, domain, and network-flow variables, and the observability audit found direct representation of only three of seven attack-chain stages. The identity and session events that define AiTM compromise were absent. The erosion of password-based authentication security is therefore not demonstrated by a single classifier score; it is demonstrated by the mismatch between modern attack behavior

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

and the telemetry used to evaluate defenses.

Password-plus-phishable-MFA architectures should be treated as transitional controls rather than complete protection against credential relay and session theft. Research and operational detection must move toward phishing-resistant authentication and integrated identity-session datasets capable of observing the full compromise sequence.

RECOMMENDATIONS

1. Prioritize phishing-resistant authentication. Organizations should migrate high-risk users and sensitive applications toward WebAuthn/FIDO2 or other cryptographically bound authentication, while retaining secure recovery and enrollment controls.

2. Treat session tokens as high-value credentials. Use short and risk-appropriate lifetimes, rotation, revocation, secure cookie settings, reauthentication for sensitive actions, and device- or channel-aware token protection where supported.

3. Integrate security telemetry across layers. Phishing indicators, identity-provider logs, endpoint/browser evidence, token events, and application actions should be correlated so that token capture and replay can be distinguished from ordinary sign-in and network traffic.

4. Avoid overstating benchmark results. Accuracy on phishing or CICIDS2017 data should be described as performance on artifact or flow labels, not as direct evidence of MFA-bypass or session-hijacking detection. Temporal holdouts, external validation, class support, and dataset-quality limitations should be reported.

5. Develop an ethical multilayer research benchmark. A future dataset should contain consented or synthetic sequences linking phishing delivery, credential submission, MFA challenge, session issuance, token replay, device context, revocation, and post-authentication behavior. Privacy-preserving identifiers can connect events without exposing real credentials or tokens.

Data and reproducibility statement. The analysis scripts, result tables, feature-importance files, figures, and machine-readable execution record accompany this manuscript. The underlying UCI and Vrbančič datasets remain available from their cited repositories. The CICIDS2017 result applies to the documented 56,661-flow sample used in this execution and should not be presented as a full-corpus result.State the general and specific objectives or purpose of conducting the study in paragraph format.

REFERENCES

1. Bonneau, J., Herley, C., van Oorschot, P. C., & Stajano, F. (2012). The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In 2012 IEEE Symposium on Security and Privacy (pp. 553-567). IEEE. https://doi.org/10.1109/SP.2012.44
2. Cybersecurity and Infrastructure Security Agency. (2022). Implementing phishing-resistant MFA. U.S. Department of Homeland Security. https://www.cisa.gov/sites/default/files/2023-01/fact-sheet-implementing-phishing-resistant-mfa-508c.pdf
3. D'Hooge, L. (n.d.-a). CIC-IDS2017 [Data set]. Kaggle. https://www.kaggle.com/datasets/dhoogla/cicids2017
4. D'Hooge, L. (n.d.-b). CIC-IDS-Collection [Data set]. Kaggle. https://www.kaggle.com/datasets/dhoogla/cicidscollection
5. Engelen, G., Rimmer, V., & Joosen, W. (2021). Troubleshooting an intrusion detection dataset: The CICIDS2017 case study. In 2021 IEEE Security and Privacy Workshops (SPW) (pp. 7-12). IEEE. https://doi.org/10.1109/SPW53761.2021.00009
6. Hannousse, A., & Yahiouche, S. (2021). Web page phishing detection [Data set]. Mendeley Data. https://doi.org/10.17632/c2gw7fy2j4.3
7. Microsoft. (2022, July 12). From cookie theft to BEC: Attackers use AiTM phishing sites as entry point to further financial fraud. Microsoft Security Blog. https://www.microsoft.com/en-

us/security/blog/2022/07/12/from-cookie-theft-to-bec-attackers-use-aitm-phishing-sites-as-entry-point-to-further-financial-fraud/

8. Microsoft. (2023, June 8). Detecting and mitigating a multi-stage AiTM phishing and BEC campaign. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2023/06/08/detecting-and-mitigating-a-multi-stage-aitm-phishing-and-bec-campaign/
9. Mohammad, R. M., & McCluskey, L. (2012). Phishing websites [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C51W2X
10. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. Computer Science Review, 17, 1-24. https://doi.org/10.1016/j.cosrev.2015.04.001
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
12. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (pp. 108-116). SCITEPRESS. https://doi.org/10.5220/0006639801080116
13. Temoshok, D., Fenton, J. L., Choong, Y.-Y., Lefkovitz, N., Regenscheid, A., Galluzzo, R., & Richer, J. P. (2025). Digital identity guidelines: Authentication and authenticator management (NIST Special Publication 800-63B-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-63B-4
14. Vrbančič, G., Fister, I., Jr., & Podgorelec, V. (2020). Datasets for phishing websites detection. Data in Brief, 33, Article 106438. https://doi.org/10.1016/j.dib.2020.106438
15. Western OC2 Lab. (n.d.). Intrusion-Detection-System-Using-Machine-Learning [Data set and code repository]. GitHub. https://github.com/Western-OC2-Lab/Intrusion-Detection-System-Using-Machine-Learning
16. World Wide Web Consortium. (2021). Web authentication: An API for accessing public key credentials - Level 2 (W3C Recommendation). https://www.w3.org/TR/webauthn-2/