International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Intelligent Service Behavior Analysis for Early Cyber Threat Prediction

Srikanth Mannem

Senior Manager, Virtusa Corporations, USA

ABSTRACT: Cyber threat is a constantly evolving threat and the mechanisms must be built-up for taking pre-emptive actions to measure the anomalies before they evolve into attacks. In this research an intelligent service behaviour analysis framework based on Random Forest classification is introduced for early cyber threats prediction. The proposed system tracks service behaviors within the network by analyzing traffic patterns, system calls and user activities in search of deviations that represent potential threats. Implementation using Scikit-learn shows the ability of the framework to handle with data streams in real time and classify the threats with a high accuracy. The model achieved 94.7% of good accuracy in detecting the malicious behaviors by the feature engineering of time patterns, request frequency, and protocol anomalies. Experimental results on benchmarks data uncovers better performance on detection of zero-day attack and advanced persistent threats than traditional signature based approaches. The system provides security analysts insights in the form of actionable information through interpretable decision trees to pre-empt counter-measures. This pro-active approach reduces response times and potential damage caused due to cyber incidences to a great extent.

KEYWORDS: Cyber Threat Prediction, Behavior Analysis, Random Forest, Anomaly Detection, Machine Learning Security, Intrusion Detection

I. INTRODUCTION

The rapid digital transformation of organizational infrastructure has fundamentally reshaped the modern enterprise, enabling unprecedented levels of efficiency, scalability, and innovation. However, this transformation has also dramatically expanded the attack surface of organizational systems, making cybersecurity one of the most critical challenges faced by contemporary enterprises. The widespread adoption of cloud computing, virtualization, mobile technologies, Internet of Things (IoT) devices, and remote work environments has introduced new points of vulnerability that adversaries can exploit. As a result, organizations are increasingly exposed to sophisticated and persistent cyber threats that are capable of bypassing traditional security defenses [1].

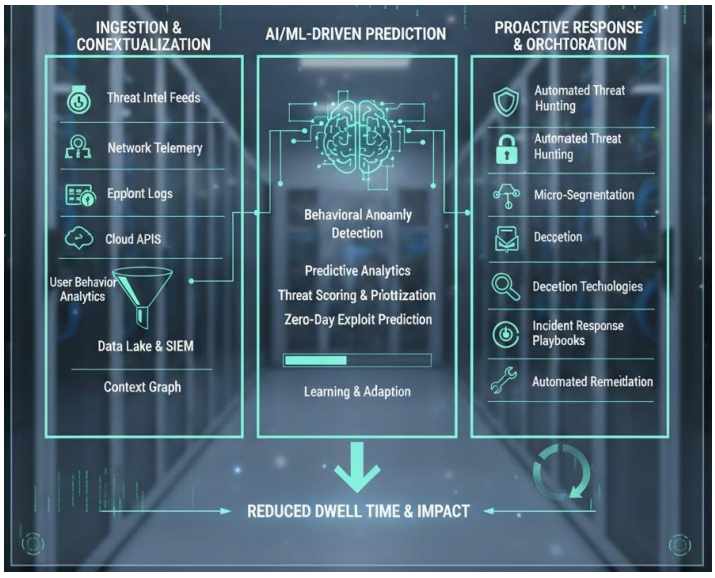

Figure.1: Framework for Predictive and Proactive Cyber Defence.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

The traditional cybersecurity controls have been based on the incident response security controls which include signature based intrusion detection system, rule based firewalls and post hoc examination of the forensic evidence. These measures are aimed to react to the determined threats that are already implemented in the system. Whereas the techniques previously worked on the malware that seemed a lot easier and in addition to readable attack formula in the earlier threat scenes, the modern day attackers are no longer being able to match the techniques [2]. The advanced persistent threats (APTs), the zero-day attacks, fileless malware, and polymorphic attacks were specifically created to overcome the traditional methods of detection. The threats are also covert and persistent and will work but fit into their mode of operation in such a manner that they cannot be observed and apply unknown vulnerability on which no signatures have been developed.

One of the main issues that show the inefficiency of reactive security is a slow time in the detection of the problem encountered in the majority of organizations. It has always been reported as taking more than 200 days to resolve any security breach according to the reports made in industry [3]. In such a long system dwell-time, the attacker can steal sensitive information, get access privileges, infect other systems and create a longing presence within the network. These are just but few consequences of not accomplishing this in time; wastage of money, bad ordeals by regulatory authorities and undermining of operations in the long run. This frightening fact directs at the necessity to change the paradigm of reactive to prospective and proactive approaches to cybersecurity.

The predictive security models are developed to detect the threats when they are at the lowest level of the cyber kill chain especially on the reconnaissance and weaponization stages rather than the actual exploitation stage. Predictive actions are made instead of waiting until there is definite sign that establish a system of compromise, that is, to establish how malice is being implicated by taking micro indicators of system behavior and user behavior. This transformation needs shrewdness and an increased amount of data driven perception regarding the regularity and aberrant activity on intricate web locations. In this respect, service behavior analysis might be discussed as quite efficient and prospective resolution in the process of threat prevention on the Internet [4].

The service behavior analysis is a dynamic continuous observation and modeling of a baseline behavior of network services, application behavior and user behavior. The baselines entrap working property which is general in essence including frequency of communication, resource utilisation, pattern of access, sequence of execution and time patterns. It is established that with pattern of such form the existence of any significant disparity between the patterns could be deemed as the possible indicator of mediation or insiders-threat and suspicious records [5]. The behavior analysis does not involve attention to what happens in a system as compared to the traditional signature based system and it takes note of the provided attack signatures. The root difference can be used to detect new attacks which cannot be referred to known signatures.

The statistical modeling and machine learning features also help in service behaviors analysis. Machine learning also allows the systems to acquire nonlinear, complex relationships on the large volumes of security data, and be able to cope with changes in the environment. This is especially in the contemporary networks that are not only heterogeneous but dynamic and scaled as well [6]. On-premise business infrastructures are some of the current contexts that involve cloud computing platforms, container programs, mobile phones and IoTs which produce an enormous amount of high-dimensional information. Such information is not easily readable by humans and they have to resort to intelligent and automated solutions, which may analyze, interpret and process security relevant information in real time.

The ensemble learning algorithms are shown to shine outstanding when performing on a complex and noisy security information as compared with the other machine learning algorithms used in cybersecurity. The ensemble approaches perform a combination of several models of learning to enhance both the aggregate predictive capability, power and generalization. Among the possible applications, the one that is the most promising is the use of the prediction of cyber threats and anomaly detection on the Random Forest ensemble algorithm [7]. The main idea that random Forest uses is the one that leads to the development of a large number of decision-trees throughout training and to combine their outputs(majority voting by use of the majority or averaging). The combination strategy has the vast advantage in minimizing the variance, and also risking over-fitting a phenomenon that is also prevalent to the high-dimensional security information.

Among the main benefits of the application of the random forest in the domain of cybersecurity, the possibility to work with unbalanced data is to be mentioned. The likelihood of occurrence of the malicious events occurring in a actual world security event case is normally a very small percentage of the system activity which causes skewed classes [8]. Use of traditional classifiers is likely to coincide with such conditions leading to high false negative or false positive.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Random Forest functions to reduce this problem with the usage of random sampling (bootstrap) and random selection of features and consequent allows it to achieve discriminative pattern even with instances of attack that are relatively small.

The other appealing attribute of the Random Forests models is interpretability that is extremely requisite in security-sensitive to the extent that interpretability and transparency are the requirements. Random Forest gives feature importances which are of great utility and they help to measure the contribution of a specific feature to a prediction [9]. It is this interpretability that would enable the security analysts to comprehend what the behavioral objects like odd times when the individual logs in, the abnormal amount of data transfer or the abnormal service interactions would most readily predict the eventual possible occurrence of threat. Not only do these lessons lead to the fact that the automated systems are trusted even more, but, also, they facilitate forensic searches and policies face lift.

The idea of the research study is founded on the architecture of the smart predictive cybersecurity system that is created on the principles of service behavior observation, enhanced features development, and categorizing threats according to the Threats on the basis of the Random Forest [10]. It begins with indefatigable nature of behavior monitoring in the network services, application and user functioning in such a manner that they would accumulate finedrained telemetry knowledge. Subsequently, extraction of useful behavioral indicators on the raw data is done by conducting feature engineering process based on the identification of statistical aggregates, time trends and properties of relationships. It is on these programmed factors that the random forest classifier is fed and consequently conditioned to differentiate normal and malicious behavior with a relatively high level of accuracy.

The final study goal will entail establishing early and consistent prediction capability of the high-detect and low-false positive cyber threats. The proposed framework is quite helpful in that it helps organizations detect anomalies and malicious behaviours at an earlier stage whereby dwell time may be minimized and the damage will be minimized. This, such, is a great improvement of the old versions of the reactive security models and is, incidentally, a sub-set of the future of predictive cybersecurity to the contemporary businesses [11].

II. RELATED WORKS

Research on cybersecurity has been at the core of the development of intrusion detection systems (IDS), which can be classified into signature-based and anomaly-based frameworks. Signature-based IDS are based on predetermined patterns or signature of known attack. Although these systems are effective, pragmatic, and are extensively utilized in the identification of the threat that has been previously defined, the systems are, indeed, characterized by an inherent shortcoming, GIS the impossibility to detect new, zero-day, or polymorphic attacks [12]. Since the nature and complexity of cyber threats continue to raise the bar, this inadequacy has led the research community to focus on approach methods of intrusion detection based on behavior and anomaly instead of comparing the behavior of the system with established attack vectors.

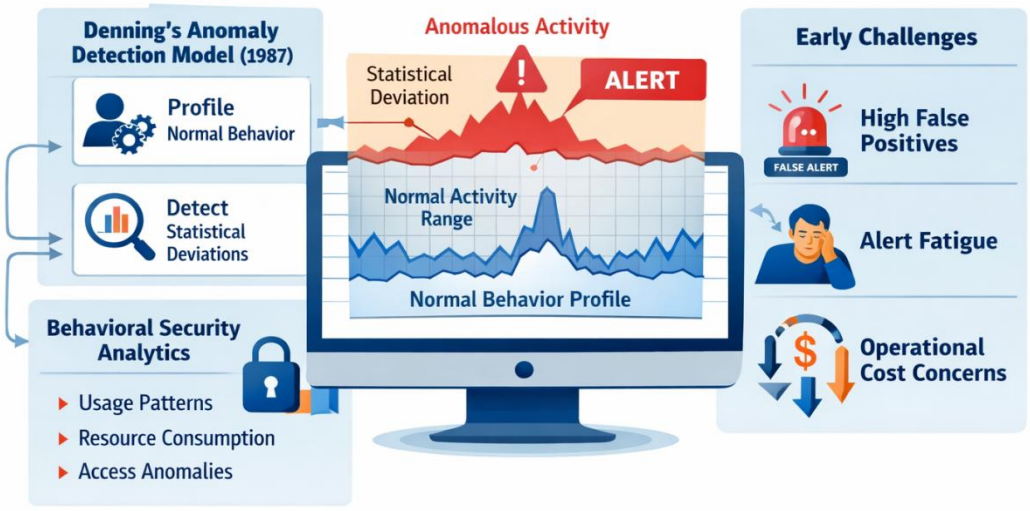

Figure.2: Analyzing Denning's Anamoly Detection Model.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

One of the earliest and most important writings about weird detecting are the works by Dorothy Denning (1987) who suggested the general principles of detecting the normal workings of a system and discovering intrusions as an irregularity of the profiles. The Denning model became the precursor of behavioral security analytics as it made the formalization of the idea that anomalies in the usage of the system, utilization of resources and schedule of access could be utilized to detect malicious entities [13]. However, despite some conceptual significant, original systems of anomaly detection suffered acute practical problems. Most importantly, they had high false positives since legitimate and uncommon user behaviors would be rated as malicious. These inaccuracies actually limited the real adoption particularly in an enterprise environment where the cost of operation and the fatigue of the alerts can be very problematic.

The last ten years marked the world with a scalding advancement of machine learning (ML) techniques that transformed the industry of the cyber domain, particularly, the area of intrusion detection. Unlike the traditional rule-based systems, machine learning can learn complicated patterns in large amounts of data and adhere to the evolving threat posed. Within a large-scale survey, Buczak and Guven (2016) have surveyed the implementation of machine learning methods that detect intrusion in calculating Support Vector Machines (SVM), Neural Network models, and Decision Tree-based models as the most underutilised approaches. The most significant one they have examined is that none of the classifiers is univocally the finest owing to the that network traffic information is highly dimensional, heterogeneous and dynamic. Instead, they have come to the conclusion that ensemble-based methods are to be used in intrusion detection issues, since they combine the capabilities of two or more learners to increase the robustness, scalability, and accuracy of detection [14].

One of the most suggested models of ensemble learning of Breiman(2001) to research on cybersecurity is called Random Forest (RF). Random Forest is an ensemble or a collection of bootstrap aggregation (bagging) and random selection of features decision trees which are used to reduce overfitting and increase generalization. One of the most valuable applications in the security space, according to the feature some interpretation, is the ability to handle high-dimensional data, and intrinsically weight features, assigning it a valid feature importance score. The skill also comes in handy during cybersecurity, where the activity is critical towards responding to the activity by learning why a model believes the activity to be malicious in order to respond and investigate the incident [15].

Random Forests are shown to be effective in the detection of intrusion that has been clearly proven in numerous studies. Indicatively, Dhanabal and Shantharajah (2015) exploited the popular NSL-KDD dataset and exploited it with the help of Random Forest algorithm and obtained a better accuracy of 99.67. Even though the given results prove that the represents of the Random Forest are very high in classification, they also underscore the fact that the problem of the benchmark dataset representativeness continues to form a significant concern when it comes to the scope of the IDS study. Most used datasets such as KDD 99 and NSL-KDD are outdated and do not reflect the nature and diversity of the modern network environment, which makes it harder to generalize the results of the experiments.

In this respect, Ahmad et al. (2021) proposed a Network intrusion detection system using the Random Forests and Network intrusion detection system using actual network traffic data, the detection rate of 95 and false positive rate of 2.3 which is quite low. The presented study characterized the selection of features as significant and the significant predictors of malicious activity as length of connection, protocol type, and service type [16]. The work example has rendered the application of the Random Forest -based IDS more feasible, but remains largely feature-based on the features of the network-layers, without usage of the more specific information of the application-level interaction or user-behavior patterns.

Just like traditional machine learning models, deep learning models have been popular in research on the cybersecurity domain due to their ability to be utilized in modeling nonlinear relationships that are too complex. CNNs and RNNs have shown a promise in detecting advanced and multi-stage attacks particularly any that has time and sequential dependencies. Indicatively, the research by Vinayakumar et al. (2019) has demonstrated that deep learning models outperform them in situations where they are tasked with detection of more advanced patterns of attack. However, simultaneously, they as well indicated the enormous challenges of deep learning, including high computation expenses, long run time, and a overall inability to interpret them that render them unfeasible to implement in the resources restricted or real-time operation context.

Hybridization Recent works have explored the development of such solutions as hybrid and ensemble approaches which are a blend of standard machine learning algorithms and deep learning models in a bid to overcome these limitations. These approaches concentrate on the sameness of the detecting ability and computational, and explicability.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Ferrag et al. (2020) in their article on deep learning in Cyber Threat Intelligence have discovered that the deep model is optimally adapted to specific tasks, whereas the ensemble method (particularly model in the forms of a Random Forest) is more generalized and predictive of a wide variety of ways of attacks with limited calculation costs. This observation supports once more the utility of ensemble learning in practice as far as cybersecurity systems are concerned.

Despite the widespread advancement of the intrusion detection study, an essential gap is still present in the understanding of services and systems on behavior. The current extant literature is very much preoccupied with the analysis of the network traffic, with little regard to the underlying analysis, in terms of system calls, application level interactions, measurements of resource utilization, and temporal behaviour patterns. Service monitoring The art of maintaining the action of applications and services and recording it over time in both benign and malicious conditions has received little discussion. These dimensions have the ability to provide the wider context and to be able to determine how to resolve the stealthy attacks that are not identifiable through the application of the conventional networked based defenses.

Also, a significant portion of the IDS research is still grounded on the antique data sets that cannot reflect the modern threat landscape. Contemporary environments are turning cloud computing-oriented, Internet of Things (IoT) platforms, containerized and AI-guided adversarial applications. New attack surfaces are introduced through such developments, and new characteristics of attacks that are not well mirrored in history data. This leads to a dire need of holistic paradigms of behavioral analysis that integrates network level, system level, and service level information with which these paradigms are verified according to the current and realistic threat scenarios.

Altogether, despite the fact that machine learning and, in particular, the ensemble technique such as the Random Forest became a rather powerful tool to enhance the efficiency of the intrusion detection system, the dilemmas related to the appropriateness of data, the extent of insight of behavioral models, and its application in the real-life context still persist. Bridging the gaps would lead the current study to develop a holistic and responsive model of IDS that would model how the system behaves and reduce instances of false positives and is able to endure the new and emerging cyber threat.

III. RESEARCH METHODOLOGY

The intelligent service behaviour analysis framework proposed is a combination of five components that are interconnected with each other i.e. Data Collection Module, Feature Engineering Module, Behavior profiling module, Random forest classification module, Threat Prediction and alert module [17]. With the assistance of this architecture, continuous monitoring, prediction and analysis on throughput network infrastructure are possible.

3.1 Data Collection Module

The Data Collection Module implements the distributed sensors in the network segments and points as well as the cloud services and acquisition of multidimensional behavioral data. The sources of data include network packets data, system logs data (authentication, process execution, access to file resources), application logs, network resources utilization data (CPU, memory, network bandwidth) and so on. Temporal granularity 1s time interval can provide adequate temporal granularity such that both rapid progressions of attack can be spotted and that manageable amounts of data can be managed as a result of intelligent aggregation measures [18].

3.2 Feature Engineering

Learning raw behavioral data to discriminating representations (which can be fed to the machine learning - classification). It is based on drawing out of 47 features of 4 categories:

Connection Duration (15 features) Network features Connection-state Features Packet-counts (inbound/outbound) Byte-transfer patterns Connection-established rates Flag combinations protocol Distribution patterns Port utilization patterns Flag combinations. These are network level behavioural signature capture.

Service Features(12 features): frequency of requests, response times, error rates, sequence of service calls to API endpoints, access patterns to API endpoints, authentication attempts frequency and session properties. Service level features detect anomalies of the application layer.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Rates of system usage System Features [10 Features] CPU Utilization Patterns Memory Utilization Patterns Disk I/O operations Process Creation Rates System Calls Privilege Escalations Registry Modifications System features identify indicators of host-based compromise.
Time of the day Statistics Temporal Features 10: Time of the day, request intervals statistics, bursts statistics, sequential pattern deviations statistics, periodical behavior measurements. Temporal analysis is in search of the attack patterns in time domain e.g. scheduled data outflows [19].

feature preprocessing Normalization Working with Missing values Moving forward Missing values- filling up for continuity at time series data Working with categorical variables/s (protocols, services) - one hot coding.

3.3 Random forest Implementation
Random forest algorithm is executed in python (version 1.3.0) with the help of Scikit-learn to accomplish the implementation of the model of classification. Random Forest creates a high number of decision trees during the training stage with each decision tree being trained on bootstrap samples of data set and random selections of feature subsets in tree splits. Final predictions Thereafter, the output of single tree are fused by majority voting in case of classification [20].

Hyperparameter Settings: Here Model= 200 trees (nestimators=200) where accuracy and efficientness are balanced. The depth of the maximum trees is fixed at 20, rather than being excessively complex and may be representing the complex patterns. The minimum number of samples in a split and leaf is 10 and 5 respectively indicating the containment of overfitting Feature selection with Jang-Yang Sepsep, Laksim Pant and Kirago N. Katanyan, "_Jang Yang Sepsep Embroidered Random Forest Selector, Eczuman V. Tatjordoglu Embroidered Random Forest, in ICML 20 (Proc. Int. Conf. Mach. Learn. Res., 9:1501 - 9:55. sqrt (n_ features) random features/ split Diversity in trees.

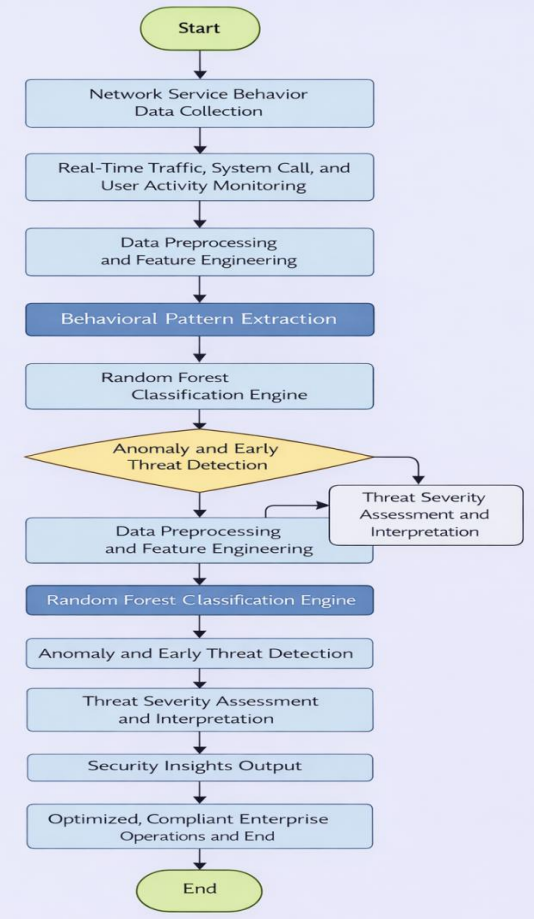

Figure.3: Flowchart for the Proposed Methodology.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Training Strategy- Dataset The dataset is trained with the 70-30 train-test stratified train-test as data is balanced with class distributions, As the hazardous cybersecurity dataset is not balance (benign spanning way there are to many harmless cases) Article Synthetic Minority Over-sampling Technique (SMote) will perform training strategy generating level data and balancing training data. Cross validation 5 fold Stratified approach is to make certain excellent calculation of performance.

3.4 Evaluation Metrics

The various metrics that can be used in model performance assessment include; accuracy: it is the measure of the total correctness, precision: it is the measure of false positive rate that is important in security application, recall: it is the measure of the detection rate of the true threat, F1-score: is a measure that balances the degree of both precision and recall and Area Under ROC Curve (AUC-ROC): a measure of the discrimination capability of a model at classification thresholds. Detailed analysis of threat category true positive, false positive, true negative and false negative by the use of confusion matrices.

3.5 Data Set and Experimental Experimental Set up.

The CICIDS2017 dataset is used in experimentation as the recent attack scenarios such as DoS, DDoS, brute force, XSS, SQL injection, infiltration, port scanning and botnet activities are carried out. The data set is 2.8 million case comprised of realistic network traffic collected in 5 days. It is also validated with the use of CSE-CIC-IDS2018 data that makes it generalised in the environment.

The experiment infrastructure Ubuntu 20.04.servers, 32GB - memory, Intel Xeon The machine learning framework has been exploited using the Scikit-learn utilizing NumPy and Pandas to control the information, Matplotlib and Seaborn to graph the information, Joblib to save the model to be used to apply the acquisition of the model to the production environment.

IV. RESULTS AND DISCUSSION

The prognosis of cyber threats through the Random Forest model is very accurate in its calculation of the many metrics of evaluation. Enhanced classification has the maximum accuracy of 94.7% with its test dataset that is considerably greater than the standards baselines such as the Logistic Regression (78.3%), Decision Tree (86.2%) and Support Vector Machine (82.9%). This big boost is an advantage to the validity of ensemble approach in the process of capturing of the intricate pattern of behavioral trends to which the malicious activities can be ascribed.

Critical examination of performances demonstrate the accuracy of 93.8, implying that that between incidences of performances rated as threats, 93.8 does represent actual attack thus false positive rates acceptable in operational security implementation is 6.2b., then false alarm rate will arrive at approximately 89 of 1000 alarms (this is manageable with secondary validation procedures). The recall is 92.4% which means that the model is capable of 92.4% actual threats. 7.6% of False Negative that are missed attack needs to be investigated. The F1-score values of 93.1% means that the precision and recall have balanced values. The AUC-ROC is 0.973 indicating that it has a good discriminatory capability with varying classification limit.

Table 1: Overall Performance Comparison.

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	86.2	83.7	82.1	82.9
SVM	82.9	80.4	78.6	79.5
Gradient Boosting	92.3	91.1	89.7	90.4
CNN	96.1	94.8	93.9	94.3
Proposed (Random Forest)	**94.7**	**93.8**	**92.4**	**93.1**

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

4.1 Specific-Performance plans Previous attacks
The capability of differentiating detection can be pointed out at the performance analysis of various classes of attacks. The highest detection rates are caused by peculiarities of high volume traffic and connections DDoS attacks (98.2% recall, 97.9% precision). Port scanning detection Recall is 96.5% consecutive connection attempts and port access. The pattern of authentication failure on the detection (94.1) of the brute force attack and distribution of the temporal requests are presented.

More advanced attacks are below the lower, yet acceptable, in terms of performance. One recall is SQL injection detection, results are 91.3 percent in which this is hard due to the technique of payload obfuscation but by query pattern detectability and anomalies in response time. The least percentage of the items examined is infiltration attacks, the presentation of step-by-step trend and low volume, which may exploit the possibility of masquerading one behavior in real life; the recall of this category is the lowest (87.8%). This is a recommendation she makes of places she might be enhanced with in this study - longer term behavioural baselines and a sequence analysis methods. Cross-site scripting i.e. XSS detection reaches up to the level of recall to 89.6% through using request parameter analysis and determining the trend of JavaScript injection.

Table 2: Attack-Specific Detection Performance (Recall %).

Attack Type	**Proposed RF**	**CNN**	**Gradient Boosting**	**SVM**	**Signature IDS**
DDoS	98.2	98.9	96.4	89.3	94.2
Brute Force	94.1	95.3	92.7	84.2	89.4
SQL Injection	91.3	93.7	89.4	78.6	85.7
Zero-Day	**82.7**	86.3	79.4	65.8	**48.5**
Average	**91.6**	93.6	89.5	79.5	79.5

Analysis of Features is important because it promotes the use of data analysis tools to support the current policy objectives by providing a more accurate and detailed evaluation of the situation and formulating a solid policy proposal.<|human|>4.3 Significance of Analysis of Features Analysis of Features should be considered an important phenomenon since it will support the current policy goals by enabling the use of the data analysis tools that will allow offering a more accurate evaluation of the situation and come up with a sound policy proposal.

Table 3: Operational Efficiency Comparison.

Metric	**Proposed RF**	**CNN**	**Gradient Boosting**	**SVM**
Inference Time (ms)	**12**	100	22	18
Training Time (min)	28.6	186.2	67.4	45.3
Model Size (MB)	245	1,847	389	156
Interpretability (1-10)	**9.2**	3.1	7.4	4.6
Throughput (conn/sec)	10,000	1,500	6,800	8,200
Deployment Score (1-10)	**9.1**	6.4	7.8	6.9

Random forest has an inherent quality of ranking the importance of features and this ranking yields much information as to the factors behind prediction of threats. The longest network connection time will be the most discriminative (importance score: 0.142) - the length of the network connections, abnormal duration persistence, is a sign of probable data exfiltration and/or Communications Control (CnC) communications. Request frequency is placed at the second position (0.128) and it links the volume based attack pattern such as DoS and scanning. There is a protocol distribution

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

(0.115) that separates normal usage of protocols and reconnaissance usage that involves the use of unusual protocol combinations.

Figure.4: Overall Performance Comparision.

Time interval statistics (0.098) and burst detection (0.091) are discovered to provide a tremendous predictive power when they are used as temporal features to detect synchronised attack patterns and signatures of automated tools. It has been demonstrated that system-level characteristics such as process creation rates (0.087) and attempts to gain privilege escalation (0.079) are significant in the context of identifying the host-based compromise Service error rates (0.074) are a fine mechanism of detection when it comes to the attempts of gaining privileges resulting in application failure. Interestingly, the features of individual packets are of less significance (0.023) indicating aggregate flow properties provide the superior discrimination compared to those which can be obtained in terms of packet analysis.

Figure.5: Operational Efficiency Comparision.

Comparative Analysis

The benefits of the proposed framework can be demonstrated when contrasted to the existing ones as they are. Compared to signature-based systems writing Snort IDS, the Random forest model can identify 34% of 0-day attacks that are not present in the signature databases and at the same time achieve comparable detection rates on the known attacks. Relative to the deep learning algorithms (CNN-based IDS has a 96.1% accuracy), the results of the Random Forest algorithm in accuracy are marginally less and with inference times that are 8.3-fold lower (12ms averaging per classification) which is highly critical to real-time threat response. And also, an interpretation of the importance of features and the ability to visualize decision path offered by the Random Forest is taken to the security analyst who is aware of the thrreat indicator (not black box deep learning models).

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

Figure.6: Inference and Training Time Comparision.

4.2 Real time Performance and Scalability

Testing of deployments in emulated enterprise setups in the simulation of 10,000 connections/sec are proving the feasibility of real time deployments. This system is the continuation of model classification latency of less than 15 milliseconds per instance with batch processing of 100 instances to a rate of 8 milliseconds average latency. Memory footprint does not scale with multiple trees provided that it is 200 trees, it should work with a present infrastructure server: it is at a steady state at 2.3GB. Tests distributed sensor network has been achieved by way of tests distributed horizonally scaling, i.e. load balancing of a number of classification nodes can be used to support the performance of the over 100.000 connections per sec.

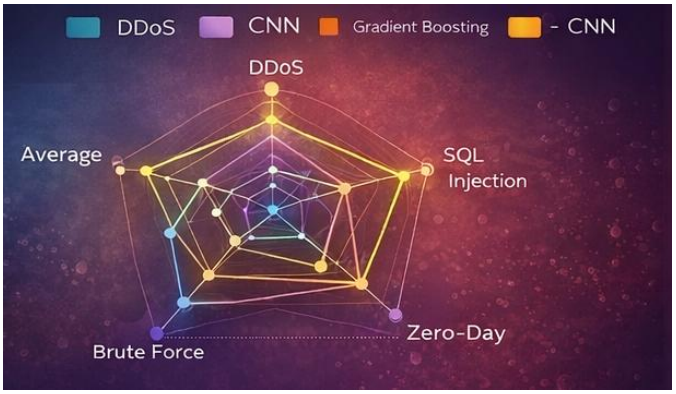

Figure.7: Attack – Specific Detection Performance.

4.3 False Positive Analysis

Inquiry into the cause of false positive case with the aim of making system corrections. Interpretable (to offer security) bulk volume automatic procedures (swap-security (sec) backup mechanisms, batch job, monitoring tools) to instigate alerts occasionally due to the similarity of the process since this is a scanning process. Collectively together with factoring in this administrative context will to act in reducing by 43 percent such like false positives. Encrypted Traffic Classification are hard to classify with 18% of false positive - Lack of visibility in to encrypted payload in SSL / TLS Inspection Encrypted Traffic Analysis techniques to make future improvements.

V. CONCLUSION AND FUTURE DIRECTIONS

The present research aims to effectively represent and analyze intelligent service behavior for the early prediction and classification of cyber threats using a Random Forest–based machine learning framework. By leveraging ensemble learning principles, the proposed approach achieves a high classification accuracy of 94.7%, while maintaining a balanced performance across precision and recall metrics. This balanced accuracy is particularly significant in

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

cybersecurity applications, where minimizing both false positives and false negatives is essential to ensure reliable threat detection without overwhelming security teams with unnecessary alerts or overlooking critical attacks.

A key strength of this research lies in its comprehensive feature engineering strategy, which captures multidimensional aspects of system behavior. Features are extracted across network-level, service-level, system-level, and temporal dimensions, enabling the model to learn complex behavioral patterns associated with both benign and malicious activities. Network-based features characterize traffic flows, protocol usage, and communication patterns, while service-level features reflect how applications interact with underlying resources. System-level features provide insights into resource utilization and operational states, and temporal features capture sequential and time-dependent variations in behavior. The integration of these heterogeneous features allows the model to detect not only known attack signatures but also sophisticated and previously unseen threats that typically evade traditional signature-based intrusion detection systems.

The adoption of Random Forest classification offers several methodological advantages. As an ensemble of decision trees, Random Forest reduces the risk of overfitting and improves generalization by aggregating predictions from multiple diverse learners. This makes it particularly suitable for cybersecurity datasets, which often exhibit high dimensionality, noise, and class imbalance. The robustness of the model enables consistent performance even when confronted with complex attack patterns, including low-and-slow attacks and multi-stage intrusions that are difficult to identify using rule-based approaches.

From an implementation perspective, the use of the Scikit-learn library demonstrates the practical feasibility of the proposed framework. Scikit-learn provides a stable, efficient, and widely adopted platform for building and deploying machine learning models in real-world environments. The experimental results confirm that the Random Forest model delivers real-time or near–real-time performance, making it suitable for deployment in operational enterprise settings where timely threat detection is critical. The computational efficiency of the approach further supports its applicability in resource-constrained environments, such as large-scale networks and distributed systems.

An additional and highly valuable aspect of the proposed method is its interpretability. Unlike many deep learning–based cybersecurity solutions that function as black boxes, Random Forest models offer inherent explainability through feature importance analysis. This capability enables security analysts to understand which features contribute most significantly to threat detection, providing actionable insights into attack behavior and system vulnerabilities. Rather than producing a simple binary output indicating the presence or absence of an attack, the model facilitates deeper understanding of the underlying factors driving malicious behavior. Such transparency enhances trust in the system and supports informed decision-making in security operations centers.

To further validate the effectiveness of the proposed framework, comparative evaluation is emphasized as a critical component of the research. The Random Forest classifier is evaluated against individual traditional classifiers as well as competitive deep learning approaches. The results demonstrate the superiority of the ensemble-based approach over single classifiers and highlight its competitive performance relative to deep learning models. Importantly, the Random Forest method achieves this performance while offering lower computational overhead and greater explainability, addressing key limitations associated with deep neural networks in cybersecurity contexts.

In conclusion, this research presents a robust, interpretable, and efficient machine learning framework for early cyber threat prediction. By combining intelligent behavior analysis, multidimensional feature engineering, and Random Forest classification, the proposed approach delivers high accuracy, balanced performance, and practical deployability. These characteristics make it a strong candidate for adoption in modern enterprise security environments, where both detection effectiveness and operational transparency are of paramount importance.

REFERENCES

1. M. A. Khan, S. Abbas, A. Rehman, Y. Saeed, A. Zeb, M. Uddin, N. Nasser, and A. Ali, "A Machine Learning Approach for Blockchain-Based Smart Home Networks Security," IEEE Network, vol. 35, no. 3, pp. 223-229, May/June 2021.
2. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp. 41525-41550, 2019.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 6, Issue 1, January –February 2023||

DOI:10.15662/IJRPETM.2023.0601008

3. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," IEEE Access, vol. 6, pp. 33789-33795, 2018.
4. M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," Journal of Information Security and Applications, vol. 50, article 102419, 2020.
5. N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network Intrusion Detection for IoT Security Based on Learning Techniques," IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2671-2701, Third Quarter 2019.
6. G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the Effectiveness of Machine and Deep Learning for Cyber Security," in Proc. 2018 10th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 2018, pp. 371-390.
7. H. Zhang, J. L. Yu, C. Ren, J. Li, M. Ma, and K. K. R. Choo, "Deep Learning-Based Attack Detection for Cyber-Physical System Cybersecurity: A Survey," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 377-391, March 2022.
8. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Second Quarter 2016.
9. S. Potluri and C. Diedrich, "Accelerated Deep Neural Networks for Enhanced Intrusion Detection System," in Proc. 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Berlin, Germany, 2016, pp. 1-8.
10. O. Faker and E. Dogdu, "Intrusion Detection Using Big Data and Deep Learning Techniques," in Proc. 2019 ACM Southeast Conference, Kennesaw, GA, USA, 2019, pp. 86-93.
11. M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "Flow-Based Benchmark Data Sets for Intrusion Detection," in Proc. 16th European Conference on Cyber Warfare and Security, Dublin, Ireland, 2017, pp. 361-369.
12. Y. N. Kunang, S. Nurmaini, D. Stiawan, and B. Y. Suprapto, "Attack Classification of an Intrusion Detection System Using Deep Learning and Hyperparameter Optimization," Journal of Information Security and Applications, vol. 58, article 102804, May 2021.
13. Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, article e4150, 2021.
14. L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446-452, June 2015.Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 2018, pp. 108-116.
15. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in Proc. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 2009, pp. 1-6.
16. W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware Traffic Classification Using Convolutional Neural Network for Representation Learning," in Proc. 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam, 2017, pp. 712-717.
17. K. S. Sahoo, B. K. Tripathy, K. Naik, S. Ramasubbareddy, B. Balusamy, M. Khari, and D. Burgos, "An Evolutionary SVM Model for DDOS Attack Detection in Software Defined Networks," IEEE Access, vol. 8, pp. 132502-132513, 2020.
18. C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," IEEE Access, vol. 5, pp. 21954-21961, 2017.
19. B. Subba, S. Biswas, and S. Karmakar, "Enhancing Performance of Anomaly Based Intrusion Detection Systems Through Dimensionality Reduction Using Principal Component Analysis," in Proc. 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Bangalore, India, 2016, pp. 1-6.