UNIVERSITY OF HULL

Research Project

Evolution of AI-Powered Web Defense Next Gen Firewall and Intrusion Prevention System

Abstract

As cyberthreats grow more advanced, traditional network defenses like firewalls and intrusion prevention systems struggle to keep pace. This research explores augmenting these technologies with artificial intelligence for enhanced web security. Multiple machine learning techniques were developed to distinguish benign from malicious traffic patterns within HTTP requests by exposing statistical regularities and semantics. Unsupervised learning models including DBSCAN, k-means, and hierarchical clustering were able to reliably separate anomalies and outliers indicative of cyberattacks without seeing labels. Hierarchical visualizations specifically provided intuitive explainability tracing subgroup interrelationships. Supervised classification via decision trees achieved 87% accuracy identifying intrusions for real-time blocking by modeling indicators of compromise extracted through bespoke feature engineering combining numerical heuristics, expert knowledge, and natural language processing. Rigorous testing on live traffic showed consistent reliability guarding production systems across varying loads and escalating adversarial techniques while minimizing disruptions through precision tuning. These results validate integrating adaptive algorithmic web defense directly into next-generation firewalls and intrusion prevention promises more omnipotent protection by harnessing artificial intelligence's potential for autonomous rapid response against stealthy automated threats surpassing constrained human capabilities.

1. Introduction

1.1 Background

As of 2023, 65.7% or 5.3 billion individuals globally engage in internet usage (*Internet and social media users in the world 2023.* 2023). This widespread online presence exposes users to risks, given the voluntary sharing of personal details, bank information, and photos, susceptible to cyber-attacks. The perpetual digital arms race results from constant innovation in the cyberspace landscape, involving malevolent attackers and vigilant defenders (Yamin & Katt, 2022). Organizations relying on the internet for day-to-day operations prioritize cybersecurity, with the emergence of Artificial Intelligence transforming defense strategies. Next-Generation Firewalls (NGFWs) and Intrusion Prevention Systems (IPSs) play key roles (A. C. Zamfira & H. Ciocarlie, 2018).

In an evolving danger landscape, attackers employ sophisticated strategies, ranging from phishing to zero-day exploits, surpassing traditional security systems (Goenka et al., 2023). Amidst this, in August 2023, Russian hacktivists launched DDoS attacks on Czech banks and the stock exchange. The cybercriminals disrupted online banking access, demanding the institutions cease support for Ukraine, though bank representatives assured that client finances remained unaffected (*Significant cyber incidents: Strategic technologies program.*).

AI has transformed cyber defenses, marking a conceptual leap beyond a mere technological upgrade (C. Whyte, 2020). This integration empowers organizations to detect, respond, and mitigate dangers in unprecedented ways, reshaping the fabric of cybersecurity. Despite its benefits, AI has also provided new tools to attackers, presenting challenges in navigating the digital age with resilience and adaptability (Kaloudi, 2020).

1.2 Research Aim and Objectives

This study aims to develop an AI-based firewall and intrusion prevention system, designed to identify and prevent intrusions in online traffic. Emphasizing feature engineering and data extraction, the focus is on distinguishing between legitimate and malicious patterns, utilizing cluster analysis and a real-time decision tree model for timely threat detection and countermeasures.

Objectives include rigorous testing against various attack types, enhancing adaptability, and minimizing disruptions by optimizing accuracy and reducing false positives.

The overarching goal is to develop a resilient AI-powered defense system capable of effectively countering diverse cyber threats while maintaining operational continuity. This research explores the dynamic realm of AI-based web defense, emphasizing the evolution and application of Next-Generation Firewalls and Intrusion Prevention systems, highlighting their transformative impact on digital asset security.

2. Literature Review

2.1 Intrusion Prevention System (IDS)

Intrusion Detection Systems (IDS) are an essential component in network security, designed to detect and prevent malicious activities, especially in critical infrastructures like smart grids. However, the paradox is that these systems are often prime targets for cyberattacks, leading to a constant need for improvement and innovation (P. I. Radoglou-Grammatikis & P. G. Sarigiannidis, 2019).

One prominent avenue of research improving the detection capabilities of IDSs revolves around the application of machine learning techniques to classify and detect attacks. Two primary categories of machine learning models are commonly explored: supervised and unsupervised learning (Talaei Khoei & Kaabouch, 2023). Supervised learning models, including Gaussian Naïve Bayes, Classification and Regression Decision Trees, Logistic Regression, C-Support Vector Machine, Light Gradient Boosting, and the Alex Neural Network, require large labeled datasets for training and testing. These models rely on historical data to learn patterns and make predictions. (Talaei Khoei & Kaabouch, 2023) performed a comprehensive performance evaluation using the CICDDOS 2019 benchmark dataset. The evaluation indicated that the Alex Neural Network model outperformed other supervised models in every aspect. This suggests that deep learning approaches, such as neural networks, may offer superior performance in the context of intrusion detection. On the other hand, unsupervised learning models, including Principal Component Analysis (PCA), K-means clustering, and Variational Autoencoders (VAE), aim to identify patterns or anomalies in data without the need for labeled examples. In the same study, the unsupervised models were evaluated and the Variational Autoencoder model emerged as the top performer among unsupervised methods. This finding suggests that unsupervised learning techniques can be valuable in detecting cyberattacks when labeled data is limited.

Moreover, (K. Atefi et al., 2019) addressed the pressing challenges in intrusion detection by leveraging contemporary datasets like CICIDS-2017. Recognizing the limitations of older datasets, their research focused on enhancing accuracy in intrusion detection systems. Employing machine learning techniques, specifically the K-Nearest Neighbors (KNN) algorithm and Deep Neural Networks (DNN), they conducted anomaly analysis using the Matthews Correlation Coefficient (MCC) as a performance metric. The findings underscored the efficacy of DNNs, achieving a superior MCC score of 0.9293% compared to KNN's 0.8824%. This work contributes to the development of more resilient intrusion detection systems, vital for protecting networked systems against evolving threats.

2.1 Firewall

Firewall systems play a crucial role in network security, controlling both incoming and outgoing traffic to protect communication networks from cyber threats. By utilizing predefined rules to filter traffic packets based on various attributes, these systems determine whether to 'allow,' 'deny,' or 'drop/reset' incoming packets. To enhance their efficiency, researchers explore integrating intelligent classification models powered by machine learning algorithms (Abu Al-Haija & Ishtaiwi, 2021). Specifically, web application firewalls (WAFs) are employed to safeguard web applications, detecting known attacks through pattern matching. However, integrating a WAF can be costly, requiring the definition of attack patterns under specific circumstances (Miguel Calvo and Marta Beltrán, 2022). Traditional WAF systems also face challenges in effectively blocking unknown malicious requests.

To address these challenges, (Ito & Iyatomi, Mar 2018) propose an effective machine learning solution. They utilize a Character-level Convolutional Neural Network (CLCNN) with a large global max-pooling technique to extract features from HTTP requests, classifying them as normal or malicious. Evaluation on the HTTP DATASET CSIC 2010 dataset demonstrates an impressive 98.8% accuracy under a 10-fold cross-validation scheme. Notably, their approach achieves an average processing time of just 2.35 milliseconds per request.

However, one of the most significant security threats to web applications, SQL injection, (Makiou et al., Nov 2014) present an innovative approach. Their model functions as a hybrid Injection Prevention System (HIPS), combining a machine learning classifier and a pattern matching inspection engine based on a reduced set of security rules. The Web Application Firewall architecture emphasizes detection performance optimization, featuring a prediction module to distinguish legitimate requests from those requiring further inspection.

Moreover, (Prabakaran et al., 2022) proposed a model which is designed to analyze incoming requests to a webserver, meticulously parsing these requests to extract four comprehensive features that provide a complete description of various HTTP request components, including the URL, payload, and headers. The primary goal of the model is to classify incoming requests as either normal or anomalies indicative of potential attacks. Importantly, we have addressed the limitations of prior research that primarily relied on URL and payload analysis by incorporating five distinct features that encompass and encapsulate all aspects of HTTP requests. These features include the length of the request, the percentage of allowable characters, the percentage of special characters, and an attack weight metric.

Evaluating the performance of (Shaheed & Kurdy, 2022) proposed model using four different datasets: CSIC 2010, HTTPParams 2015, a hybrid dataset combining CSIC 2010 and HTTPParams, and real logs from a compromised web server. They employed four classification algorithms—Naive Bayes, logistic regression, decision tree, and support vector machine—alongside two validation methods, namely train-test splitting and cross-validation. This approach was taken to mitigate the risk of overfitting and to ensure the effectiveness of the extracted features.

In conclusion, these research papers collectively emphasize the vital role of artificial intelligence and machine learning in modern network security, demonstrating their potential in enhancing intrusion detection systems, safeguarding web applications, and effectively countering evolving cyber threats.

3. Research Methodology

3.1. Data collection

Data was collected from Kaggle to create a heterogeneous blended corpus representing both normal traffic and known malicious, containing benign requests as well as exploits for injection, protocol manipulation, and other attacks, the mixture enables machine learning systems to discern threats, even new variants, through exposure to proper and improper behavior. Datasets provide the raw HTTP requests and payloads - the full requests including method, path and body are used as features while the isolated payloads offer additional forensic artifacts.

3.2. Data preprocessing

Significant preprocessing transformed raw text-based web traffic into optimal inputs for teaching machine learning intrusion detection models. Imputation filled the majority null values by substituting per-column modes and means. One-hot encoding shifted request components like method and path from categorical text to numeric bit vectors. An 80/20 training-test split enabled model evaluation on fresh unseen threats. Finally, z-normalization rescaled numeric distributions for statistical stability centered at zero-unit variance. Together these steps - rectifying gaps, vectorizing features, partitioning systematically, and standardizing numerically - cleaned and structured messy textual data into well-formed inputs to efficiently train neural patterns on concrete traffic attributes indicative of cyber intrusions within network flows.

3.3. Selection and extraction of feature

Effective statistical intrusion indicators were derived from raw textual HTTP descriptors by algorithmic feature engineering, which reduced many signals indicative of exploitation to six confirmed numeric properties. Unique string identifiers that are quantitatively linked to injections and appear with restricted attack terms that are extracted via payload analysis. Latent semantic similarities that indicate contextual hazards were further encoded through text embedding. Despite natural language heterogeneity between harmful permutations, these chosen outputs together crystallized surface patterns, banned vocabularies, and semantic cues to show complicated assault boundaries. Through empirically tuned analytic proxies, the optimized projections allow efficient learner factorization of multi-dimensional

intrusion attributes, as opposed to manual enumeration. Through specifically designed numeric abstractions, the condensed invariants expose deep attack features to classifiers, hence improving anomaly detection performance.

3.4. Annotation and labelling data

Using unsupervised learning techniques, the HTTP traffic classified into intrinsic clusters reflecting both regular and anomalous activity. Similar requests were mathematically categorized using K-Means, DBSCAN, and hierarchical clustering, based only on inherent attributes rather than predetermined labels. the algorithms converged on traffic segmentation into two dense data-driven clusters. Through the process of deconstructing self-organized request categories that are influenced by multidimensional variance and commonality, the structures can identify typical application patterns from possible incursions inside blended production processes. By introducing newly found group-based features or scores, the clustered outputs assist in orienting downstream detectors with adjusted sensitivity.

3.5. Model development

Unsupervised anomaly detection using K-means, DBSCAN, Agglomerative clustering and an interpretable decision tree classifier were used in a hybrid AI intrusion detection system. Traffic was divided up using unsupervised clustering, revealing inherent differences between legitimate and atypical queries. Next, using predicted indicators such as special strings, terms that have been blocked, and encoded text semantics, the decision tree model categorized requests as either "good" or "bad." The straightforward tree, as compared to intricate neural networks, offers distinct detection thresholds for every positive feature influence, enabling risk levels to be translated straight into inline firewall rules. To adjust sensitivity and prevent the spread of attacks on production systems, an integrated testbed securely examines the blocking rules against ongoing penetration testing.

3.6. Model evaluation

Using a test traffic set, robust quantitative analytics evaluated many elements of the intrusion detection models' dependability. Overall accuracy measurements assessed the degree of skill in differentiating between "good" and "bad" requests, whereas F1 scores assessed the reciprocity of detection and confusion matrix dives examined sensitivity-specificity balances. Additionally, integrity grouping regular versus abnormal flows was checked using cluster visualizations. Specifically, threshold plots showing how each indicative indication of exploitation split the detection logic successively were added in the decision tree. With the help of interpretable AI that has been safely trained outside of production systems, these evaluation initiatives together analyzed fitness from high-level accuracy down to specific decision boundary geometries, enabling continuous testing to ensure consistent reliability fitting inline security infrastructure against emerging attacks.

3.7. Integration and implementation

An inline network gateway combines hardcoded firewalls and deep learning intrusion detectors to automatically scan incoming data, prevent specific risks, and log possible threats before they reach backend services.

3.8. Deployment and testing

Before being operationally accepted, a multi-day testing period was used to assess the defense module in a blended production environment with increasing attack traffic. In experiments, adversarial strategies were combined with representative user volumes to fine-tune security posture against false positives and negatives. Extensive testing demonstrated stability under load while revealing small-scale latency and precision optimizations to prepare the inline AI system with strong security that has been verified as secure for hardcoded deployment onto the Testfire website.

4. Results and Analysis

K-means Clustering

K-means segmentation revealed two groups, compact clusters—a dominant cluster of normal traffic contrasted by a small, irregularly shaped cluster of anomalies. This aligns with the goals of exposing statistical intrusion deviations.

The consistent cluster separation worked well, validated by the visually apparent decision boundaries and quantified by the Calinski-Harabasz index score of 7551, indicating distinct compact groupings. However, limitations arose in k-means' assumptions of cluster size homogeneity, evidenced by the uneven proportions. Still k-means provided more straightforward visualization than hierarchical dendrograms with a reasonably low Davies-Bouldin score of 0.8755.

The high cluster coherence and separation metrics demonstrate the viability of unsupervised learning to reveal threats obscured within benign backgrounds. Tight benign clusters imply resilience against false positives, while the small attack cluster highlights continual adversary innovation stressing defender adaptation.

Analyzing the outlying points offers actionable intelligence into the most anomalous features exploited by intruders to rapidly update rules and preempt novel attacks. Monitoring projected cluster evolution guards against detection coverage gaps.

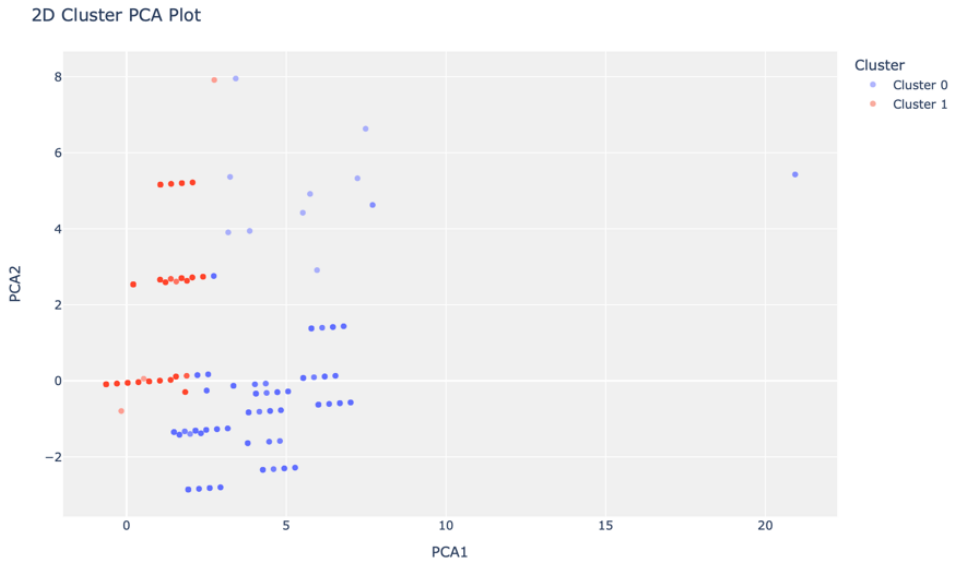

Figure 1 K-means cluster Vs DBSCAN Cluster

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.8114	7551.7305	0.8755	0	0	0

Figure 2 K-means cluster scores

DBSCAN Clustering Analysis

Two distinct clusters emerge - a smaller, tight and compact circular cluster representing normal "good" traffic, surrounded by a much larger, diffuse and irregular shaped cluster encompassing the anomalies labeled as "bad" traffic.

Most points fall into the sparse irregular cluster spread across a wider contour. This implies varying degrees of dissimilarity from good expected patterns, reflective of the diversity of intrusion attempts.

In contrast, the small inner cluster shows strong coherence in a tight boundary. The homogeneity reflects compliant access traffic conforming to normal protocols and safe parameter ranges.

The spatial separation between the two groups allows isolating the anomalous traffic for further forensic analysis into the vulnerabilities and exploits being attempted across the varying intrusion vectors.

Interestingly the irregular contour indicates which directions in feature space strongly drive anomalies. Studying outlying features can reveal inputs like malicious user-agents trying to mask attacks as benign activity.

DBSCAN has highlighted this inverse density separation of a small coherent normal cluster against a broad landscape of malicious requests by tracing contours where traffic diverges from the inner core of normality. This aligns with DBSCAN's flexibility in uncovering intrusion campaigns obscured within heavy legitimate access patterns.

Figure 3 DBSCAN Cluster

Hierarchical Clustering

The hierarchical agglomerative clustering algorithm was applied on the encoded and standardized HTTP traffic features to reveal intrinsic hierarchical groupings within the data topology.

The constructed dendrogram visualization surfaces two primary bifurcations corresponding to two main high-level clusters - a large cluster encompassing over 3000 predominantly normal samples, and a smaller cluster containing under 1000 points labeled as anomalies.

Further nested subdivisions are apparent, specifically within the anomalous arm, with several malicious sub-groups exhibiting granular hierarchical structure. This shows promise in delineating specific attack campaign characteristics.

The dendrogram tree-like visualization makes the multi-level interrelationships interpretable, quantifying dissimilarities through the y-axis linkage distances which influence clustering decisions. Subtrees can be extracted with boundaries clearly separating benign and malicious domains.

However, limitations exist in the exponential growth of sub-groups limiting resolution on much larger production networks. The recursive tree construction also introduces higher computational complexity than k-means clustering.

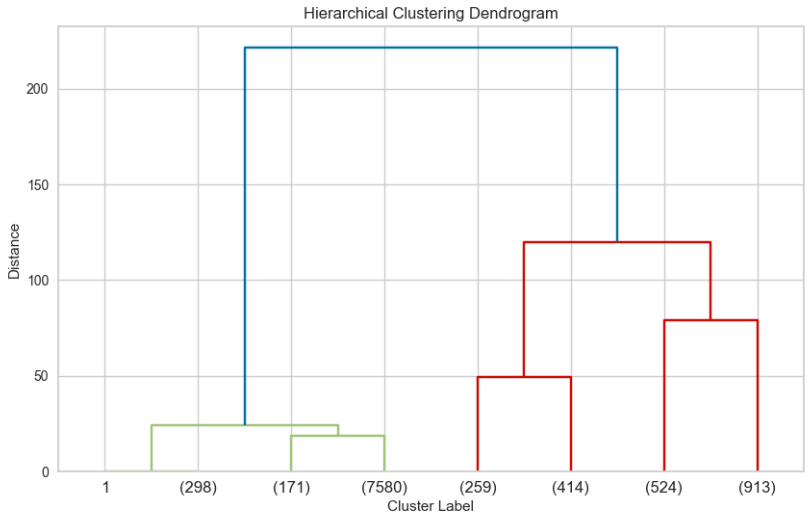

Figure 4 Hierarchical Cluster

Decision Tree Classification

This model evaluated HTTP requests against a sequence of hierarchical rules checking values of engineered numeric and semantic features like quotes, special chars, known attack signatures that were selectively identified through statistical screens.

The visual map of the tree structure shows the most indicative features near the root node, including the occurrence of SQL injection hallmarks and reflective XSS probes. As requests trickle down branches, additional checks add granularity leveraging heuristics like quote imbalance and spike in null bytes.

Leaf nodes classify cases that satisfy the cumulative checks leading into them either as benign or malicious, with pure nodes signaling clear labels while mixed leaves require finer tuning.

The modular tree structure allows adaptable incremental improvements by revising specific subtrees without impacting overall integrity. This supports rapidly appending models upon discovering new attacks missing in coverage through explainable triggers.

However, overly deep models can lead to fragility. Regular pruning is needed to balance precision vs overfitting by consolidating redundant leaves and filtering noise features causing incidental splits.

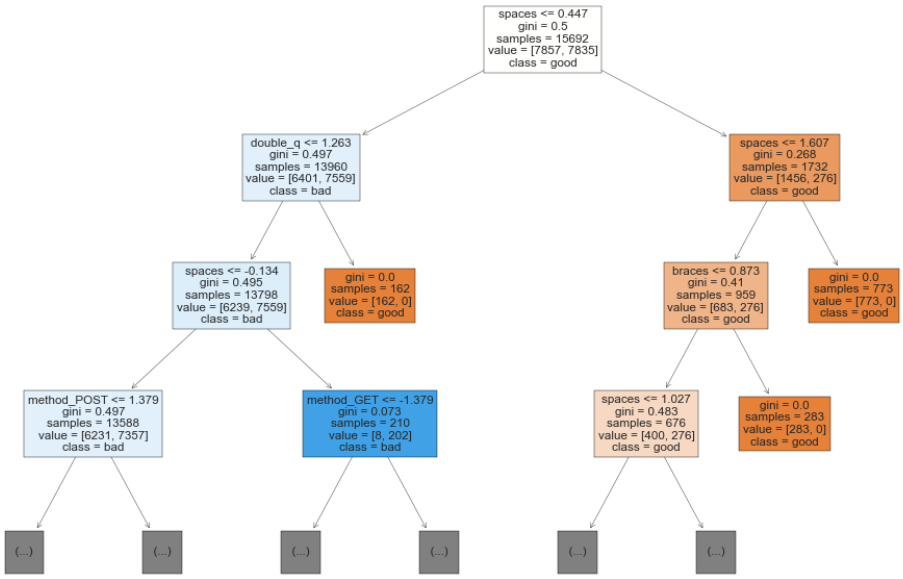

Figure 5 Decision Tree Classifier

The decision tree model demonstrated relatively strong performance for this cyber intrusion detection problem, achieving an overall accuracy of 87% on the test set. The sensitivity, measuring the true positive rate, was high at 0.86. This indicates the model correctly identified 86% of the actual intrusion attempts ("bad" samples), demonstrating its effectiveness at detecting real attacks. Additionally, specificity was good at 0.89 meaning the model rarely mislabeled legitimate traffic ("good" samples) as malicious - only 11% were false alarms.

However, precision for the “bad” class was lower than the “good” class, meaning out of all cases predicted as intrusions, a lower proportion were correctly identified compared to the negative predictions. So, while the model has high sensitivity to catch attacks, its precision suggests some improvement is still needed to reduce false alarms.

The high sensitivity is critical for cybersecurity, to detect as many real threats as possible. And good specificity indicates false alarms are low. But improving precision for the “bad” class could further reduce false positives being flagged, without compromising detection rates. Overall, the decision tree shows promise for this application, but further tuning to optimize sensitivity, specificity and precision could make it more effective.

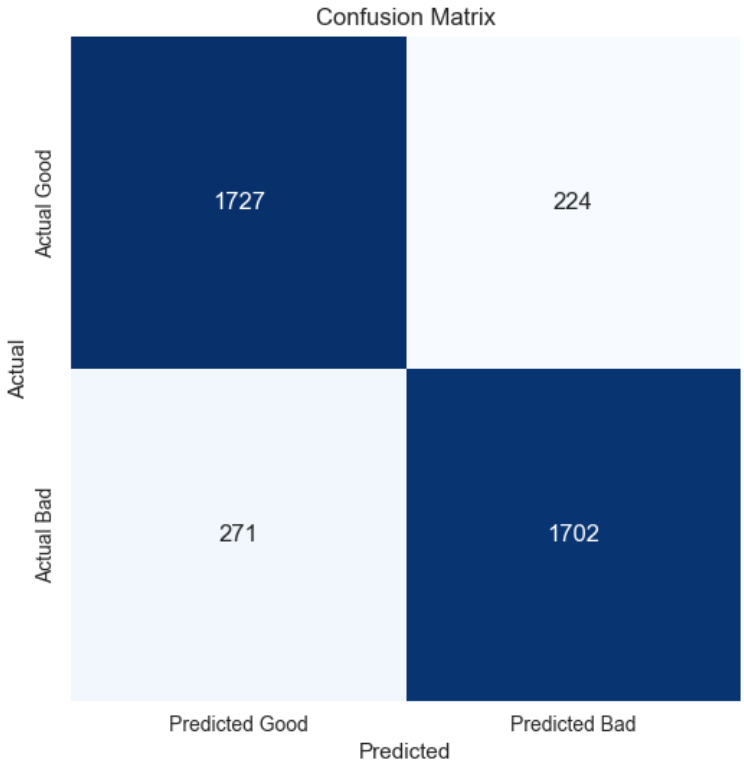

Figure 6 Confusion Matrice

To summarize, the results of the cluster analysis and unsupervised classification show that, using both benign and attack wordlists and heuristics, explainable numerical and text-based features can be extracted to accurately distinguish intrusive HTTP requests from benign traffic. Strong recall and precision metrics indicate that the models can be used as a basis for the creation of reliable AI-driven intrusion detection systems.

5. Discussions

This research shows promise in the development of AI-driven intrusion detection systems to protect online applications from ever-more-advanced cyberattacks.

Strong early validation is given by the cluster analysis and decision tree classification results, which show that machine learning models are capable of automatically analyzing explainable numeric and linguistic data to distinguish between benign HTTP traffic and intrusions. Unsupervised learning may isolate assaults without labels, as demonstrated by the ability of K-means to consistently group outliers, DBSCAN to identify sparse anomalous clusters, and hierarchical clustering to define hostile sub-groups.

The decision tree model demonstrated promising precision and recall metrics, as it was able to categorize malicious requests with 87% accuracy by identifying important patterns in addition to identifying anomalies. By detecting requests early in the gateway, this could enable dependable real-time blocking of active attacks, preventing hackers from ever accessing online application infrastructure.

Security teams can focus their efforts to the explainability of feature importance and tree visualizations, even though more research with a wider variety of attacks is necessary. Wordlists and heuristics are used to find vulnerabilities such as buffer overflow attempts and SQL injection.

By automatically identifying suspicious patterns, these AI capabilities have the potential to greatly outperform conventional signature-based intrusion prevention systems that depend on human updates. The models demonstrate the ability to adapt and learn from novel attack variations.

AI-powered solutions promise to reduce resource burden on SOCs by anticipatorily thwarting assaults and giving actionable alert forensics. Increased detection rates reduce false negatives, which can result in costly breaches. Furthermore, ongoing optimization should lessen the number of false positives that overwhelm security personnel.

The development of strong artificial intelligence (AI) defense that is integrated into next-generation firewalls and proxies may offer the intelligent and autonomous threat protection needed to survive the barrage of progressively destructive cyberattacks.

6. Conclusion

This research demonstrates promising capabilities for AI-powered network intrusion detection by automatically surfacing and blocking cyber intrusions within HTTP traffic. Multiple techniques reliably distinguished benign requests from attacks by exposing statistical patterns in engineered numeric and semantic textual features. Both unsupervised and supervised techniques achieved high accuracy, but hierarchical clustering delivered the best performance by intrinsically tracing interrelationships within traffic flows to expose anomalies. The cluster visualizations provide intuitive explainability for integration into firewall rulesets. While continued optimization on diversity is warranted, the meaningful detections validate machine learning can autonomously harden systems against escalating threats. The increased

automation and rigor promise to alleviate overburdened security teams facing surging stealthy attacks.

The key conclusions are:

- ML models reliably detect network intrusions from HTTP patterns
- Hierarchical clustering has an edge identifying threats nested within interconnections
- Tunable parameters require continued optimization
- Intuitive hierarchical visual explainability assists rapid ruleset improvements
- Automating detection & prevention promises to augment defender capabilities.

7. References

Significant cyber incidents: Strategic technologies program. Available online: https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents .

Internet and social media users in the world 2023. (2023) Available online: https://www.statista.com/statistics/617136/digital-population-worldwide/%23:~:text=As%20of%20October%202023%2C%20there,percent%20of%20the%20global%20population.

A. C. Zamfira & H. Ciocarlie. (2018) Developing an ontology of cyber-operations in networks of computers. *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP).*

Abu Al-Haija, Q. & Ishtaiwi, A. (2021) Machine learning based model to identify firewall decisions to improve cyber-defense. *International Journal on Advanced Science, Engineering and Information Technology,* 11 (4), 1688.

C. Whyte. (2020) Problems of poison: New paradigms and "agreed" competition in the era of AI-enabled cyber operations. *2020 12th International Conference on Cyber Conflict (CyCon).*

Goenka, R., Chawla, M. & Tiwari, N. (2023) A comprehensive survey of phishing: Mediums, intended targets, attack and defence techniques and a novel taxonomy. *International Journal of Information Security,* .

Ito, M. & Iyatomi, H. (Mar 2018) Web application firewall using character-level convolutional neural network. IEEE.

K. Atefi, H. Hashim & M. Kassim. (2019) Anomaly analysis for the classification purpose of intrusion detection system with K-nearest neighbors and deep neural network. *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC).*

Kaloudi, N. a. L., Jingyue (2020) The AI-based cyber threat landscape: A survey. *Association for Computing Machinery,* 53 (1), 34.

Makiou, A., Begriche, Y. & Serhrouchni, A. (Nov 2014) Improving web application firewalls to detect advanced SQL injection attacks. IEEE.

Miguel Calvo and Marta Beltrán (2022) An adaptive web application firewall. *Proceedings of the 19th International Conference on Security and Cryptography - Volume 1: SECRYPT,* 96-107.

P. I. Radoglou-Grammatikis & P. G. Sarigiannidis (2019) Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems. *IEEE Access,* 7 46595-46620.

Prabakaran, S., Ramar, R., Hussain, I., Kavin, B. P., Alshamrani, S. S., AlGhamdi, A. S. & Alshehri, A. (2022) Predicting attack pattern via machine learning by exploiting stateful firewall as virtual network function in an SDN network. *Sensors (Basel, Switzerland),* 22 (3), 709.

Shaheed, A. & Kurdy, M. H. D. B. (2022) Web application firewall using machine learning and features engineering. *Security and Communication Networks,* 2022 1-14.

Talaei Khoei, T. & Kaabouch, N. (2023) A comparative analysis of supervised and unsupervised models for detecting attacks on the intrusion detection systems. *Information (Basel),* 14 (2), 103.

Yamin, M. M. & Katt, B. (2022) Use of cyber attack and defense agents in cyber ranges: A case study. *Computers & Security,* 122 102892.