

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2026

Volume 41, Pages 87-100

ICRETS 2026: International Conference on Research in Engineering, Technology and Science

Early Prediction of Tactical Transitions in Cyber Attack Processes Through Network Flows

Ozge Tas
Kastamonu University

Abstract: Cyber attacks are becoming increasingly complex, comprehensive, and dynamic. This makes it crucial not only to detect malicious activity but also to predict how an attack may progress in its early stages. This study predicts MITRE ATT&CK tactical transitions using Zeek-derived network flows from the UWF-ZeekData24 dataset (Elam et al., 2025). The core objective is to examine whether the next tactic can be inferred from time-ordered network behavior by treating attack activity as a temporal process. The analysis uses 1,916,757 raw flow records, removes duplicate records from the main analytical pipeline, aggregates flows into fixed-duration windows, and constructs a leakage-controlled next_tactic target. Static machine learning models, temporal deep learning architectures, and a first-order Markov baseline are evaluated under chronological train-validation-test partitioning. The best static models, Random Forest and XGBoost, reached test Macro-F1 = 0.4996, Balanced Accuracy = 0.6667, and Top-2 Accuracy = 0.9973 in the 10-minute window configuration. The best temporal model, LSTM with sequence length 3, matched but did not exceed this performance, while the Markov baseline achieved the same Top-1 and Top-2 accuracy. A reviewer-motivated ablation further compared raw-flow-only, tactical-context-only, and full feature settings to clarify what signal drives the prediction. Early prediction analysis showed that useful next-tactic ranking remained possible when only 10% of the attack process was observed. The findings indicate that ATT&CK-based tactical forecasting from Zeek flow summaries, optionally assisted by current tactical context, can support proactive early warning for cyber attack progression modeling, although minority tactic recovery remains constrained by severe class imbalance.

Keywords: Cyber attack prediction, MITRE ATT&CK, Tactical transition prediction, Zeek network flows, Early warning systems

Introduction

Traditional intrusion detection systems usually frame the analytical task as a binary distinction between benign and malicious traffic. This framing remains useful, but it is limited for operational security centers because it describes the current state of an incident rather than the likely direction of adversarial behavior. As attacks become more complex and dynamic, security teams need early indicators of what an adversary may do next so that defensive playbooks can be prioritized before the full attack process unfolds.

The MITRE ATT&CK framework provides a structured vocabulary for representing adversarial tactics and techniques. In this study, ATT&CK is used not only as a labeling scheme but also as a temporal structure for modeling tactical movement across an attack process. This perspective shifts the analytical question from whether traffic is malicious to which tactical stage is likely to follow the currently observed behavior.

Zeek network telemetry is well suited to this objective because it summarizes connection behavior, service and protocol patterns, duration, packet counts, byte counts, and destination diversity without requiring payload inspection. When such flow records are aggregated into time windows, they can represent behavioral states that are compatible with tactical transition analysis. At the local-network level, Tas (2025) showed that protocol-specific weaknesses such as ARP spoofing can be examined through simulated attack architectures and anomaly

- This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 Unported License, permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

© 2026 Published by ISRES Publishing: www.isres.org

analysis, underscoring the value of behavioral network evidence for identifying malicious activity. At the modeling level, Tas (2025) emphasized that target leakage, training-test contamination, preprocessing leakage, and disregard of temporal order can inflate predictive performance. Together, these observations support the present study's use of Zeek-derived behavioral summaries, chronological partitioning, and explicit leakage control when forecasting ATT&CK tactical transitions.

This study aims to predict early-stage MITRE ATT&CK tactical transitions using network flows obtained from the UWF-ZeekData24 dataset. The contribution is a leakage-controlled and chronologically evaluated next-tactic prediction framework for cyber attack progression modeling. Unlike classical intrusion detection studies, the proposed approach offers a proactive early warning view of attack evolution rather than a reactive malicious/benign decision. The framework is positioned as a ranked decision-support layer rather than a fully autonomous replacement for analyst judgment.

The study addresses three research questions: Can Zeek-derived window features predict the next MITRE ATT&CK tactic with operationally meaningful performance? Do temporal deep learning models improve over static models and a Markov baseline? How early can tactical transition prediction remain informative when only a partial attack process is observed?

Literature Review

Recent cybersecurity analytics increasingly emphasize behavioral context rather than isolated malicious/benign labels. The MITRE ATT&CK framework is central to this shift because it organizes adversary behavior into tactics, techniques, and procedures that can be used to reason about attack progression (MITRE, 2026). ATT&CK-oriented analytics also support threat-informed defense by connecting observed events to adversarial intent, campaign stage, and response planning. Recent systematization studies further show that ATT&CK is now used across detection engineering, threat intelligence, risk analysis, and incident response, while also identifying the need for stronger operational evaluation of ATT&CK-based methods (Al-Sada et al., 2023; Roy et al., 2023).

Prior work on ATT&CK-based cyber threat intelligence has shown that unstructured threat reports and security observations can be mapped to tactics and techniques, creating a bridge between low-level telemetry and higher-level adversary behavior (Husari et al., 2017). More recent work has extended this mapping problem to NIDS rules and analyst-facing alert explanations, indicating that ATT&CK labels can support SOC triage when combined with machine learning or language-model-assisted enrichment (Daniel et al., 2024). This literature motivates the present study's focus on tactical transitions. Instead of treating ATT&CK labels as static categories, the study models them as ordered states whose next step may be predicted from current network behavior and tactical context.

Network-flow intrusion detection research provides the second foundation of the study. Public IDS datasets and Zeek-like telemetry have been widely used to evaluate machine learning models, but dataset realism, temporal dependence, and class imbalance remain recurring concerns (Sharafaldin et al., 2018; Ring et al., 2019). Recent work using MITRE ATT&CK-labeled data has also examined RandomForestNN-style classification under adversarial AI black-box conditions, showing that ATT&CK-labeled network traffic continues to be an active testbed for robust machine learning research (Mink et al., 2026). For this reason, the present design uses chronological splitting, leakage control, class-level reporting, and Top-k metrics rather than relying on raw accuracy alone.

The methodological literature further supports the use of balanced metrics and temporal baselines. In imbalanced domains, dominant classes can inflate accuracy while rare classes remain poorly recovered (Branco et al., 2016; He & Garcia, 2009). For future-state prediction, chronological validation is preferable to random splitting because random partitions can leak temporal structure into evaluation (Bergmeir & Benitez, 2012). Markov models remain useful as interpretable transition baselines, while LSTM-type architectures provide a comparison point for longer-range sequence dependence (Hochreiter & Schmidhuber, 1997; Rabiner, 1989).

A large portion of machine-learning-based intrusion detection research still evaluates models as if the operational objective were a single final label. Surveys and comparative studies of machine learning and deep learning in cybersecurity show that classification accuracy has often been reported as the primary success indicator, even when the target distribution is highly skewed and when the temporal structure of attacks is not represented explicitly (Apruzzese et al., 2018; Buczak & Guven, 2016; Xin et al., 2018). This convention is useful for

benchmarking, but it is less aligned with security operations, where analysts usually need to know what is unfolding, what may happen next, and which response path should be prioritized.

The cyber kill-chain perspective provides one conceptual basis for this shift from detection to progression analysis. Kill-chain models describe attacks as staged processes, and ATT&CK provides a more granular and operationally maintained vocabulary for representing adversarial behavior within those processes (Hutchins et al., 2011; MITRE, 2026). For network telemetry, this means that the analytical unit should not be only a connection or a flow record. It can also be a temporal window that summarizes the behavior of an attack stage and supports inference about the subsequent stage.

Recent deep learning studies have shown that sequence models can be effective for cyber intrusion tasks when meaningful temporal order is preserved (Ferrag et al., 2020). Recurrent models such as LSTM and GRU are designed to model temporal dependencies, while temporal convolutional networks and Transformer architectures provide alternative ways of representing ordered observations (Bai et al., 2018; Cho et al., 2014; Vaswani et al., 2017). However, these architectures do not automatically guarantee better operational value. If transitions are low-entropy or strongly dependent on the current state, simpler baselines may match complex models. This is why the present study compares temporal deep learning with both static machine learning and a Markov transition baseline.

Evaluation design is especially important in this setting. Time-dependent security data violate the assumptions behind ordinary random train-test partitioning, because nearby observations may share campaign-specific structure, infrastructure, and tactical state. Temporal validation is therefore a methodological requirement rather than a cosmetic choice (Bergmeir & Benitez, 2012). Dataset shift is another concern: if the distribution of tactics changes between training and testing periods, aggregate accuracy can overstate operational reliability (Moreno-Torres et al., 2012). Model selection bias is also possible when many algorithms are compared on a narrow evaluation design (Cawley & Talbot, 2010). The present study explicitly reports class distribution shift and treats model comparison as descriptive rather than as a statistically definitive ranking.

The study also draws from the literature on imbalanced learning and diagnostic metrics. When a dominant class accounts for most examples, conventional accuracy can remain high even if minority classes are ignored. Precision-recall based reasoning, macro-averaged metrics, balanced accuracy, and class-level reports are more appropriate in such conditions (Branco et al., 2016; Saito & Rehmsmeier, 2015). The Matthews correlation coefficient can also provide a more informative summary than accuracy in skewed binary settings (Chicco & Jurman, 2020), although the present paper emphasizes multiclass Macro-F1 and Top-k accuracy because the primary task is next-tactic ranking.

This study is positioned at the intersection of these strands. It extends Zeek-based intrusion analytics from binary detection toward ATT&CK-based next-tactic forecasting. Its contribution is not only model comparison, but also an evaluation design that asks whether early, ranked tactical suggestions can be produced under leakage-controlled and chronologically realistic conditions. In this sense, the study contributes an ATT&CK-based next-tactic prediction framework that transforms Zeek network flows into early warning signals for cyber attack progression modeling. The study therefore treats tactical transition prediction as a bridge between low-level network-flow features and higher-level security operations decision support.

Method

Dataset and Preprocessing

The empirical analysis uses the UWF-ZeekData24 dataset, an enterprise MITRE ATT&CK-labeled network attack traffic dataset released by the University of West Florida Cyber Analytics Research Group (Elam et al., 2025; University of West Florida, 2026). In this project, the data were stored as seven parquet partitions. The loading report shows 1,916,757 network flow records and 27 columns loaded successfully with the Polars engine. No file-level loading failure or schema mismatch was detected. Duplicate control identified 18,144 duplicate records, which were excluded from the main modeling pipeline, leaving 1,898,613 records for analytical processing.

The tactical distribution is highly imbalanced. Credential Access dominates the attack records, followed by Reconnaissance, Defense Evasion, Initial Access, and Exfiltration. At the next_tactic target level, Reconnaissance constitutes the largest window-level target class. Because of this imbalance, the analysis does not rely on accuracy alone and emphasizes Macro-F1, Balanced Accuracy, Top-k Accuracy, and per-class results.

Table 1. Dataset and modeling summary

Dataset characteristic	Value
Raw flow records	1,916,757
Loaded parquet files	7
Duplicate records removed	18,144
Records retained	1,898,613
Valid 10-minute transition windows	2,432
Final predictors	23
Temporal split	70% / 15% / 15%

Table 2. Literature positioning

Theme	Relevance to this study	Representative sources
ATT&CK framework	Adversary behavior is represented as tactics and techniques; this supports behavioral, phase-aware security analytics.	MITRE, 2026
ATT&CK analytics and CTI extraction	Recent ATT&CK-oriented work motivates mapping alerts and reports to tactics, techniques, and TTP sequences rather than using isolated labels.	Husari et al., 2017; MITRE, 2026
Network-flow intrusion datasets	Flow-level IDS research emphasizes dataset realism, temporal split discipline, and careful interpretation of high accuracy under skewed labels.	Sharafaldin et al., 2018; Ring et al., 2019
Imbalanced classification	Macro-F1, balanced accuracy, and class-level reports are necessary when dominant classes can inflate accuracy.	Branco et al., 2016; He & Garcia, 2009
Temporal and transition modeling	Chronological validation and Markov/sequence baselines are appropriate when the prediction target is a future tactical state.	Bergmeir & Benitez, 2012; Rabiner, 1989

Research Design

The study follows a quantitative predictive research design. Its purpose is not to explain the complete causal mechanism of adversarial behavior, but to test whether time-ordered Zeek flow summaries and tactical context can provide an early ranked estimate of the next MITRE ATT&CK tactic. The design is empirical because all modeling decisions are evaluated against observed windows derived from the UWF-ZeekData24 dataset. It is also temporal because observations are ordered by `window_start` and evaluated forward in time. The unit of analysis is the fixed-duration attack-behavior window rather than the individual raw flow. This choice is central to the research design. Individual flows are often too narrow to represent tactical behavior, whereas windows can summarize traffic volume, service diversity, byte and packet asymmetry, and tactical context. A window-level design also makes it possible to define `next_tactic` as a future tactical state and to evaluate early prediction before the whole attack process has been observed.

The dependent variable is `next_tactic`. For each active attack window, the next different tactic in the subsequent ordered windows is assigned as the prediction target. Windows without a later tactical transition are removed from the main next-tactic modeling dataset. This target construction operationalizes attack progression as a transition problem. It differs from conventional intrusion detection because the model does not simply ask whether traffic is malicious. It asks which tactical stage is likely to follow the current state. The research design includes three model families to separate three possible sources of predictive signal. Static machine learning models test whether a single window summary contains enough information for next-tactic prediction. Temporal deep learning models test whether short sequences of windows add useful information beyond the current summary. The Markov baseline tests whether the transition process is mostly explained by the current tactical state. This three-way comparison is important because a complex model is not automatically preferable when a simple transition model captures the same structure.

Window Construction and Leakage Control

Raw flows were aggregated into fixed-duration time windows. The principal configuration used 10-minute windows because preliminary sensitivity analysis indicated that this window length best balanced behavioral stabilization and alert latency. The resulting feature set contains 23 predictors: 18 numerical and 5 categorical variables. These variables summarize flow volume, duration statistics, packet and byte asymmetries, service and protocol structure, destination diversity, and current tactical context.

Table 3. Research design matrix

Design element	Configuration	Rationale
Research type	Quantitative, empirical, predictive modeling study	The study tests whether time-ordered Zeek windows can forecast the next ATT&CK tactic.
Unit of analysis	10-minute attack-behavior window	Raw flows are converted into comparable temporal observations.
Primary target	next_tactic	The target represents the next different MITRE ATT&CK tactic after the current tactical state.
Main predictors	Zeek behavioral summaries and tactical context	Network-flow statistics are combined with current and previous tactical state variables.
Validation logic	Chronological 70/15/15 split	Evaluation follows a forward-in-time design and avoids random temporal leakage.
Model families	Static ML, temporal DL, Markov baseline	The design compares feature-based learning, sequence learning, and transition probability modeling.
Primary metrics	Macro-F1, balanced accuracy, Top-k accuracy	Metrics were selected to avoid over-interpreting raw accuracy under severe class imbalance.
Operational question	Early ranked next-tactic warning	The output is intended to support SOC triage through a short list of likely next tactics.

Table 4. Validity controls

Validity concern	Control or interpretation
Label leakage	Ground-truth labels, next_tactic, identifiers, raw timestamps, and raw IPs were excluded from the feature matrix.
Temporal leakage	Train, validation, and test periods are ordered chronologically, with no random shuffling.
Class imbalance	Macro-F1, balanced accuracy, Top-k accuracy, and per-class F1 are reported instead of accuracy alone.
Tactical-context dependency	The paper frames results as tactical-context-assisted prediction and states the need for an upstream ATT&CK mapping component.
Model comparison	Findings are interpreted descriptively because the project uses one chronological split rather than repeated folds.
Binary auxiliary task	Binary detection is not treated as the main contribution because the test period is single-class.

A strict leakage-control strategy was applied because label leakage can artificially inflate predictive performance in data mining experiments (Kaufman et al., 2012). Ground-truth label fields, the dependent variable, raw timestamps, unique identifiers, raw IP addresses, and future-derived fields were excluded from predictors. The target next_tactic was constructed from the subsequent tactical state, and no_transition windows were removed. Thus, only features observable at scoring time were retained.

The tactical-context variables require a careful operational interpretation. In the offline experiment, tactic_current, tactic_previous, and tactic_changed_flag represent the currently inferred tactical state at the window level. In a deployed system, these variables would need to be supplied by an upstream ATT&CK mapping component, analyst-confirmed alert stream, or a current-tactic classifier. Therefore, the full-feature results are interpreted as tactical-context-assisted next-tactic prediction, while the ablation experiment separately tests whether Zeek-derived flow summaries alone can preserve the same ranking behavior.

The intended SOC workflow is as follows: Zeek logs are aggregated into behavioral windows; an upstream ATT&CK mapping or current-tactic classifier assigns the current tactical state when available; the forecasting layer ranks likely next tactics; and analysts use the Top-k output to prioritize playbooks, log sources, and defensive controls. This architecture keeps the model in a decision-support role and avoids treating next-tactic prediction as an automatic operational verdict.

Table 5. Next-tactic target distribution

Next tactic	Windows	Percent	Rank
Reconnaissance	2089	85.8964	1
Credential Access	248	10.1974	2
Defense Evasion	51	2.097	3
Initial Access	44	1.8092	4

Overall Research Workflow

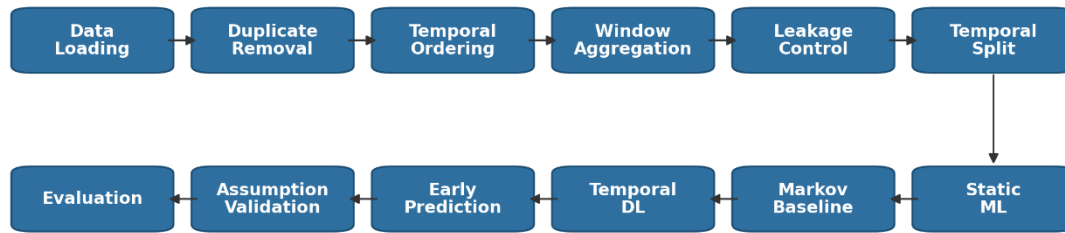


Figure 1. Research workflow

Data Processing Pipeline



Figure 2. Processing pipeline

Experimental Design

The modeling data were divided chronologically rather than randomly, following the principle that temporal prediction should be evaluated forward in time rather than by random mixing of past and future observations (Bergmeir & Benitez, 2012). The training period spans 2024-02-28 04:00:00 to 2024-03-20 20:30:00, the validation period spans 2024-03-20 20:40:00 to 2024-03-23 09:10:00, and the test period spans 2024-03-23

09:20:00 to 2024-03-27 22:10:00. The corresponding split sizes are 1,702 training windows, 364 validation windows, and 366 test windows.

Three model families were compared. Static models included Logistic Regression, Random Forest (Breiman, 2001), LightGBM, XGBoost (Chen & Guestrin, 2016), and MLP. Temporal deep learning models included LSTM, GRU, TCN, and Transformer architectures with sequence lengths of 3, 5, and 10. A first-order Markov chain with Laplace smoothing was used as an interpretable transition baseline. After the initial reviewer-style audit, an additional ablation experiment was added. This experiment compared three feature settings under the same chronological split: raw-flow-only features, tactical-context-only features, and the full feature set. The purpose was to distinguish whether prediction was driven by Zeek behavioral summaries, current tactical state, or their combination.

Results and Discussion

Next-Tactic Prediction

Table 6 reports test performance for static machine learning models. Random Forest and XGBoost achieved the strongest results, each reaching Accuracy = 0.9973, Macro-F1 = 0.4996, Balanced Accuracy = 0.6667, and Top-2 Accuracy = 0.9973. LightGBM and MLP were close, whereas Logistic Regression lagged behind, especially in Macro-F1 and Top-2 Accuracy.

The high accuracy should be interpreted cautiously. The test set is dominated by Reconnaissance and contains very few minority-class examples. Therefore, Macro-F1 and per-class F1 provide a more honest assessment of balanced multiclass recovery. Operationally, the results support ranked next-tactic recommendation, but they do not show uniform success across all ATT&CK tactics.

Table 6. Static model test results

Model	Accuracy	Macro-F1	Balanced Acc.	Top-2	Top-3
Logistic Regression	0.4973	0.4003	0.4766	0.9426	1
Random Forest	0.9973	0.4996	0.6667	0.9973	1
Lightgbm	0.9918	0.4988	0.6646	0.9973	1
Xgboost	0.9973	0.4996	0.6667	0.9973	1
Mlp	0.9945	0.4964	0.6591	0.9973	1

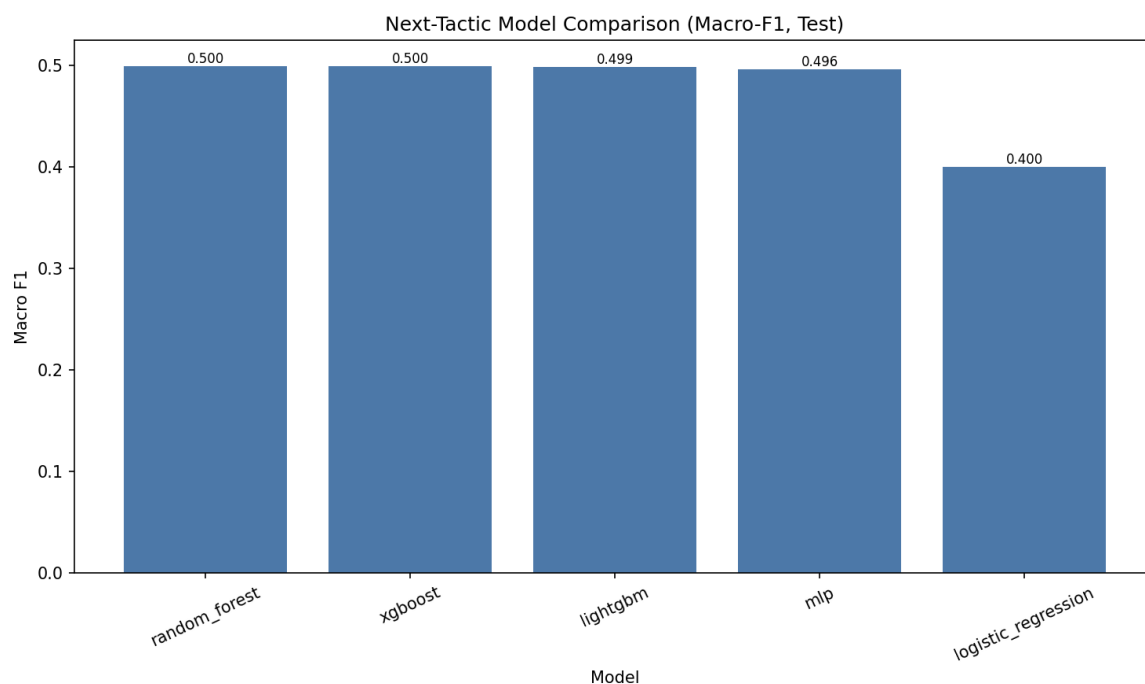


Figure 3. Static model comparison

The contrast between Accuracy and Macro-F1 is one of the most important findings of the study. Accuracy suggests that the task is almost solved, but Macro-F1 shows that the solution is uneven across tactics. This difference is expected in a transition dataset where one or two tactical paths dominate the chronology. For an operational SOC, the implication is that the model may be valuable as a candidate-ranking tool, but it should not be used as an autonomous rare-tactic detector without additional balancing or confirmation mechanisms.

The static model results also show that nonlinear models are more appropriate than a linear baseline for this feature space. Logistic Regression performed substantially worse than tree-based models on Macro-F1 and Top-2 Accuracy. This does not mean that Logistic Regression is unusable; rather, it indicates that the interaction between flow-volume features, diversity features, service categories, and tactical context is not well captured by a simple linear boundary. The strong performance of Random Forest and XGBoost is consistent with their ability to handle heterogeneous feature types and nonlinear splits.

Feature Contribution Analysis: Flow Features and Tactical Context

The reviewer-motivated ablation analysis clarifies the source of predictive signal. The raw-flow-only Random Forest model matched the full-feature Random Forest on the test split, reaching Accuracy = 0.9973, Macro-F1 = 0.4996, Balanced Accuracy = 0.6667, and Top-2 Accuracy = 0.9973. This result shows that the study's title is not merely supported by tactical-context variables; Zeek-derived behavioral summaries alone can reproduce the dominant transition pattern in this dataset.

At the same time, the context-only condition should be interpreted cautiously. XGBoost with only tactical-context variables matched the full model, whereas Random Forest with the same variables performed poorly in Accuracy and Macro-F1 while still retaining high Top-k ranking. This model-family sensitivity indicates that the transition structure is low-entropy and state-dependent, but also that direct class prediction remains unstable under severe imbalance.

Overall, the ablation strengthens the paper's framing. The framework should not be presented as uniformly accurate multiclass prediction for all tactics. It is better described as an early next-tactic ranking approach in which Zeek window summaries and tactical context can each encode dominant transition regularities, while rare tactic recovery remains unresolved.

Table 7. Feature contribution analysis of flow and tactical-context features

Feature set	Model	Features	Accuracy	Macro-F1	Balanced Acc.	Top-2
Raw-flow-only	Random Forest	20	0.9973	0.4996	0.6667	0.9973
Raw-flow-only	XGBoost	20	0.9945	0.4992	0.6656	0.9973
Context-only	Random Forest	3	0.123	0.2516	0.3344	0.9973
Context-only	XGBoost	3	0.9973	0.4996	0.6667	0.9973
Full model	Random Forest	23	0.9973	0.4996	0.6667	0.9973
Full model	XGBoost	23	0.9973	0.4996	0.6667	0.9973

Temporal and Markov Models

Temporal deep learning did not produce a meaningful improvement over the best static models. The best temporal configuration was LSTM with sequence length 3, which matched the best static Macro-F1 and Top-2 Accuracy. This suggests that, at 10-minute resolution, much of the useful temporal information is already captured by the current window summary and tactical context variables.

The Markov baseline reinforces this interpretation. The first-order Markov chain achieved test Top-1 Accuracy = 0.9973 and Top-2 Accuracy = 0.9973, matching the strongest machine learning models on ranking-oriented metrics. Thus, the transition structure in this dataset appears strongly state-dependent. The practical implication is that interpretable transition baselines should be retained in operational systems rather than treated as weak reference models.

This result is valuable because it prevents an overstatement of deep learning. In many cybersecurity studies, the inclusion of LSTM, GRU, TCN, or Transformer architectures can create the impression that sequence complexity is necessary. The present results show a more cautious picture. At the selected temporal resolution, the next tactic is largely determined by the current tactical state and the dominant transition pattern in the dataset. Complex

temporal architectures may become more useful in higher-entropy campaigns, multi-actor environments, or datasets where tactical transitions are less repetitive.

The Markov result also improves interpretability. A transition matrix can be inspected by analysts, compared with threat intelligence expectations, and monitored for drift over time. If a deployed system observes that real transition probabilities deviate from the learned matrix, this could indicate a new campaign pattern or a change in adversary behavior. In this sense, the Markov baseline is not only a benchmark. It is also a diagnostic artifact for operational monitoring.

Table 8. Temporal model test results

Model	Seq.	Accuracy	Macro-F1	Balanced Acc.	Top-2
LSTM	3	0.9973	0.4996	0.6667	0.9973
GRU	5	0.9863	0.498	0.6625	0.9973
LSTM	10	0.9808	0.4972	0.6604	0.9973
TRANSFORMER	3	0.9616	0.4948	0.6531	0.9973
TRANSFORMER	10	0.956	0.4936	0.651	0.9973

Table 9. Markov baseline performance

Model	Split	N	Top-1	Top-2	Top-3
Markov chain	Test	366	0.9973	0.9973	1.0000

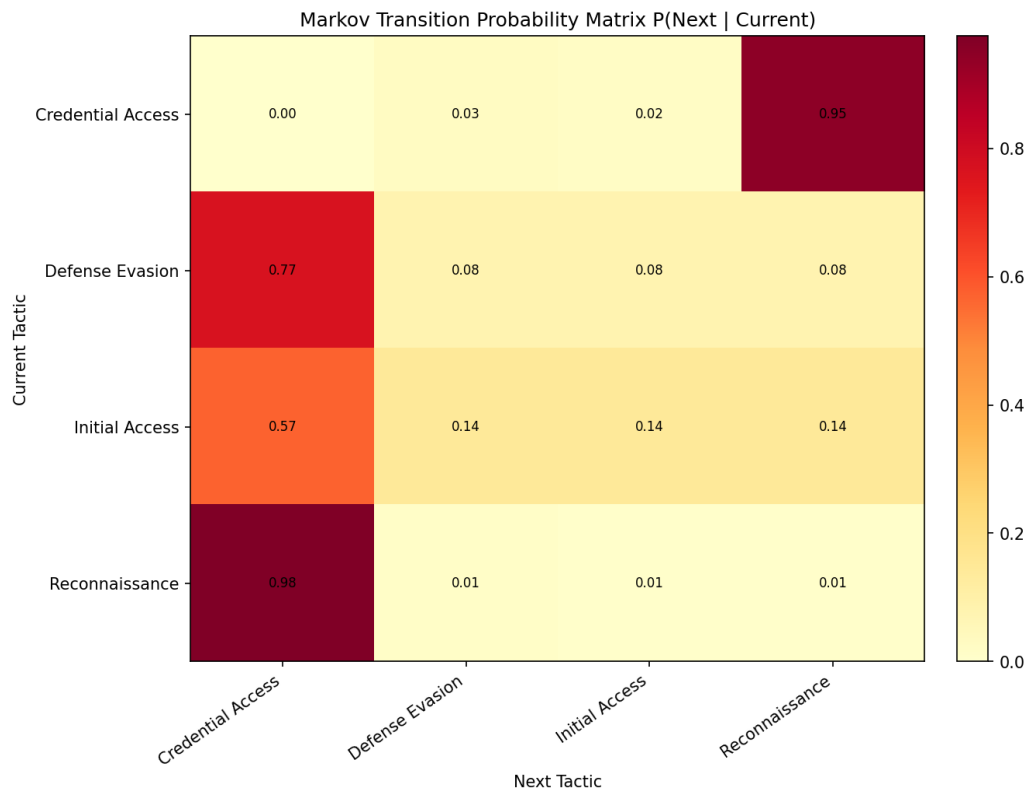


Figure 4. Transition probability heatmap

Early Prediction

A central goal of the study was to determine whether tactical transition prediction remains informative before the whole attack process is observed. Table 10 shows that at only 10% observation, the Random Forest model achieved Accuracy = 0.9178, Macro-F1 = 0.4879, Balanced Accuracy = 0.6667, and Top-2 Accuracy = 0.9178. Macro-F1 remained stable as observation increased, while accuracy and Top-2 performance generally improved. This result supports the proactive early warning premise of the study. The model can generate useful ranked next-tactic suggestions early in the attack process. However, the stability of Macro-F1 also indicates that observing more of the attack does not solve the minority-class bottleneck by itself.

The early prediction experiment should be interpreted as a partial-observability simulation. It does not prove that every SOC can obtain the same performance at the first minutes of a live attack. It shows that, under the present feature construction and tactical-context assumption, meaningful next-tactic ranking can be retained even when only a small portion of the post-training attack horizon is used for evaluation. This distinction is important because early warning systems must operate under uncertainty, delayed labels, and incomplete context.

The practical value of early ranking lies in decision support. A Top-2 tactical candidate list can help analysts prioritize playbooks, decide which logs to inspect next, and prepare controls associated with likely follow-on tactics. This is different from issuing a single deterministic prediction. Under imbalance and uncertainty, a ranked output is more realistic because it communicates the model's short list of likely tactical states rather than pretending that the next stage is known with full certainty.

Table 10. Early prediction results

Observed %	N	Accuracy	Macro-F1	Balanced Acc.	Top-2
10	73	0.9178	0.4879	0.6667	0.9178
25	182	0.9066	0.4866	0.5	0.9396
50	365	0.9534	0.4935	0.5	0.9781
75	547	0.9689	0.4956	0.5	0.9872
100	730	0.9753	0.4965	0.5	0.9849

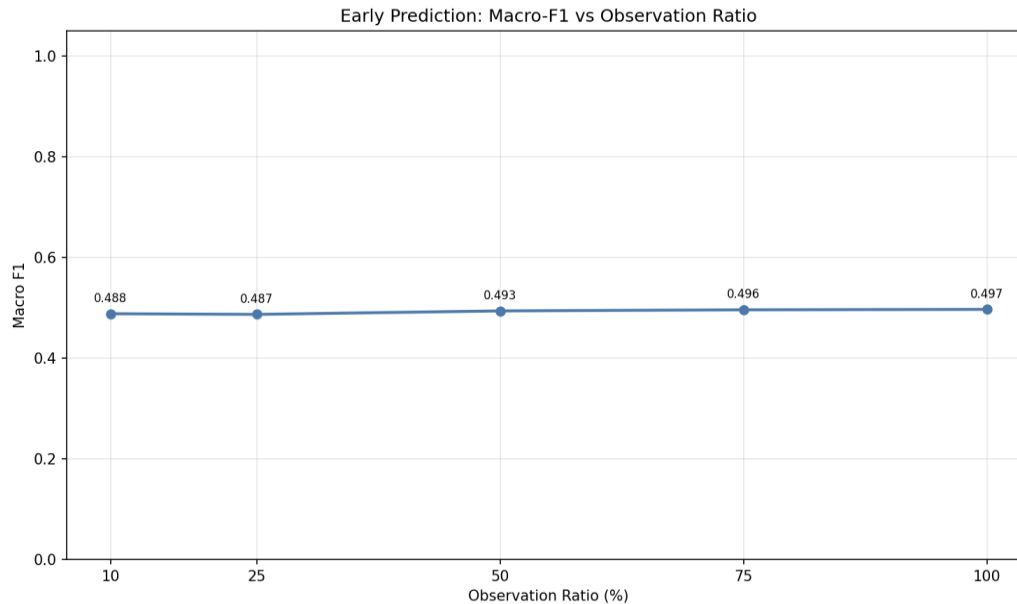


Figure 5. Early prediction curve

Window Size and Class-Level Behavior

Window-size sensitivity analysis showed that the 10-minute configuration provided the best overall balance. One-minute windows were too noisy, five-minute windows improved but remained weaker, and 30-minute windows reduced the number of test examples substantially. The 10-minute setting achieved the highest Macro-F1 while maintaining near-perfect Top-2 Accuracy. Class-level evaluation reveals the main limitation of the framework. Credential Access and Reconnaissance were recovered almost perfectly in the test split, whereas Initial Access had only one test example and was not recovered. Defense Evasion had no test support. Therefore, the proposed approach is strongest as an operational ranked-warning mechanism for dominant transitions, while rare tactic recovery requires additional data and imbalance-aware modeling.

The window-size result also provides an operational design recommendation. Very short windows can support low-latency alerting, but they may not accumulate enough behavioral evidence to represent a tactical state. Very long windows can stabilize behavior, but they reduce sample count and delay response. The 10-minute setting appears to balance these competing requirements for this dataset. This should not be generalized blindly to every network; rather, it provides an empirical starting point for SOC deployments that use Zeek-like telemetry.

The per-class results clarify why minority-aware design remains necessary. If an organization is most concerned with rare but high-impact stages such as Exfiltration or Defense Evasion, the current model would require additional data, synthetic augmentation, cost-sensitive learning, or hierarchical modeling. A future version could first predict a broader tactical family and then refine the prediction into a tactic or technique. Such a hierarchy may reduce the damage caused by extremely sparse leaf classes.

Table 11. Window-size comparison

Window	N	Accuracy	Macro-F1	Balanced Acc.	Top-2
1	2210	0.3937	0.2841	0.2935	0.7231
5	726	0.7755	0.3481	0.3837	0.8967
10	366	0.9973	0.4996	0.6667	0.9973
30	36	0.9444	0.4857	0.5	0.0556

Table 12. Per-class next-tactic results

Class	Precision	Recall	F1	Support
Credential Access	1	1	1	44
Defense Evasion	0	0	0	0
Initial Access	0	0	0	1
Reconnaissance	0.9969	1	0.9984	321

Interpretability and Assumption Checks

Permutation importance analysis identified `tactic_current` as the most important predictor in the full-feature model. Its disruption produced the only meaningful reduction in Macro-F1, whereas traffic-volume, byte-ratio, packet-ratio, and service-diversity features provided secondary support. The ablation analysis qualifies this result: although `tactic_current` dominates the full model's internal importance ranking, raw Zeek-flow summaries can still reproduce the dominant test-set transition pattern when tactical-context variables are removed. Assumption checks confirmed chronological split ordering and leakage-control integrity, but they also showed class-distribution shift and severe imbalance. These checks are important because they prevent overly optimistic interpretation of accuracy and clarify why Macro-F1, Balanced Accuracy, Top-k Accuracy, and per-class reporting are emphasized throughout the study.

The interpretability findings change how the contribution should be stated. The paper should not claim that all tactics are accurately predicted under balanced conditions. Instead, the evidence supports a combined view: Zeek-derived behavioral summaries and tactical-context variables both capture dominant transition regularities, but rare-tactic recovery remains limited. The model can be understood as a forecasting layer that operates after, or alongside, ATT&CK-oriented enrichment.

The assumption checks also strengthen the paper by making weaknesses explicit. The failed or partially satisfied assumptions are not hidden; they are incorporated into the interpretation. This is a better scientific position than reporting only the strongest metrics. In particular, the lack of a valid binary test regime and the absence of some minority tactics in the test set are treated as constraints on inference rather than as minor technical details.

Conclusion

This study examined whether Zeek-derived network flow behavior can be used to predict the next MITRE ATT&CK tactic before a cyberattack process fully unfolds. Using the UWF-ZeekData24 dataset, a leakage-controlled and chronological next_tactic prediction pipeline was developed. The strongest static models reached Macro-F1 = 0.4996, Balanced Accuracy = 0.6667, and Top-2 Accuracy = 0.9973 on the 10-minute window configuration. A new ablation analysis showed that raw-flow-only Random Forest performance matched the full-feature setting, while tactical-context-only performance was model-sensitive.

The results show that next-tactic prediction is feasible as a ranked early warning task. Temporal deep learning models did not outperform the best static models, and a first-order Markov baseline matched the strongest Top-k performance. This indicates that the tactical transition process in this dataset is largely state-driven at the chosen aggregation level.

The main contribution is therefore not simply higher classification accuracy, but a reframing of network intrusion analytics from reactive detection to proactive ATT&CK-based tactical forecasting and attack progression modeling. For security operations centers, the framework can support earlier prioritization of defensive actions by offering a short list of likely next tactics.

Recommendations

Future studies should validate the framework on multiple Zeek corpora, mixed attack-normal test periods, and higher-entropy attack campaigns. Rare tactics such as Defense Evasion and Initial Access require cost-sensitive learning, targeted resampling, hierarchical tactic-technique modeling, or additional attack scenarios. Operational deployments should retain a Markov baseline for interpretability and calibration, while using machine learning models to exploit richer behavioral features when transition entropy is higher.

Limitations

The study is limited by its dependence on a single dataset and a campaign-specific transition structure. The chronological test split contains few or no examples of some minority tactics, which restricts conclusions about rare transitions. The auxiliary binary detection experiment is not treated as a main contribution because the test period contains only normal traffic. Finally, feature importance results explain model behavior but should not be interpreted as causal evidence.

A further limitation concerns the tactical-context variables. Because `tactic_current` and `tactic_previous` are strong predictors, the full-feature framework assumes that current tactical state can be obtained before next-tactic scoring. This is plausible in a SOC pipeline that already maps alerts to ATT&CK tactics, but it is a dependency rather than a raw Zeek-flow property. The added raw-flow-only ablation reduces this concern for the present dataset, yet an end-to-end deployment should still include a validated current-tactic mapping component.

Model comparison is also limited by the single chronological split. This design is operationally realistic and avoids random temporal leakage, but it does not support strong statistical claims about algorithm superiority of the kind normally expected in multi-dataset classifier comparisons (Demsar, 2006). Repeated rolling-origin validation or cross-campaign testing would be needed before claiming that one model family is generally superior.

Data Availability Statement

The study uses the UWF-ZeekData24 dataset obtained from the University of West Florida Network Traffic Flow Datasets portal. The dataset provider states that the articles and datasets are distributed under the Creative Commons Attribution (CC BY) license. Derived analytical outputs are stored in the project workspace and include preprocessing scripts, modeling scripts, tabular summaries, and visual outputs used to prepare this manuscript. Dataset redistribution should follow the original dataset provider's access and licensing conditions.

Scientific Ethics Declaration

* The author declares that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the author.

* No human participants, personal intervention, or human-subject experimental procedure were involved in this study.

Funding

* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

* The author declares that there is no conflict of interest

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Research in Engineering, Technology and Science (www.icrets.net) held in Batumi/Georgia on July 09-12, 2026.

* The study is based on analytical outputs generated from the UWF-ZeekData24 dataset and the author's experimental pipeline.

References

- Al-Sada, B., Sadighian, A., & Oligeri, G. (2023). MITRE ATT&CK: State of the art and way forward. *arXiv preprint arXiv:2308.14016*. Retrieved from arXiv
- Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. In *2018 10th International Conference on Cyber Conflict (CyCon)* (pp. 371–390). IEEE.
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*. Retrieved from arXiv
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), 1–50.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176.
- Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY: Association for Computing Machinery.
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6.
- Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics.
- Daniel, N., Kaiser, F. K., Giladi, S., Sharabi, S., Moyal, R., Shpolyansky, S., Murillo, A., Elyashar, A., & Puzis, R. (2024). Labeling NIDS rules with MITRE ATT&CK techniques: Machine learning vs. large language models. *arXiv preprint arXiv:2412.10978*. Retrieved from arXiv
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
- Elam, M., Mink, D., Bagui, S. S., Plenkers, R., & Bagui, S. C. (2025). Introducing UWF-ZeekData24: An enterprise MITRE ATT&CK labeled network attack traffic dataset for machine learning/AI. *Data*, 10(5), 59.
- Ferrag, M. A., Maglaras, L., Moschogiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Husari, G., Al-Shaer, E., Ahmed, M., Chu, B., & Niu, X. (2017). TTPDrill: Automatic and accurate extraction of threat actions from unstructured text of CTI sources. In *Proceedings of the 33rd Annual Computer Security Applications Conference* (pp. 103–115). New York, NY: Association for Computing Machinery.
- Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*, 1, 80–106.

- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21.
- Mink, D., Simpson, A., Bagui, S. S., & Bagui, S. C. (2026). RandomForestNN classification for adversarial AI black-box techniques on MITRE ATT&CK labeled data. *Electronics*, 15(12), 2598.
- MITRE. (n.d.). *MITRE ATT&CK*. Retrieved from MITRE ATT&CK
- Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1), 521–530.
- Paxson, V. (1999). Bro: A system for detecting network intruders in real-time. *Computer Networks*, 31(23–24), 2435–2463.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
- Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*, 86, 147–167.
- Roy, S., Panaousis, E., Noakes, C., Laszka, A., Panda, S., & Loukas, G. (2023). SoK: The MITRE ATT&CK framework in research and practice. *arXiv preprint arXiv:2304.07411*. Retrieved from arXiv
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432.
- 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 (ICISSP)* (pp. 108–116).
- Tas, O. (2025a). An examination of the role of the address resolution protocol in computer networks and the ARP spoofing threat. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics*, 35, 1–9.
- Taş, Ö. (2025b). Tahmin modellemesinde veri sızıntısı: Sistematik kaynaklar, tespit stratejileri ve metodolojik etkileri. In Ç. Karamaşa (Ed.), *İstatistik alanında araştırma yöntemleri ve yenilikçi uygulamalar* (pp. 57–67). Ankara, Turkey: Platanus Publishing.
- University of West Florida. (n.d.). *UWF network traffic flow datasets*. Retrieved from UWF Network Traffic Flow Datasets
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008).
- Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6, 35365–35381.

Author(s) Information

Ozge Tas

Kastamonu University, Catalzeytin Vocational School, Department of
Computer Technologies Kastamonu/Catalzeytin, Kastamonu, Türkiye
ORCID iD: <https://orcid.org/000-0001-7220-5054>

*Contact e-mail: otas@kastamonu.edu.tr

* Corresponding author(s) contact email

To cite this article:

Tas, O. (2026). Early prediction of tactical transitions in cyber attack processes through network flows. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM)*, 41, 87-100.