IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Forecasting Cyber Storms: A Comparative Analysis Of Random Forests And Deep Neural Networks For Predicting Cyber Attacks

**S Tamil Selvi[1], Anisha Thangakani S***

Assistant Professor, Department of computer Science, Vidhya sagar women's college

**Abstract:** In recent years, the exponential growth of data has revolutionized the landscape of information security, necessitating advanced data analysis systems tailored for big data. Traditional techniques struggle to cope with the sheer volume, variety, and velocity of data generated across networks, making cyber attack detection increasingly challenging. Intrusion Detection Systems (IDS) leveraging big data technology offer a promising avenue for accurate and efficient analysis.

This paper presents a novel approach utilizing Spark Chi SVM model for cyber attack detection. Leveraging ChiSqSelector for feature selection and employing a Support Vector Machine (SVM) classifier on the Apache Spark big data platform, the model demonstrates high performance, reduced training time, and efficiency for processing big data. Through experimentation with the KDD99[10] dataset, a comparison between Chi SVM and Chi logistic regression classifiers underscores the superiority of the SVM model.

Additionally, the paper introduces a cutting-edge technology incorporating deep learning models for constructing a balanced representation of imbalanced raw datasets. By leveraging deep neural networks (DNN)[7] and decision tree (DT)[8] classifiers, this approach significantly enhances network attack detection. Validation using real-world critical infrastructure datasets showcases a remarkable improvement, with a 10% higher F1 score compared to conventional classifiers like Random Forests (RF)[6] and standard DNNs[7]. Achieving precision rates of 95.86% and 99.67% on datasets from gas pipeline and safe water treatment facilities respectively, this research paves the way for more accurate and effective cybersecurity measures.

**Keywords:** Skin lesion, HOG, GLCM, CNN, classification.

## I. Introduction

Critical infrastructure systems are vital components of modern societies, encompassing a blend of cyber and physical elements to facilitate their operations efficiently. Among these systems, industrial plants serve as the linchpin, driving essential functions across sectors like smart grids, oil and gas, aerospace, and transportation. Integral to the operation of these plants are Industrial Control Systems (ICS), which assume responsibility for monitoring and controlling various processes. However, the integration of Internet of Things (IoT) technology into ICS infrastructure introduces a new dimension of vulnerability, potentially opening avenues for cybercriminals to exploit system weaknesses and orchestrate malicious cyberattacks. The inception of Stuxnet marked a turning point in the realm of cybersecurity, representing the first documented instance of a cyberattack specifically targeting industrial control systems. Since then, awareness surrounding the vulnerabilities inherent in ICS has grown significantly, prompting ongoing efforts to bolster cybersecurity measures and mitigate potential risks. Despite these endeavors, challenges persist in detecting and thwarting cyberattacks within ICS environments. A common approach to addressing cybersecurity threats in ICS involves leveraging feature engineering techniques. However, these methods are inherently complex and demand advanced learning capabilities, posing significant hurdles for implementation and efficacy. Moreover, reliance on feature engineering may not adequately account for the evolving nature of cyber threats, necessitating a more dynamic and adaptable approach to detection and response. Moving forward, the development of more sophisticated and efficient methods for detecting cyberattacks in ICS is imperative. This entails exploring innovative approaches that transcend the limitations of traditional feature engineering and embrace emerging technologies such as machine learning and artificial intelligence. By harnessing the power of these advanced techniques, researchers and practitioners can enhance the resilience of critical infrastructure systems against cyber threats, ensuring the continued functionality and security of essential operations in an increasingly interconnected world.

## II. Related work

Hadis et al. and Venkata et al. [1] address the increasing need for accurate and efficient real-time state estimation in complex power systems by developing a massively parallel dynamic state estimator using Graphics Processing Units (GPUs). Their approach, which employs a lateral two-level Extended Kalman Filter (EKF) method, integrates data from Supervisory Control and Data Acquisition (SCADA) systems and Phasor Measurement Units (PMUs), with predictions for buses lacking PMUs based on historical data. Compared to a multithreaded CPU-based implementation, their GPU-based estimator demonstrates up to a 15-fold speed-up in processing for a 4992-bus system, highlighting its superior performance in managing large-scale power systems.

Shorfuzzaman et al. [2] addresses the growing concern of cyber attacks within Internet of Things (IoT) networks, which are increasingly vulnerable due to their widespread use and the autonomous operation of connected devices. As these attacks evolve, timely and intelligent network-based security solutions become crucial to prevent potential system failures. Although various machine learning techniques have been proposed for network intrusion detection, there has been limited focus on identifying malicious attacks specifically in IoT networks. To address this gap, Shorfuzzaman proposes an effective intrusion detection system (IDS) utilizing ensemble methods, including bagging and boosting, along with a feed-forward artificial neural network. The proposed models are evaluated using the UNSW-NB15 dataset, which contains simulated IoT sensor data, and their performance is assessed through a 5-fold cross-validation technique. The results demonstrate the models' effectiveness, highlighting their ability to detect unforeseen IoT cyberattacks with a small set of automatically selected optimal features from the dataset.

Rishabh Verma et al. and Bharti Thakur et al. [3] propose the development of a machine learning (ML) model for predicting cyberattacks using the UNSW-NB15 dataset. ML, a subset of artificial intelligence (AI), is instrumental in detecting cyberattacks by analyzing patterns in existing data to identify similarities between previous and new attacks. The paper evaluates four different classifiers: Random Forest (RF), Decision Tree (DT)[8], Logistic Regression (LR), and Artificial Neural Network (ANN). The ANN model comprises an input layer with 10 neurons, two hidden layers with 15 and 10 neurons respectively, and an output layer with 1 neuron. The study employs 9-fold cross-validation to determine the best accuracy for each classifier. The evaluation metrics include accuracy, precision, recall, F1-score, mean squared error (MSE), true positive rate (TPR), and false positive rate (FPR). Among the classifiers tested, Random Forest outperformed the others, achieving an accuracy of 95.167%, a TPR of 96.252%, and an FPR of 6.749%.

Martin Husák et al. and Jana Komárková et al. [4] present a comprehensive survey of prediction and forecasting methods in cybersecurity. They explore four key tasks: attack projection and intention recognition, which involves predicting an attacker’s next move or intentions; intrusion prediction, which focuses on forecasting imminent cyber attacks; and network security situation forecasting, which aims to project the overall cybersecurity status of a network. The survey reviews both discrete models, such as attack graphs, Bayesian networks, and Markov models, and continuous models, including time series and grey models. These methods often share theoretical foundations and can be complementary. Additionally, the paper examines machine learning and data mining techniques, which have recently gained attention for their potential in the dynamic field of cybersecurity. The survey also emphasizes the practical usability of these methods and discusses challenges related to their evaluation.

## III. EXISTING METHOD

The utilization of big data techniques in the realm of cyberattack detection has garnered significant attention from researchers in recent years. With the proliferation of complex cyber threats, traditional methods for processing vast amounts of data have become increasingly intricate. As a result, there has been a concerted effort among scholars to harness big data machine learning techniques to develop cyberattack detection systems that are both swift and precise.

In this section, we delve into the works of researchers who have leveraged big data machine learning methodologies, particularly cluster machine learning technology, to address cyberattack detection challenges. One notable approach involves the application of the k-means method, implemented within Spark's machine learning library, to discern the nature of network traffic as either indicative of an attack or normal operation. Central to this proposed method is the utilization of the KDD Cup 1999[10] dataset for both training and testing purposes. Furthermore, the authors have employed feature selection techniques to identify pertinent features for enhancing the efficacy of the detection system.

Despite the promising aspects of this approach, certain drawbacks warrant consideration. Firstly, the research endeavors to surmount several challenges inherent in existing cyberattack detection methodologies. Specifically, it seeks to address issues related to imbalanced raw datasets, a common impediment in this domain. The proposed model introduces a novel approach to creating a balanced representation of the data, thereby mitigating the effects of data imbalance and enhancing the robustness of the detection system. Additionally, the research advocates for an in-depth approach, emphasizing a thorough understanding of the underlying data and intricate nuances of cyber threats.

However, it is imperative to acknowledge some limitations of the proposed method. Notably, the use of multiple auto-encoders (AE) in the detection process, while theoretically sound, can incur significant computational overhead, rendering the approach time-consuming. This highlights a trade-off between detection accuracy and computational efficiency, which necessitates careful consideration in practical implementations.

In summary, the research discussed herein represents a significant contribution to the field of cyber-attack detection by harnessing big data machine learning techniques. While the proposed method exhibits promising capabilities in addressing existing challenges, further refinement is warranted to mitigate computational complexities and enhance scalability for real-world applications.

## IV. DATASET DESCRIPTION

The KDD99[10] dataset stands as a cornerstone in the realm of intrusion detection and cybersecurity research, offering a simulated environment reflective of military networks and comprising diverse network traffic data. This dataset serves as a pivotal benchmark for the development and evaluation of intrusion detection systems (IDS) and anomaly detection algorithms. Encompassing normal traffic alongside a spectrum of attacks like Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), and Probing, it presents a rich landscape for model training and testing.

However, leveraging the KDD99[10] dataset presents formidable challenges. Class imbalance is a notable concern, with certain attack types being disproportionately represented compared to normal traffic. Moreover, redundant features within the dataset can impede classifier performance. Despite these challenges, recent research endeavors have yielded innovative strategies to mitigate these obstacles and bolster the efficacy of intrusion detection systems utilizing KDD99. These strategies often involve adept feature selection or extraction techniques to streamline dimensionality and enhance model generalization.

Additionally, the adoption of ensemble methods and deep learning architectures has emerged as a potent avenue for capturing intricate data patterns.

In essence, the KDD99[10] dataset remains an indispensable asset for propelling advancements in intrusion detection and cybersecurity. Its continued relevance fuels the development of robust and accurate detection systems, essential for fortifying network infrastructures against an array of evolving threats.

## V. Proposed Method

In this section, we present the proposed version, outlining the equipment and strategies utilized within the method. Illustrated by the Spark Chi SVM model diagram, our approach involves sending new expressions deep into the ensemble for analysis. Specifically tailored for Industrial Control System (ICS)[5] environments, our proposed attack detection model employs a combination of a Deep Neural Network (DNN)[7] and a Decision Tree (DT)[8] classifier to effectively identify cyber threats.

Evaluation of our proposed method is conducted through 10-fold cross-validation on genuine ICS datasets, demonstrating its superiority over conventional techniques. We compare its performance against various classifiers such as Random Forest (RF)[6], DNN[7], and AdaBoost[9], affirming its efficacy across different scenarios. Notably, our approach is designed to be versatile, adaptable to existing systems for seamless integration.

One critical challenge addressed in our methodology is the potential misclassification of new malicious data, particularly when trained directly with severe imbalance. To mitigate this, we propose an ensemble deep representation learning model based on Stacked Autoencoder (SAE), enhancing the overall performance of the detection system.

Furthermore, leveraging machine learning, our cybersecurity system can analyze and learn from patterns, thereby fortifying defenses against similar attacks and dynamically responding to evolving threat behaviors. This empowers cybersecurity teams with the capability to preemptively thwart threats and swiftly counteract attacks in real-time, ensuring robust protection of critical infrastructures.

### A. Collection of data

The process of data collection in the realm of cybersecurity involves systematically gathering information on cyberattacks from diverse sources, which serves as the foundation for constructing machine learning models. This entails acquiring a comprehensive dataset comprising various features pertinent to cyber attacks. The pivotal focus at this stage lies in meticulously selecting a subset from the extensive pool of available data for analysis and model development. Ideally, machine learning endeavors commence with a copious amount of data, where the desired outcomes are already known—a concept referred to as labelled data. This labelled data provides a clear understanding of the expected results, thereby facilitating the formulation and evaluation of effective predictive models for cybersecurity purposes.

### B. Preprocessing the data

In the realm of data pre-processing, three critical steps pave the way for effective analysis: formatting, cleaning, and sampling. Formatting entails transforming data from its original structure into a more manageable format, facilitating ease of analysis. This step is indispensable when dealing with data in proprietary file formats or when transitioning between relational databases and flat files. Cleaning, on the other hand, addresses the issue of missing or inadequate data instances. By replacing or eliminating such occurrences, researchers ensure the integrity and reliability of their analyses. Finally, sampling offers a strategic approach to dealing with large volumes of data. By selecting representative subsets, researchers can streamline computational resources, reduce processing time, and expedite exploration and testing of ideas. These three steps collectively form the backbone of data pre-processing, enabling researchers to extract meaningful insights and draw reliable conclusions from complex datasets

### C. Feature Extraction

The subsequent phase involves feature extraction, which is a process following attribute reduction. Unlike feature selection, which merely prioritizes existing attributes based on their predictive significance, feature extraction actively modifies attributes. This modification entails linearly combining the original attributes to generate new ones, referred to as features. Subsequently, classifier algorithms are employed to train our models, utilizing the acquired labeled dataset. Evaluation of the models is conducted using the remaining labeled data. The pre-processed data underwent categorization utilizing various machine learning techniques, with a particular emphasis on random forest classifiers for their suitability in the context.

### D. Evaluating the model

In the model development process, model evaluation is a crucial step aimed at selecting the most suitable model for depicting data and predicting future performance. To avoid overfitting, it's imperative not to evaluate model performance using training data alone, as this can lead to overly optimistic results. Instead, techniques such as Hold-Out and Cross-Validation are employed in data science to assess models effectively, using a separate test set unseen during model training.

Various algorithms are utilized in this process, including Deep Neural Networks, Decision Trees, Random Forests, and AdaBoost[9]. Deep Neural Networks, inspired by the human brain, consist of interconnected layers of nodes for data processing across tasks like image recognition and natural language processing. Decision Trees[8], on the other hand, recursively split data into subsets based on significant features, presenting results in a tree-like structure for easy interpretation. Despite their simplicity, Decision Trees may suffer from overfitting, necessitating techniques like pruning and ensembling.

Random Forests, an ensemble learning technique, employ multiple decision trees to improve prediction accuracy and stability, although they can be challenging to interpret due to their complexity. AdaBoost[9], another popular ensemble method, iteratively trains weak models to focus on misclassified examples, demonstrating versatility in handling diverse datasets but may be sensitive to noisy data and computationally intensive during training.

The proposed approach involves starting with a cyber attack dataset, filtering it based on analysis requirements, and preprocessing it for further use. The dataset is then divided into training and testing sets. Following this, the classification algorithms are trained using the training data and evaluated on the testing data, yielding accuracy metrics to gauge model performance accurately. This comprehensive methodology ensures the selection of robust models capable of effectively analyzing and predicting cyber attack data.

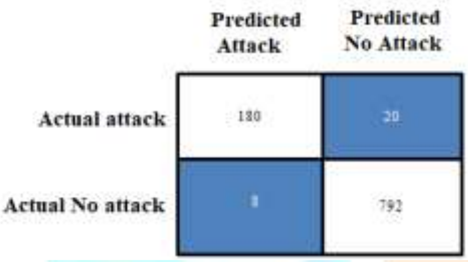
## VI. RESULT & DISCUSSION

### Description about Experimental Setup:

| Hardware & Software | Specification |
|---|---|
| RAM | 8.00 GB (7.45 GB usable) |
| Processor | AMD A6-9220e RADEON R4, 5 COMPUTE CORES 2C+3G 1.60 GHz |
| Hard Disk | 1TB |
| Operating System | windows |
| Language | Python |
| Front End | Spyder IDE |
| Back End | TensorFlow |

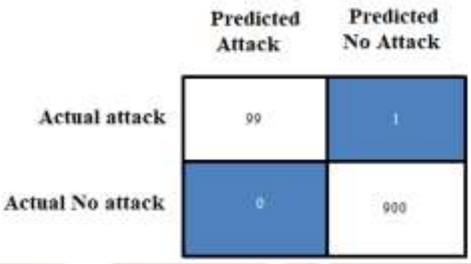
**Table 1.** Hardware & software Specification

### Confusion Matrix:



| | Predicted Attack | Predicted No Attack |
|---|---|---|
| Actual attack | 180 | 20 |
| Actual No attack | 8 | 792 |

**Fig 1.** Confusion Matrix for Random Forest



| | Predicted Attack | Predicted No Attack |
|---|---|---|
| Actual attack | 99 | 1 |
| Actual No attack | 0 | 900 |

**Fig 2.** Confusion Matrix for DNN

### Result:

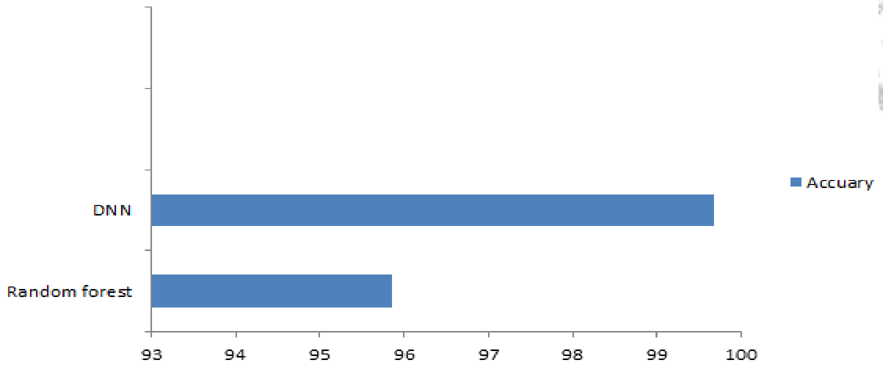


**Fig 3.** Plot for Random Forest and DNN

### Conclusion:

In this proposed research, a novel technology integrating representative deep learning models and innovative construction techniques is introduced to address the challenge of handling imbalanced datasets effectively. The approach involves achieving a balanced representation of the raw data, which is often skewed in real-world scenarios. By employing advanced deep learning architectures, the model generates new expressions from the data, enhancing its ability to discern intricate patterns and features within the dataset. Specifically, the model utilizes deep neural network (DNN)[7] and decision tree (DT)[8] classifiers to detect network attacks, a critical task in safeguarding infrastructural systems. To evaluate the efficacy of the proposed methodology, two distinct datasets sourced from real critical infrastructure systems are utilized. The results demonstrate significant performance improvements over conventional classifiers, with a notable 10% increase in the F1 score. Specifically, the proposed model achieves an impressive accuracy rate of 95.86% and 99.67% on the gas pipeline dataset and the safe water treatment dataset, respectively. These results underscore the superiority of the proposed approach over traditional classification techniques such as random forests (RF)[6] and standard deep neural networks (DNN)[7]. Overall, this research contributes a robust and effective methodology for enhancing the detection of network attacks in critical infrastructure systems, offering considerable advancements in cybersecurity and system reliability.

## References

[1] H. Karimipour et V. Dinavahi, `` Parallel dynamic state estimation based on extended Kalmanfilter,`` IEEE Trans. Clever Réseau, vol. 6, no. 3, p. 1539–1549, May 2015.

[2] Shorfuzzaman, Mohammad. "Detection of cyber attacks in IoT using tree-based ensemble and feedforward neural network." 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2020.

[3] Verma, Rishabh, and Bharti Thakur. "Machine Learning Techniques for the Prediction of Cyber-Attacks." 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). IEEE, 2023.

[4] Husák, Martin, et al. "Survey of attack projection, prediction, and forecasting in cyber security." IEEE Communications Surveys & Tutorials 21.1 (2018): 640-660.

[5] Koay, Abigail MY, et al. "Machine learning in industrial control system (ICS) security: current landscape, opportunities and challenges." Journal of Intelligent Information Systems 60.2 (2023): 377-405.

[6] Biau, Gérard. "Analysis of a random forests model." The Journal of Machine Learning Research 13.1 (2012): 1063-1095.

[7] Li, Guanpeng, et al. "Understanding error propagation in deep learning neural network (DNN) accelerators and applications." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2017.

[8] Suthaharan, Shan, and Shan Suthaharan. "Decision tree learning." Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning (2016): 237-269.

[9] Hu, Weiming, Wei Hu, and Steve Maybank. "Adaboost-based algorithm for network intrusion detection." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 38.2 (2008): 577-583.

[10] https://www.kaggle.com/datasets/galaxyh/kdd-cup-1999-data