

UNIVERSITI TEKNOLOGI MARA

**BIG DATA ANALYTICS APPROACH
FOR INTRUSION DETECTION
USING RANDOM FOREST AND
SUPPORT VECTOR MACHINE**

MOHAMAD ARIFFIN BIN MOHAMAD TAHAIR

BACHELOR OF COMPUTER SCIENCE (Hons.)

JULY 2026

SUPERVISOR APPROVAL

BIG DATA ANALYTICS APPROACH FOR INTRUSION DETECTION USING RANDOM FOREST AND SUPPORT VECTOR MACHINE

By

MOHAMAD ARIFFIN BIN MOHAMAD TAHAIR

2024832444

This proposal was prepared under the supervision of the project supervisor, Mr. NORIZUANDI BIN IBRAHIM. It was submitted to the Faculty of Computer and Mathematical Sciences and was accepted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science (Hons.)

Approved by

.....

NORIZUANDI BIN IBRAHIM

Project Supervisor

JULY 19, 2026

ACKNOWLEDGEMENT

First and foremost, all praise and gratitude are due to Allah SWT for His countless blessings, guidance, strength, and mercy throughout this Final Year Project. Without His will and grace, this research would not have reached completion.

My deepest thanks go to my supervisor, Mr. Norizuandi, for his continuous guidance, invaluable advice, constructive criticism, and steady encouragement throughout this research. His expertise, patience, and dedication contributed greatly to the completion of this project, and his insightful feedback has improved both the quality of the work and my own understanding of the subject.

I am also sincerely grateful to my Final Year Project lecturer, Dr. Liew Chin Ying, for her guidance, knowledge, constructive comments, and continuous support over the course of this subject. Her dedication gave me valuable insight into how academic research is conducted and has strengthened this thesis considerably.

My appreciation extends to all lecturers and staff of the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Sarawak Branch, for their guidance, support, and encouragement throughout my academic journey. What they taught me made this study possible.

I am deeply grateful to my beloved parents, my family, and my friends for their endless love, prayers, patience, understanding, and moral support through a demanding journey. Their belief in me was my greatest source of motivation while this research took shape.

Finally, I wish to thank everyone who contributed to this work, whether directly or indirectly. Every kindness, every word of encouragement, and every form of help has been sincerely appreciated and will be remembered.

ABSTRACT

The increasing volume and complexity of network traffic have created significant challenges for conventional Intrusion Detection Systems (IDS), particularly in processing large-scale data and detecting malicious activities efficiently. This study develops a Big Data Analytics-based IDS by integrating Apache Spark with machine learning techniques for network intrusion detection. Random Forest (RF) and Support Vector Machine (SVM) were implemented as the classification algorithms, while the CICIDS2017 benchmark dataset was used for training and evaluation. The system was developed using Python, Apache Spark, Scikit-learn, and Streamlit following the Agile Software Development methodology. Network traffic data were preprocessed using Apache Spark to validate the required features and remove missing, invalid, and infinite values before classification. The models were evaluated using accuracy, macro precision, macro recall, macro F1-score, classification reports, and confusion matrices.

The experimental results showed that RF achieved the best overall performance, with an accuracy of 99.05%, Macro Precision of 24.91%, Macro Recall of 24.92%, and Macro F1-score of 24.92%. In comparison, SVM achieved 77.05% accuracy, 15.47% Macro Precision, 18.34% Macro Recall, and 16.20% Macro F1-score. The preprocessing module also successfully reduced 445,909 uploaded records to 445,645 valid records by removing 264 invalid records before prediction. The developed web-based system provides dataset uploading, security analytics, threat prediction, model comparison, and automated security reporting. Overall, the findings demonstrate that RF provided better classification performance than SVM within the developed IDS, while the integration of Apache Spark and Streamlit enabled an end-to-end platform for network traffic processing, intrusion detection, analysis and reporting.

TABLE OF CONTENTS

CONTENT	PAGE
SUPERVISOR APPROVAL.....	iii
STUDENT DECLARATION	iv
ACKNOWLEDGEMENT	v
ABSTRACT	vi
LIST OF TABLES.....	xi
LIST OF FIGURES	xii
CHAPTER 1	14
1.1 Background of the study	14
1.2 Problem statement.....	15
1.3 Aim of study.....	16
1.4 Objectives of study.....	16
1.5 Scope of study	16
1.6 Significance of the study	18
1.7 Conclusion	18
CHAPTER 2	20
2.1 Intrusion Detection System	20
2.1.1 Host-Based Intrusion Detection Systems (HIDS).....	23
2.1.2 Network-Based Intrusion Detection Systems	24
2.1.3 Signature-Based Intrusion Detection System.....	25
2.1.4 Anomaly-Based Intrusion Detection System	26
2.1.5 Comparison of the types of Intrusion Detection System (IDS)	26