

AUTOMATIC METHOD TO CLASSIFY CYBER CRIME INCIDENT USING ARTIFICIAL INTELLIGENCE AND DEEP LEARNING APPROACHES

¹Singalla Parameswararao, ²Mrs.B.Bhagya Lakshmi

¹Student, Dept. of Master of Computer Applications, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

²Asst.Prof, Dept. of Computer Science & Engineering, Amrita Sai Institute of Science and Technology, Paritala, Andhra Pradesh, 521180, India.

ABSTRACT

Understanding the landscape The analysis of cyber incident data is vital to understanding the changing threat landscape. In times of continuing cyber-ruthlessness and numerous hacking events, the proposed work aims to explore the versatility of cyber-attacks and to devise counter weapons. Rather than assume these attacks are placed at random, we suggest that it is reasonable to model the interarrival times between hacking attacks and the number of accounts breached using stochastic processes that reflect the auto-correlation properties of these events. In order to handle the complexity of the data, we utilize deep learning algorithms, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). These approaches support solid analysis, unlocking latent patterns and trends in cyber events. By using these insights, the project will help us improve our understanding of cybersecurity while building a foundation for a set of proactive responses to developing threats.

Keywords: *Cybercrime Classification, Artificial Intelligence, Deep Learning, Incident Detection, Cybersecurity, Neural Networks, Threat Vector Analysis, Anomaly Detection, Cyber Incident Response, Intrusion Detection System, Training and Testing Dataset.*

1. INTRODUCTION

1.1 Introduction of Project

An information breach occurs when sensitive or classified data is accessed, transmitted, or utilized by an unauthorized individual. This includes incidents such as theft or loss of digital media, posting information online without proper security measures, or transferring data to unauthorized systems or entities. Despite technological advancements in cybersecurity, data breaches remain a significant challenge. Understanding the evolution of these incidents is crucial for mitigating their impact, with avenues such as insurance playing a potential role in damage control. (Ali et al., 2021; PARAMESWARARAO et al., 2021)However, accurately assessing cyber risk for insurance purposes is currently limited by the lack of precise data breach metrics. In our study, we demonstrate the importance of modeling data breach incidents using stochastic processes rather than traditional distributions. By doing so, we show that these models can effectively predict both the timing and size of breaches. Our research highlights the presence of a dependency between breach occurrence and size, which is crucial for accurate prediction. Neglecting this dependency can lead to inaccurate results. We believe that our findings will encourage further research into alternative risk mitigation strategies, providing valuable insights for insurance companies, government agencies, and regulators. (Valdez, 2014)Deep understanding of data breach risks is essential for developing effective strategies to safeguard sensitive information in today's digital landscape.(Rangel, 2019)

1.2 INTRODUCTION OF DOMAIN

Deep Learning represents a transformative application of artificial intelligence that has revolutionized statistical data processing and analysis . While emerging from the broader field of Machine Learning, which requires human guidance and emphasizes generalization for handling unseen data, Deep Learning has established itself as a distinct discipline within Computer Science that offers innovative approaches to data analysis .(Sinha, Pandey and Pattnaik, 2018; Alom et al., 2019; Mishra, Reddy and Pathak, 2021)

The fundamental distinction between traditional statistical methods and Deep Learning lies in their underlying approaches to model construction. Statistical techniques typically rely on probabilistic models to describe observed data within a defined class of related models. In contrast, Deep Learning techniques, while solving similar optimization problems, transcend these constraints by not being restricted to probabilistic frameworks. This flexibility represents a significant advantage, particularly in the era of Big Data, where data sources are increasingly complex and multifaceted. (Najafabadi et al., 2015; PARAMESWARARAO et al., 2021)

Deep Learning techniques can be categorized into two primary classes: supervised and unsupervised learning. Supervised learning, exemplified by both traditional statistical methods like logistic regression and modern approaches like Support Vector Machines (SVM), operates on labeled training data to predict outcomes or classify new instances. While logistic regression employs probability theory to optimize coefficients through maximum likelihood estimation, SVM utilizes geometric principles to identify optimal hyperplanes for classification in the predictor space. (Yu and Kim, 2012; Kubát, 2015; Yang, Molefyane and Lin, 2023)

Unsupervised learning, on the other hand, focuses on discovering inherent patterns within data without predetermined response variables. Traditional statistical methods like Principal Component Analysis (PCA) seek to summarize high-dimensional data through orthogonal subspaces, while machine learning techniques such as cluster analysis aim to uncover natural grouping structures within datasets.

The emergence of Deep Learning offers statistical agencies and researchers a broader spectrum of flexible analytical methods better suited to modern data sources. As data complexity continues to increase, the ability to process and analyze information without strict probabilistic model constraints becomes increasingly valuable. This paradigm shift necessitates a careful evaluation of traditional versus modern approaches to determine the most effective methodologies for future analytical needs. (Nowell, 2011; Rosendo et al., 2022)

This evolution in data analysis techniques represents a crucial juncture in the field of statistics and machine learning, where the integration of Deep Learning approaches with traditional statistical methods may offer the most comprehensive solutions for contemporary data challenges. (Panch, Szolovits and Atun, 2018; Fan et al., 2021; Bhadra et al., 2024)

2. LITERATURE REVIEW

2.1 Detection of Port Scan Attempt Using Comparative Study of Deep Learning and Support Vector Machine Techniques(Sowmya and Anita, 2023)

Recent advances in computer and communication technologies have brought tremendous opportunities and new threats such as cyber terrorism. In order to mitigate the threats posed by this problem, intrusion detection systems (IDS) play a vital role. We assess deep learning and support vector machine (SVM) for port scan attempt detection on the CICIDS2017 dataset. Results show that deep learning can achieve an accuracy of 97.80%, which is much higher than that obtained by SVM (69.79%).(Sowmya and Anita, 2023)

2.2 Monitoring Cyber-Attacks with aCRPS-Based Approach144(Hesham et al., 2024)

Cyber-attacking continues to be a big threat to network security, which calls for efficient anomaly detection methods. In this paper, two statistical methods are proposed for detecting TCP SYN flood attacks based on the continuous ranked probability score (CRPS). The CRPS-Shewhart and CRPS-EWMA methods are obtained by combining CRPS with Shewhart and exponentially weighted moving average (EWMA) control charts. These methods have been tested against the 1999 DARPA intrusion detection datasets and shown to be effective in detecting anomalies.(Hesham et al., 2024)

2.3 A classification of malicious traffic for intrusion detection systems(Wang et al., 2024)

As network attacks become more and more common and the key to accurate intrusion detector system is to have systematical knowledge of attack type. In this paper, a detailed characterization of network attacks is proposed to facilitate the creation of more specific IDS and focused databases.(Wang et al., 2024)

2.4 Simplistic Way to Keep the Cyber Criminals at Bay and to Detect Cyber Crimes(Momand, Jan and Ramzan, 2023)

Cybercrime has become increasingly common with the rapid penetration of the internet in the daily activities of people, and these activities are generally challenging to detect and to identify blame. This paper presents a framework which combines automatic and

manual detection methodologies to trace cybercrimes and produce the evidence to convict the converts for the societal menace cause by these crimes.(Momand, Jan and Ramzan, 2023)

3. PROJECT DESCRIPTION

3.1 Problem Definition

Clearly states the problem the project aims to solve.

There are several challenges in the development of intelligent systems of the near future:

- ⇒ Performance and scalability issues caused by the system.
- ⇒ Abnormal system messages may be generated accidentally that leaks scanning behavior.
- ⇒ A lack of proper security exposes the systems to risk.
- ⇒ Time consumption is high for detection and response processes.
- ⇒ Gullibility of users and how common man is gullible.

3.2 Objective of the Project

Cybercrime has grown to be a major and changing concern given the Internet's explosive development. Cybercrime is illicit activity carried out via computers or networks, either with an eye towards the network itself—such as hacking and system destruction—or as a tool for crimes such fraud and identity theft. Cybercrimes span widely from virus attacks to financial fraud, internet frauds, illegal access, and cyberstalking. Defining cybercrime is difficult given its ever shifting character and the variety of offences involved. Though only a small percentage of transactions take place online, phishing emails alone create billions in illegal earnings and online credit card theft accounts for half of the \$4 billion lost yearly. According to polls, most businesses have had security breaches—often resulting in financial losses. These numbers underline how urgently Internet crime has to be addressed and decreased.

3.3 Scope of the Project

The fast speed and amount of internet activity in the linked world of today make cyberspace quite vulnerable to attacks. Automatic, intelligent solutions are required for real-time threat identification and response since conventional security approaches are inadequate. This project is to investigate how artificial intelligence, especially bio-inspired computing, could solve problems related to cybercrime.

The main project scopes are:

1. Analyse how precisely artificial intelligence forecasts cybercrime events.
2. Decide and examine the causes of cybercrimes.
3. Make sure the created techniques apply on several datasets.

4. COMPUTATIONAL ENVIRONMENT

4.1 Software Specification

Details software tools, frameworks, and languages used (e.g., Python, Java, MySQL).

Component	Details
Operating System	Windows 10
Coding Language	Python
Technology	Internet of Things (IoT)

Table 1: Software Specifications

5. FEASIBILITY STUDY

A feasibility study assesses a project's viability by calculating costs, benefits, and resource requirements, because you don't want to end up with a burden. It evaluates stakeholder requirements via techniques such as questionnaires, users' observations and document analysis. The analysis evaluates alternative options in four aspects: economic (cost and benefit), technical (available resource and technology), social (impact on stakeholders), and operational (implementation and its difficulty). The aim here is to find the best match for the institution."

6. SYSTEM ANALYSIS

6.1 Existing System

Designed for processing grid-like data—especially photos and videos—a convolutional neural network (CNN) is a specialised deep learning model.(Redhu et al., 2024) CNNs, unlike conventional neural networks, find hierarchical patterns using shared-weight topologies and convolutional layers, therefore lowering pre-processing requirements and increasing efficiency. Motivated by the human visual cortex, they shine in tasks including segmentation, object detection, and image categorisation.(Guerra et al., 2024)

Important characteristics include:

1. Translation invariance—that which recognises patterns independent of position.
2. Local receptive fields (with an eye towards discrete input data zones).
3. Sharing parameters helps to lower overfitting and computational expense.

Because CNNs learn spatial hierarchies, they are being applied in financial analysis, NLP, and medical imaging. Their effectiveness in feature extraction distinguishes them over hand-engineered techniques.

6.1.1 Existing System Disadvantages

Highlights limitations or issues with the existing system.

1. Less precision in forecasting the cases of cyber hacking.
2. Not user-friendly paradigm.
3. More labour-intensive procedure
4. More Computational Cost
5. Insufficient criteria
6. Not applicable for every dataset

6.2 Proposed System

Especially appropriate for processing sequential input, including natural language, where the number of steps, words, or items is not fixed, a recurrent neural network (RNN) is a sort of artificial neural network whereby connections between units form a directed cycle.(Ghojogh and Ghodsi, 2023) RNNs differ from feedforward networks in that they include past inputs in current computations, therefore capturing temporal influences and carrying a hidden state that functions as memory.(Das et al., 2023)

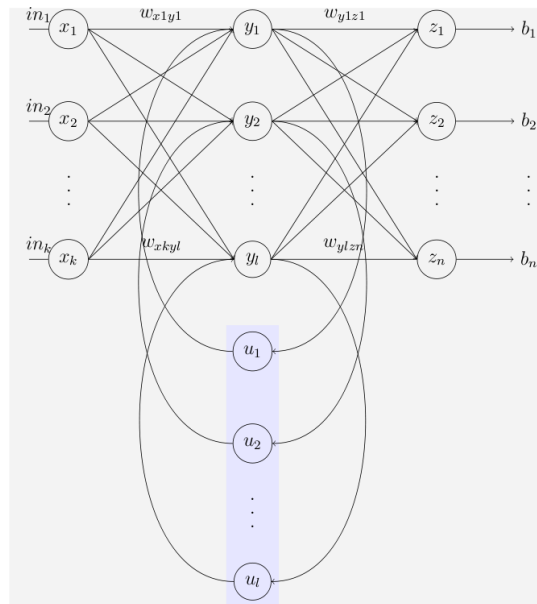


Fig 1: Neural network with layers x, y, z, b , and recurrent u , with weights w

Because RNNs can handle input sequences with variable-length, they are widely applied in handwriting recognition, machine translation, and speech recognition. Standard RNNs regretfully cannot learn long-range dependencies since they are impacted by the vanishing gradient problem. Developed to address this issue were more complex models such GRUs (Gated Recurrent Units) and LSTMs (Long Short-Term Memory). (Redhu et al., 2024)

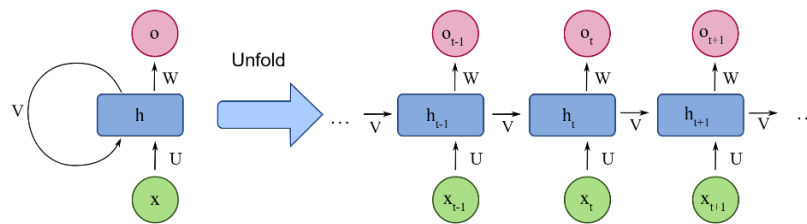


Fig 2: RNN Unrolling: Depicts inputs (x), hidden states (h), and outputs (o) over time.

Essential traits:

- ⇒ Time-dependent procedures (perform as information is handled).
- ⇒ Memory—hidden state stores past data.
- ⇒ arbitrary sequential job input and output durations.

6.2.1 Proposed System Advantages

- Elevated production efficiency.
- Intuitive user interface.
- Reduced time expenditure.
- Applicable to all datasets.
- The early forecast of criminal activities is feasible.

7. SYSTEM DESIGN

7.1 System Architecture

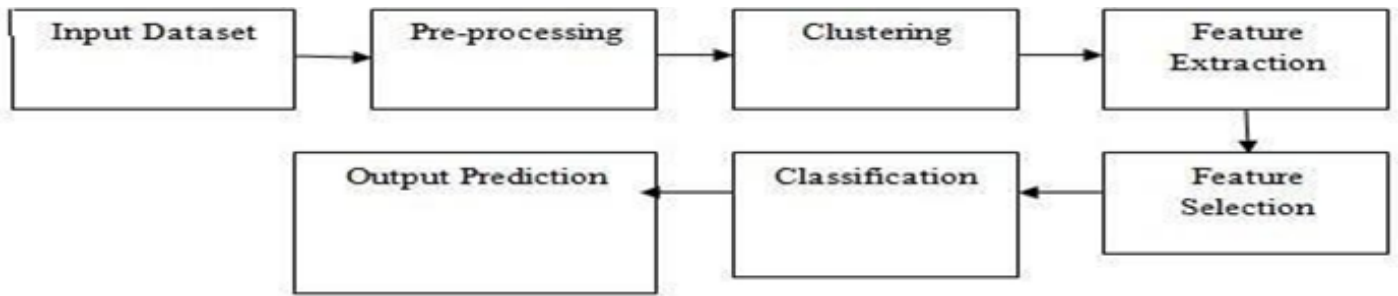


Fig 3: The System Architecture

7.1.1 Class Diagram

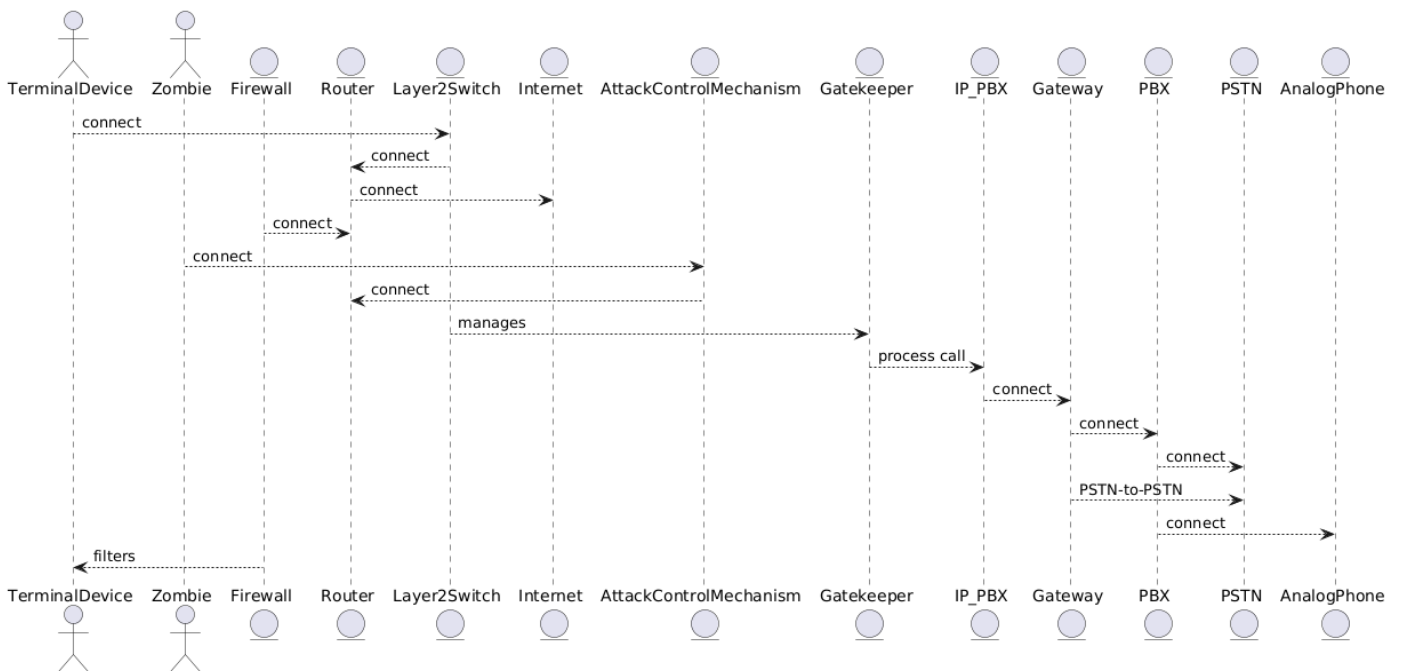


Fig 4: Class-Diagram

7.1.2 Component Diagram

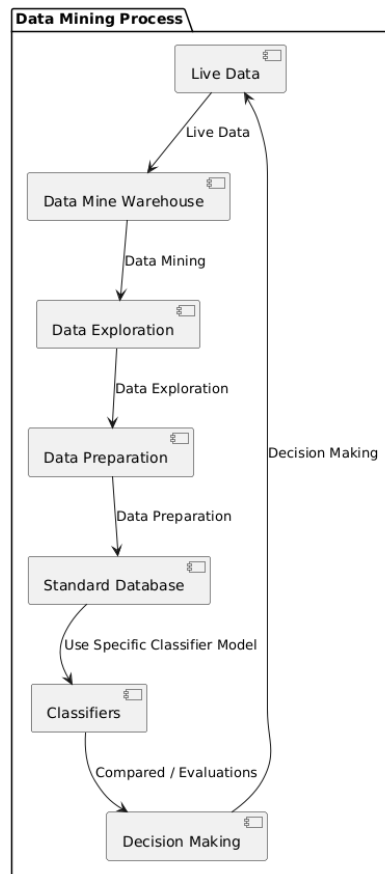


Fig 5: Component-Diagram

8. SYSTEM IMPLEMENTATION

8.1 Module Description

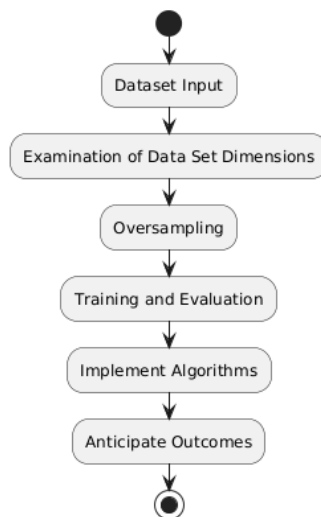


Fig 6: Module:Description-Flowchart

8.2 Comparative Study

Our suggested system employs Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for analysis, resulting in enhanced accuracy and reduced noise relative to current systems. This illustrates that our system surpasses existing methods.

8.3 Flowchart

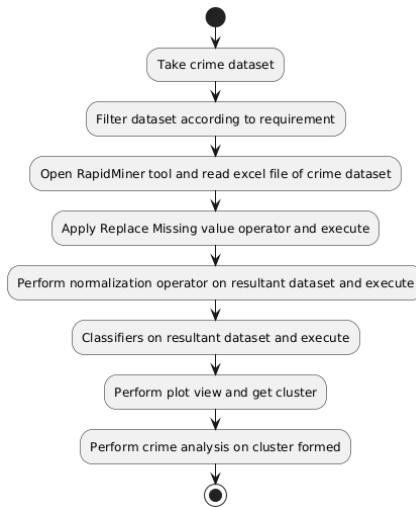


Fig 7: Comparative:Study-Flowchart

9. TESTING

9.1 Functional Testing

Item	Description
Valid Input	Identified valid input classes must be accepted.
Invalid Input	Identified invalid input classes must be rejected.
Functions	Identified functions must be exercised.
Output	Identified output classes must be exercised.
Systems/Procedures	Interfacing systems or procedures must be invoked.

Table 2: Functional-Testing

11. RESULTS

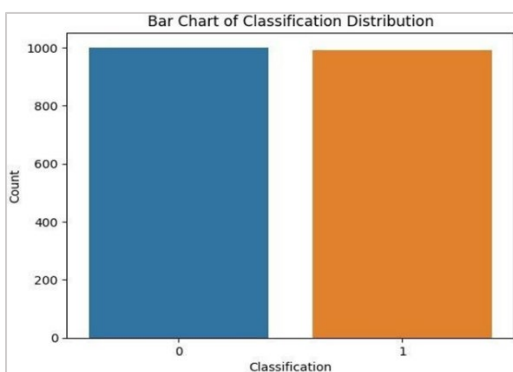


Fig 8: Bar-chart: Classification Distribution

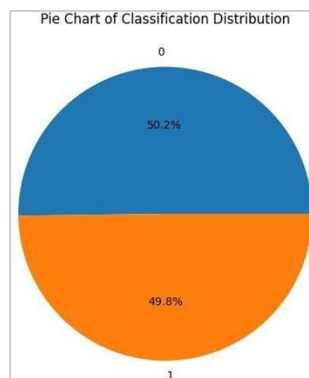


Fig 9: Pie-chart: Classification Distribution

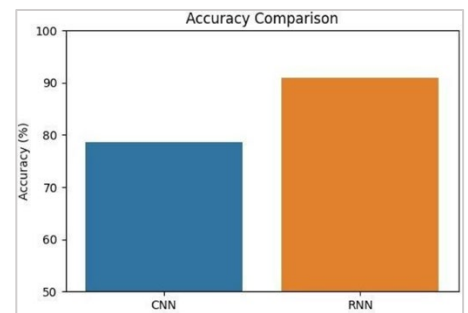


Fig 10: Accuracy: Comparission

12. CONCLUSION

The increase in regular data leaks around the world emphasises the increasing threat to vital infrastructure. Vulnerabilities rise as important information systems grow and as hackers get more advanced. Though such a road calls for careful thought of implications, these attacks could be regarded as acts of terrorism, necessitating action under the Internal Security Act. The secret is to fortify security by means of a combination of technology, qualified personnel, prudence, and sound legal systems. Future initiatives should

concentrate on evaluating fresh hazards to important infrastructure and guaranteeing appropriate legal and policy protections for any area judged vital.

ACKNOWLEDGMENT

I am thankful to the Management of Amrita Sai Institute of Science and Technology for giving me an opportunity to work with his project.

I would like to thank **Dr. M. Sasidhar**, Principal, Amrita Sai institute of science and technology, for his constant encouragement and support during the progress of this work.

I am deeply grateful to **Dr. P. Chiranjeevi**, Professor and Head of the Department, for his valuable guidance and consistent support during the course of the project.

A special note of thanks to my internal guide, **Mrs. B.Bhagya Lakshmi (M.Tech)**, for her exceptional guidance, constant motivation, and continuous encouragement, which played a crucial role in the successful completion of this project.

SINGALLA PARAMESWARARAO

REFERENCES

1. Ali, R.F. et al. (2021) "Information Security Behavior and Information Security Policy Compliance: A Systematic Literature Review for Identifying the Transformation Process from Noncompliance to Compliance," *Applied Sciences*, 11(8), p. 3383. doi:10.3390/app11083383.
2. Alom, M.Z. et al. (2019) "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics*, 8(3), p. 292. doi:10.3390/electronics8030292.
3. Bhadra, A. et al. (2024) "Merging two cultures: Deep and statistical learning," *Wiley Interdisciplinary Reviews Computational Statistics*. Wiley. doi:10.1002/wics.1647.
4. Das, S. et al. (2023) "Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research," in *Neuromethods*. Humana Press, p. 117. doi:10.1007/978-1-0716-3195-9_4.
5. Fan, J. et al. (2021) "Modern Data Modeling: Cross-Fertilization of the Two Cultures," *Observational Studies*, 7(1), p. 65. doi:10.1353/obs.2021.0023.
6. Ghogh, B. and Ghodsi, A. (2023) "Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and Survey," *arXiv (Cornell University) [Preprint]*. doi:10.48550/arxiv.2304.11461.
7. Guerri, M.F. et al. (2024) "Deep learning techniques for hyperspectral image analysis in agriculture: A review," *ISPRS Open Journal of Photogrammetry and Remote Sensing*. Elsevier BV, p. 100062. doi:10.1016/j.ophoto.2024.100062.
8. Hesham, M.E. et al. (2024) "Evaluating Predictive Models in Cybersecurity: A Comparative Analysis of Machine and Deep Learning Techniques for Threat Detection," 45, p. 33. doi:10.1109/imsa61967.2024.10652833.
9. Kubát, M. (2015) *An Introduction to Machine Learning*, Springer eBooks. Springer Nature. doi:10.1007/978-3-319-20010-1.
10. Mishra, R., Reddy, G.Y.S. and Pathak, H. (2021) "The Understanding of Deep Learning: A Comprehensive Review," *Mathematical Problems in Engineering*. Hindawi Publishing Corporation, p. 1. doi:10.1155/2021/5548884.

11. Momand, A., Jan, S.U. and Ramzan, N. (2023) “A Systematic and Comprehensive Survey of Recent Advances in Intrusion Detection Systems Using Machine Learning: Deep Learning, Datasets, and Attack Taxonomy,” *Journal of Sensors*, 2023, p. 1. doi:10.1155/2023/6048087.
12. Najafabadi, M.M. et al. (2015) “Deep learning applications and challenges in big data analytics,” *Springer Science+Business Media*, 2. doi:10.1186/s40537-014-0007-7.
13. Nowell, L. (2011) “The science of visual analysis at extreme scale,” *Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE*, 7868, p. 786802. doi:10.1117/12.881434.
14. Panch, T., Szolovits, P. and Atun, R. (2018) “Artificial intelligence, machine learning and health systems,” *Journal of Global Health*, 8(2). doi:10.7189/jogh.08.020303.
15. PARAMESWARARAO, S. et al. (2021) “AUTOMATIC METHOD TO CLASSIFY CYBER CRIME INCIDENT USING ARTIFICIAL INTELLIGENCE AND DEEP LEARNING APPROACHES.”
16. Rangel, A. (2019) “Why enterprises need to adopt ‘need-to-know’ security,” *Computer Fraud & Security*, 2019(12), p. 9. doi:10.1016/s1361-3723(19)30127-7.
17. Redhu, A. et al. (2024) “Deep learning-powered malware detection in cyberspace: a contemporary review,” *Frontiers in Physics*. *Frontiers Media*. doi:10.3389/fphy.2024.1349463.
18. Rosendo, D. et al. (2022) “Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review,” *Journal of Parallel and Distributed Computing*, 166, p. 71. doi:10.1016/j.jpdc.2022.04.004.
19. Sinha, R.K., Pandey, R. and Pattnaik, R. (2018) “Deep Learning For Computer Vision Tasks: A review,” *arXiv (Cornell University)*. *Cornell University*. doi:10.48550/arXiv.1804.03928.
20. Sowmya, T. and Anita, E.A.M. (2023) “A comprehensive review of AI based intrusion detection system,” *Measurement Sensors*. *Elsevier BV*, p. 100827. doi:10.1016/j.measen.2023.100827.
21. Valdez, E.A. (2014) “Empirical investigation of insurance claim dependencies using mixture models,” *European Actuarial Journal*, 4(1), p. 155. doi:10.1007/s13385-014-0088-x.
22. Wang, L. et al. (2024) “Incorporating Gradients to Rules: Towards Lightweight, Adaptive
23. Provenance-based Intrusion Detection,” *arXiv (Cornell University)* [Preprint]. doi:10.48550/arXiv.2404.14720.
24. Yang, C., Molefyane, T. and Lin, Y. (2023) “The Forecasting of a Leading Country’s Government Expenditure Using a Recurrent Neural Network with a Gated Recurrent Unit,” *Mathematics*, 11(14), p. 3085. doi:10.3390/math11143085.
25. Yu, H. and Kim, S. (2012) “SVM Tutorial — Classification, Regression and Ranking,” in *Springer eBooks*. *Springer Nature*, p. 479. doi:10.1007/978-3-540-92910-9_15.