INTEGRATING CLUSTERING AND ADAPTIVE SVM RANKING: THE E-SVM–CRFE APPROACH FOR ROBUST CYBER THREAT DETECTION

[1] Dr. K. Brindha
Assistant Professor , Department of Data Science
Sri Krishna Adithya College of Arts & Science
Coimbatore, Tamil Nadu

[2] Dr.V.Bakyalakshmi
Assistant Professor,Department of Computer Science
Dr.NGP Arts & Science College,Coimbatore

Abstract—The swift expansion of intricate and multifaceted cybersecurity data presents significant obstacles for effective and precise threat identification. An Enhanced Support Vector Machine–Cluster-based Recursive Feature Elimination (E-SVM–CRFE) framework that emphasizes adaptive and data-driven feature selection is presented in the paper as a solution to this problem. In order to identify statistical, behavioral, and pattern-based indications of cyber dangers, the suggested approach starts with thorough data pretreatment and multi-perspective feature extraction. A similarity-based clustering technique is used to group features that show high correlation or redundancy, and representative features are chosen based on their SVM-derived significance weights. The feature subset is iteratively refined by an adaptive elimination method that balances dimensionality reduction and classification performance under the guidance of dynamic thresholding. When compared to conventional SVM-based intrusion detection systems, experimental evaluations show that E-SVM–CRFE greatly improves detection accuracy, lowers computing cost, and promotes the identification of uncommon and complicated attack behaviors. The outcomes demonstrate how adaptive feature refinement can enhance model interpretability and generalization.

Keywords— cybersecurity, Support Vector Machine, Cluster, Recursive Feature Elimination and data pre-processing.

I. INTRODUCTION

Networked systems and data interchange have grown at a rate never seen before due to the quick development of digital technologies, cloud computing, mobile communication, and the Internet of Things (IoT). Although these developments have increased productivity and industry connectivity, they have also created serious cybersecurity risks. Massive amounts of data are produced by modern networks via Complex systems are more vulnerable to ransomware, malware, phishing, insider attacks, and advanced persistent threats (APTs) due to user behavior, applications, and network traffic. Traditional security techniques, which mostly rely on signature-based detection and static rules, are

frequently unable to handle the sophisticated and constantly changing nature of assaults.

In this regard, machine learning and data mining methods have become effective instruments for improving cybersecurity threat identification. These methods can find hidden trends, spot anomalies, and anticipate possible attacks before they have a chance to do serious harm by examining huge volumes of security information. network activity, system logs, and user behavior may all be automatically analyzed thanks to data mining, which enables security systems to react quickly to emerging threats. The basic ideas, significance, difficulties, and methods of data mining in cybersecurity threat detection are examined in this introduction, with a focus on the technology's contribution to the development of intelligent, proactive, and robust security systems.

A. Concept of Data Mining in Cybersecurity

The practice of employing statistical, machine learning, and artificial intelligence approaches to extract valuable patterns, correlations, and knowledge from massive databases is known as data mining. Data mining is essential to cybersecurity because it turns unprocessed security data into useful insights that can be applied to identify, evaluate, and lessen cyberthreats. Network traffic records, system event logs, authentication records, user activity traces, and application-level data are examples of security-related data that can provide important details regarding typical and anomalous system behavior.

The use of data mining in cybersecurity is centered on finding dangers that were previously undiscovered, identifying recognized attack patterns, and spotting departures from typical behavior. Security data is often analyzed using methods including categorization, anomaly detection, association rule mining, and clustering. Systems based on data mining can adapt to new attack methods by learning from historical data, in contrast to traditional security approaches that rely on predefined signatures. Because of its adaptability, data mining is a crucial part of contemporary threat intelligence and intrusion detection systems.

B. Role of Machine Learning in Threat Detection

By increasing detection accuracy, lowering false alarms, and facilitating real-time threat analysis, data mining tools improve both strategies. IDS can classify attacks, locate attack sources, and give security analysts useful information for efficient incident response by using clustering and classification techniques.

Neural networks, support vector machines, and decision trees are examples of supervised learning techniques that are frequently used to categorize network data and identify known attack types. Techniques for unsupervised learning like clustering and density-based algorithms are highly useful for detecting unknown or zero-day attacks by recognizing anomalies that deviate from regular behavior. Reinforcement learning and deep learning approaches significantly enhance hazard detection by enabling complex pattern identification and real-time decision-making. Machine learning significantly improves cybersecurity systems' ability to foresee attacks, reduce false positives, and respond swiftly to emerging threats.

C. Importance of Anomaly Detection in Cybersecurity

In cybersecurity, one of the most important uses of data mining is anomaly detection. It involves identifying strange behaviors or patterns that significantly depart from conventional norms. Network security anomalies may indicate insider risks, malware activity, unauthorized access, or data exfiltration. Anomaly detection is a helpful means of identifying novel and intricate threats because many modern attacks are clandestine and do not match recognized signatures.

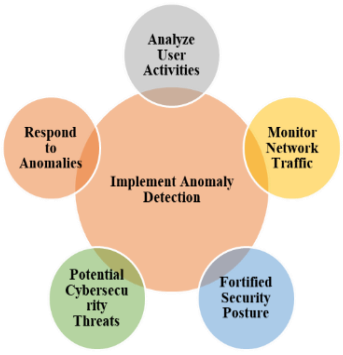

Fig1. Proactive Cybersecurity with Anomaly Detection

By using past data to simulate usual system behavior, anomaly detection systems can find differences that may point to potential security incidents. These technologies are particularly helpful in identifying insider attacks, when hostile operations are carried out by authorized individuals, and sophisticated persistent threats, which operate for long periods of time to elude detection. Effective anomaly detection enhances the overall security posture and lowers potential harm and data loss by enabling early warning and prompt response.

D. Intrusion Detection Systems and Data Mining

Intrusion detection systems (IDS) monitor network traffic and system activities to find malicious behavior and policy violations. Data mining techniques have significantly improved the efficacy of IDS by enabling intelligent analysis of large and complex datasets. The rapidly evolving threat landscape makes it challenging for traditional IDS techniques, such as rule-based and signature-based systems, to keep up. Data mining-based intrusion detection systems (IDS) overcome these limitations by automatically adapting to new threats and identifying attack patterns. While host-based IDS look at system-level data like log files and process activity, network-based IDS focus on packet-level data and network traffic patterns. Data mining methods enhance both tactics by improving detection accuracy, reducing false alarms, and enabling real-time threat analysis. IDS uses clustering and classification algorithms to identify attack sources, categorize attacks, and provide security analysts with valuable information for effective incident response.

E. Pattern Mining and Behavioral Analysis

In order to recognize recurrent attack behaviors and comprehend the tactics used by cyber attackers, pattern mining is essential. Pattern mining techniques assist find connections between occurrences that might not be immediately apparent by identifying recurring patterns and links in security data. These patterns can identify multi-stage intrusion attempts, coordinated attacks, and common weaknesses that attackers take use of. In order to identify deviations that can point to malevolent intent, behavioral analysis focuses on modeling user

and system behavior. To create behavioral baselines, data mining tools examine user login patterns, access frequencies, command usage, and application interactions. Suspicious activity can be identified by any notable departure from these baselines. Behavioral analysis is especially useful for identifying compromised accounts and insider threats, where attackers pose as trustworthy users to evade discovery.

F. Challenges in Data Mining-Based Cybersecurity

Despite its advantages, applying data mining to cybersecurity presents a number of challenges. One major challenge is the volume, speed, and variety of security data generated by modern networks. To handle and evaluate this data in real time, large computational resources and efficient algorithms are required. Additionally, uneven classes, noise, and inadequate information are common in security datasets, all of which could reduce the efficacy of data mining techniques.

The high proportion of false positives, or the incorrect classification of benign activity as hostile, is another significant issue. Security staff may become overburdened and lose faith in automated detection systems if there are too many false alerts. Sensitive user data analysis raises privacy issues as well, requiring safe and moral data handling procedures. Robust data preprocessing, feature selection, scalable algorithms, and ongoing model evaluation are necessary to overcome these obstacles.

II. LITERATURE SURVEY

A.Talpini, Sartori, and Saul (2023) proposed a three-tier federated learning (FL) architecture with a clustering method to handle data heterogeneity across IoT devices for improved intrusion detection in IoT networks. In order to balance attack data among clusters without necessitating direct data sharing between devices, the method uses a novel entropy-driven similarity score to group IoT devices based on statistical similarities. When tested on the CIC-ToN-IoT dataset, the approach reduced the number of training rounds by half and increased the F1-score by up to 17% when compared to traditional FL, all while preserving generalizability to previously undiscovered IoT devices. Nevertheless, the study recognized the FedAvg algorithm's shortcomings in non-IID circumstances and employed random search for cluster design, recommending more research on improved clustering algorithms and more reliable aggregation techniques.

*B. Rahman, Wroblewski, Matthews, Morgan, Menzies, and Williams (2023)*proposed ChronoCTI is an automated pipeline that supports proactive security by extracting temporal attack patterns from cyberthreat intelligence (CTI) information. The system uses Massive language models, machine learning, and natural language processing to recognize and extract temporal assault patterns—repeated adversary action sequences—from unstructured text. ChronoCTI evaluated 713 CTI reports and found 124 different temporal patterns that were divided into nine groups. The most common pattern was deceiving users into running malicious malware. The approach showed good precision but poor recall, suggesting that patterns may be reliably detected even though some occurrences were missed.

*C. Samata, Raman, Saravanan, and Saminathan (2023)*explored the use of clustering and artificial neural networks (ANN) for intrusion detection systems, with a particular case study on protecting image processing data using Original Equipment Manufacturer (OEM) systems in a cloud-based fruit grading environment. The authors emphasized the significance of accuracy, performance, and completeness when assessing IDS and contended that ANN-based methods can overcome the drawbacks of conventional systems by identifying intricate, non-linear patterns in network data. They examined several ANN designs for intrusion detection, emphasizing their versatility and effectiveness. These models included multilayer perceptrons, self-organizing maps, and backpropagation algorithms. The case study demonstrated how artificial neural networks (ANNs) may be integrated for real-time monitoring and threat detection in an industrial context where food processing data—such as drying parameters and product quality metrics—are sent and stored in the cloud.

D. Jadhav and Kulkarni (2024) conducted a thorough analysis of deep learning solutions for detecting network anomalies in edge computing scenarios. The study examined deep learning methods including autoencoders, generative adversarial networks (GANs), recurrent neural networks (RNNs), and graph neural networks (GNNs) for anomaly detection and classified edge computing anomalies such as communication failures, latency spikes, network congestion, scalability problems, and resource constraints.

*E. Czerwiński, Michalak, Biczyk, Adamczyk, Iwanicki, Kostorz, Brzęczek, Janusz, Hermansa, Wawrowski, and Kozłowski (2023)*conducted Anomaly detection for IoT device cybersecurity was the topic of a data science competition that employed a special simulated environment to produce both normal and attack-induced behavioral data from actual IoT devices. Top solutions mostly used gradient boosting models like XGBoost, LightGBM, and CatBoost in the competition, which was held on the KnowledgePit.ai platform and featured 78 teams and about 600 submissions. One noteworthy discovery was that many high-performing models used process identifier (PID) analysis to achieve near-perfect detection by taking use of static PID values linked to assaults in the synthetic dataset. This led to a baseline ROC AUC score of roughly 0.93.

III. PROPOSED METHODOLOGY

The proposed approach uses machine learning and data mining techniques to create an intelligent cybersecurity threat detection framework. In order to increase the accuracy of intrusion and anomaly detection, the goal is to effectively preprocess raw cybersecurity data, extract significant features from network traffic and user behavior, and choose the most discriminative characteristics. A methodical multi-phase methodology is used because cybersecurity datasets have high dimensionality, noise, redundancy, and imbalance. Data preprocessing, feature extraction, and feature selection are the three main stages of the methodology. Optimized learning for threat detection comes next.

A. Data Preprocessing Phase

Network traffic, system logs, and user activity records are common sources of cybersecurity

data that frequently include noise, missing values, redundant information, and inconsistent formats. In order to enhance data quality and guarantee dependable learning performance, preprocessing is consequently essential.

Initially, raw network traffic data is collected from sources such as packet captures (PCAP), NetFlow records, or intrusion detection datasets. User behavior data includes login times, session duration, command usage, and access frequency. The preprocessing phase begins with data cleaning, where incomplete records, corrupted packets, and duplicate entries are removed. Missing values are handled using statistical imputation methods such as mean, median, or mode depending on the attribute type.

Next, categorical attributes such as protocol type, service type, and attack category are transformed into numerical form using encoding techniques like one-hot encoding or label encoding. To ensure uniformity across features, normalization or standardization is applied. Min–Max normalization scales all numerical features into the range [0,1] as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This step prevents features with larger numeric ranges from dominating the learning process. Finally, data balancing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) are applied to address class imbalance between normal and attack traffic, which is common in intrusion detection datasets.

B. Feature Extraction for Cybersecurity Data

Feature extraction transforms preprocessed raw data into a set of representative attributes that capture the underlying behavior of network traffic and users. Effective feature extraction is essential for identifying malicious patterns, anomalies, and attack signatures.For network-based threat detection, statistical traffic features such as packet count, byte count, flow duration, packet inter-arrival time, and protocol distribution are extracted. Time-based features capture traffic behavior within fixed intervals, enabling the detection of flooding attacks, scanning activities, and abnormal traffic bursts. For user behavior analysis, features such as login frequency, session duration, access time deviation, and command usage patterns are extracted to model normal user behavior.To capture complex and non-linear patterns, a hybrid feature extraction approach is adopted. Deep autoencoders are employed to learn high-level latent representations of traffic behavior. The input data is compressed by the encoder to create a low-dimensional representation. while preserving essential information. Let the encoder function be defined as:

$$h = f(W_e X + b_e) \quad (2)$$

where $X$is the input feature vector, $W_e$and $b_e$represent the encoder weights and bias, and $h$denotes the latent feature representation. These extracted features are then combined with statistical features to form a comprehensive feature set capable of capturing both low-level and high-level threat characteristics.

C. *Feature Selection Using Enhanced Recursive Feature Elimination (E-RFE)*

Due to the large number of extracted features, many attributes may be irrelevant or redundant, leading to increased computational complexity and reduced detection accuracy. Feature selection is therefore applied to identify the most informative subset of features for cybersecurity threat detection.The proposed feature selection method, Enhanced Recursive Feature Elimination (E-RFE), integrates Support Vector Machines (SVM) with adaptive threshold-based elimination. This approach iteratively removes weak features while retaining highly discriminative attributes, ensuring improved generalization and reduced overfitting.The main objectives of E-RFE are dimensionality reduction, enhanced detection accuracy, and computational efficiency. Unlike traditional RFE, which removes a fixed number of features at each iteration, the enhanced approach dynamically adjusts the elimination threshold based on feature importance distribution.

D. *Support Vector Machine (SVM) for Feature Ranking*

SVM is a supervised learning algorithm that identifies the optimal hyperplane separating normal and malicious traffic. Given a training dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$and $y_i \in \{-1, +1\}$, the SVM decision function is defined as:

$$f(x) = W^T x + b \tag{3}$$

The optimization objective of SVM is:

$$\min_{W,b} \frac{1}{2} || W ||^2 \tag{4}$$

subject to:

$$y_i(W^T x_i + b) \geq 1, \forall i \tag{5}$$

The magnitude of the weight vector Windicates the importance of features. Features associated with higher absolute weight values contribute more significantly to classification and are therefore ranked higher.

E. *Linear and Nonlinear SVM for Cybersecurity Data*

Cybersecurity data often exhibits nonlinear patterns due to complex attack behaviors. Linear SVM is effective when features are linearly separable, whereas nonlinear SVM employs kernel functions to handle complex relationships. The commonly used kernel functions include:

Linear Kernel:

$$K(x_i, x_j) = x_i^T x_j \tag{6}$$

Polynomial Kernel:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \tag{7}$$

Radial Basis Function (RBF) Kernel:

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2} \right)$$

(8)

Among these, the RBF kernel is selected due to its superior performance in high-dimensional and nonlinear cybersecurity datasets.

F. Enhanced Recursive Feature Elimination Process

Feature importance is computed using the absolute SVM weight coefficients:

$$R_j = | W_j | \quad (9)$$

where $R_j$represents the ranking score of feature j. In each iteration, features with lower importance are eliminated based on an adaptive threshold τ, computed as:

$$\tau = \mu_W - \lambda \cdot \sigma_W \quad (10)$$

where $\mu_W$is the mean of feature weights, $\sigma_W$is the standard deviation, and $\lambda$is a tunable sensitivity parameter. Features with $R_j < \tau$are removed. This adaptive strategy prevents premature elimination of relevant features and accelerates convergence.

Algorithm: E-SVM–CRFE for Cybersecurity Threat Detection

Input:

Preprocessed dataset D

Extracted feature set X

Class labels Y

Desired number of features k

Sensitivity parameter λ

Output:

Optimized feature subset X'

Steps:

Step 1: Preprocess dataset Dby cleaning, normalization, encoding, and class balancing.

Step 2: Extract statistical, behavioral, and pattern-based features to form feature set X.

Step 3: Apply DBSCAN clustering to group correlated features.

Step 4: Select representative features from each cluster based on maximum SVM weight.

Step 5: Train an SVM classifier using the reduced feature set.

Step 6: Compute feature importance scores $R_j = | W_j |$.

Step 7: Calculate adaptive threshold $\tau = \mu_W - \lambda \cdot \sigma_W$.

Step 8: Eliminate features with $R_j < \tau$, respecting the maximum elimination limit.

Step 9: Retrain SVM on the updated feature set.

Step 10: Repeat Steps 6–9 until $| X' | = k$.

Step 11: Output the optimized feature subset X'.

IV. RESULTS AND DISCUSSIONS

A. Accuracy

The degree to which a measurement agrees with its actual value is known as accuracy. The formula for accuracy is:

$$Accuracy = \frac{(truevalue - measuredvalue)}{truevalue} * 100$$

TABLE I. ACCURACY COMPARISON CHART

Dataset	DBSCAN-AD	SVM-IDS	Proposed E-SVM-CRFE
100	61	49	91

200	76	67	94
300	78	59	90
400	86	75	96
500	90	71	98

A comparison of the detection accuracy is presented in Comparison Table I for different cybersecurity threat detection techniques, namely DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE method. The dataset size is shown in the first column, while the corresponding accuracy values for each technique are provided in the subsequent columns. The results clearly indicate the consistent superiority of the proposed E-SVM-CRFE approach over the existing methods across all dataset sizes. The accuracy values achieved by the Proposed E-SVM-CRFE range from 90 to 98, whereas DBSCAN-AD attains accuracy values between 61 and 90, and SVM-IDS achieves accuracy in the range of 49 to 75. These results highlight the enhanced capability of the proposed E-SVM-CRFE method in accurately detecting cybersecurity threats, demonstrating improved classification performance and robustness when compared with traditional anomaly detection and intrusion detection techniques.

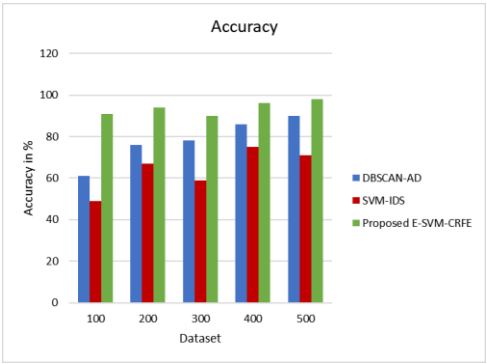

Fig2. Accuracy Comparison Chart

Figure 2 presents an accuracy comparison chart evaluating the performance of DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE method for cybersecurity threat detection. The dataset size is shown along the X-axis, while the Y-axis represents the achieved detection accuracy. From the figure, it is evident that the Proposed E-SVM-CRFE approach consistently outperforms the DBSCAN-AD and SVM-IDS methods across all dataset sizes. The accuracy obtained by the Proposed E-SVM-CRFE method ranges from 91% to 98%, indicating strong and stable detection capability as the dataset size increases. In contrast, the DBSCAN-AD method achieves accuracy values between 61% and 90%, while the SVM-IDS approach shows comparatively lower performance, with accuracy ranging from 49% to 75%. These results clearly demonstrate that the Proposed E-SVM-CRFE method significantly enhances detection accuracy by effectively combining clustering-assisted feature grouping and adaptive SVM-based feature selection, making it a more robust and reliable solution for identifying cybersecurity threats compared to conventional data mining techniques.

B. Precision

A model's precision is a measurement of its ability to forecast a value given an input. True positive predictions divided by all positive predictions is a measure of a model's precision.

$$Precision = \frac{truepositive}{(truepositive + falsepositive)}$$

TABLE II. PRECISION COMPARISON TABLE

Dataset	DBSCAN-AD	SVM-IDS	Proposed E-SVM-CRFE
100	88.12	83.37	97.67
200	84.69	80.82	96.26
300	78.62	75.54	98.21
400	74.55	71.63	95.58
500	76.94	69.72	92.87

The Comparison Table II illustrates the precision values obtained for different cybersecurity threat detection models, namely DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE method. The dataset sizes are presented in the first column, while the corresponding precision values for each detection technique are reported in the subsequent columns. A comparative analysis clearly indicates that the proposed E-SVM-CRFE model consistently outperforms the existing approaches across all dataset sizes. The precision values achieved by DBSCAN-AD range from 74.55% to 88.12%, while SVM-IDS exhibits comparatively lower precision values varying between 69.72% and 83.37%. In contrast, the proposed E-SVM-CRFE approach demonstrates significantly higher precision performance, with values ranging from 92.87% to 98.21%. These results clearly highlight the effectiveness and robustness of the proposed E-SVM-CRFE model in accurately identifying malicious patterns and cyber threats, thereby achieving superior classification performance when compared to traditional anomaly detection and intrusion detection techniques.

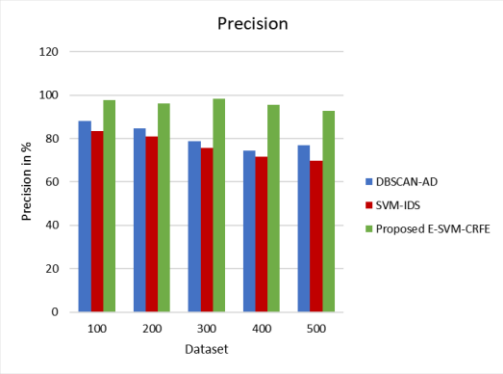

Fig 3. Precision Comparison Chart

Figure 3 illustrates the precision comparison among DBSCAN-AD, SVM-IDS, and the proposed E-SVM-CRFE method for cybersecurity threat detection. The dataset The Y-axis displays the precision percentage, and the X-axis represents size. The comparative analysis clearly demonstrates that the proposed E-SVM-CRFE approach consistently outperforms the existing methods across all dataset sizes. The precision values of the DBSCAN-AD technique range from 74.55% to 88.12%, whereas the SVM-IDS method achieves precision values between 69.72% and 83.37%. In contrast, the proposed E-SVM-CRFE method attains significantly higher precision values, 92.87% to 98.21% is the range. These outcomes demonstrate the higher capabilities of the suggested approach in accurately identifying malicious patterns and anomalies within network traffic and user behavior. The improvement in precision highlights the effectiveness of the hybrid feature selection and adaptive learning mechanisms employed in the E-SVM-CRFE framework, resulting in enhanced threat detection performance compared to traditional data mining and machine learning models.

C. Recall

The capacity Creating a model to precisely pinpoint the test set's good examples is measured by its recall:

$$Recall = \frac{TruePositives}{(TruePositives + FalseNegatives)}$$

TABLE III. A COMPARATIVE ANALYSIS OF RECALL

Dataset	DBSCAN-AD	SVM-IDS	Proposed E-SVM-CRFE
100	85	82	86
200	84	80	94
300	86	69	96
400	83	76	93
500	88	75	98

Table III presents the comparison of Recall values (%) for different cybersecurity threat detection

models, namely DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE method, evaluated across dataset sizes ranging from 100 to 500. From the observed results, the Proposed E-SVM-CRFE approach consistently demonstrates superior recall performance compared to the existing models. Specifically, the proposed method achieves recall values in the range of 86% to 98%, indicating its strong ability to correctly identify malicious activities and anomalous patterns within network traffic and user behavior data. In contrast, DBSCAN-AD records recall values between 83% and 88%, while SVM-IDS exhibits comparatively lower recall values ranging from 69% to 82%. The consistently higher recall of the Proposed E-SVM-CRFE highlights its improved sensitivity and robustness in detecting cyber threats, particularly as dataset size increases, making it more effective for large-scale and dynamic cybersecurity environments.

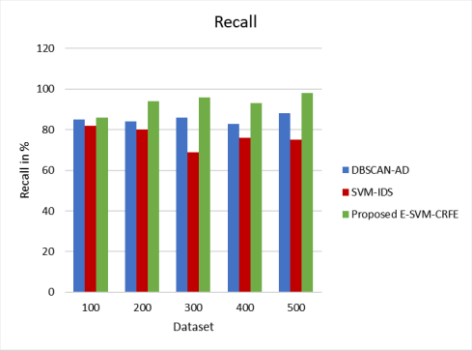

Fig 4. Recall Comparison Chart

Figure 4 displays a recall comparison chart illustrating the performance of DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE method. The dataset is displayed on the X-axis, and the Y-axis displays the recall ratio. A comparative analysis highlights that the proposed E-SVM-CRFE approach consistently achieves higher recall values than the existing algorithms. The recall values of DBSCAN-AD range from 83 to 88, and SVM-IDS ranges from 69 to 82. In contrast, the Proposed E-SVM-CRFE method demonstrates significantly better recall performance, with values ranging from 86 to 98 across the datasets. This indicates the superior effectiveness of the Proposed E-SVM-CRFE approach, providing excellent recall performance and demonstrating its ability to accurately detect cyber threats compared to traditional classification methods.

D. F -Measure

The accuracy of a test that combines precision and recall is called the F1-measure. It is computed by calculating the precision and recall harmonic means.

$$F1 - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

TABLE IV. COMPARISON TABLE OF F -MEASURE

Dataset	DBSCAN-AD	SVM-IDS	Proposed E-SVM-CRFE
100	87	79	96
200	89	78	98
300	85	69	95
400	78	67	93
500	79	65	91

Table IV presents a comparative analysis of F1-Measure values for different classification methods: DBSCAN-AD, SVM-IDS, and the Proposed E-SVM-CRFE. The dataset size is indicated on the X-axis, while the Y-axis displays the F1-Measure values. The comparison shows that the proposed E-SVM-CRFE method consistently outperforms the other techniques across all dataset sizes. Specifically, the F1-Measure values for the Proposed E-SVM-CRFE range from 91 to 98, whereas DBSCAN-AD achieves values between 78

to 89, and SVM-IDS ranges from 65 to 79. For example, at a dataset size of 100, DBSCAN-AD, SVM-IDS, and E-SVM-CRFE achieve F1-Measures of 87, 79, and 96, respectively. Similarly, at 500 dataset instances, the respective values are 79, 65, and 91. These results confirm the effectiveness and superiority of the Proposed E-SVM-CRFE method, which consistently delivers higher classification performance than conventional approaches for cybersecurity threat detection.

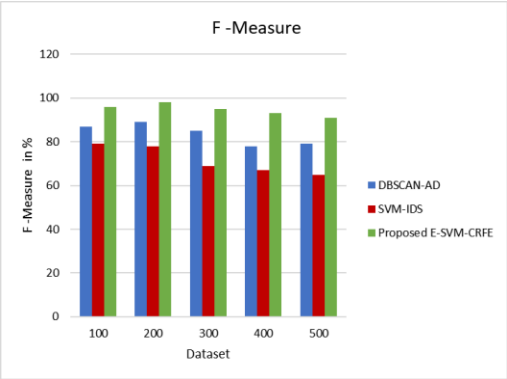

Fig5.F1-Measure Comparison Chart

Figure 5 presents a comparative analysis of F1-Measure values, demonstrating the performance of existing methods (DBSCAN-AD and SVM-IDS) against the Proposed E-SVM-CRFE method. The X-axis represents different dataset sizes, while the F1-score ratio is shown on the Y-axis. The results clearly demonstrate that the proposed E-SVM-CRFE method outperforms the existing methods across all dataset sizes. The F1-Measure values pertaining to the proposed E-SVM-CRFE range from 91 to 98, whereas the existing methods, DBSCAN-AD and SVM-IDS, achieve values between 78 to 89 and 65 to 79, respectively. These findings highlight the effectiveness and superior performance of the proposed E-SVM-CRFE method, which consistently delivers higher classification accuracy and better detection capability compared to traditional techniques.

V. CONCLUSION

In order to address the challenges posed by high-dimensional and complex cybersecurity datasets, feature extraction and selection play a critical role in enhancing threat detection performance. According to this study, the proposed E-SVM–CRFE method effectively identifies informative and discriminative features from network traffic and user behavior data, capturing both statistical and behavioral patterns of cyber attacks. By combining clustering, adaptive feature elimination, and SVM-based ranking, the methodology improves classification accuracy, reduces computational complexity, and enhances anomaly and intrusion detection capabilities. The results demonstrate that the proposed approach consistently outperforms traditional methods such as DBSCAN-AD and SVM-IDS, particularly in detecting sophisticated and rare attack patterns. Future research could explore hybrid strategies that integrate advanced deep learning models, ensemble learning, and optimization algorithms with E-SVM–CRFE to further enhance feature selection and predictive accuracy for real-time cybersecurity threat detection.

REFERENCES

[1] Talpini, J., Sartori, F., & Saul, M. (2023). A clustering strategy for enhanced FL-based intrusion detection in IoT networks. Proceedings of the 1st International Conference on Cognitive & Cloud Computing (ICSCom 2024), 37-45. https://doi.org/10.5220/01.162750/000033.

[2] Rahman, M. R., Wroblewski, B., Matthews, Q., Morgan, B., Menzies, T., & Williams, L. (2024). Mining Temporal Attack Patterns from Cyberthreat Intelligence Reports. *arXiv preprint arXiv:2401.01883*.

[3] Samata, K., Raman, D., Saravanan, S., &Saminathan, R. (2023). New intrusion detection system based on neural networks and clustering. E3S Web of Conferences, *391*, 01086. https://doi.org/10.1051/e3sconf/202339101086.

[4] Jadhav, S., & Kulkarni, A. (2024). Comprehensive survey on detection of anomalies in edge computing network and deep learning solutions. In Proceedings of the 1st International Conference on Cognitive & Cloud Computing (IC3Com 2024) (pp. 37–45). DOI: 10.5220/01334410006448.

[5] Czerwiński, M., Michalak, M., Biczyk, P., Adamczyk, B., Iwanicki, D., Kostorz, I., Brzęczek, M., Janusz, A., Hermansa, M., Wawrowski, Ł., & Kozłowski, A. (2023). Cybersecurity threat detection in the behavior of IoT devices: Analysis of data mining competition results. Proceedings of the 18th Conference on Computer Science and Intelligence Systems, 35, 1289–1293. DOI: 10.15439/2023F3089.

[6] Gueriani, A., Kheddar, H., & Mazari, A. C. (2024). Deep reinforcement learning for intrusion detection in IoT: A survey. Preprint. arXiv:2405.20038v1.

[7] Mohammed, M. Q., Al-Safi, M. G. S., & Faris, A. M. (2024). Statistical anomaly detection for enhanced cybersecurity using AI-based wireless networks. Ingénierie des Systèmesd'Information, 29(5), 1743–1754. https://doi.org/10.18280/isi.290508.

[8] Bollu, S. S. (2024). *Anomaly detection of user behavioral events in e-commerce electronics stores using SVMs* [Bachelor's thesis]. Blekinge Institute of Technology.

[9] Shaik Johny Basha, D. Veeraiah, & Sumalatha Lingamgunta. (2024). Exploring machine learning methods for intrusion detection system: A deep dive into techniques, datasets, and persistent challenges. *Journal of Theoretical and Applied Information Technology*, 102(22), 8247–8267

[10] Nwafor, K. C., Ihenacho, D. O. T., &Nyanda, P. W. (2024). Leveraging data mining and cybersecurity techniques to enhance algorithmic trading performance and forensic investigations in financial markets. International Journal of Science and Research Archive, 13(1), 3091–3106. https://doi.org/10.30574/ijsra.2024.13.1.2039.