УДК: 004.41

Bakundukize E.P., PhD student
Plekhanov Russian University of Economics
Russia, Moscow

MACHINE LEARNING-BASED DETECTION OF COVERT ADVANCED PERSISTENT THREATS IN ENTERPRISE SYSTEMS

Abstract: *The human factor frequently serves as the primary vector for the propagation of cyber threats within enterprise environments. While technical infrastructure typically operates as a cohesive, deterministic mechanism where anomalies can be isolated and mitigated using diagnostic tools, investigating covert (stealth) attacks necessitates a fundamentally novel system component. Modern enterprises and the industrial sector critically require intelligent defense and detection systems for covert threats, predicated on machine learning algorithms. The detection of such stealthy incursions demands a comprehensive approach encompassing feature extraction, holistic component analysis, high-precision predictive modeling, and the generation of actionable recommendations. This paper addresses the challenges associated with constructing knowledge bases from historical vulnerability data in corporate settings. We formalize the phenomenon of diagnostic information feature saturation and highlight the inherent risks of neural network overfitting. Furthermore, robust data processing methodologies and their algorithmic implementations are evaluated. Analyzing the statistical distribution of enterprise attack detection and contextualizing the human factor from a historical perspective are integral to modeling the manifestation of covert threats, serving as a primary criterion for vulnerability identification. The methodologies and findings presented herein establish a robust mathematical and empirical framework for transforming raw telemetry into structured knowledge. By systematically analyzing an enterprise's historical data across specific criteria and applying data mining to*

diagnostic information, critical constituent values are isolated. Ultimately, this research addresses applied problems necessitating the enhancement of internal and external parameter analysis to systematically uncover and mitigate covert Advanced Persistent Threats (APTs).

Keywords: *Data Processing, Information Security, Big Data, Data Mining, Machine Learning, System Analysis, Covert Cyber Attacks, Anomaly Detection.*

Бакундукизе Э.П., аспирант
Российский экономический университет имени Г.В. Плеханова
Россия, г. Москва

ОБНАРУЖЕНИЕ СКРЫТЫХ ПРОДВИНУТЫХ ПОСТОЯННЫХ УГРОЗ В КОРПОРАТИВНЫХ СИСТЕМАХ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ

Аннотация: *Человеческий фактор часто является основным фактором распространения киберугроз в корпоративной среде. В то время как техническая инфраструктура обычно работает как единый детерминированный механизм, где аномалии могут быть выявлены и устранены с помощью диагностических инструментов, для расследования скрытых атак требуется принципиально новый системный компонент. Современные предприятия и промышленный сектор остро нуждаются в интеллектуальных системах защиты и обнаружения скрытых угроз, основанных на алгоритмах машинного обучения. Обнаружение таких скрытых вторжений требует комплексного подхода, включающего извлечение признаков, целостный анализ компонентов, высокоточное прогнозирующее моделирование и разработку практических рекомендаций. В этой статье рассматриваются проблемы, связанные с созданием баз знаний на основе исторических данных об уязвимостях в корпоративных условиях.*

Мы формализуем феномен насыщения диагностической информацией и подчеркиваем риски, связанные с переобучением нейронных сетей. Кроме того, проводится оценка надежных методологий обработки данных и их алгоритмической реализации. Анализ статистического распределения случаев обнаружения атак на предприятии и учет человеческого фактора с исторической точки зрения являются неотъемлемой частью моделирования проявления скрытых угроз и служат основным критерием выявления уязвимостей. Представленные здесь методологии и результаты исследований обеспечивают надежную математическую и эмпирическую основу для преобразования необработанной телеметрии в структурированные знания. Систематический анализ исторических данных компании в соответствии с определенными критериями и применение интеллектуального анализа данных для получения диагностической информации позволяют идентифицировать критические компоненты. В конечном счете, это исследование направлено на решение прикладных задач, требующих улучшенного анализа внутренних и внешних параметров для систематического выявления и устранения скрытых продвинутых постоянных угроз (APT).

Ключевые слова: *Обработка данных, Информационная безопасность, Большие данные, интеллектуальный анализ данных, Машинное обучение, Системный анализ, Скрытые кибератаки, обнаружение аномалий.*

The analysis of historical data necessitates the continuous, intelligent processing of multidimensional parameters derived from vast datasets.

Such analytical measures are systematically implemented to identify the most vulnerable aspects of an enterprise’s production cycle, strategic planning, and economic infrastructure.

Accumulated technical specifications of products, chronologies of events documented by technical specialists, and operational logs collectively constitute this

historical data.

Typically, this data is classified and recorded by registration sensors and network telemetry devices. However, depending on their primary sector of production, many companies neglect systematic data aggregation. Lacking scalable cloud infrastructure and dedicated departments for implementing innovative Big Data solutions, they remain unequipped to store and process large arrays of diagnostic information.

To engineer an effective predictive Recommender System (RS), it is imperative to extract both qualitative and quantitative features from raw diagnostic logs. The formulation of predictive rules relies heavily on the temporal sequencing of this data. Organizations specializing in cybersecurity software provide services—either locally or via distributed cloud platforms—that heavily leverage machine learning (ML) architectures. Similarly, major corporations whose core operations are inherently tied to data processing excel in this domain.

Conversely, traditional enterprises rarely prioritize the implementation of rigorous data analysis. They miss critical opportunities to dynamically classify operational deficiencies or conduct supplementary telemetry collection to gauge systemic vulnerabilities. Instead, decision-making relies heavily on heuristic intuition. This paper proposes a formalized methodology for the processing and analytical evaluation of diagnostic data, focusing on the information systems of industrial enterprises and corporations.

In the majority of instances, enterprises fail to maintain comprehensive historical records of security events. High-quality analysis requires spatial-temporal data—that is, comprehensive diagnostic information regarding the internal environment of the investigated systems over time ^{t}.

Let the raw enterprise telemetry be denoted as a feature matrix $X \in R^{N \times M}$, where N represents the number of recorded network events and N denotes the dimensionality of the diagnostic features. For a developing predictive model to yield high-fidelity results, it is crucial to isolate a feature subspace $X' \cup X$ wherein the

primary vectors of vulnerabilities can be described with maximal variance and minimal noise.

This objective is heavily complicated by the sheer absence of continuous data streams. Furthermore, highly specialized expert assessments of system states often present significant integration challenges.

The most prevalent anomalies encountered in historical datasets include missing values (undefined parameter cells), anomalous outliers, uninformative duplicate entries, and formatting inconsistencies.

To ensure that forecasting algorithms generate reliable probability distributions, it is strictly necessary to perform preliminary data sanitization. The creation and operational maintenance of a predictive model mandate the use of historical data collected at uniform time intervals Δ_t, the rigorous normalization of the feature space, and the application of advanced data mining techniques.

To resolve the challenge of insufficient diagnostic information, rigorous mathematical data cleaning methodologies must be applied.

The utilization of statistical techniques enables the imputation of missing criteria within X. For a given feature vector x_j, missing values can be estimated utilizing proportional dependencies. The Pearson correlation coefficient r_{xy} is deployed to evaluate the linear dependency between features x and y, facilitating the optimal reconstruction of numerical parameters:

$$r_{xy} \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - y)^2}}$$

Where $\bar{x}$ and $\bar{y}$ denote the sample means. Features exhibiting a correlation threshold $|r_{xy}| > 0.85$ are considered highly collinear, and dimensionality reduction is subsequently applied to mitigate feature saturation.

The investigation of characteristics directly impacts the neural network training process. Over-saturating the model with redundant features ($X \in R^{N \times M}$ where $M \gg N$) exponentially increases the risk of overfitting. By examining the

interrelationships among data categories via correlation matrices, the algorithm maintains focus on the core predictive manifold, significantly mitigating the propagation of erroneous classifications (false positives) from the perceptron's output layer.

Consequently, the primary objective is to construct a robust dataset mapping the spatial data parameters of the research object to a binary classification space $Y \in \{0,1\}$, where 1 denotes a covert threat and 0 denotes benign operations.

The deployment of machine learning methods for the analysis of knowledge bases facilitates the optimization of enterprise information systems. A multi-algorithmic approach is proposed to maximize the probability of identifying covert attacks.

To identify nascent threat clusters without labeled data, Kohonen Self-Organizing Maps (SOM) are utilized. The SOM maps the high-dimensional input space R^M onto a low-dimensional (typically 2D) grid of neurons. For an input vector $x(t)$, the winning neuron c (Best Matching Unit) is identified using the Euclidean distance:

$$c = \arg\min_j ||x(t) - W_j(t)||$$

The synaptic weight vectors W_j are then updated iteratively according to the learning rate $\alpha(t)$ and the neighborhood function $h_{cj}(t)$:

$$W_j(t+1) = W_j(t) + \alpha(t)h_{cj}(t)[x(t) - W_j(t)]$$

Following clustering and feature extraction, supervised learning algorithms execute the final threat classification.

10; 11

To empirically optimize the neural network variants, the Binary Cross-Entropy (BCE) loss function is minimized during backpropagation:

$$\tau_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log_{(\widehat{y_i})} + (1 - y_i)\log(1 - (\widehat{y_i})]$$

Where y_i is the ground truth label and $\widehat{y_i}$ is the predicted probability of a covert

threat.

The integrated sequence of operations is formalized in Algorithm 1.

Algorithm 1: Intelligent Covert Threat Detection Pipeline

Input: Raw historical telemetry matrix X_{raw}, Labels Y_{raw} (if available).

Output: Classified threat status $\hat{Y}$ and System Recommendations.

Phase 1: Preprocessing

$$X_{clean} \leftarrow \text{ImputeMissing}(X_{raw})$$

Compute Correlation Matrix $R = \{r_{xy}\}$

$$X_{reduced} \leftarrow \text{RemoveCollinearFeatures}(X_{clean}, \text{threshold} = 0.85)$$

Phase 2: Unsupervised Feature Extraction

Initialize SOM weights W

For each epoch$t \in [1, T]$ do:

Update BMU c and weights $W_j(t)$

Extract cluster topologies C from SOM.

Phase 3: Supervised Classification

Train Ensemble Model (Random Forest / Neural Network) on$(X_{reduced}\theta C, Y)$

Minimize τ_{BCE} via gradient descent.

$$\hat{Y} \leftarrow \text{Predict}\ (X_{test})$$

Return $\hat{Y}$

To advance the predictive model, the principal components of the historical data knowledge base have been established.

These specific parameters provide a comprehensive descriptive representation of characteristics actively present within the internal enterprise environment.

Table 1 illustrates the presence or absence of specific systemic events tracked across three critical temporal stages. The degree of manifestation is codified; affirmative markers (+) indicate a high density of occurrence within the information system modules as isolated by the ML algorithms.

Table 1.

Temporal Evolution of Covert Threat Parameters

Threat Indicator / Characteristic	Baseline State (t0)	Pre-Modification (t−1)	Active Incursion (tattack)	Post-Modification (t+1)
Anomalous File I/O Velocity	+	–	–	+
Unauthorized Protocol execution	+	+	+	–
Sudden Dept/Unit Closure (Logical)	+	–	–	–
Access Rights Violations (Escalation)	-	+	+	–
Security Perimeter Fluctuation	+	+	–	+
Network Traffic Asymmetry	-	-	+	+

To avoid the risks of neural network overfitting [14], the isolated features were categorized into distinct technical domains, systematically narrowing the R^M feature space (Table 2).

Table 2.

Feature Engineering and Categorization Matrix

Feature Category	Data Type	Key Metrics Analyzed	ML Preprocessing Technique
Network Telemetry	Continuous	Packet size, Inter-arrival time, Bandwidth spikes	Min-Max Normalization, PCA
User Access Logs	Categorical	Login frequency, Geolocation, Privilege escalation	One-Hot Encoding, Frequency limits
System Diagnostics	Time-Series	CPU load, Memory dumps, Kernel panics	Moving Average Smoothing
Human Factor (HR)	Discrete	Employee turnover, Security policy violations	Ordinal Encoding

The feature identification system is architected on binary responses. The affirmative or negative status of values measures the magnitude of the impact over

a specific time interval. Establishing a high-quality dataset demands a rigorous focus on the specific nature of the vulnerabilities. Table 3 presents the comparative empirical performance of the integrated algorithms utilized within the analytical framework.

Table 3.

Representative Performance of Evaluated Classifiers

Algorithm	Accuracy	Precision	Recall	F1-Score	Computational Complexity
Decision Tree (DT)	0.884	0.852	0.860	0.856	$O(M \cdot$
Random Forest (RF)	0.941	0.935	0.920	0.927	$O(K \cdot M \cdot$
Neural Network (MLP)	0.958	0.941	0.955	0.948	$O(E \cdot N \cdot$
SOM + RF (Proposed)	0.962	0.950	0.948	0.949	Hybrid formulation

Integrating neural networks as an intelligent predictive methodology ensures a high probability of detecting the secondary effects of the human factor and uncovering the digital footprints left by malicious actors. The analytical criteria established unlock the full computational potential of the applied algorithms.

Not all enterprises possess a well-documented history of adverse operational events; consequently, the active search for covert attacks is frequently hindered by empirical data scarcity. To circumvent this limitation and refine the machine learning algorithms, advanced mathematical data processing, statistical imputation, and the intelligent mining of historical data were rigorously applied/

Currently, commercial Security Information and Event Management (SIEM) systems rely heavily on deterministic, signature-based comparisons with previous incidents

However, the environment of prolonged, covert Advanced Persistent Threats (APTs) requires a probabilistic, multi-tiered situational assessment. Relying on disparate, unintegrated security tools is fundamentally mathematically insufficient.

Through the rigorous formalization of data mining methodologies and the application of sophisticated classification (e.g., Random Forests, Neural Networks) and unsupervised clustering (SOM) algorithms, the formulated matrix of historical data significantly elevates the reliability of the predictive model. The proposed algorithmic pipeline demonstrates that continuous modernization of analytical methods—specifically dimensionality reduction, collinearity filtering, and rigorous loss optimization—is strictly necessary.

The practical significance of this research lies in its capacity to translate theoretical mathematical models and algorithmic architectures into a cohesive, deployable enterprise defense system. Such a system guarantees high reliability and computational performance under conditions of systemic vulnerability, dynamically mapping the future trajectory of threat detection within complex, large-scale corporate information systems.

References:

1. Albanese, M. Automated Cyber Situation Awareness Tools and Models for Improving Analyst Performance / M. Albanese, H. Cam, S. Jajodia // Cybersecurity Systems for Human Cognition Augmentation / eds. R. Pino, A. Kott, M. Shevenell. – Cham : Springer, 2014. – (Advances in Information Security ; Vol. 61). – DOI: 10.1007/978-3-319-10374-7_3.

2. Bacciotti, A. Stability and Control of Linear Systems / A. Bacciotti. – Cham : Springer, 2019. – 189 p. – (Studies in Systems, Decision and Control). – DOI: 10.1007/978-3-030-02405-5.

3. Brink, H. Real-World Machine Learning / H. Brink, J. Richards, M. Fetherolf. – Manning Publications, 2017. [Рус. пер.: Бринк Х., Ричардс Дж., Феферолф М. Машинное обучение. – СПб. : Питер, [б. г.]. – 336 с. – ISBN 978-5-496-02989-6].

4. Burnashev, R. A. Research on the Development of Expert Systems Using Artificial Intelligence / R. A. Burnashev, R. G. Gabdrahmanov, I. F. Amer [et

al.] // Advances in Intelligent Systems and Computing. – 2020. – Vol. 1051. – P. 233–242. – DOI: 10.1007/978-3-030-30604-5_21.

5. Burkov, A. The Hundred-Page Machine Learning Book / A. Burkov. – Andriy Burkov, 2019. [Рус. пер.: Бурков А. Машинное обучение без лишних слов. – СПб. : Питер, 2020. – 192 с. – ISBN 978-5-4461-1560-0].

6. Witten, I. H. Data Mining: Practical Machine Learning Tools and Techniques / I. H. Witten, E. Frank. – 2nd ed. – Morgan Kaufmann Publishers, 2005. – 560 p. – ISBN 0-12-088407-0.

7. Dey, R. Stability and Stabilization of Linear and Fuzzy Time-Delay Systems: A Linear Matrix Inequality Approach / R. Dey, G. Ray, V. E. Balas. – Cham : Springer, 2017. – 267 p. – (Intelligent Systems Reference Library ; Vol. 141). – DOI: 10.1007/978-3-319-70149-3.

8. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, J. Friedman, R. Tibshirani. – New York : Springer, 2001. – 536 p. – (Springer Series in Statistics). – DOI: 10.1007/978-0-387-21606-5.

9. Хейсти, Т. Основы статистического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование / Т. Хейсти, Р. Тибширани, Дж. Фридман. – 2-е изд. – М. : Диалектика-Вильямс, 2020. – 770 с. – ISBN 978-5-907144-42-2.

10. Зыков, С. В. Основы проектирования корпоративных систем / С. В. Зыков. – М. : Изд. дом ВШЭ, 2012. – 431 с. – ISBN 978-5-7598-0862-6.

11. Зыков, С. В. Технология интеграции гетерогенного контента в корпоративных информационных системах / С. В. Зыков // Вопросы кибербезопасности. – 2015. – № 4. – С. 48–52.

12. Luisi, J. Pragmatic Enterprise Architecture: Strategies to Transform Information Systems in the Era of Big Data / J. Luisi. – 1st ed. – Morgan Kaufmann, 2014. – 372 p. – ISBN 978-0128005026.

13. Ou, X. Quantitative Security Risk Assessment of Enterprise Networks

/ X. Ou, A. Singhal. – New York : Springer, 2011. – 28 p. – (SpringerBriefs in Computer Science). – DOI: 10.1007/978-1-4614-1860-3.

14. Xu, Z. Graph Enhanced Memory Networks for Sentiment Analysis / Z. Xu, R. Vial, K. Kersting // Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017 / eds. M. Ceci, J. Hollmén, L. Todorovski [et al.]. – Cham : Springer, 2017. – (Lecture Notes in Computer Science ; Vol. 10534). – DOI: 10.1007/978-3-319-71249-9_23.

15. Jones, A. Risk Management for Computer Security: Protecting Your Network and Information / A. Jones, D. Ashenden. – 1st ed. – Butterworth-Heinemann, 2005. – 296 p.

16. Hinkel, G. NMF: A Multi-platform Modeling Framework / G. Hinkel // Theory and Practice of Model Transformation. ICMT 2018 / eds. A. Rensink, J. Sánchez Cuadrado. – Cham : Springer, 2018. – (Lecture Notes in Computer Science ; Vol. 10888). – DOI: 10.1007/978-3-319-93317-7_10.

17. Шелухин, О. И. Сетевые аномалии: обнаружение, локализация, прогнозирование / О. И. Шелухин. – М. : Горячая линия – Телеком, 2019. – 448 с. – ISBN 978-5-9912-0756-0.

18. Chollet, F. Deep Learning with Python / F. Chollet. – Manning Publications, 2018. [Рус. пер.: Шолле Ф. Глубокое обучение на Python. – СПб. : Питер, [б. г.]. – 400 с. – ISBN 978-5-4461-0770-4]