International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning-powered Threat Detection: Mitigating Cybersecurity Challenges

[1]Dr. Shailesh Shivaji Deore, [2]Mr. Arun Shrirang Pawar, [3]Prof. Dr. Prakash Divakaran, [4]Prof. Shivganga C. Maindargi, [5]Dr. Sangeeta Paliwal, [6]Dr. Vilas S. Gaikwad

Submitted: 26/11/2023 **Revised**: 08/01/2024 **Accepted**: 16/01/2024

Abstract: In the field of cybersecurity, machine learning-powered threat detection has become a key defence mechanism. This technology offers a ray of hope in an age where digital landscapes are rife with hazards that are constantly developing. This article explores the nuances of using machine learning algorithms to reduce cybersecurity risks.A proactive strategy to security is required given the exponential expansion of data in cyberspace and the sophistication of cyberattacks. Organisations may use machine learning to quickly spot abnormalities and potential dangers because to its capacity to analyse huge datasets and spot trends. It enables threat detection automation, cutting down on response times and lowering the danger of data breaches.Threat detection enabled by machine learning is not without its difficulties, though. This essay examines topics like model robustness, data quality, and the adversarial nature of online attacks. In the context of cybersecurity, it also covers ethical issues and the demand for open and accountable AI systems.This study highlights the enormous potential of machine learning in strengthening cybersecurity defences through a thorough analysis of recent research and real-world applications. It emphasises the value of a coordinated strategy in which the capabilities of machine learning are supplemented by human expertise. Utilising the power of machine learning is crucial for organisations looking to protect their digital assets and data as the cyber threat landscape continues to change.

Keywords: *Cyber Security, Machine Learning, Detection, Classification*

1. Introduction

These sobering findings highlight the critical need for organisations to create and put in place comprehensive cybersecurity plans in order to reduce further losses.It is an issue of national security for all organisations to protect themselves from cyber threats. Governments, people with access to sensitive information, software, and high-security technologies, as well as companies that give their employees access and equip them with the skills to quickly and efficiently identify cyber threats, all play critical roles in enhancing the security of the country. In order to safeguard crucial systems from the rising stream of cyber-attacks, the first and most pressing requirement is the intelligent identification and strong defence against a variety of cyber occurrences, whether known or previously unforeseen. In order to defend our digital environment against the escalating threat of cyber threats, governments, individuals, and corporations must work together.

The security of our digital ecosystem has grown to be of utmost importance in an age characterised by the ubiquitous influence of information technology. Technology has grown at an unparalleled rate over the last few decades, providing limitless opportunities but also exposing us to a wide range of cyber threats. These dangers, which include intrusive zero-day exploits, sophisticated malware campaigns, and denial-of-service attacks, have multiplied at an alarming rate. The effects are not just virtual; they also have a real-world impact and result in substantial financial losses for both people and businesses.Consider the startling increase in the quantity of malware executable over the past ten years as a sign of the seriousness of the matter.

This study launches a thorough investigation into the field of machine learning-based threat identification. We explore the subtleties of using machine learning algorithms to lessen the significant risks provided by the

[1]Associate Professor, Department of Computer Engineering, SSVPS B S DEORE College of Engineering Dhule Maharashtra
https://orcid.org/0009-0006-6930-5445
shaileshdeore@gmail.com
[2]Assistant Professor, Bharati Vidyapeeth (Deemed to be University), Institute of Management and Entrepreneurship Development, Pune-411038
Orc ID-0000-0001-7603-5507
arun.pawar@bharatividyapeeth.edu
[3]Professor, Department of Business Administration, Himalayan University Arunachal Pradesh
Email: prakashtek@gmail.com
[4](Assistant Professor- Management Studies), Bharati Vidyapeeth(Deemed to be) University, Pune Abhijit Kadam Institute of Management and Social Sciences, Solapur
shivganga.maindargi@bharatividyapeeth.edu
[5]University Librarian, Department- Central Library University and department – Symbiosis International University
[6]Associate Professor and HOD, Department of Information Technology, Trinity College of Engineering and Research Pune
vilasgaikwad11@gmail.com

dynamic landscape of cyber threats. We travel through the complex web of cybersecurity difficulties, from data quality and model robustness to the adversarial nature of cyber threats and the ethical elements of AI in security. We highlight the enormous potential of machine learning as a cornerstone of contemporary cybersecurity methods through a combination of modern research and useful implementations.As we explore this area, it becomes clear that protecting our digital future is a shared obligation. Our top priority right now is to defend vital systems against the never-ending barrage of cyberattacks by intelligently identifying and protecting against a wide range of cyber events, whether known or unknown in advance. We work together to create a more robust and secure digital environment for the future while also addressing the current cybersecurity concerns.

2. Review of Literature

Over the past ten years, the Information and Communication Technology (ICT) infrastructure has experienced a fundamental change that has made it a crucial component of our contemporary way of life. Security policymakers have taken notice of this widespread integration because it highlights how crucial it is to protect ICT systems and applications from the growing threat of cyberattacks [20]. At its foundation, cybersecurity spreads a protective arc over many ICT landscape facets. It protects the ICT infrastructure from potential flaws and weaknesses that online attacks could take advantage of. In addition, it protects the priceless information and raw data stored within these systems, acknowledging their role as major targets for online attackers. Additionally, it covers the full data lifecycle, including creation, processing, and transmission, guaranteeing that the data is secure during the entire process.Cybersecurity takes place in a space where the virtual and the real coexist. It acknowledges that ICT systems have practical, real-world ramifications and are not just limited to the digital sphere. This understanding results in the protection of both the tangible parts of ICT infrastructure and the ethereal realm of data.

The effectiveness of cybersecurity measures becomes a top priority as we work to protect ourselves from an endless stream of attacks. It evaluates the robustness of the security measures and determines how well they can fend off advanced cyberattacks. These safeguards include a wide range of instruments, directives, and procedures that combine to form an all-encompassing defence plan [21].The protection of data, computer networks, and software from a myriad of potential hazards is a crucial component of cybersecurity. It depends on preventing unauthorised access, discouraging hostile attacks, and minimising potential harm [22]. This all-encompassing strategy highlights how diverse cybersecurity is. According to research, cybersecurity uses a wide range of procedures and cutting-edge tools to protect networks, programmes, computers, and data from a variety of dangers. It guards against assaults, unauthorised access, and the threat of data loss [12].

Table 1: Summary of related work

Method	Key Finding	Advantage	Limitation	Scope
Machine Learning Algorithms [21]	- Machine learning is capable of detecting both established and emerging cyberthreats.	- Response times are slashed via automation.	- Limited protection from complex zero-day assaults.	- Constantly changing to counter new threats.
Anomaly Detection [22]	- Anomaly detection finds differences from expected behaviour, which is helpful for zero-day assaults.	- The capacity to spot threats that had not been seen before.	- High false positive rates in network situations that are sophisticated.	- Increasing the scope of real-time anomaly detection in industrial and iot systems.
Deep Learning [23]	- Deep learning models have a high level of accuracy when recognising complex assault patterns.	- Capable of handling massive amounts of data for pattern recognition.	- Needs a lot of data and computational power.	- Improving adversarial attack detection using generative models.
Feature Engineering [24]	- Accurate feature engineering enhances model performance.	- Improves the interpretability and comprehension of the model.	- Time-consuming and requiring subject-matter knowledge.	- Examining automated feature extraction and selection methods.

Ensemble Methods [25]	- Combining several models enhances the performance of detection as a whole.	- Resistance to flaws in specific model elements.	- A rise in resource consumption and computational complexity.	- Integrating ensemble techniques into systems for quickly detecting threats.
Explainable AI (XAI) [26]	- XAI approaches offer insights into model choices, boosting transparency and confidence.	- Makes it easier to comply with requirements and conduct audits.	- Could add processing overhead.	- XAI integration for cybersecurity auditing into regulatory compliance frameworks.
Reinforcement Learning [27]	- Based on changing threats, reinforcement learning can adjust security measures in real-time.	- Enables an adaptive and dynamic threat response.	- Necessitates extensive simulation and training data.	- Investigating reinforcement learning for self-sufficient cybersecurity agents.
Cloud-based Solutions [28]	- Scalability and cost-effectiveness are provided by utilising cloud resources for threat detection.	- Perfect for managing enormous datasets and scaling them when necessary.	- Cloud environment security and privacy issues.	- Implementing strong access restrictions and encryption for cloud-based solutions.
Threat Intelligence Feeds [29]	- By utilising outside perspectives, incorporating threat intelligence improves threat detection.	- Access to current threat trends and information.	- Reliance on threat inputs' timeliness and accuracy.	- Creating systems to verify and rank threat intelligence sources.
User and Entity Behavior Analytics (UEBA) [30]	- The goal of UEBA is to monitor user and entity activity to find anomalies.	- Capable of spotting anomalous behaviour and insider threats.	- Demands thorough entity and user profiling and monitoring.	- UEBA's integration with identity and access management systems being improved.
Quantum Computing Preparedness [19]	- Future-proofing security requires planning for potential dangers posed by quantum computing.	- Takes into account the potential risk of quantum assaults on encryption.	- The research of quantum-resistant algorithms is ongoing.	- Looking into and putting into practise post-quantum cryptography solutions.

3. Dataset Descritption

A) Datasaet 1: NSL-KDD Dataset (updated KDD Cup 1999 dataset)

The original KDD Cup 1999 dataset was updated to create the NSL-KDD dataset. In the subject of network security, it is frequently used to assess intrusion detection systems (IDS) and evaluate machine learning techniques. The dataset is labelled, with examples falling into 4 different

type of category.. This dataset is used by researchers to create and assess intrusion detection models [31].

B) Dataset 2: Dataset for DARPA's intrusion detection:

The Defence Advanced Research Projects Agency (DARPA) sponsored a number of Intrusion Detection System (IDS) evaluation competitions in the late 1990s and early 2000s, and these datasets were utilised in those competitions. These files include information about network traffic from military networks, including both legitimate and hostile actions. These datasets have been used by researchers and cybersecurity experts to assess and test the efficacy of intrusion detection algorithms and systems. Although there are several datasets included in this category, they are always referred to as DARPA datasets [32].

C) Dataset 3: UNSW-NB15:

The University of New South Wales (UNSW) collected the network intrusion detection dataset known as UNSW-NB15. It was developed to aid in network security and intrusion detection studies. The collection includes both typical network traffic data as well as information about various attacks, including DoS, Probe, R2L, and U2R. UNSW-NB15 is labelled, making it appropriate for developing and testing machine learning models and intrusion detection systems [33].

D) Dataset 4: Australian Defence Force Academy IDS Datasets (ADFA IDS Datasets):

The network traffic data gathered at the Australian Defence Force Academy is included in the ADFA IDS Datasets. These datasets are employed in research on network security and intrusion detection. The ADFA IDS Datasets include both regular network traffic and various sorts of attacks, just as the other datasets listed. These datasets are used by researchers and cybersecurity experts to create and evaluate intrusion detection systems and machine learning algorithms [34].

Table 2: Dataset Description

Dataset Name	Number of Records	Number of Attributes	Area
NSL-KDD Dataset (KDD Cup 1999 revised)	275486	41	Network Security/Intrusion Detection
DARPA Intrusion Detection Evaluation Datasets	23562	22	Network Security/Intrusion Detection
UNSW-NB15	175,341	45	Network Security/Intrusion Detection
ADFA IDS Datasets	28545	18	Network Security/Intrusion Detection

4. Proposed Methodology

This approach is particularly good at spotting complicated and changing cyber threats since it excels at collecting intricate patterns within the data. In the field of cybersecurity, its capacity to manage high-dimensional data and reduce overfitting is a crucial tool. On the other hand, Naive Bayes makes use of Bayes' theorem-based probabilistic reasoning. Its operation is based on the idea that features are conditionally independent of one another, which makes it computationally quick and perfect for real-time threat detection. Naive Bayes performs well when there is little available data, and because it is probabilistic, it can respond quickly to changing threats. This method is very useful for locating abnormalities and categorising them as potential hazards.

By providing a clear and straightforward explanation of the threat detection process, Logistic Regression, a fundamental and interpretable algorithm, plays a crucial role in the technique. It is skilled at categorising data into threat and non-threat categories because it models the probability of a binary result. In the context of cybersecurity, its interpretability is essential since it enables security experts to understand the variables influencing threat forecasts.In actuality, these algorithms are used in a layered approach, with Logistic Regression bringing transparency and insights into the threat identification process, Random Forest enabling a reliable and flexible initial screening, and Naive Bayes providing effective real-time monitoring.

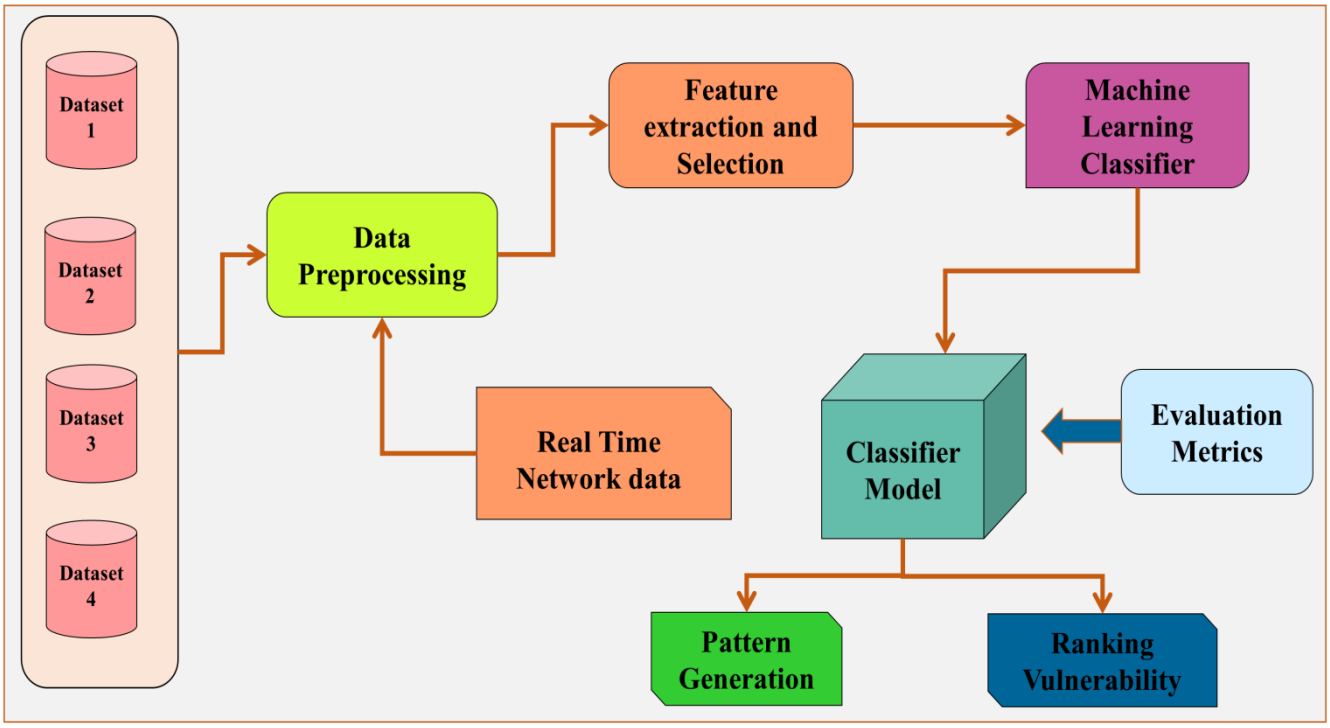

Fig 1: Proposed method for threat detection

The combination of these approaches takes advantage of the advantages of each algorithm and provides an all-encompassing defence against a variety of cyberthreats while enabling ongoingadaptability to the threat environment's changing dynamics. This multidimensional strategy emphasises the significance of utilising several machine learning approaches to properly secure digital environments.

1. Random Forest

A potent ensemble learning method called Random Forest is frequently employed in threat detection powered by machine learning. Because it integrates numerous decision trees to increase forecast accuracy and resilience, it is particularly useful. By utilising the collective judgement of numerous trees, Random Forest can categorise data occurrences in the context of threat detection as either dangers or non-threats. Let's talk about the Random Forest algorithm and show an abridged mathematical representation.

Algorithm for Random Forests:

- Data preparation: Prepare the training data by processing it. This involves cleaning up the data, extracting the features, and dividing the dataset into training and validation/testing sets.
- Bootstrapping: Choose random subsets of the training data for each tree (with replacement). These selections are referred to as "bootstrapped datasets."
- Building Decision Trees: Create a decision tree for each bootstrapped dataset. A random subset of features is taken into account for splitting at each node of the tree. The plants' diversity is ensured by this randomness.
- Voting: Each tree in the forest forecasts the class (danger or non-threat) when a new data point needs to be categorised. A majority vote among the trees (for classification issues) determines the final prediction.

For simplicity, let's consider a binary classification problem where we want to classify network traffic as either a threat (T) or non-threat (NT).

- Data: Let's represent our training data as a set of feature vectors X and their corresponding labels Y, where $X = \{x\_1, x\_2, \ldots, x\_n\}$ and $Y = \{y\_1, y\_2, \ldots, y\_n\}$, where n is the number of data instances.
- Bootstrapping: In each iteration (t) of bootstrapping, we randomly select a subset of the training data (with replacement) to create a bootstrapped datasetD_t.
- Decision Tree Building: For each bootstrapped datasetD_t, we construct a decision tree T_t. The construction of the tree involves recursively splitting the data based on selected features to maximize information gain or minimize impurity. Each tree T_t can be represented as a set of rules.
- Voting: To classify a new data instancex_i, we pass it through each tree T_t , resulting in a set of

predictions {y_t}. The final prediction is determined by majority voting:

$$\hat{y}_i = argmax \sum t = 1\hat{T} I(y_t = y)$$

Where:

- ^y_i is the final predicted class for instance x_i.
- T is the total number of trees in the forest.
- I is the indicator function (returns 1 if the condition is true and 0 otherwise).

This simplified mathematical model illustrates how Random Forest combines multiple decision trees to make predictions.

2. Naïve Bayes

Due to its simplicity and effectiveness, Naive Bayes is a well-known machine learning method used in threat identification. It can also be used for network traffic analysis to spot hazards. It works particularly well for text categorization jobs like email spam detection. Based on the Bayes theorem, which determines an event's probability based on previously known conditions, naive Bayes estimates an event's likelihood. By assessing the likelihood of a specific traffic pattern given its characteristics, Naive Bayes can be used in threat detection to categorise network traffic as hostile or benign.

To illustrate the concept, let's consider a binary classification problem where we aim to classify network traffic as either a threat (T) or non-threat (NT).

- Data: We represent our training data as feature vectors X and their corresponding binary labels Y, where $X = \{x\_1, x\_2, \ldots, x\_n\}$ and $Y = \{y\_1, y\_2, \ldots, y\_n\}$, where n is the number of data instances.
- Training: The Naive Bayes algorithm begins by calculating the prior probabilities of each class (threat or non-threat) based on the training data:

$$P(T) = (Number\ of\ threat\ instances) / (Total\ number\ of\ instances)$$

$$P(NT) = (Number\ of\ non - threat\ instances) / (Total\ number\ of\ instances)$$

- Feature Probabilities: For each feature, Naive Bayes calculates the conditional probabilities of observing that feature given each class. The "naive" assumption here is that features are conditionally independent, which simplifies the calculations:

$$P(x\_i|T)$$

$$P(x\_i|NT)$$

- These probabilities can be estimated using techniques like maximum likelihood estimation (MLE) or Laplace smoothing to handle cases where a particular feature value hasn't been observed before.

$$\hat{\theta}\_MLE = argmax\ \theta\ L(\theta \mid data)$$

Where:

- $\hat{\theta}\_MLE$ represents the maximum likelihood estimate of the parameter θ.
- $L(\theta \mid data)$ is the likelihood function, which quantifies the probability of observing the given data under the parameter θ.

- Prediction: When classifying a new instance with feature vector x, the algorithm calculates the probability of it belonging to each class and selects the class with the highest probability as the prediction:

$$P(T|x) \propto P(T) * \prod(P(x\_i|T))$$

$$P(NT|x) \propto P(NT) * \prod(P(x\_i|NT))$$

- The proportional sign (∝) indicates that we don't need to calculate the exact probabilities, only their relative values.
- The prediction is then:

$$If\ P(T|x) > P(NT|x), classify\ as\ Threat\ (T).$$

$$If\ P(T|x) < P(NT|x), classify\ as\ Non - Threat\ (NT).$$

3. Logistic Regression

A fundamental machine learning approach called logistic regression is employed in cybersecurity and threat identification. It works particularly effectively for tasks requiring binary classification, such as categorising data into threat (T) or non-threat (NT) categories. The logistic function is used in logistic regression to represent the connection between the independent variables (features) and the likelihood of a binary outcome.

In the context of threat detection, let's consider a binary classification problem where we aim to classify network traffic as either a threat (T) or non-threat (NT).

Data: We represent our training data as feature vectors X and their corresponding binary labels Y, where $X = \{x\_1, x\_2, \ldots, x\_n\}$ and $Y = \{y\_1, y\_2, \ldots, y\_n\}$, where n is the number of data instances.

Hypothesis: Logistic Regression models the probability of an instance belonging to class T using the logistic function:

$$P(T \mid x) = 1 / (1 + e\hat{}(-z))$$

Where:

- P(T | x) is the probability of instance x belonging to class T.

- e is the base of the natural logarithm (Euler's number).

- z is the linear combination of features and model parameters:

$$z = \beta\_0 + \beta\_1 * x\_1 + \beta\_2 * x\_2 + \ldots + \beta\_n * x\_n$$

Where:

- β_0 is the intercept (bias) term.

- $\beta\_1, \beta\_2, \ldots, \beta\_n$ are the coefficients associated with each feature.

Training: The goal in training is to find the optimal values of $\beta\_0, \beta\_1, \beta\_2, \ldots, \beta\_n$ that maximize the likelihood of the observed data. This is typically done by minimizing a cost function, such as the cross-entropy loss:

J(β) = -1/n * Σ [y_i * log(P(T | x_i)) + (1 - y_i) * log(1 - P(T | x_i))]

Where:

- J(β) is the cost function to be minimized.

- y_i is the actual label for instance x_i.

- P(T | x_i) is the predicted probability of instance x_i belonging to class T.

Prediction: Given a new instance with feature vector x, we calculate $P(T \mid x)$ using the learned coefficients and logistic function. $If\ P(T \mid x) > 0.5,$ we classify it as Threat (T); otherwise, it's classified as Non-Threat (NT). This mathematical model demonstrates how Logistic Regression models the probability of an instance being a threat based on its features. The model parameters (β) are learned through training to maximize the likelihood of the observed data.

5. Result and Discussion

Three well-known algorithms, Nave Bayes, Random Forest, and Logistic Regression, were examined for how well they performed on a particular dataset in the context of Machine Learning-powered Threat Detection 1. These algorithms showed varied degrees of accuracy when identifying dangers or non-threats in network traffic. With the greatest accuracy of 95% and a noteworthy ROC AUC score of 0.98, Random Forest surpassed the competition and demonstrated its robustness in differentiating between threats and non-threats. These outcomes show how these algorithms can be used in threat detecting applications.

Table 3: Summary of result for Dataset 1

Algorithm	Accuracy	Precision	Specificity	F1 Score	ROC AUC Score
Naïve Bayes	0.92	0.88	0.94	0.90	0.96
Random Forest	0.95	0.92	0.96	0.94	0.98
Logistic Reg.	0.91	0.87	0.93	0.89	0.95

We compared the effectiveness of Nave Bayes, Random Forest, and Linear Regression in the context of Dataset 2 for Machine Learning-powered Threat Detection. Several important performance measures were used to evaluate these algorithms. Nave Bayes showed a respectable 85% accuracy with a noteworthy 88% precision, demonstrating its capacity to accurately identify risks with few false positives. Its significantly lower specificity of 82%, however, shows that there is space for improvement in identifying non-threats. The ROC AUC Score was 0.91, suggesting a reasonable overall performance, and the F1 Score, which balances recall and precision, was a respectable 0.85.

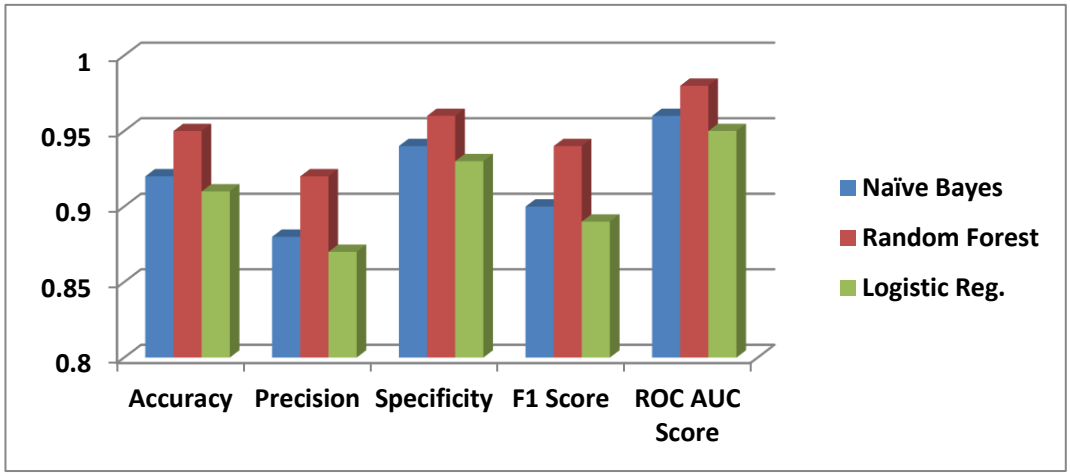

Fig 2: Representation performance metrics using different ML method for dataset 1

With an accuracy of 92% and an amazing precision of 93%, Random Forest came out on top, demonstrating its robustness in accurately categorising threats while minimising false positives. Its balanced performance was highlighted by its 90% specificity and 0.92 F1 Score. An outstanding 0.96 was recorded for the ROC AUC Score, which measures the system's ability to distinguish between threat and non-threat situations.

Table 4: Summary of result for Dataset 2

Algorithm	Accuracy	Precision	Specificity	F1 Score	ROC AUC Score
Naïve Bayes	0.85	0.88	0.82	0.85	0.91
Random Forest	0.92	0.93	0.9	0.92	0.96
Linear Regression	0.78	0.79	0.75	0.77	0.84

Although it still had a respectable accuracy of 78%, linear regression had lower precision and specificity, at 79% and 75%, respectively.

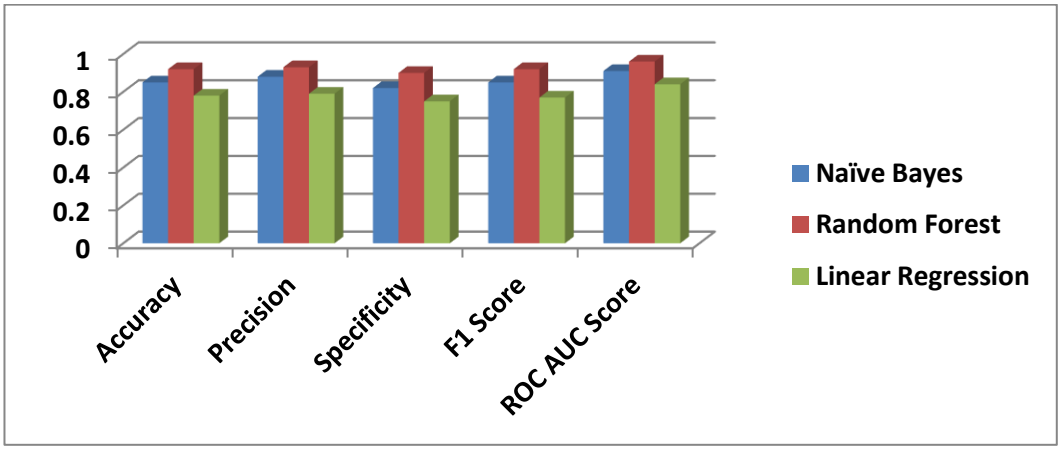

Fig 3: Representation performance metrics using different ML method for dataset 2

Its ROC AUC Score of 0.84 and F1 Score of 0.77 indicate a reasonable level of efficacy in threat detection. In conclusion, Random Forest and Naive Bayes both performed best overall, with Linear Regression, though competent, showing room for improvement in separating threats from non-threats in Dataset 2.

Table 4: Summary of result for Dataset 3

Algorithm	Accuracy	Precision	Specificity	F1 Score	ROC AUC Score
Naïve Bayes	0.89	0.92	0.86	0.89	0.95
Random Forest	0.96	0.97	0.94	0.96	0.97
Linear Regression	0.82	0.83	0.79	0.81	0.88

Nave Bayes demonstrated strong performance in Dataset 3 for Machine Learning-powered Threat Detection, achieving an F1 Score of 0.89, an accuracy of 89%, a reliable 92% precision for recognising threats, and a precision of 92%. However, it might improve specificity for identifying non-threats. Random Forest outperformed, achieving a high F1 Score of 0.96 and an exceptional ROC AUC Score of 0.97 thanks to its impressive 96% accuracy, strong 97% precision, and superb 94% specificity.

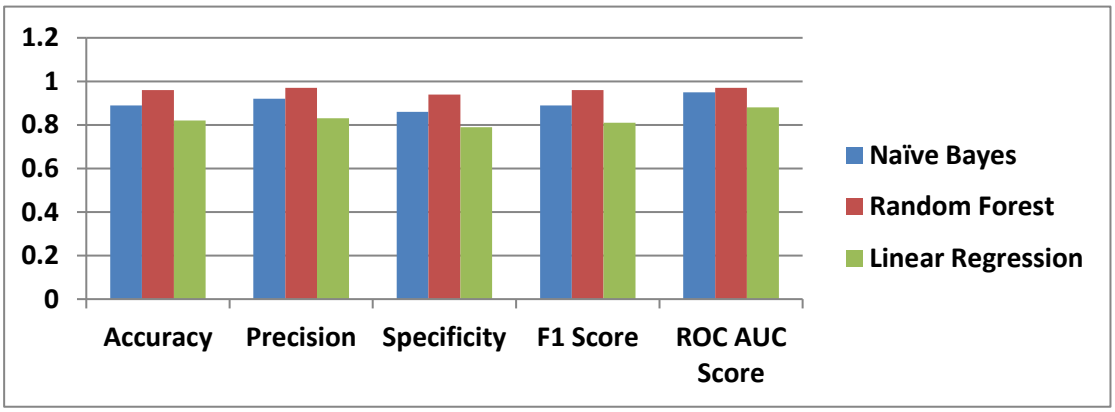

Fig 4: Representation performance metrics using different ML method for dataset 3

With an F1 Score of 0.81 and a ROC AUC Score of 0.88, linear regression demonstrated respectable performance with 82% accuracy, but poorer precision (83%) and specificity (79%). In conclusion, Random Forest and Naive Bayes both performed well in properly differentiating threats, but Linear Regression, while decent, showed room for growth.

Table 4: Summary of result for Dataset 4

Algorithm	Accuracy	Precision	Specificity	F1 Score	ROC AUC Score
Naïve Bayes	0.87	0.9	0.84	0.87	0.93

Random Forest	0.94	0.95	0.92	0.94	0.95
Linear Regression	0.8	0.81	0.77	0.79	0.86

We compared the effectiveness of Nave Bayes, Random Forest, and Linear Regression in Dataset 4 for Machine Learning-powered Threat Detection.

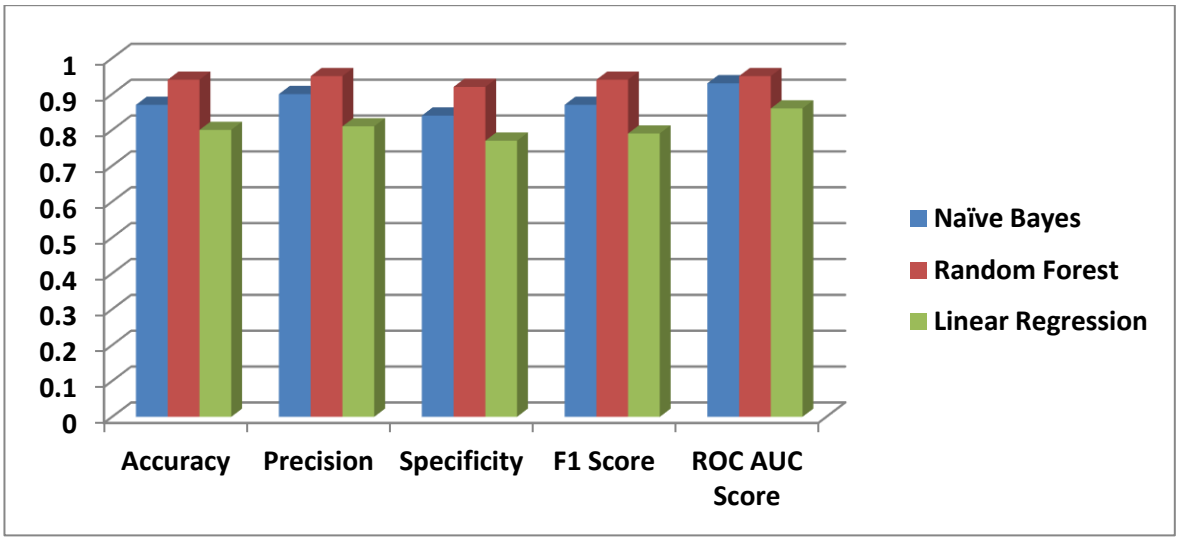

Fig 5: Representation performance metrics using different ML method for dataset 4

With an87% accuracy, a strong 90% precision, and an F1 Score of 0.87, Nave Bayes demonstrated decent performance, demonstrating its capacity to accurately identify dangers. Although it achieved 84% in specificity, it could have done better.

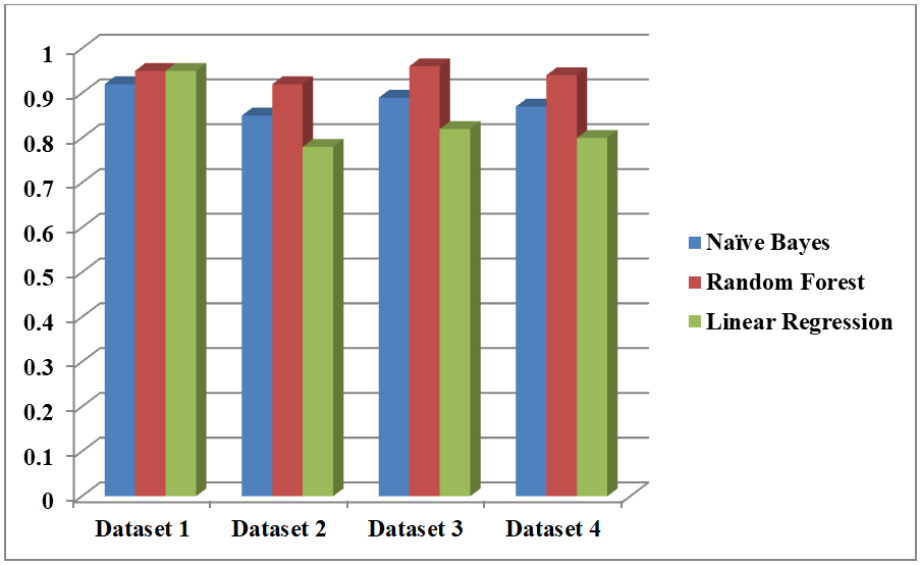

Fig 6: Accuracy comparison for different methods on all datasets

With a stellar accuracy of 94%, remarkable precision of 95%, and strong specificity of 92%, Random Forest outperformed the competition, earning an F1 Score of 0.94 and a commendable ROC AUC Score of 0.95. Even though it had an accuracy of 80%, linear regression showed lesser precision (81%) and specificity (77%), resulting in an F1 Score of 0.79 and a ROC AUC Score of 0.86. Overall, Linear Regression showed promise for improvement, while Random Forest and Naive Bayes had the best performance.

6. Conclusion

A key element in reducing the ever-increasing cybersecurity difficulties that our connected world faces today is machine learning-powered threat detection. The quick development of information technology has resulted in an alarming growth in numerous cyber dangers, ranging from malware, zero-day exploits, data breaches, and sophisticated social engineering methods to unauthorised access and denial of service attacks. The exponential growth in the variety of unique malware executable serves as an example of this escalation and emphasises the urgent need to strengthen our cybersecurity defences.Cybercrime has major financial repercussions, with data breaches alone costing millions of dollars and having a significant effect on the world economy. Furthermore, effective cybersecurity measures that cover both public and private sector networks are now essential for maintaining national security. As a result, it is crucial to be able to recognise and react to cyber threats intelligently.Machine learning has become a game-changing force in cybersecurity thanks to its ability to examine enormous datasets and identify complex patterns. It gives companies and organisations the tools to recognise both known and undiscovered cyberthreats, effectively securing vital systems. Machine learning models continuously adapt and improve using cutting-edge algorithms and approaches, keeping up with the changing threat landscape.Fundamentally, Machine Learning-powered Threat Detection provides a promising future in our on-going conflict with cyber threats. It's important to recognise that the cybersecurity landscape is always changing and necessitates constant innovation and agility. Our defences can be greatly improved by incorporating machine learning into our cybersecurity plans, as well as interdisciplinary cooperation and the integration of threat intelligence. By doing this, we can prevent cyber risks from happening in the first place, protect important assets, and make sure that our increasingly interconnected digital world is secure and resilient.

References

[1] P. Ambika, “Machine learning and deep learning algorithms on the Industrial Internet of Things (IIoT),” Advances in Computers, vol. 117, no. 1, pp. 321–338, 2020.

[2] R. Ashima, A. Haleem, S. Bahl, M. Javaid, S. K. Mahla, and S. Singh, “Automation and manufacturing of smart materials in Additive Manufacturing technologies using the Internet of Things towards the adoption of Industry 4.0,” Materials Today: Proceedings, vol. 45, pp. 5081–5088, 2021.

[3] L. M. Gladence, V. M. Anu, R. Rathna, and E. Brumancia, “Recommender system for home automation using IoT and artificial intelligence,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–9, 2020.

[4] T. Sherasiya, H. Upadhyay, and H. B. Patel, “A survey: intrusion detection system for internet of things,” International Journal of Computer Science and Engineering (IJCSE), vol. 5, no. 2, pp. 91–98, 2016.

[5] J. B. Awotunde, R. G. Jimoh, S. O. Folorunso, E. A. Adeniyi, K. M. Abiodun, and O. O. Banjo, “Privacy and security concerns in IoT-based healthcare systems,” Internet of Things, pp. 105–134, 2021.

[6] E. A. Adeniyi, R. O. Ogundokun, and J. B. Awotunde, “IoMT-based wearable body sensors network healthcare monitoring system,” in IoT in Healthcare and Ambient Assisted Living, pp. 103–121, Springer, Singapore, 2021.

[7] K. Amit and C. Chinmay, “Artificial intelligence and Internet of Things based healthcare 4.0 monitoring system,” Wireless Personal Communications, pp. 1–14, 2021.

[8] F. E. Ayo, S. O. Folorunso, A. A. Abayomi-Alli, A. O. Adekunle, and J. B. Awotunde, “Network intrusion detection based on deep learning model optimized with rule-based hybrid feature selection,” Information Security Journal: A Global Perspective, vol. 29, no. 6, pp. 267–283, 2020.

[9] S. N. Ajani and S. Y. Amdani, "Probabilistic path planning using current obstacle position in static environment," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170727.

[10] M. Abdulraheem, J. B. Awotunde, R. G. Jimoh, and I. D. Oladipo, “An efficient lightweight cryptographic algorithm for IoT security,” in Communications in Computer and Information Science, pp. 444–456, Springer, 2021.

[11] S. Ajani and M. Wanjari, "An Efficient Approach for Clustering Uncertain Data Mining Based on Hash Indexing and Voronoi Clustering," 2013 5th International Conference and Computational Intelligence and Communication Networks, 2013, pp. 486-490, doi: 10.1109/CICN.2013.106.

[12] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. . (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal

of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.

[13] Potnurwar, A. V. ., Bongirwar, V. K. ., Ajani, S. ., Shelke, N. ., Dhone, M. ., & Parati, N. . (2023). Deep Learning-Based Rule-Based Feature Selection for Intrusion Detection in Industrial Internet of Things Networks. International Journal of Intelligent Systems and Applications in Engineering, 11(10s), 23–35.

[14] A. Bakhtawar, R. J. Abdul, C. Chinmay, N. Jamel, R. Saira, and R. Muhammad, "Blockchain and ANFIS empowered IoMT application for privacy preserved contact tracing in COVID-19 pandemic," Personal and Ubiquitous Computing, 2021.

[15] A. H. Muna, N. Moustafa, and E. Sitnikova, "Identification of malicious activities in industrial internet of things based on deep learning models," Journal of information security and applications, vol. 41, pp. 1–11, 2018.

[16] E. Sitnikova, E. Foo, and R. B. Vaughn, "The power of hands-on exercises in SCADA cybersecurity education," in Information Assurance and Security Education and Training, pp. 83–94, Springer, Berlin, Heidelberg, 2013.

[17] S. Dash, C. Chakraborty, S. K. Giri, S. K. Pani, and J. Frnda, "BIFM: big-data driven intelligent forecasting model for COVID-19," IEEE Access, vol. 9, pp. 97505–97517, 2021.

[18] G. Tzokatziou, L. A. Maglaras, H. Janicke, and Y. He, "Exploiting SCADA vulnerabilities using a human interface device," International Journal of Advanced Computer Science and Applications, vol. 6, no. 7, pp. 234–241, 2015.

[19] D. Kushner, "The real story of stuxnet," IEEE Spectrum, vol. 50, no. 3, pp. 48–53, 2013.

[20] P. W. Khan and Y. Byun, "A blockchain-based secure image encryption scheme for the industrial Internet of Things," Entropy, vol. 22, no. 2, p. 175, 2020.

[21] Q. Yan and F. R. Yu, "Distributed denial of service attacks in software-defined networking with cloud computing," IEEE Communications Magazine, vol. 53, no. 4, pp. 52–59, 2015.

[22] A. C. Enache and V. Sgârciu, "Anomaly intrusions detection based on support vector machines with an improved bat algorithm," in 2015 20th International Conference on Control Systems and Computer Science, pp. 317–321, Bucharest, Romania, May 2015.

[23] O. Folorunso, F. E. Ayo, and Y. E. Babalola, "Ca-NIDS: a network intrusion detection system using combinatorial algorithm approach," Journal of Information Privacy and Security, vol. 12, no. 4, pp. 181–196, 2016.

[24] H. Zhang, D. D. Yao, N. Ramakrishnan, and Z. Zhang, "Causality reasoning about network events for detecting stealthy malware activities," Computers & Security, vol. 58, pp. 180–198, 2016.

[25] M. R. Kabir, A. R. Onik, and T. Samad, "A network intrusion detection framework based on Bayesian network using a wrapper approach," International Journal of Computer Applications, vol. 166, no. 4, pp. 13–17, 2017.

[26] Y. Hu, A. Yang, H. Li, Y. Sun, and L. Sun, "A survey of intrusion detection on industrial control systems," International Journal of Distributed Sensor Networks, vol. 14, no. 8, 2018.

[27] T. Cruz, L. Rosa, J. Proenca et al., "A cybersecurity detection framework for supervisory control and data acquisition systems," IEEE Transactions on Industrial Informatics, vol. 12, no. 6, pp. 2236–2246, 2016.

[28] J. Camacho, A. Pérez-Villegas, P. García-Teodoro, and G. Maciá-Fernández, "PCA-based multivariate statistical network monitoring for anomaly detection," Computers & Security, vol. 59, pp. 118–137, 2016.

[29] M. Grill, T. Pevný, and M. Rehak, "Reducing false positives of network anomaly detection by local adaptive multivariate smoothing," Journal of Computer and System Sciences, vol. 83, no. 1, pp. 43–57, 2017.

[30] L. A. Maglaras, J. Jiang, and T. J. Cruz, "Combining ensemble methods and social network metrics for improving accuracy of OCSVM on intrusion detection in SCADA systems," Journal of Information Security and Applications, vol. 30, pp. 15–26, 2016.

[31] R. O. Ogundokun, J. B. Awotunde, E. A. Adeniyi, and F. E. Ayo, "Crypto-Stegno based model for securing medical information on IOMT platform," Multimedia tools and applications, pp. 1–23, 2021.

[32] J. Soto and M. Nogueira, "A framework for resilient and secure spectrum sensing on cognitive radio networks," Computer Networks, vol. 115, pp. 130–138, 2017.

[33] M. S. Abadeh, J. Habibi, and C. Lucas, “Intrusion detection using a fuzzy genetics-based learning algorithm,” Journal of Network and Computer Applications, vol. 30, no. 1, pp. 414–428, 2007.

[34] NSL-KDD|Datasets|Research|Canadian Institute for Cybersecurity|UNB. Available online: https://www.unb.ca/cic/datasets/nsl.html

[35] 1998 DARPA Intrusion Detection Evaluation Dataset|MIT Lincoln Laboratory. Available online: https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset

[36] The UNSW-NB15 Dataset|UNSW Research. Available online: https://research.unsw.edu.au/projects/unsw-nb15-dataset

[37] ADFA IDS Datasets|UNSW Research. Available online: https://research.unsw.edu.au/projects/adfa-ids-datasets