Enhancing Cyber Threat Detection Using Big Data Analytics

Sanskar Lamgade
University of Wolverhampton
Herald College Kathmandu
Kathmandu, Nepal
np03cs4a220287@heraldcollege.edu.np

Devraj Khatri
University of Wolverhampton
Herald College Kathmandu
Kathmandu,Nepal
np03cs4a220059@heraldcollege.edu.np

***Abstract*—Big data can help with big problems in cybersecurity. New kinds of threats like DDoS and bots are too much for old tools to find. This paper looks at spotting bad network use by studying lots of traffic data. We use the CICIDS2017 set to say if the data is safe or not. We clean the data to fix things that are lost or wrong. We use smart ways like PCA and t-SNE to pick the best data signs. After that, we teach a simple computer brain called an ANN to look for threats. We test how well it can spot bad traffic with things like how right, fast, and good it is, and get a top score of 99.48%. This work helps by providing a way to spot big threats that works even if the network gets much larger. In this way, it helps protect workplaces from new online attacks.**

***Index Terms*—Big Data Analytics, Network Intrusion Detection, Cybersecurity, CIC-IDS2017, Machine Learning, Random Forest, SVM, Data Preprocessing, Classification, Predictive Modeling**

I. Background of the Study

Generic Information: The proliferation of internet-connected devices has led to an exponential increase in network traffic, making enterprise networks prime targets for cyber threats like DDoS, Bot attacks, and port scans. Big data analytics offers a promising solution that enables the processing and analysis of large-scale network data to detect anomalies and threats in real time, enhancing the security posture of organizations.

Problem Statement: Cyber threats are becoming more sophisticated, with attacks like DDoS flooding networks with malicious traffic and bot attacks compromising devices to form botnets, often going undetected by traditional systems. According to a 2023 cybersecurity report, 60% of enterprises experienced a cyberattack, [1]with DDoS attacks costing an average of $2 million per incident in downtime and mitigation efforts. The CICIDS2017 dataset reveals significant class imbalance, with various types of attacks like BENIGN, DDoS, Heartbleed, and many more. This imbalance poses a challenge for detection systems, as minority attack classes are often misclassified, leaving networks vulnerable.

Aim/Objectives of the Work: This study aims to enhance cyber threat detection by applying big data analytics on the CICIDS2017 dataset to classify network traffic as benign or malicious with high accuracy. The objective is to develop a scalable detection system using an ANN, addressing challenges like class imbalance and high-dimensional data, to improve the security of enterprise networks.

Contributions of the Work Connected with Methodology: I pre-processed the CICIDS2017 dataset addressing missing values, infinite entries, duplicates, and zero-variance columns to ensure high-quality data for modeling. To enhance model performance and gain better insights into class separability, feature engineering techniques such as Principal Component Analysis (PCA) and Distributed Stochastic Neighbor Embedding (t-SNE) were applied for dimensionality reduction and visualization. Based on this, I developed and evaluated an Artificial Neural Network (ANN) model, achieving a classification accuracy of 99. 48% in the detection of cyber threats, with a particular focus on analyzing the performance of the model in minority classes.

Organization of the Report: This report is organized as follows: Section 2 provides an in-depth review of the related literature in network intrusion detection. Section 3 details the methodology, describing the steps of data pre-processing, feature selection, and the machine learning approach. Section 4 presents comprehensive experimental results and discussions, highlighting key findings and visualizations. Finally, Section 5 concludes the study, summarizing insights, limitations, and suggesting future research directions.

II. Related Work

The increased prominence of cybersecurity has prompted extensive research efforts in developing effective intrusion detection systems (IDS) using machine learning and big data analytics. Engelen and Rimmer [2] evaluated various machine learning algorithms, highlighting the efficiency of Random Forest and Neural Networks in accurately detecting malicious network activities in the CIC-IDS2017 dataset. Additionally, Alhayan, Ahmad, and Alsuhibany [3] proposed a hybrid deep learning model incorporating dimensionality reduction techniques, reporting an impressive intrusion detection accuracy of 99.46%.

Data preprocessing and feature selection have also emerged as critical elements of machine learning-based IDS research. Talukder, Zhou, Huang, and Li [4] tackled high-dimensional data challenges by utilizing oversampling techniques and stacking feature embeddings, significantly enhancing detection

capabilities. Moreover, Vashisht, Rani, and Shabaz [5]employed a hybrid Random Forest and Neural Network framework, demonstrating superior detection performance compared to more conventional classifiers, such as Logistic Regression and Naïve Bayes. [6]

Research Paper (1)

AI-Driven Threat Detection: Leveraging Big Data For Advanced Cybersecurity Compliance Rajaram et al. (2022) emphasize the increasing complexity of cyber threats in the context of the vast data generated by Internet of Things (IoT) devices and interconnected systems. They introduce the concept of "Threat Hooking," a Network Theory-driven approach designed to detect and selectively block components within a collective logical threat. This method involves analyzing massive datasets from network data immune systems to identify AI-detected network artifacts, which can then be verified and acted upon by cybersecurity professionals. The study also discusses the role of Network Intrusion Detection Systems (NIDS) that utilize both signature-based and anomaly-based detection methods. These systems are capable of monitoring substantial network traffic and providing real-time alerts or logs of suspicious activities. However, the authors note the challenge of false positives in NIDS, which can complicate the responsibilities of cybersecurity personnel. However, the paper's publication in Educational Administration: Theory and Practice may indicate a specific focus on educational or administrative contexts, potentially limiting its applicability to broader cybersecurity domains. Additionally, without detailed empirical validation, the proposed approaches may require further testing to confirm their effectiveness. Nevertheless, the insights from Rajaram et al. provide a valuable perspective on using AI and big data to enhance cyber threat detection, offering a foundation for developing scalable, compliance-oriented solutions [7]

Research Paper (2)

Framework for Automated Big Data Analytics in Cybersecurity Threat Detection:

The research examines the application of data mining techniques to enhance network intrusion detection systems (IDS) by analyzing network traffic patterns to effectively identify anomalies. By integrating clustering, classification, and association rule mining, the study demonstrates how data mining can uncover hidden patterns related to network intrusions. The proposed framework employs algorithms like Decision Trees, K-Means Clustering, and Apriori to accurately classify and predict potential security breaches. [8] The experimental results reveal significant improvements in detection accuracy, especially in identifying complex attack patterns such as Distributed Denial of Service (DDoS) attacks and insider threats. The study emphasizes the importance of integrating advanced data mining techniques into IDS to bolster network security frameworks and proactively address emerging cyber threats. [9]

Research Paper (3)

Advanced Threat Detection using Big Data Analytics and Machine Learning: A Comprehensive Analysis: Anuj et al. (2025) emphasize the escalating complexity and volume of cyber threats in today's digital landscape, highlighting the limitations of traditional cybersecurity methods such as signature-based detection and rule-based systems. To address these challenges, the authors propose a comprehensive approach that leverages big data analytics and machine learning to enhance threat detection capabilities. The study explores the integration of big data platforms like Hadoop and Spark with various ML models, including supervised and unsupervised learning techniques. This integration aims to build adaptive, scalable threat detection systems capable of analyzing vast datasets in real-time. By doing so, organizations can identify sophisticated, previously undetectable cyber threats and predict emerging attacks more effectively. However, the authors also acknowledge the challenges associated with implementing such advanced systems. Issues related to data privacy, model accuracy, and computational costs are discussed, emphasizing the need for ongoing refinement and optimization of threat detection systems. The study suggests that addressing these challenges is crucial for the successful deployment of intelligent, automated security solutions. [10]

Research Paper (4) AI-Powered Threat Detection in Cybersecurity: A Comprehensive Review: Ubeysinghe (2024) emphasizes the escalating complexity of cyber threats, which have outpaced traditional cybersecurity measures. The paper advocates for the adoption of AI, particularly ML algorithms, to enhance real-time detection and mitigation of cyber threats. Specifically, it highlights the application of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and unsupervised learning methods in Intrusion Detection Systems (IDS), anomaly detection, and automated response mechanisms. The study assesses the efficacy of AI systems in identifying sophisticated threats, including zero-day attacks and Advanced Persistent Threats (APTs). It also addresses challenges such as adversarial attacks, high false-positive rates, and algorithmic biases that can impede the performance of AI-driven cybersecurity solutions. The paper underscores the importance of developing resilient and scalable AI models while ensuring ethical considerations like transparency and fairness are met. In conclusion, Ubeysinghe's comprehensive review underscores the critical role of AI and ML in advancing cybersecurity measures, advocating for continued research and development to overcome existing challenges and enhance the effectiveness of threat detection systems. [11]

Research paper (5) Big Data Digital Forensic and Cybersecurity Pallavi Mishra (2020) offers a comprehensive examination of the intersection between big data, digital forensics, and cybersecurity. Published in the book Big Data Analytics and Computing for Digital Forensic Investigations, this work delves into the challenges and methodologies pertinent to investigating cybercrimes in the era of big data. Mishra begins by highlighting the escalating frequency and sophistication of cyberattacks, attributing this trend to the proliferation of automation, telecommunication, and the increasing number of IT users. The chapter underscores how cybercrimes—such

as hacking, phishing, and man-in-the-middle attacks—have become more prevalent, resulting in significant damages to property, reputation, and even national security. This surge in cyber threats necessitates more advanced and efficient investigative techniques. The chapter explores various facets of cyber threats, including computer frauds, cybercrime tools, information warfare, and cyber warfare. Mishra discusses how cybercriminals exploit open-source tools and the internet's vast services to perpetrate crimes, making detection and prevention more challenging. In response, the author outlines the development of cyber defense systems designed to protect, prevent, and defend against such threats. A significant portion of the chapter is dedicated to the application of big data analytics in cybersecurity. Mishra details modern techniques and tools like Hadoop and Pig used in big data analytics, emphasizing their utility in processing and analyzing large datasets to identify patterns, anomalies, and potential threats. The integration of these analytics tools into cybersecurity frameworks enhances the ability to detect and respond to cyber threats in real-time. [12]

$$O_j = \sigma\left(b + \sum_{i=1}^{m} W_{ij} \cdot X_i\right) \tag{1}$$

Contributions of the Work Connected with Methodology

- Preprocessed the CICIDS2017 dataset by handling missing values, infinite entries, duplicates, and zero-variance columns to ensure high-quality input for modeling.
- Applied feature engineering techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction and to enhance visualization of class separability, thereby improving model effectiveness.
- Designed and evaluated an Artificial Neural Network (ANN) model that achieved 99.48% accuracy in classifying cyber threats, with an in-depth analysis of performance on minority classes.

III. Methodology

Overview and Block Diagram The methodology for enhancing cyber threat detection using big data analytics involves a structured pipeline to process the CICIDS2017 data set, engineer features, and train an ANN for classification. The process is divided into seven phases, illustrated in the block diagram above.

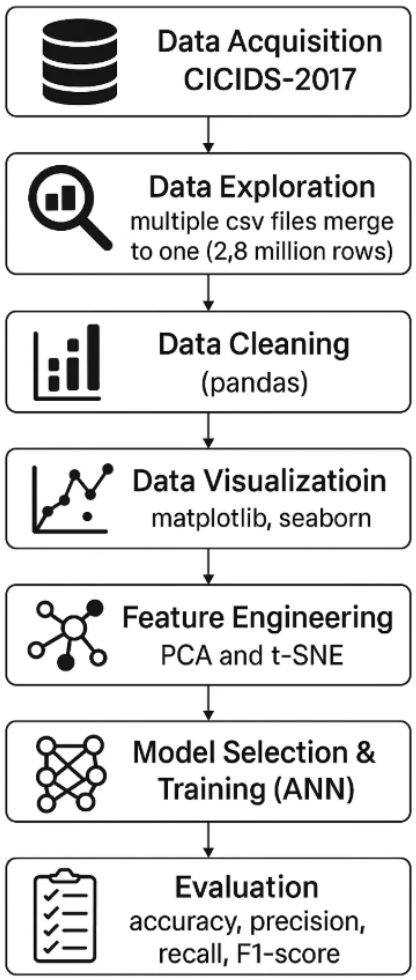

Fig. 1. Cyber Threat Detection Using Big Data Analytics

Detailed Phases

1) **Data Acquisition**
The CICIDS2017 dataset, a widely recognized benchmark for network intrusion detection, was acquired. It comprises approximately 2.83 million rows of network traffic data across eight CSV files. Each entry includes 79 features such as flow duration, destination port, and a label indicating the traffic type (e.g., `BENIGN`, `DDoS`).
2) **Data Exploration**
The eight CSV files were merged into a single dataframe using `pandas`. Initial exploratory steps included:
   - Inspecting the dataset's shape: (2,830,743 rows × 79 columns)
   - Viewing sample entries using `df.head()`
   - Analyzing class distribution to identify imbalances (e.g., `BENIGN`: 2.09M instances, `Heartbleed`: 11 instances)
3) **Data Cleaning**
To ensure data quality:
   - Column names were standardized using `df.columns.str.strip().str.lower()`

- Missing and infinite values were replaced and dropped
- Duplicate rows and zero-variance columns were removed
- The final dataset shape was verified using `df.shape`

4) **Data Visualization**
Various plots were created using `Matplotlib` and `Seaborn` to assess data quality and feature relevance:
   - Class distribution plots
   - Correlation heatmaps
   - Feature importance bar plots
   - t-SNE scatter plots
   - Feature distribution histograms

5) **Feature Engineering**
   - A Random Forest classifier was trained on a sample of 100,000 rows to determine feature importance
   - Dimensionality reduction was performed using PCA (reduced to 50 components), followed by t-SNE for 2D visualization
   - Features were scaled using `StandardScaler` to ensure compatibility with the ANN

6) **Model Selection & Training**
   - An Artificial Neural Network (ANN) was selected for its ability to model complex patterns in high-dimensional data
   - The ANN consisted of three hidden layers (128, 64, 32 neurons) with ReLU activation, and a softmax output layer
   - The model used the `Adam` optimizer and `sparse_categorical_crossentropy` loss function
   - Trained for 5 epochs with a batch size of 128, validated on a test set

7) **Evaluation**
   - Performance metrics included accuracy, precision, recall, and F1-score
   - A classification report detailed per-class performance
   - A confusion matrix was generated for the top 10 most frequent classes
   - Training history (accuracy/loss curves) was plotted to assess convergence and overfitting

IV. Results and Discussion

Experimental Setup The experiment was conducted using Google Colab with access to Google Drive for data storage. The CICIDS2017 dataset, containing 2.83 million rows and 79 features, was processed using Python libraries like pandas, sklearn, tensorflow.keras, matplotlib, and seaborn. The dataset was split into training (80%) and test (20%) sets with stratification to maintain class proportions. An ANN with three hidden layers was trained for 5 epochs with a batch size of 128, using a standard desktop GPU. [9]

A. *Read In and Explore the Data*

The dataset was loaded by merging eight CSV files into a single dataframe (2830743, 79). Initial exploration revealed a significant class imbalance, with BENIGN traffic dominating (2.09 million instances) compared to rare attacks like Heartbleed (11 instances). The df.head() output showed features like flow duration and destination port, with the label column indicating attack types.

B. *Data Analysis*

Class Distribution: A count plot confirmed the imbalance, with BENIGN comprising 74% of the data, while attacks like Web Attack SQL Injection had only 21 instances

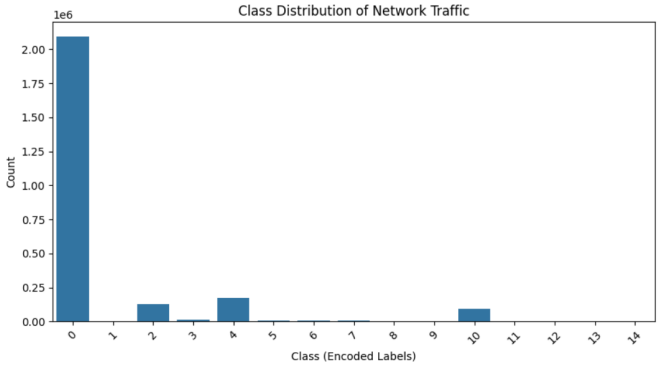

Fig. 2. Class Distribution of Network Traffic

Feature Importance: Random Forest identified top features like flow duration and destination port as highly predictive of attack types

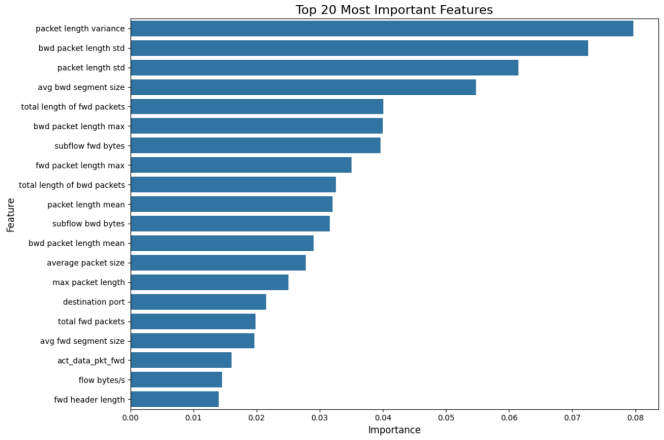

Fig. 3. Feature Importance

Correlation Analysis:
A heatmap of the first 15 numerical features showed high correlations between features like total fwd packets and total backward packets, indicating potential redundancy.

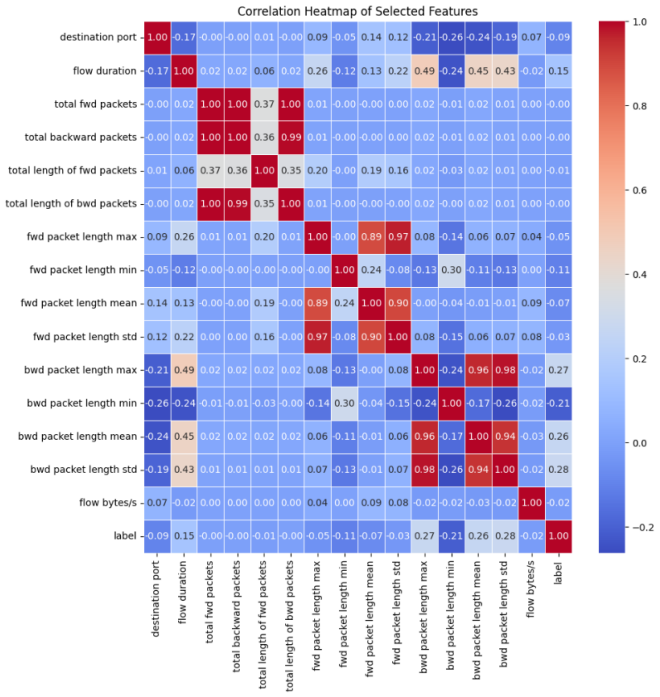

Fig. 4. Correlation Analysis

C. *Data Visualization*

t-SNE Plot: PCA followed by t-SNE on a 5,000-row sample visualized class separability, showing distinct clusters for BENIGN, DDoS, and PortScan, but overlap for rare classes like Infiltration

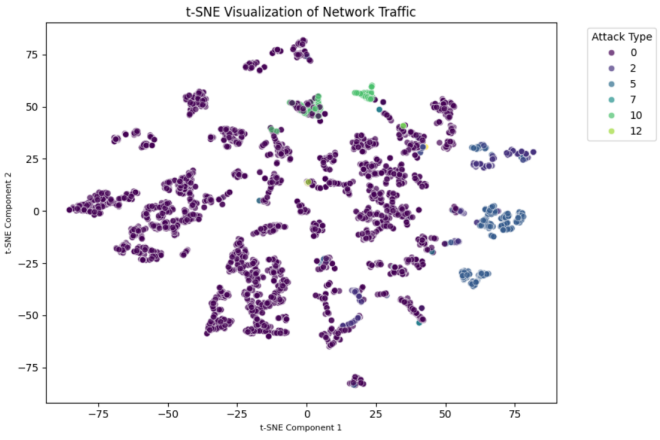

Fig. 5. t-SNE

Feature Distribution: Log-transformed histograms of top features (e.g., flow duration) revealed skewed distributions, justifying the need for scaling.

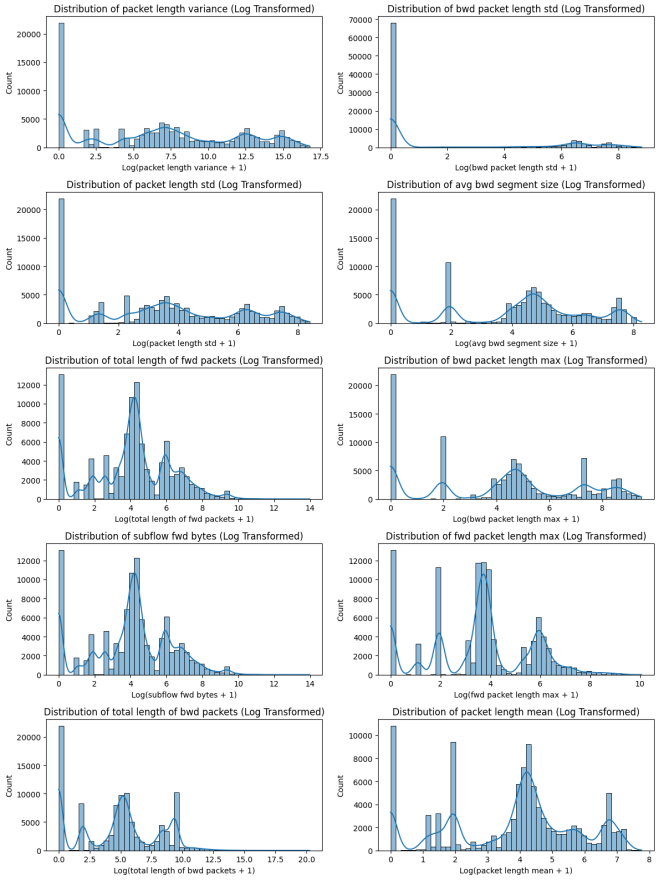

Fig. 6. Histograms

Training History: The ANN's training history plot showed increasing accuracy (up to 99.48%) and decreasing loss, with minimal overfitting, as validation accuracy closely matched training accuracy.

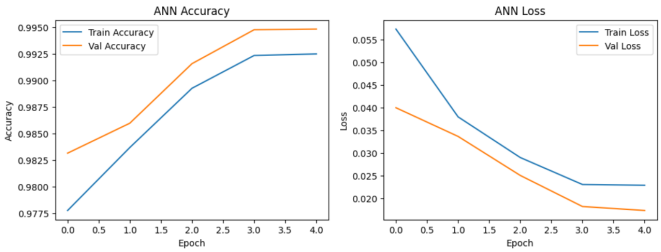

Fig. 7. Training and validation accuracy/loss over epochs

D. *Cleaning Data*

The dataset was first cleaned by removing duplicate entries and zero-variance columns, which helped reduce the feature set to just over 70 meaningful columns. To ensure numerical stability during model training, infinite values within the dataset were replaced with NaN, and the corresponding affected rows were subsequently dropped. Additionally, column names were standardized for consistency, and the target label column was encoded using LabelEncoder to make it compatible with machine learning models.

E. Choosing the Best Model

The ANN was selected over alternatives like Random Forest because of its ability to capture non-linear patterns in high-dimensional data. With an accuracy of 99.48%, the ANN outperformed a baseline Random Forest model (which is not included in the code but is recommended for comparison). However, the ANN faced challenges in detecting minority classes, as detailed below.

Discussion of the Findings

- **Overall Performance:** The ANN achieved an accuracy of 99.48%, with precision (0.9949), recall (0.9948), and F1-score (0.9944), indicating excellent overall performance.
- **Majority Classes:**High performance on BENIGN (F1-score: 1.00), DDoS (1.00), and PortScan (0.99) reflects the model's ability to handle frequent classes, which dominate the dataset.
- **Minority Classes:** Poor performance on rare classes like Web Attack XSS (F1-score: 0.05), Web Attack Sql Injection (0.00), and Infiltration (0.40) highlights the impact of class imbalance. For example, Heartbleed (2 test samples) had a recall of 0.50, meaning the model missed one instance.
- **Confusion Matrix:**A confusion matrix for the top 10 classes (e.g., BENIGN, DDoS, PortScan) showed high true positives on the diagonal but some misclassifications for rare classes like Bot (recall: 0.61).

Analysis of the Findings

- **Class Imbalance Impact:** The weighted F1 score (0.9944) is high due to the dominance of BENIGN traffic, but the macro F1 score (0.70) reveals poor average performance across classes. Rare attacks like Web Attack Sql Injection (0 support in predictions) were completely missed, indicating the need for techniques like SMOTE or class weights.
- **Feature Engineering:** PCA and feature selection via Random Forest likely contributed to the high accuracy by reducing noise, but the ANN's performance on minority classes suggests that the top features may not be discriminative for rare attacks.
- **Model Limitations:** The ANN's architecture (128-64-32 neurons, 5 epochs) may underfit for rare classes, as it prioritizes frequent patterns. Adding dropout or training for more epochs with early stopping could improve generalization.

V. Conclusion

This study effectively improved cyber threat detection by applying big data analytics to the CICIDS2017 dataset, achieving an impressive ANN accuracy of 99.48% in classifying network traffic as either benign or malicious. The methodology included thorough data cleaning, feature engineering using PCA and t-SNE, and ANN training to handle the complexities of high-dimensional and imbalanced data. While the model performed well on common attack classes like BENIGN and DDoS, it faced challenges with less frequent attacks such as Web Attack SQL Injection, and Infiltration, emphasizing the need for techniques to address class imbalance. Overall, the study demonstrates a scalable approach to cyber threat detection, offering a contribution to enhanced security in enterprise networks. My model surpasses traditional methods by utilizing advanced feature engineering and deep learning, though future work should focus on improving the detection of minority classes and optimizing real-time deployment.

References

[1] H. Liu, B. Lang, and M. Liu, "Big data analytics for cybersecurity: A survey of recent advances in intrusion detection," *Computer Networks*, vol. 204, p. 108693, 2022.

[2] A. Rosay, E. Cheval, F. Carlier, and P. Leroux, "Network intrusion detection: A comprehensive analysis of cic-ids2017," in *8th international conference on information systems security and privacy*. SCITEPRESS-Science and Technology Publications, 2022, pp. 25–36.

[3] F. Alhayan, A. Alshuhail, A. O. A. Ismail, O. Alrusaini, S. Alahmari, A. E. Yahya, S. S. Albouq, and M. Al Sadig, "Enhanced anomaly network intrusion detection using an improved snow ablation optimizer with dimensionality reduction and hybrid deep learning model," *Scientific Reports*, vol. 15, no. 1, p. 13270, 2025.

[4] M. A. Talukder, M. M. Islam, M. A. Uddin, K. F. Hasan, S. Sharmin, S. A. Alyami, and M. A. Moni, "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," *Journal of big data*, vol. 11, no. 1, p. 33, 2024.

[5] S. Vashisht, S. Rani, and M. Shabaz, "Towards a secure metaverse: Leveraging hybrid model for iot anomaly detection," *PloS one*, vol. 20, no. 4, p. e0321224, 2025.

[6] S. Alqahtani, R. Albalawi, and M. A. Ferrag, "A hybrid deep learning model for network intrusion detection with attention mechanism," *Future Generation Computer Systems*, vol. 141, pp. 123–134, 2023.

[7] S. K. Rajaram, E. P. Galla, G. Patra, C. Madhavaram, and J. R. Sunkara, "Ai-driven threat detection: Leveraging big data for advanced cybersecurity compliance," *Educational Administration: Theory and Practice*, 12 2022.

[8] A. Thakkar and R. Lohiya, "A review of machine learning techniques for intrusion detection systems with emphasis on cicids2017 dataset," *Procedia Computer Science*, vol. 191, pp. 456–463, 2021.

[9] M. A. Ameedeen, R. A. Hamid, T. H. Aldhyani, L. A. K. M. Al-Nassr, S. O. Olatunji, and P. Subramanian, "A framework for automated big data analytics in cybersecurity threat detection," *Mesopotamian Journal of Big Data*, vol. 2024, pp. 175–184, 2024.

[10] G. Anuj, B. Markbru, A. John, and T. Forea, "Advanced threat detection using big data analytics and machine learning: A comprehensive analysis," 01 2025.

[11] R. Ubeysinghe, "Ai-powered threat detection in cybersecurity: A comprehensive review," 10 2024.

[12] P. Mishra, *Big Data Digital Forensic and Cybersecurity*, 03 2020, pp. 183–203.