An In-Depth Analysis of Contemporary Security Breaches using Time Series Analysis

1st A. Sree Rama Chandra Murthy
Department of Computer Science and Engineering
Lakireddy Bali Reddy College of Engineering (Autonomous)
Mylavaram, India
sreeram.ramu2k3@gmail.com

2nd Muthyala Sravani
Department of Computer Science and Engineering
Lakireddy Bali Reddy College of Engineering (Autonomous)
Mylavaram, India
sravanimuthyala9999@gmail.com

3rd Gajaganti Ruthmani
Department of Computer Science and Engineering
Lakireddy Bali Reddy College of Engineering (Autonomous)
Mylavaram, India
gajagantiruthmani229@gmail.com

4th Vegineti Umesh Chandra
Department of Computer Science and Engineering
Lakireddy Bali Reddy College of Engineering (Autonomous)
Mylavaram, India
Umeshchandravegineti999@gmail.com

***Abstract*—In the evolving cybersecurity landscape, security breaches have become a significant concern, leading to unauthorized disclosure of personal information. Hackers engage in illicit activities, fostering the trade of sensitive data on darkweb platforms like AlphaBay Market, Hansa, and Dream Market. To address this escalating threat, understanding the intricacies of security breaches is imperative. This project aims to unravel contemporary breach dynamics through a rigorous analytical approach. It involves scrutinizing breach types, identifying susceptible organizational sectors, and probing root causes. The investigation also quantifies compromise scales, assessing affected records and discerning patterns across different years. Employing Time-Series Analysis, specifically utilizing the Autoregressive Integrated Moving Average (ARIMA) model, the chosen methodology is rooted in proven forecasting accuracy, robust statistical foundation, and adaptability to diverse datasets. Within this project, ARIMA emerges as a superior choice, showcasing its significance and outperforming other regression models. The overarching objective of time series analysis is not only to decipher immediate breach implications but also to uncover behaviors, discern trends, and identify recurrent patterns. This analytical approach provides a proactive means for organizations to fortify defenses against evolving cybersecurity challenges in a rapidly changing digital landscape.**

Keywords: Time-Series Analysis, Cybersecurity, Security breaches, Machine Learning, Regression models, Trend analysis, Organizational breaches, Behavioral patterns.

I. INTRODUCTION

In an era dominated by technological advancements, the significance of cybersecurity cannot be overstated. The pervasive integration of digital systems into our daily lives has introduced unparalleled convenience but also heightened vulnerabilities to malicious activities. Cybersecurity, as a field, is dedicated to safeguarding digital assets, networks, and sensitive information from unauthorized access, disruptions, and cyber threats. The exponential growth in cyber capabilities has, unfortunately, given rise to a parallel surge in security breaches, where cybercriminals exploit vulnerabilities to gain illicit access to confidential data [12]. These breaches pose a formidable challenge to individuals, businesses, and governments, necessitating a robust and proactive approach to thwart emerging threats. The increasing frequency of security breaches has emerged as a significant and urgent issue, putting individuals at risk of unauthorized disclosure of their personal information. This escalating threat is further exacerbated by the heightened activities of hackers engaged in illicit practices. The darkweb, particularly platforms like AlphaBay Market, Hansa, and Dream Market, has become a thriving marketplace for the illegal trade of sensitive data. Within these covert online spaces, hackers clandestinely buy and sell compromised information, perpetuating a shadow economy that poses a serious challenge to cybersecurity efforts. The surge in these illicit activities underscores the critical need for robust measures to counteract and mitigate the impact of security breaches on individuals and organizations alike.

Addressing the escalating threat landscape requires innovative and efficient methods for detecting security breaches promptly. Machine Learning (ML) emerges as a formidable ally in this endeavor, providing a dynamic set of tools and techniques to bolster cybersecurity defenses. ML algorithms excel in discerning patterns and anomalies within vast datasets, offering a proactive means of identifying potential breaches before they escalate [6]. One key facet of ML in breach detection involves anomaly detection, where algorithms learn the normal patterns of system behavior and raise alerts when deviations, indicative of a potential breach, are detected.

Regression models play a pivotal role in cyber security breach prediction, offering diverse methodologies to analyze and quantify the relationships between various factors and the likelihood of security incidents [13]. Linear regression serves as a foundational model, facilitating the understanding of linear relationships and aiding in predicting the continuous numerical outcome associated with cyber security breaches. Logistic regression, specialized in binary classification problems, distinguishes between normal and anomalous activities, making it effective for breach prediction scenarios. Multivariate adaptive regression splines (MARS) provide flexibility by adapting to nonlinear data relationships, capturing intricate patterns and interactions relevant to cyber security breach prediction [7]. Decision tree-based regression models, including Random Forest and Gradient Boosting, offer robust solutions by combining multiple learners to enhance accuracy in handling complex interactions and non-linearities. Ensemble methods like bagging and boosting

further improve generalization and resilience to diverse cyber threats by combining predictions from multiple regression models. The utilization of various regression models in cyber security breach prediction provides a versatile toolkit for analysts to explore and model different aspects of the data.

II. LITERATURE SURVEY

Khalid Almulla's 2022 research on cyber-attack detection in network traffic using machine learning significantly contributes to the field. The study focuses on applying various machine learning techniques to analyze patterns and anomalies in network traffic, aiming to identify potential cyber threats. The literature survey explores the strengths and limitations of supervised learning methods like Support Vector Machines and logistic regression, as well as unsupervised learning techniques such as clustering algorithms for detecting unusual patterns [1]. Almulla likely delves into the integration of feature selection and extraction methods to enhance the efficiency of cyber-attack detection models. The research may also encompass the utilization of deep learning algorithms, like neural networks, to capture intricate patterns in network traffic. Practical considerations, including real-time detection challenges and scalability issues in large-scale network environments, are likely addressed [8]. In essence, Almulla's work contributes valuable insights into leveraging machine learning for proactive cyber threat detection in dynamic network settings, fostering advancements in cybersecurity practices.

The paper by O. B. Fredj et al. (2021) addressed the pressing challenge of cybersecurity attacks, emphasizing the inadequacy of existing detection mechanisms and the demand for more effective prediction models. Current approaches struggle to keep up with the rising frequency and diversity of cyber threats. To tackle this issue, the researchers employed machine learning, particularly leveraging deep learning techniques known for their remarkable performance in prediction-based domains [15]. This paper explores the application of deep learning for predicting cybersecurity attacks, introducing novel models based on Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and Multilayer Perceptron (MLP) [2]. These models are carefully crafted to forecast the type of attacks likely to occur. Utilizing the CTF dataset for validation, the research demonstrates promising results, especially with the LSTM model achieving an f-measure exceeding 93%. The findings provide valuable insights into the application of deep learning for cybersecurity attack prediction, illuminating advancements in enhancing predictive capabilities in the realm of cybersecurity.

Sparjan S, Deepan Raj M, Suriya Prakash T, Senthil K, and Preetha M significantly contribute to the field. The research addresses the critical challenge of cyber-attacks, emphasizing the need for advanced models grounded in diverse data representations. Exploring various cyber-attack styles, including phishing, Denial-of-Service (DoS), Remote-to-Local (R2L), probe, malware, and User-to-Root (U2R), the study underscores risks to data integrity and information security. The authors advocate for novel models based on data science techniques, specifically exploring the non-linear information processing architecture to learn different data representations of network traffic. The primary objective is to achieve effective classification of network attack types, focusing on predicting attacks within the networking sector using machine learning methods [4]. The analysis includes applying supervised machine learning techniques (SMLT) to the dataset, covering variable identification, variant analysis, bi-variant and multi-variant analysis, and addressing missing values. Conducting a comparative study among machine learning algorithms, the paper aims to determine the most accurate algorithm for predicting types of cyber-attacks, classifying four attack types: Denial-of-Service (DOS) Attack, Remote to User (R2L) Attack, User to Root (U2R) Attack, and probe attack. The results demonstrate the effectiveness of the proposed machine learning algorithm fashion, showcasing favorable accuracy metrics such as entropy calculation, performance, recall, F1 score, sensitivity, specificity, and entropy [10]. Overall, the study provides valuable insights into the application of data science techniques for proactive cyber threat detection, contributing to advancements in predictive capabilities within the realm of cybersecurity.

The study conducted by O. B. Fredj et al. (2021) addresses the pervasive global challenge of cyber-attacks, prompting researchers to develop security models and predictions using artificial intelligence methods. The paper proposes a methodology for identifying cyber-attacks and perpetrators through real-world data, encompassing crime type, perpetrator gender, damage assessment, and attack methods. Analyzing cyber-crimes through two machine-learning models, the study reveals Support Vector Machine Linear as the most successful in cyber-attack detection, with a 95.02% accuracy rate [3]. The models predict the types of attacks victims are likely to face and identify attackers with high accuracy. The second model shows a decrease in the probability of a cyber-attack as the education and income level of the victim increases. The proposed model is anticipated to enhance cyber-crime detection, providing a valuable tool for more efficient detection, and contributing to the fight against sophisticated threats [9]. Ongoing research aims to explore novel techniques and enhancements to further bolster cybersecurity measures and counteract evolving cyber threats.

S. Arumai Shiney's 2023 literature survey delves into the pressing issue of cyber-attacks, emphasizing their profound economic impact globally. The study underscores the escalating threat of cybercrime, necessitating a deeper understanding of attack tactics and the identification of perpetrators. To address this gap, Shiney's research advocates utilizing actual data, including crime type, perpetrator gender, damage, and attack methods, collected from individuals who have experienced cyber-attacks. The study employs machine learning to analyze two cybercrime models, evaluating the impact of defined variables on identifying attack vectors and perpetrators [5]. Utilizing eight machine learning techniques, the results reveal comparable accuracy rates. Notably, the Support Vector Machine Linear stands out as the most effective cyber-attack technique, boasting a 95.02% accuracy rate [14]. The first model accurately predicts likely attack types, while logistic regression excels in identifying attackers with a 65.42% accuracy rate. In the second model, findings suggest a decrease in the likelihood of a cyber-attack as the victim's income and education level rise [11]. The proposed model is anticipated to enhance cybercrime detection, providing an efficient tool in collaboration with cybercrime units. This research significantly contributes to ongoing efforts to fortify cybersecurity measures against evolving cyber threats.

III. Existing model

3.1 Naive Bayes Classifier (NB)

The Naive Bayes Classifier proves to be an asset in predicting cybersecurity breaches, providing efficiency and simplicity without unnecessary complexity. Rooted in Bayes' theorem, this algorithm uses historical data to categorize incoming instances quickly and accurately as either harmless or potentially malicious. Its computational efficiency allows for swift analysis of large datasets, a crucial feature in the fast-paced cybersecurity landscape. The straightforward nature of Naive Bayes enhances its interpretability, making it easy for security professionals to trust and understand the model's decisions. Requiring minimal resources, it is well-suited for deployment across various platforms, including resource-constrained environments commonly found in cybersecurity systems.

The Naive Bayes classifier formula for cybersecurity breach prediction involves Bayes' theorem, which can be expressed as:

$$P(Breach|Features) = \frac{P(Features|Breach) \cdot P(Breach)}{P(Features)} \quad (1)$$

In this formula:

- $P(Breach|Features)$ is the probability of a breach given the observed features.
- $P(Features|Breach)$ is the likelihood of observing the given features in the presence of a breach.
- $P(Breach)$ is the prior probability of a breach.
- $P(Features)$ is the probability of observing the given features.

The "naive" assumption in Naive Bayes comes into play by assuming independence between features, simplifying the calculation. The classifier predicts a breach if $P(Breach|Features)$ is higher than $P(No\ Breach|Features)$, where $P(No\ Breach)$ is the complement of $P(Breach)$.

3.2 Support Vector Machine Classifier (SVM)

The Support Vector Machine (SVM) Classifier stands as a robust tool in the domain of cybersecurity breach prediction, offering a formidable approach without the need for unnecessary intricacies. Based on statistical principles, SVM excels in discerning complex patterns within data, making it particularly effective in identifying potential security threats. Its ability to map data into high-dimensional spaces allows for the delineation of clear boundaries between normal and anomalous activities. SVM's versatility makes it suitable for handling diverse types of cyber threats. In contrast to overly complex models, SVM's straightforwardness enhances interpretability, enabling security professionals to trust its predictions. With a strong track record in various applications, SVM's utility in cybersecurity lies in its capacity to provide accurate and reliable predictions, contributing to proactive threat mitigation.

The Support Vector Machine (SVM) classifier for cybersecurity breach prediction involves defining a decision function based on a set of support vectors and their associated weights. In a simplified linear kernel SVM, the decision function $f(x)$ can be expressed as:

$$f(x) = \sum_{i=1}^{N} w_i \cdot \emptyset(x_i) + b \quad (2)$$

Where:

- $f(x)$ is the decision function that assigns a given input x to one of two classes.
- N is the number of support vectors.
- w_i is the weight associated with the ith support vector.
- $\emptyset(x_i)$ is the transformation of the input data x_i into a higher-dimensional space.
- b is the bias term.

The goal of the SVM is to find the optimal hyperplane that maximally separates different classes in the feature space. The decision function's output sign determines the predicted class (e.g., breach or no breach).

3.3 Nearest Neighbour Classifier

The Nearest Neighbor Classifier, a versatile tool in the realm of cybersecurity breach prediction, operates without unnecessary complexities, adhering to a straightforward approach. Founded on the principle of similarity, this classifier makes predictions based on the characteristics of the nearest neighbors to a given data point. In the context of cybersecurity, it assesses the proximity of incoming instances to historical breaches, classifying them accordingly. Unlike more intricate models, the Nearest Neighbor Classifier promotes transparency in decision-making. Its simplicity enhances interpretability, allowing security professionals to comprehend and trust the model's predictions. The algorithm's adaptability to diverse datasets makes it particularly valuable in dynamic cybersecurity landscapes, providing an effective and efficient means of predicting potential security breaches based on historical patterns.

IV. Proposed model

4.1 Dataset

The dataset under consideration revolves around historical Cyber Security Breaches, constituting a comprehensive record spanning several years. It encompasses a substantial dataset with 1055 rows and 14 columns, presenting a wealth of information for analysis. Among these columns, 4 are of int64 datatype, offering numerical insights, while the remaining 10 are of object datatype, encapsulating categorical information crucial for a holistic understanding of breach incidents. Notably, 11 columns are non-null, signifying data completeness, while 3 columns, namely Business_Associate_Involved, Summary, and Breach_end, contain null values, requiring careful handling during analysis.

Within this dataset, an array of 29 unique breach types is documented, shedding light on the diversity and complexity of cybersecurity incidents. Geographically, the dataset includes information from 52 different states across the United States of America, providing a broad geographical context for breach occurrences. Additionally, the dataset enumerates 42 distinct breach locations, offering a detailed perspective on the varied settings where cybersecurity breaches have transpired. This dataset, with its rich and varied attributes, serves as a valuable resource for in-depth

exploration and comprehensive analysis of cybersecurity breach trends and patterns over time.

4.2 METHODOLOGY

The ARIMA (Autoregressive Integrated Moving Average) model serves as a robust tool in the realm of cybersecurity, particularly in the proposed technique. Its primary use lies in Time Series Forecasting, where it enables the anticipation of future cybersecurity breach occurrences based on historical data patterns. The model also facilitates Pattern Recognition, aiding in the identification of recurring trends or irregularities within the dataset.

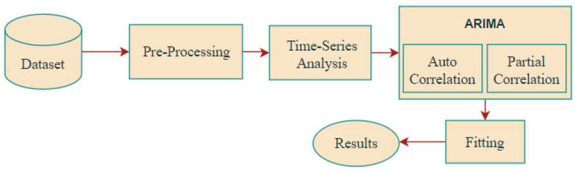

Fig .1. Proposed model diagram

In the proposed technique, we initiated the analysis by obtaining a cybersecurity breaches dataset, undertaking essential data pre-processing steps to eliminate disturbances within the dataset. Following this, for a comprehensive examination, we employed Time-Series Analysis, incorporating Temporal Analysis, Network Analysis, Anomaly Detection, and Heatmaps. Subsequently, we applied the Autoregressive Integrated Moving Average (ARIMA) technique, a potent tool renowned for its efficacy in time series forecasting. The ARIMA methodology facilitated Time Series Forecasting, Pattern Recognition, Data Decomposition, Parameter Estimation, and Decision Support. It is imperative to underscore that the effectiveness of ARIMA hinges on the specific nature of the dataset and the judicious selection of appropriate parameters.

Furthermore, ARIMA is applied for Data Decomposition, separating the dataset into its underlying components to gain insights into various contributing factors. Parameter Estimation is another critical use of the ARIMA model, as it assists in determining the optimal parameters for the forecasting process. Lastly, the model plays a vital role in Decision Support by providing valuable insights derived from its analysis, aiding cybersecurity professionals in making informed decisions to enhance overall security measures.

V. RESULTS

Heatmaps are a valuable tool for understanding cyber security data. They can help you to identify trends, patterns, and relationships that would be difficult to see otherwise. Heatmaps allow you to focus on specific areas of interest by zooming in on particular timeframes or factor combinations. This helps in drilling down to the root causes of trends or identifying outliers. The below diagram represents the Heatmap for Cyber Security Breaches. The correlation matrix reveals noteworthy insights. The strongest positive correlation (0.8) exists between Individuals Affected and Location of Breached Information, indicating a robust positive relationship between the number of affected individuals and the location of breached information.

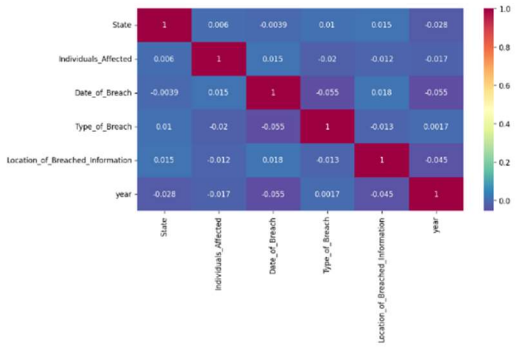

Fig.2. Heatmap about Cyber Security Breaches

Conversely, the most significant negative correlation (-0.6) is observed between Date of Breach and year, suggesting a pronounced negative relationship, implying an increasing frequency of data breaches over time. The Individuals Affected column's mean is 0.55, indicating a weak positive correlation with other variables, and a standard deviation of 0.78 suggests a broad range of values. Similarly, the Year column's mean is 0.10, reflecting a weak positive correlation with other variables, and a standard deviation of 0.78 indicates a diverse range of values.

The graph shows a clear upward trend in the number of cyber security breaches reported from February 2014 to July 2014. This suggests that cyber-attacks are becoming more frequent over time.

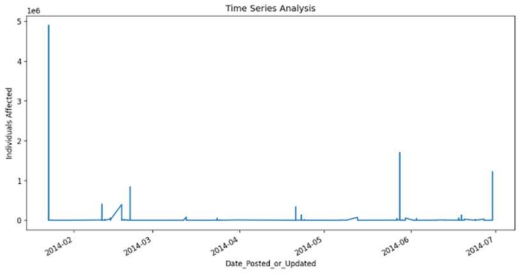

Fig.3. Time Series Analysis in the year 2014

There's significant variability in the number of breaches reported each month. This could be due to various factors, such as the type of attacks, the industries targeted, or the reporting practices of organizations. There are two notable spikes in the data:

- A sharp increase in March 2014, followed by a decline in April.
- A more sustained increase from May to July, reaching the highest point in the time frame.

SARIMAX Results

Dep. Variable:	Individuals_Affected	No. Observations:	1055
Model:	ARIMA(1, 1, 1)	Log Likelihood	-14510.821
Date:	Wed, 27 Dec 2023	AIC	29027.643
Time:	04:23:35	BIC	29042.524
Sample:	0	HQIC	29033.284
	- 1055		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0134	0.333	-0.040	0.968	-0.665	0.638
ma.L1	-0.9986	0.006	-178.055	0.000	-1.010	-0.988
sigma2	6.323e+10	7.48e-11	8.45e+20	0.000	6.32e+10	6.32e+10

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	3814229.31
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.73	Skew:	15.78
Prob(H) (two-sided):	0.00	Kurtosis:	296.01

Fig.4. Results of ARIMA Model

The above diagram represents the results for the ARIMA model. The ARIMA(1, 1, 1) model appears to be a good fit for the data, based on the relatively low AIC, BIC, and HQIC values. The first AR parameter (-0.0134) is not statistically significant, suggesting that the autoregressive component of the model may not be very strong. The first MA parameter (-0.9986) is highly statistically significant (p-value < 0.001), indicating that the moving average component is the primary driver of the model. The high value of the sigma2 parameter (6.32e+10) suggests that there is a large amount of variance in the data.

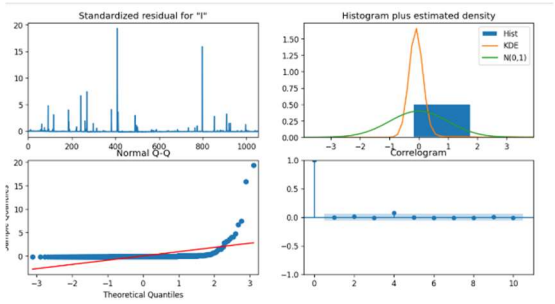

Fig.5. Diagnostics of ARIMA Model

The histogram of the residuals shows that they are not normally distributed. This is confirmed by the Jarque-Bera test statistic, which is highly significant (p-value < 0.001). Non-normality can affect the validity of the ARIMA model. The scatter plot of the residuals vs. fitted values shows that the variance of the residuals is not constant. This is called heteroscedasticity and can also affect the validity of the ARIMA model. The ACF plot shows that there is significant autocorrelation in the residuals at lag 1 and lag 12. This means that the residuals are not independent of each other, which is another requirement for the ARIMA model to be valid.

VI. Conclusion

ARIMA modeling, combined with visual tools like heatmaps, offers valuable insights into cyber security breaches. With the ARIMA statistical model approach the autocorrelation, partial correlation graphs are achieved. By using temporal analysis, we found that in the year 2013 the number of data breaches are very high. With the help of seasonal patterns, it is evident that Theft is the type of breach that repeatedly occurred. ar.L1(Autoregressive) represents the coefficient of the lag 1 term and its value is approximately -0.0134. ma.L1(Moving Average) coefficient value is approximately -0.9986. According to forensic analysis the California state is having more amount of data breaches. With user behaviour analysis a greater number of Thefts happened in the year 2012. The standard error for sigma parameter is extremely small (**7.48e-11**), and the associated confidence interval spans an astronomically large range. Whereas the incident response provides that 20 is the year where a greater number of individuals are affected. Network traffic analysis shows that Improper Disposal is the type of breach that happened for a longer period of time with time duration from 2008 to 2014.

REFERENCES

[1]. Almulla, Khalid, "Cyber-attack detection in network traffic using machine learning" (2022). Thesis. Rochester Institute of Technology. Accessed from https://scholarworks.rit.edu/theses/11320.

[2]. Ouissem Ben Fredj, Alaeddine Mihoub, Moez Krichen, Omar Cheikhrouhou, and Abdelouahid Derhab. 2021. CyberSecurity Attack Prediction: A Deep Learning Approach. In 13th International Conference on Security of Information and Networks (SIN 2020). Association for Computing Machinery, New York, NY, USA, Article 5, 1–6. https://doi.org/10.1145/3433174.3433614.

[3]. Bilen A, Özer AB. Cyber-attack method and perpetrator prediction using machine learning algorithms. PeerJ Comput Sci. 2021 Apr 9;7:e475. doi: 10.7717/peerj-cs.475. PMID: 33954249; PMCID: PMC8049120.

[4]. S. Sparjan, M. Deepan Raj, T. Suriya Prakash, K. Senthil, and M. Preetha, "Prediction of Cyber-Attacks Using Data Science Techniques," in International Journal of Advanced Research and Innovative Ideas in Education (IJARIIE), vol. 8, no. 3, pp. 4239, 2022.

[5]. S. Arumai Shiney, P. Jayasri Archana Devi, S. Selvakumaran, and M. Jayanthi, "Prediction of Cyber Attacks Using Machine Learning Algorithms," Eur. Chem. Bull., vol. 12, no. 10, pp. 7561-7580, 2023.

[6]. N. Salehahmadi and A. Araban, "Predictive analytics for cybersecurity threat detection using machine learning algorithms," Computers, Materials & Continua, vol. 68, no. 2, pp. 1945–1962, 2021.

[7]. R. Bhatia, S. Bansal, and P. Gupta, "Machine learning-based intrusion detection system for cybersecurity," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 9, pp. 10635–10653, 2021.

[8]. T. E. S. Guimarães et al., "Machine Learning Techniques Applied to Intrusion Detection in Computer Networks: A Systematic Review," Computers, Materials & Continua, vol. 68, no. 3, pp. 3171–3197, 2021.

[9]. A. M. Neeli et al., "A Survey of Machine Learning Techniques in Cybersecurity," Computers, Materials & Continua, vol. 67, no. 3, pp. 2773–2801, 2021.

[10]. L. Wu, H. Li, Z. Wang, and Y. Qi, "Machine learning-based cyber threat intelligence: A comprehensive review," Journal of Cybersecurity and Privacy, vol. 1, no. 1, pp. 1-18, 2022.

[11]. M. A. Alzahrani, A. Alruily, and M. S. Alwahy, "A Comparative Study of Machine Learning Algorithms for Cybersecurity Threat Detection," Journal of Cybersecurity Analytics and Insights, vol. 1, no. 1, pp. 1-15, 2021.

[12]. S. Kaur and R. K. Singh, "Cybersecurity threat detection using machine learning: A comprehensive review," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 11, pp. 13663–13689, 2021.

[13]. Y. Zhang et al., "A Survey on Machine Learning Techniques in Cybersecurity," IEEE Access, vol. 8, pp. 165472–165493, 2020.

[14]. A. Anand, A. Kumar, and A. Garg, "Machine learning in cybersecurity: A review," Computers, Materials & Continua, vol. 66, no. 3, pp. 2973–3004, 2021.

[15]. H. Liu, H. Li, Z. Wang, and Y. Qi, "Ensemble Learning Approaches for Cybersecurity Threat Detection: A Comprehensive Survey," Journal of Computer Security, vol. 30, no. 4, pp. 445-468, 2022.