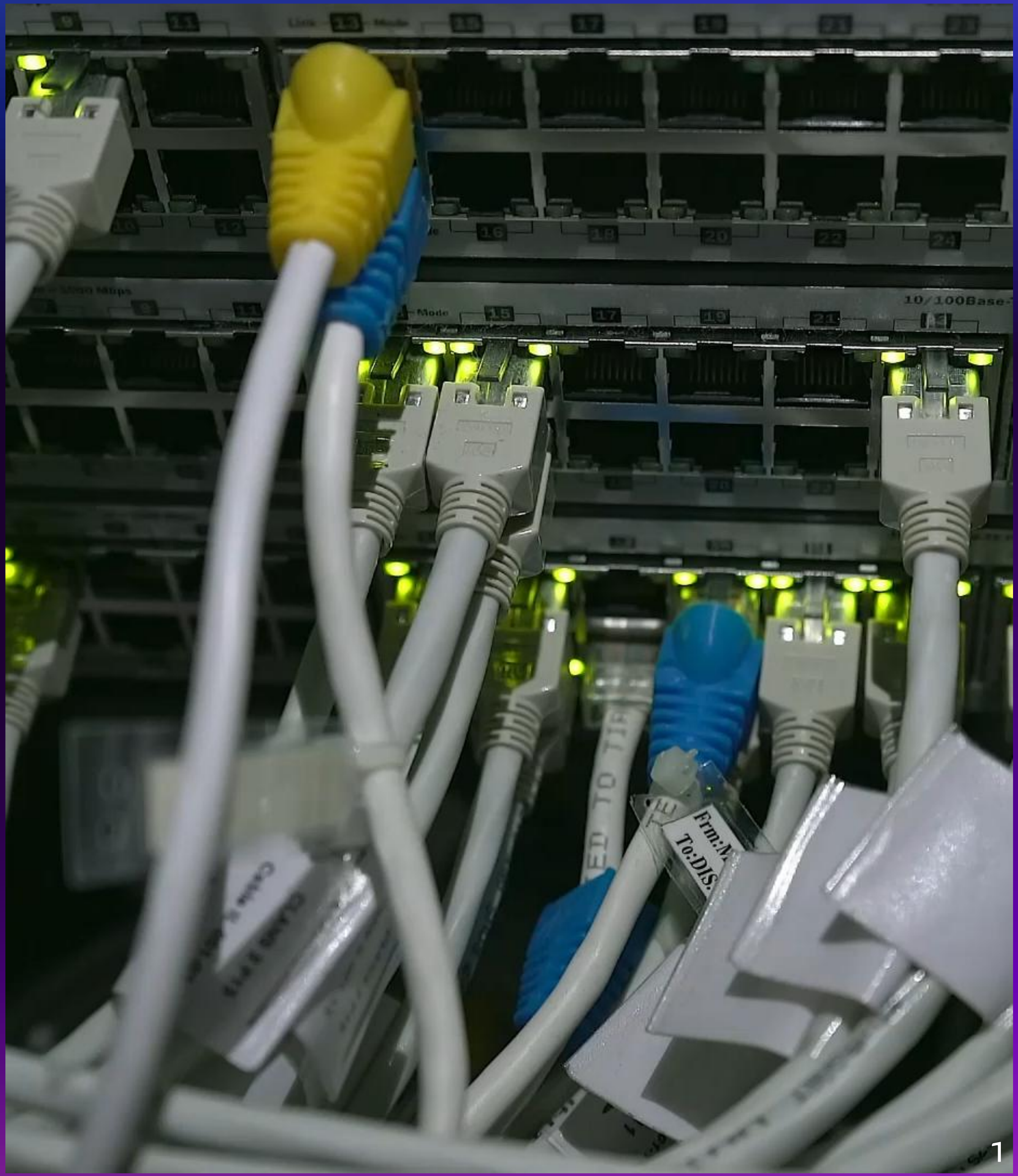
NETWORK INTRUSION DETECTION

PRESENTATION

PRESENTED BY :
MOHAMED AMER



AGENDA

Introduction

Background

Machine learning model

Implementation

Performance Analysis

Key Findings & Challenges

Conclusions

INTRODUCTION TO

NETWORK INTRUSIONS

WHY NETWORK SAFETY MATTER?

- Networks are the critical infrastructure for every enterprise and organization
- The rapid evolution of AI and technology increases the security risks
- Data privacy and confidentiality is a growing concern to everyone

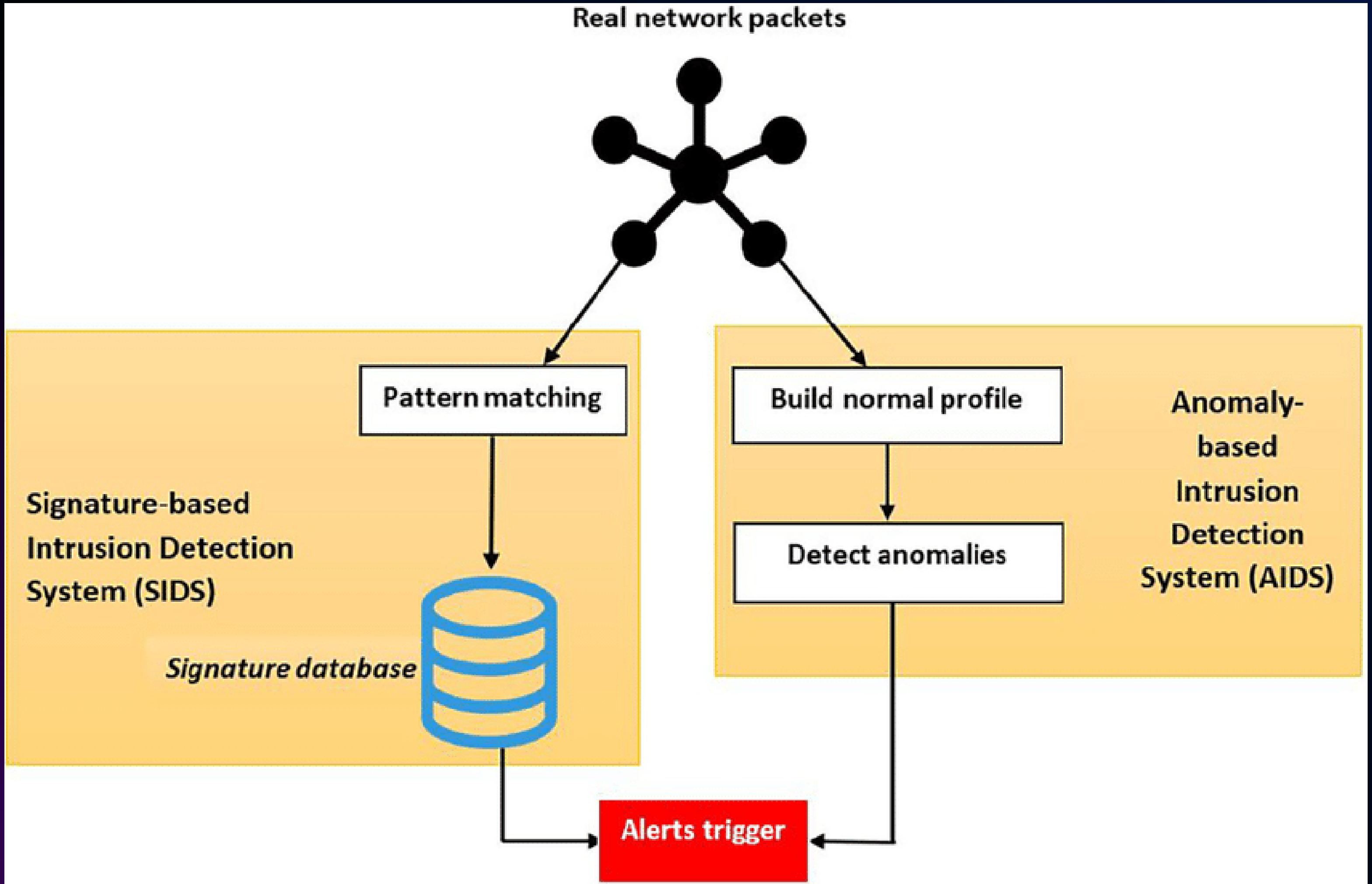
INTRUSION DETECTION SYSTEMS (IDS)

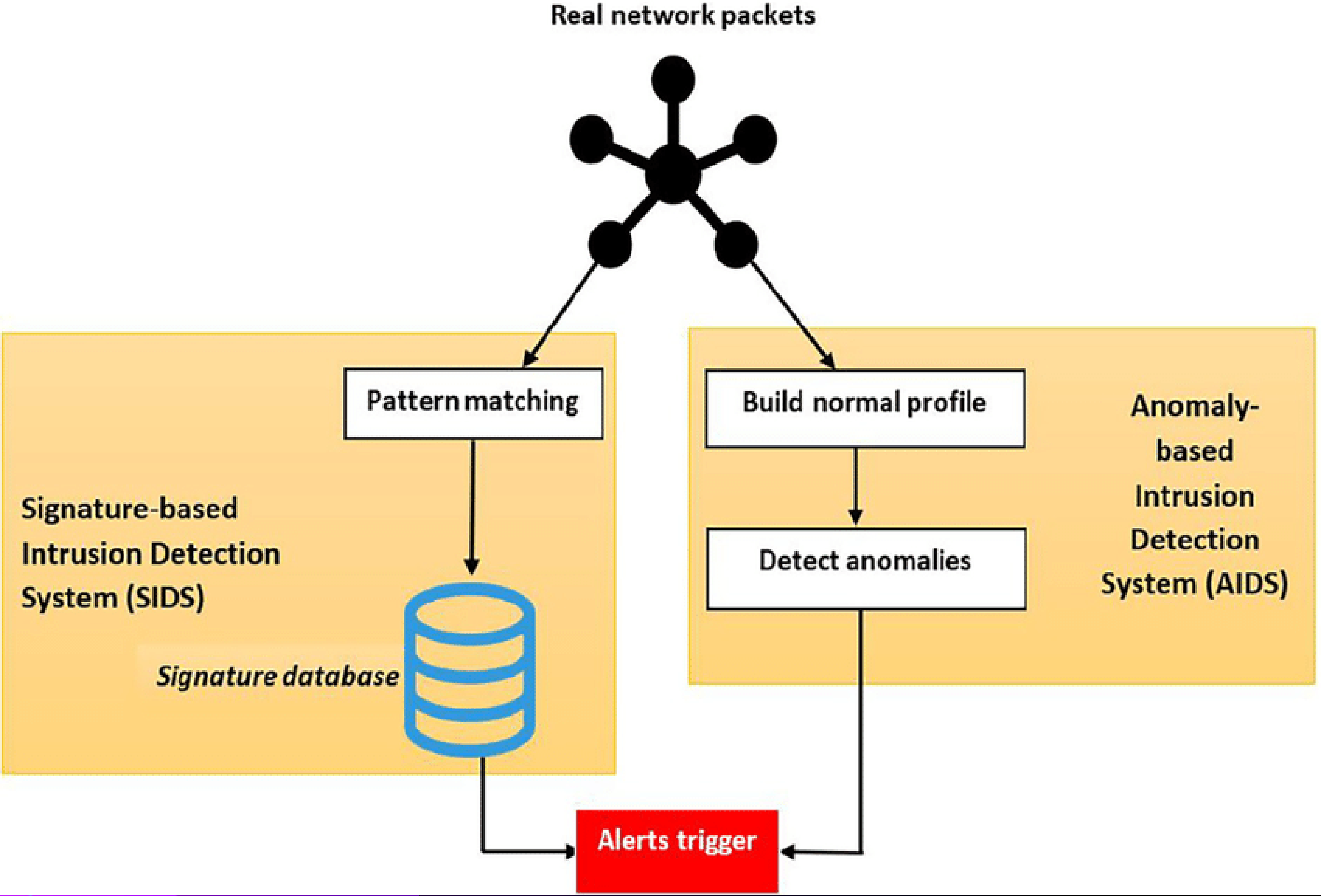
- Goal: Detect any unusual activity in the network
- Challenge: Real-time detection with a very high accuracy
- Solution: Machine learning approach
- Advantage: leveraging data abundancy to efficiently train machine learning models

BACKGROUND

EXPLORING TYPES OF IDS AND DATASET

TYPES OF IDS


Real network packets
Pattern matching
Signature-based Intrusion Detection System (SIDS)
Signature database
Build normal profile
Detect anomalies
Anomaly-based Intrusion Detection System (AIDS)
Alerts trigger



Anomaly-based intrusion detection system:

- Defines profiles of normal user behavior
- Detects deviations from normal patterns [2, 3]

Signature based IDS

- Defines unique signatures for known attacks
- Stores signatures in a database
- Matches network activity against signatures [2, 3]

LIMITATIONS OF SIDS AND AIDS

Signature-Based IDS (SIDS):

- Fails to detect new types of attacks
- Requires a huge extensive database containing signatures of known attacks
- Requires high computational requirements [3,4]

Anomaly-Based IDS (AIDS)

- Difficulty distinguishing normal vs. abnormal
- IoT devices complicate profile definition [3, 4]

Advantages:

- Can detect novel and new attack types
- Adaptive to changing patterns [2]

DATASETS

KDD1999 AND NSL-KDD

KDD 1999 CUP DATASET

Characteristics:

- Raw TCP/IP traffic capture
- 41 total features (3 qualitative + 38 quantitative
- Binary target variable
- Acquisition from Simulated attacks on U.S Air Force LAN [1]

Features categories:

Basic TCP Features
Extracted from basic TCP Connection behavior

Content Features
Inspected payload and content of connection

Time-based Features
Connections to same host in past 2 seconds

Host-based Features
Same as time-based but with a larger time window

KDD 1999 CUP DATASET

Types Of Network attacks simulated [1]

Denial of Service (DoS)
Overloading computing and memory resources

User to Root (U2R)
Authentic account access then exploits vulnerability to gain root access

Remote to Local (R2L)
Sending unauthorized packets to a machine to gain access

Probing Attack
Attempting to gather network information to breach security and gain access.

ISSUES IN DATASET

Major Problems [1, 5]:

- Redundant Records
- Class imbalance
- Lack of labelled validation dataset

Modified KDD Data set (NSL-KDD)

- Removed redundant rows
- Tailored for better performance
- Includes binary and multi-class labels
- Includes proper validation dataset with labels

ISSUES IN DATASET

	Original Records	Distinct Records	Reduction Rate
Attacks	3,925,650	262,178	93,32%
Normal	972,781	812,814	16,44%
Total	4,898,431	1,074,992	78,05%

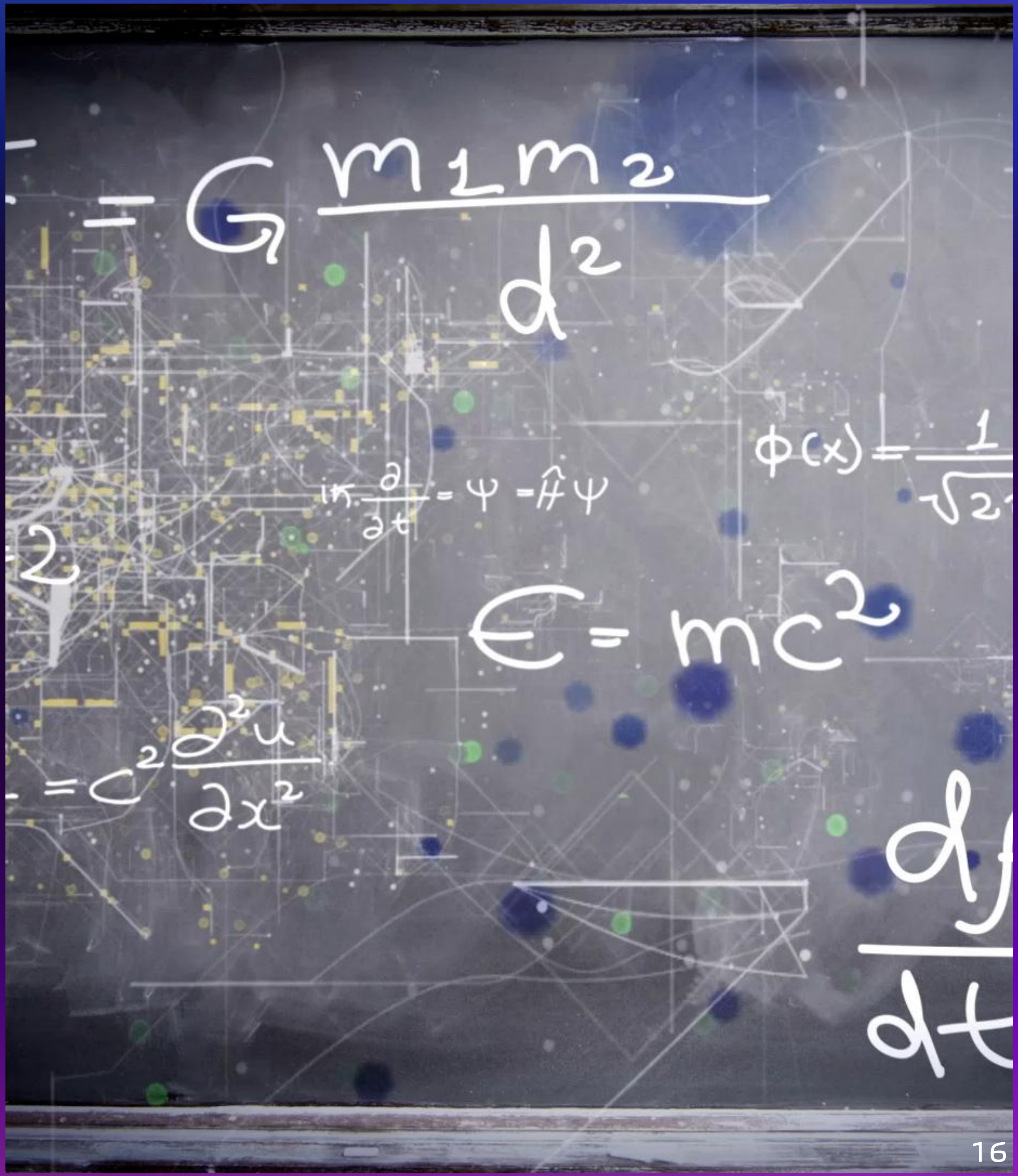
Statistics of Redundant records in KDD Train Set

DATA PREPROCESSING

- Removed always zero features
- Applied Hot Encoding for categorical vriables
- Used principal component analysis *PCA) f
- Scaled the data and dropped low importance features

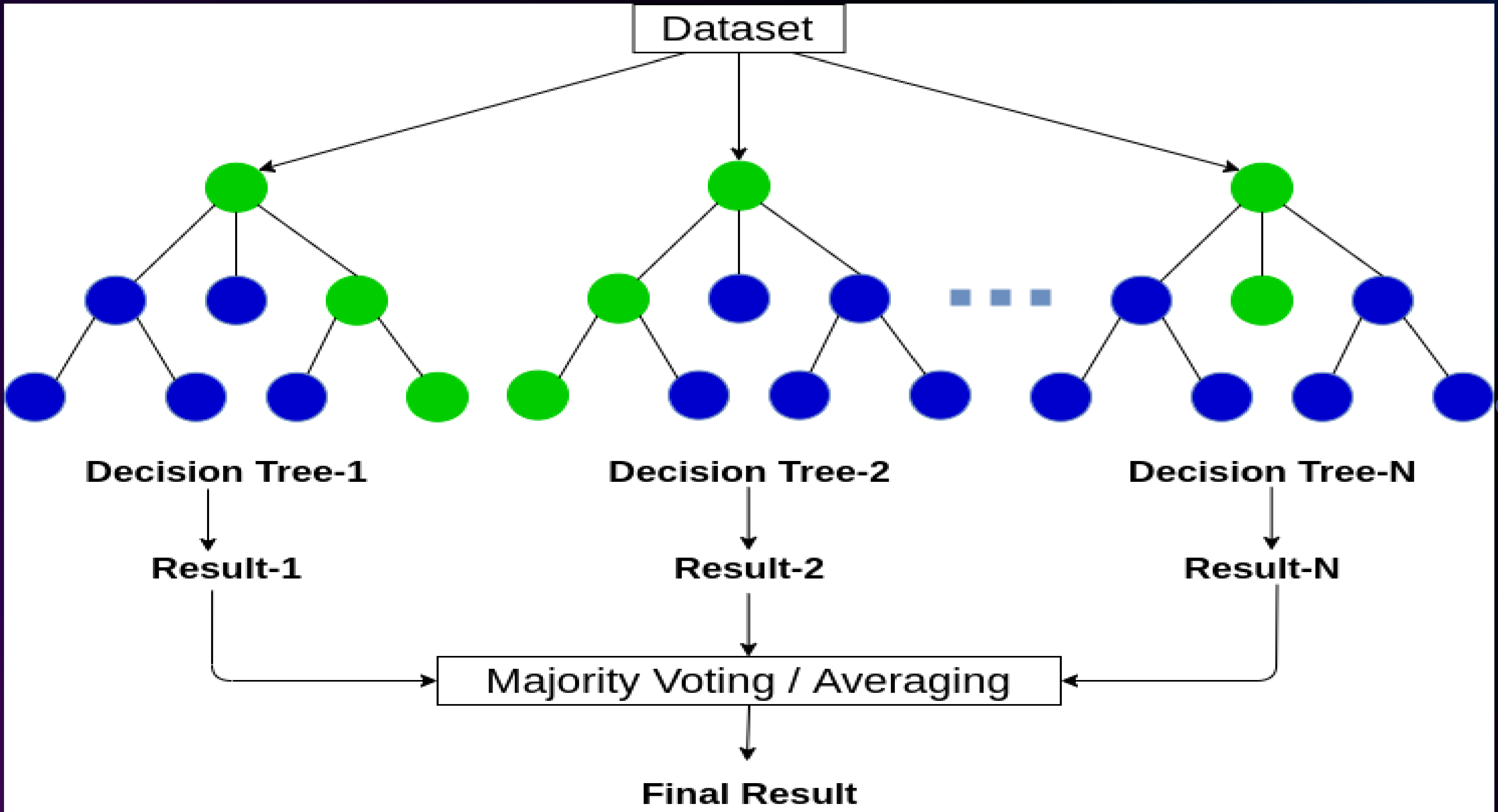
MACHINE LEARNING

MODEL



RANDOM FOREST CLASSIFIER

Algorithm overview:
Random Forest combines multiple decision trees to create a robust classifier.



PERFORMANCE ANALYSIS

Binary Splitting:
$R_1(j,s) = \{x \mid x_j < s\}$
$R_2(j,s) = \{x \mid x_j \geq s\}$

Gini Index (Purity Measure):
$G(t) = 1 - \Sigma_{k=1}{}^{K} p_k{}^2$

- Uses **Recursive binary** splitting to partition data
- **Gini Index** to measure the purity of the split

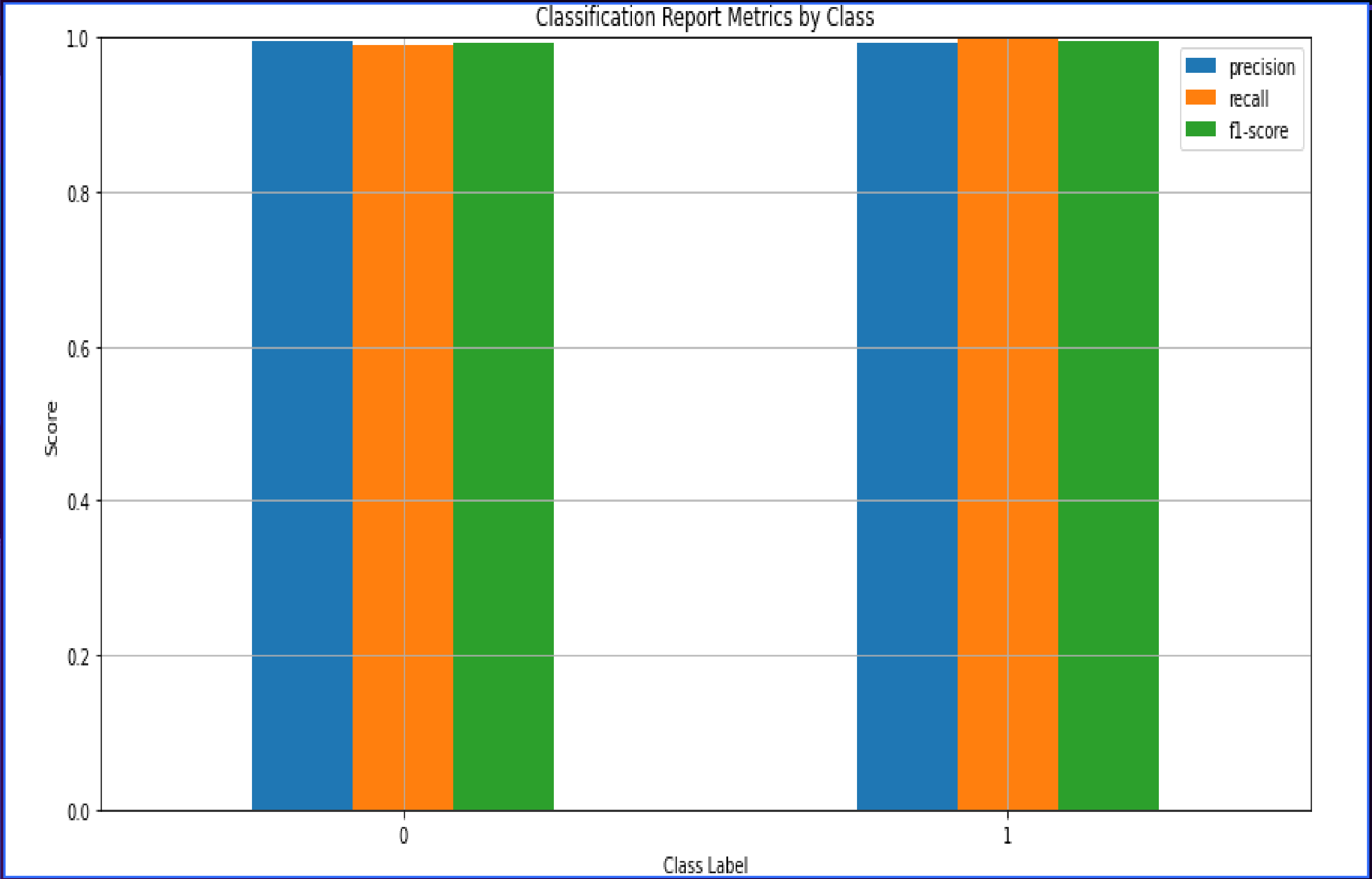
Goal: Find best parameters (j,s) that would result in the purest split

Challenge: Easy to overfit when trees become too deep

PERFORMANCE ANALYSIS

PERFORMANCE ANALYSIS

- KDD 1999 Dataset
  - Splitting the training data due to lack of labels in the validation dataset

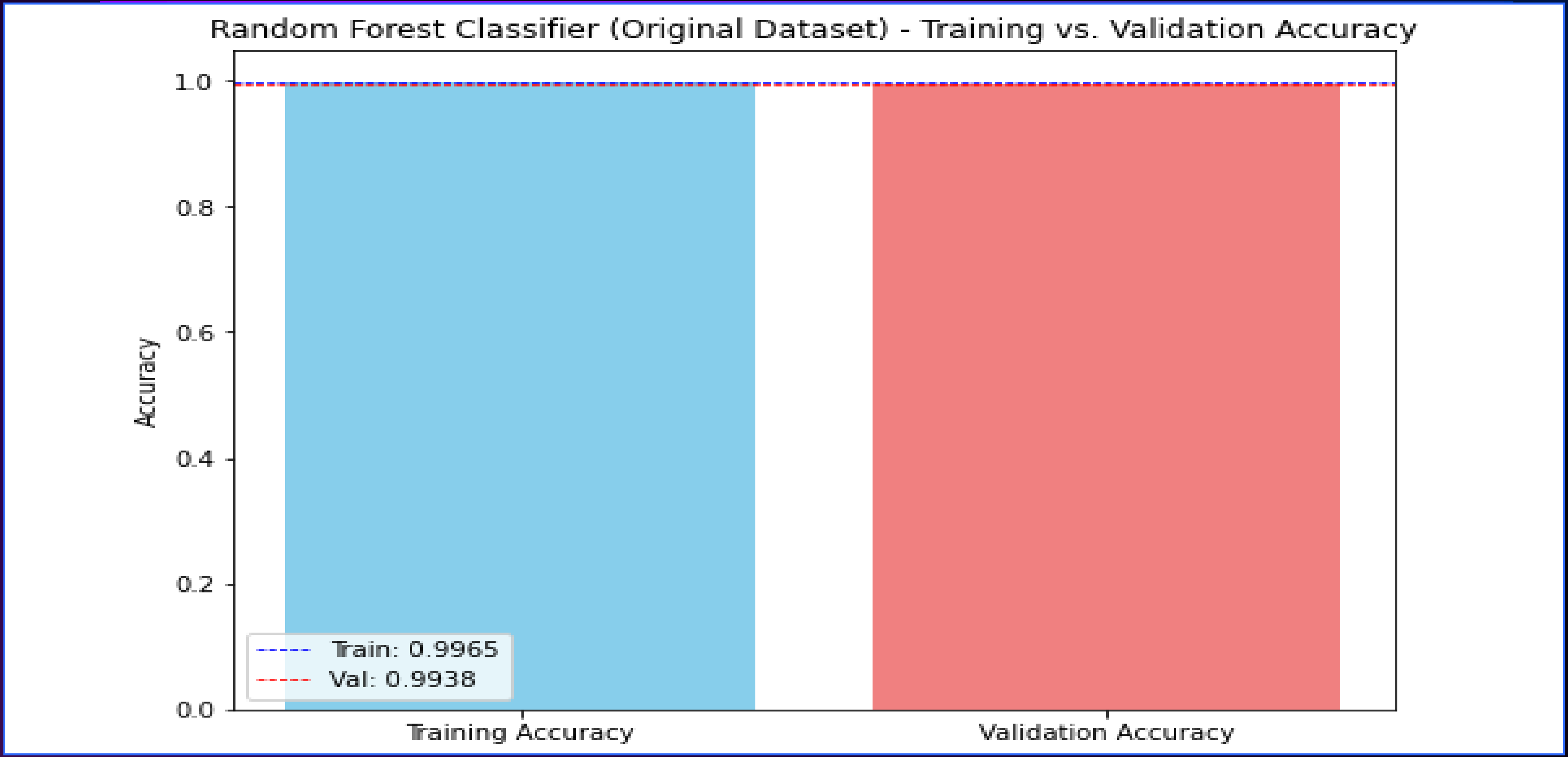


Probable issues:

- High accuracy on both training and validation does not guarantee real world accuracy
- Model likely memorizing data patterns as data is similar due to redundancy

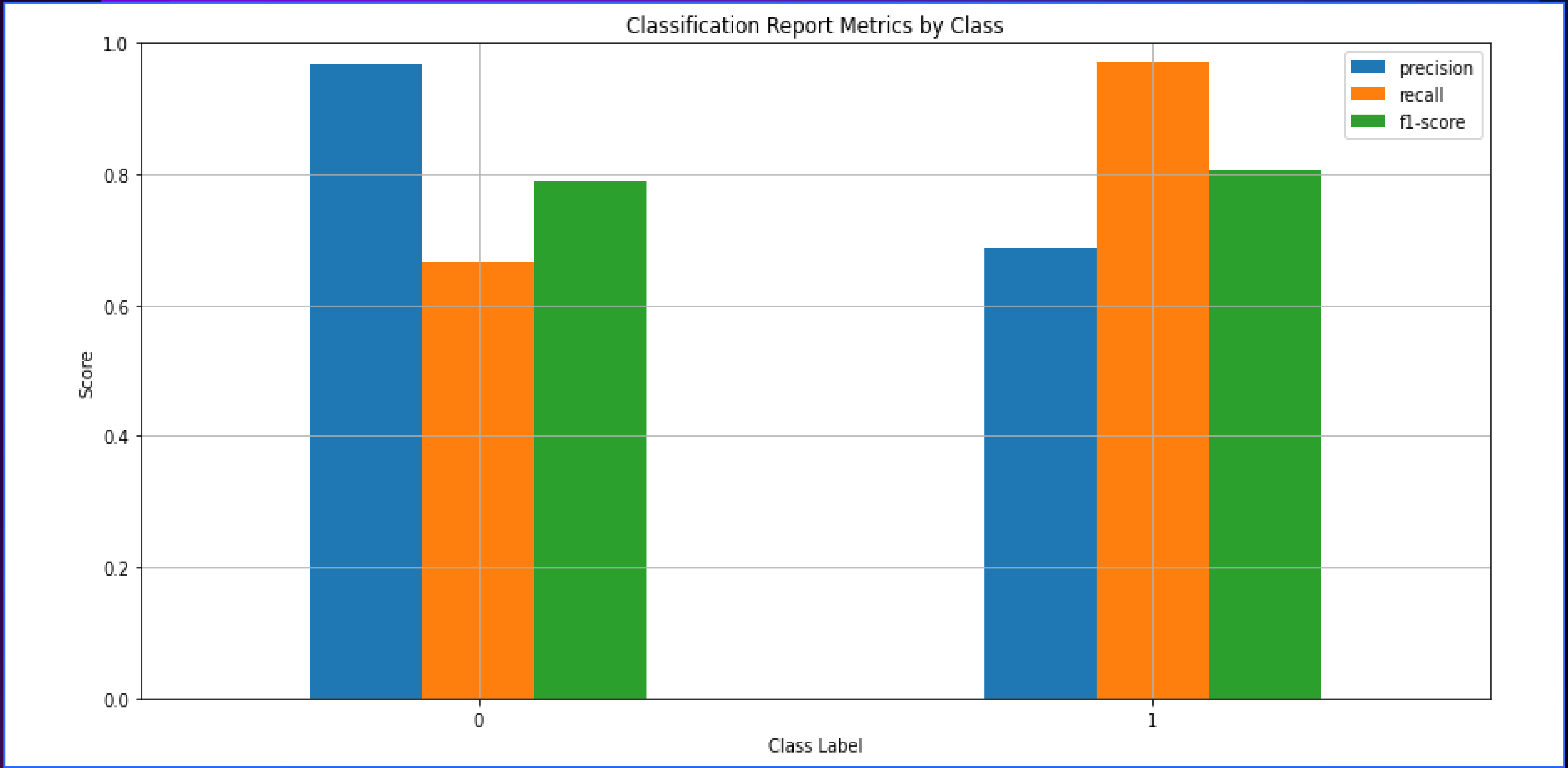
PERFORMANCE ANALYSIS

- KDD 1999 Dataset



PERFORMANCE ANALYSIS

- NSL- KDD 1999 Dataset
  - Using the validation set provided



PERFORMANCE ANALYSIS

NSL- KDD 1999 Dataset

Using the validation set provided

Initial Results when splitting the training dataset:

- Was comparable to the original provided data

Model Optimization:

- **Hyperparameter Tuning**
- **PCA**
- **Data Scaling**

Issues to be addressed:

The model showed significant overfitting with training accuracy at 99% while validation remained at 79,9% after extensive tuning

PERFORMANCE ANALYSIS

NSL- KDD 1999 Dataset
Using the validation set provided

Can it be used:

✓ Strong Attack Detection

- Excellent at identifying actual network attacks
- Low false negative rate for attacks

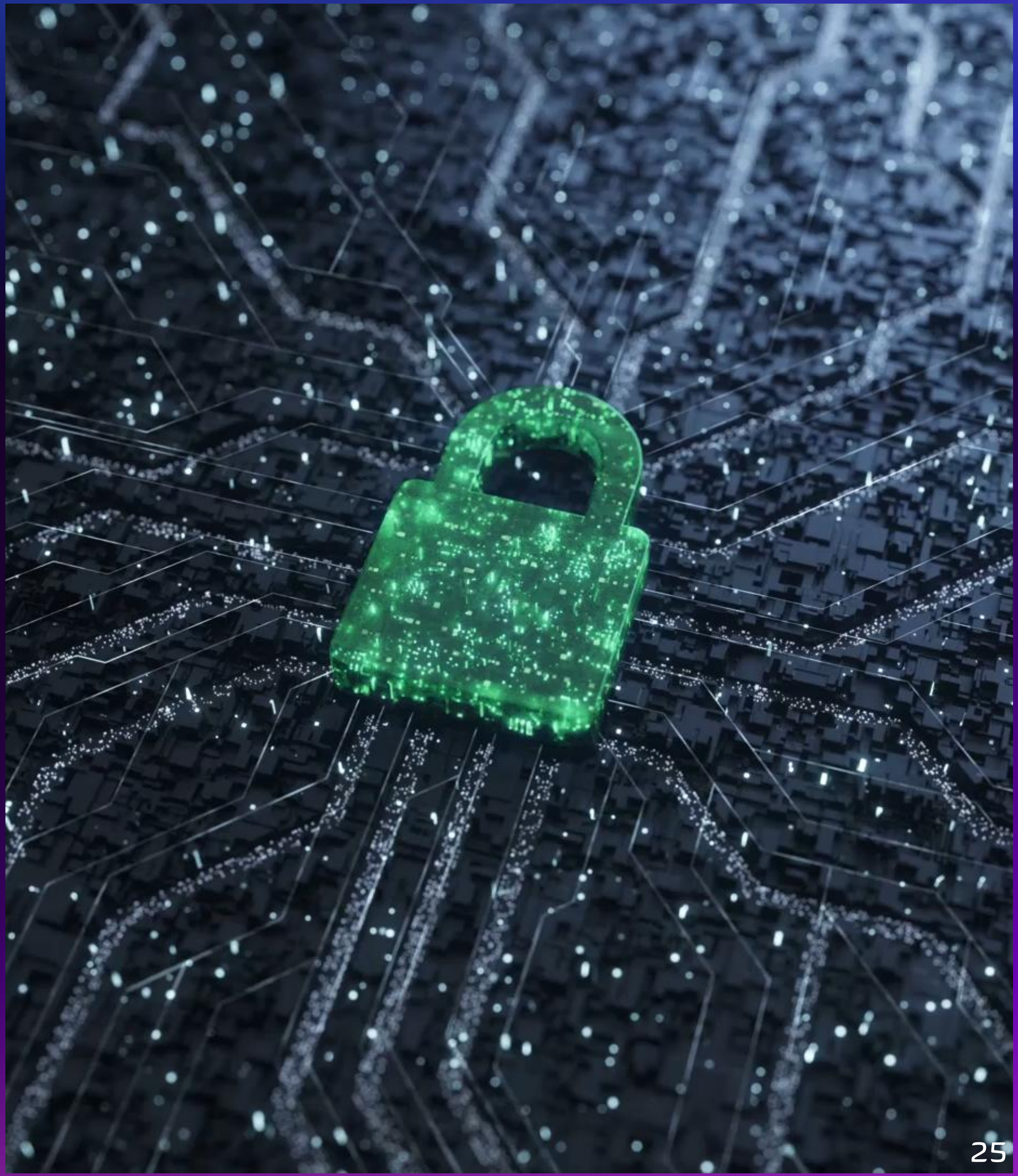
✘ Poor Normal Traffic Recognition

- High false positive rate
- Would flag about 31% of normal connections as probable attacks

Implication:
The High false positive rate would make it impractical for deployment without additional filtering mechanisms

KEY FINDINGS

AND CHALLENGES



KEY FINDINGS AND CHALLENGES

Dataset quality is critical:
Dataset quality can introduce bias and misleading results

Technical Challenges:
- Overfitting
- Class imbalance
- Generalization
- Feature engineering

Main conclusions?
- **Trade-off : Security vs usability balance**
- **Validation importance: proper validation data is crucial for realistic performance assessment**
- **The need for more clean appliable data for networks intrusion**
- **More Investigation into machine learning algorithms and advanced techniques**

THANK YOU

Mohamed Amer

Hochschule Hamm-Lippstadt

Autonomous Systems A

Summer Semester

REFERENCES

[1] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 2009, pp. 1-6, doi: 10.1109/CISDA.2009.5356528. keywords: {Testing;Intrusion detection;Data security;Statistical analysis;Computer security;Computer aided manufacturing;Learning systems;Computational intelligence;Computer networks;Application software},

[2] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F.Ahmad, ”Network intrusion detection system: A systematic study ofmachine learning and deep learning approaches,” Trans. EmergingTelecommun. Technol., vol. 32, no. 1, e4150, 2021. [Online]. Available:https://doi.org/10.1002/ett.4150

[3] W. Ma, ”Analysis of anomaly detection method for Internet of Thingsbased on deep learning,” Trans. Emerg. Telecommun. Technol., vol. 31,no. 6, e3893, 2020. [Online]. Available: https://doi.org/10.1002/ett.3893

[4] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S.Guizani, ”Internet-of-Things-based smart cities: Recent advances andchallenges,” IEEE Commun. Mag., vol. 55, no. 9, pp. 16–24, Sep. 2017.[Online]. Available: https://doi.org/10.1109/MCOM.2017.1600514

[5] A. R. Tapsoba and T. Fr´ed´eric OUEDRAOGO, ”Evaluation of super-vised learning algorithms in binary and multi-class network anomaliesdetection,” 2021 IEEE AFRICON, Arusha, Tanzania, United Republicof, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570886.keywords: Training;Supervised learning;Support vector machineclassification;Predictive models;Prediction algorithms;Featureextraction;Classification algorithms;Intrusion Detection System(IDS);Supervised Learning Algorithms (SLA);Recursive FeatureElimination (RFE);AUC - ROC Curve;NSL-KDD,