UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Enhancing IoT Security with Explainable AI-Powered Intrusion Detection System

Sani Abdullahi Sani[1*], Dr Ibidun Obagbuwa[2]

[1]School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.

[2]Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley, Northern Cape, South Africa.

Motivation

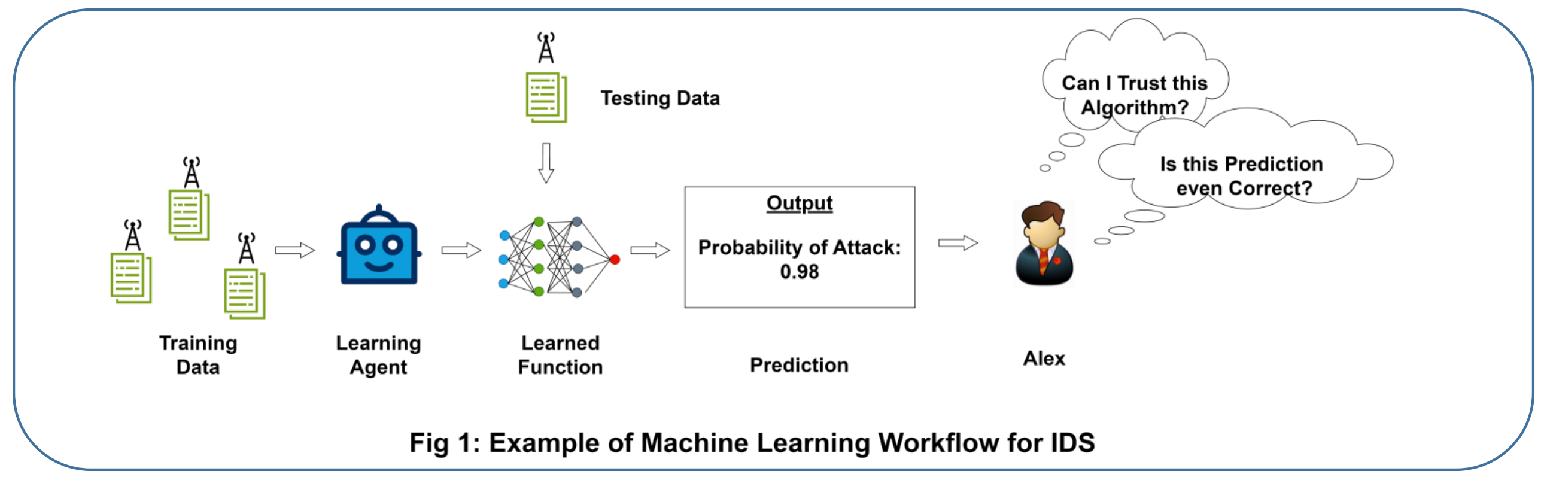

Testing Data
Can I Trust this Algorithm?
Is this Prediction even Correct?
Output
Probability of Attack: 0.98
Training Data
Learning Agent
Learned Function
Prediction
Alex

Fig 1: Example of Machine Learning Workflow for IDS

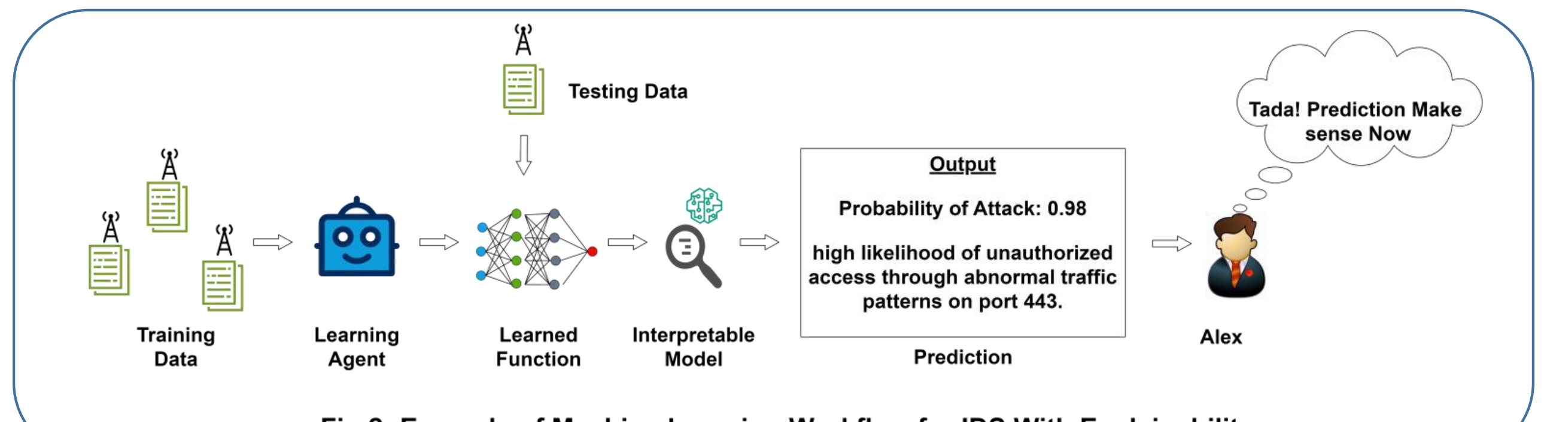

Testing Data
Tada! Prediction Make sense Now
Output
Probability of Attack: 0.98
high likelihood of unauthorized access through abnormal traffic patterns on port 443.
Training Data
Learning Agent
Learned Function
Interpretable Model
Prediction
Alex

Fig 2: Example of Machine Learning Workflow for IDS With Explainability

- The **Internet of Things (IoT)** emerges as a transformative technological paradigm.
- According to **Cisco**, the global count of IoT devices is anticipated to hit around **30 billion** by **2030.**
- However, its **heterogeneity** and **dependency** on other technologies like **Fog** and **Cloud** make it **vulnerable**.
- Robust **security** measures are **imperative**, but **conventional** solutions are inadequate due to **resource constraints**.
- **AI-Based Intrusion Detection Systems (IDS)** show great promise but lack **transparency**.

Aim

- To develop an Explainable AI-powered IDS to enhance security in IoT networks

Objectives

- To design and implement IDS using DT, RF, XGBoost, MLP, and CNN classifiers
- To investigate feature engineering methods and analyze how
- features affect the accuracy of IDS
- To compare the predictive performance of white box models and black box models
- To integrate XAI techniques, thereby providing transparent and interpretable explanations for model decisions

Methodology

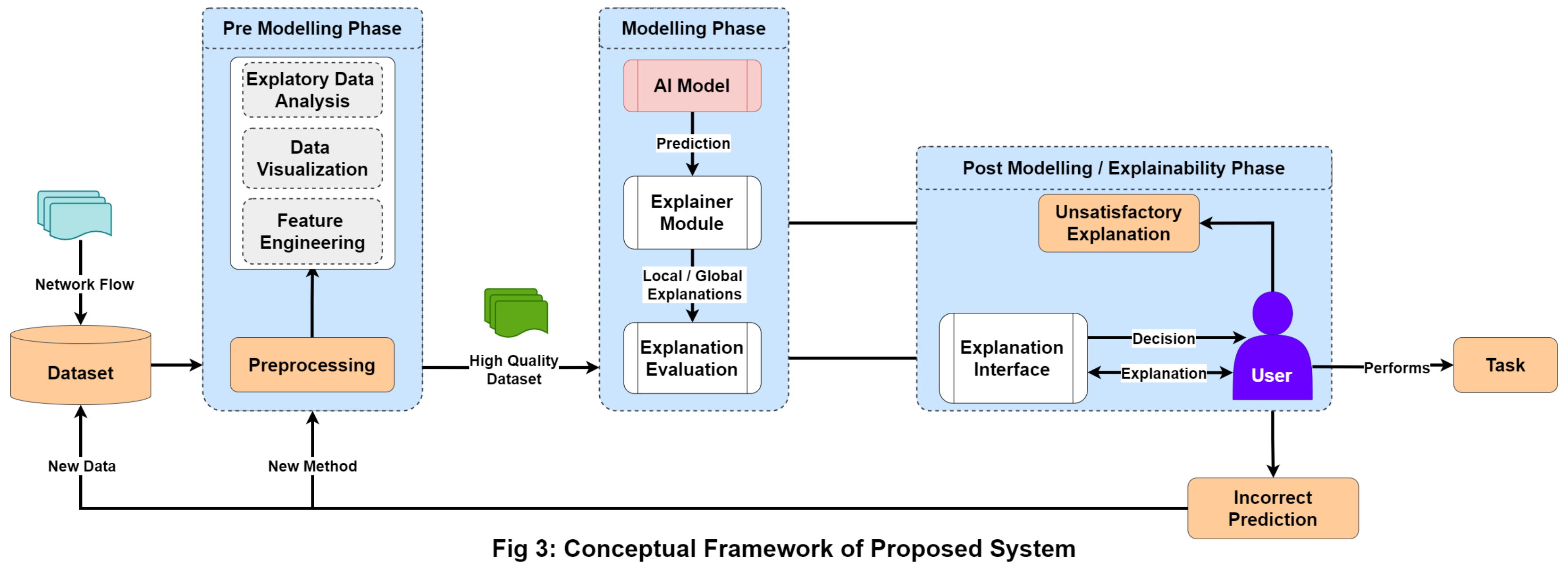


Fig 3: Conceptual Framework of Proposed System

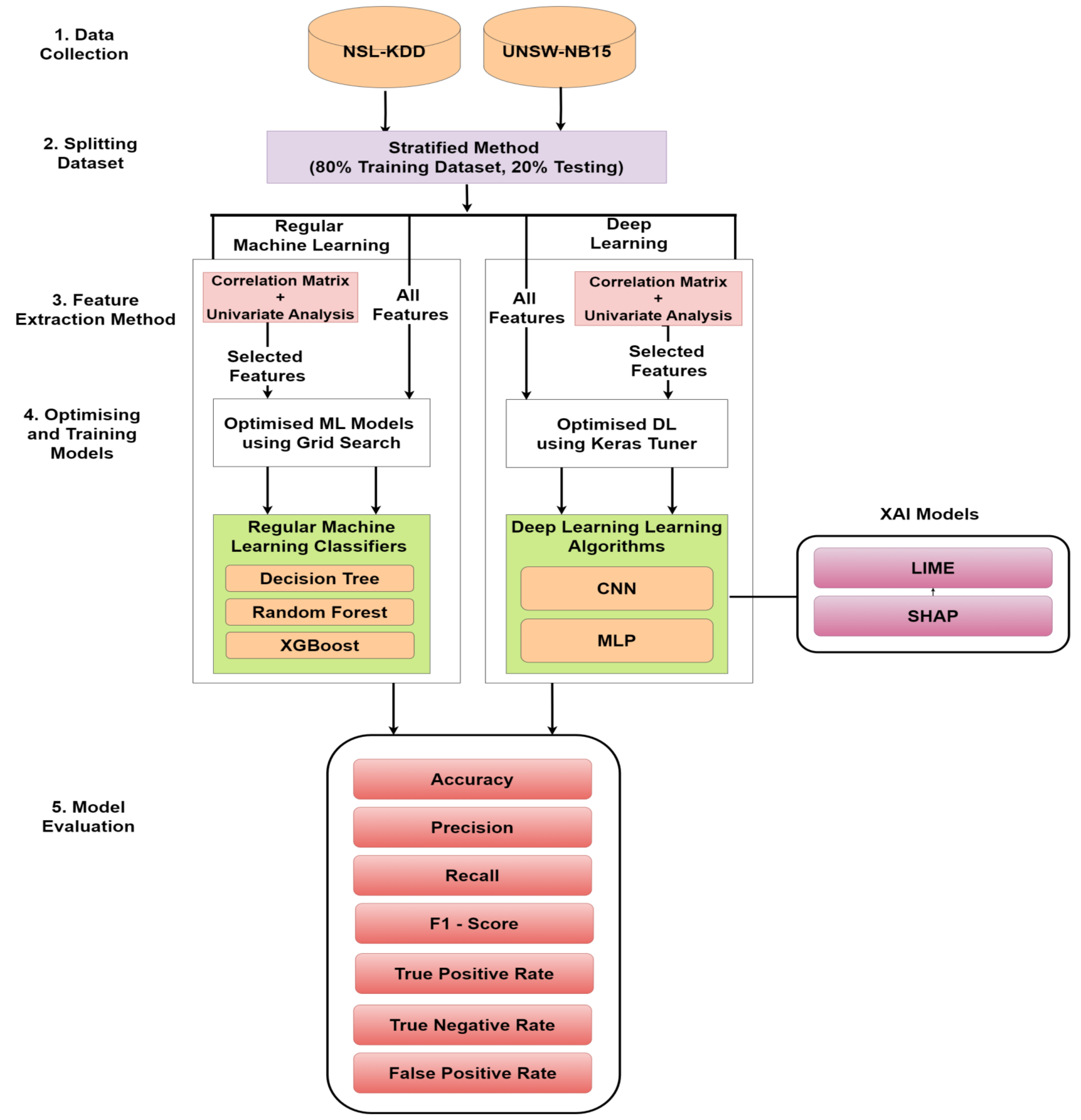
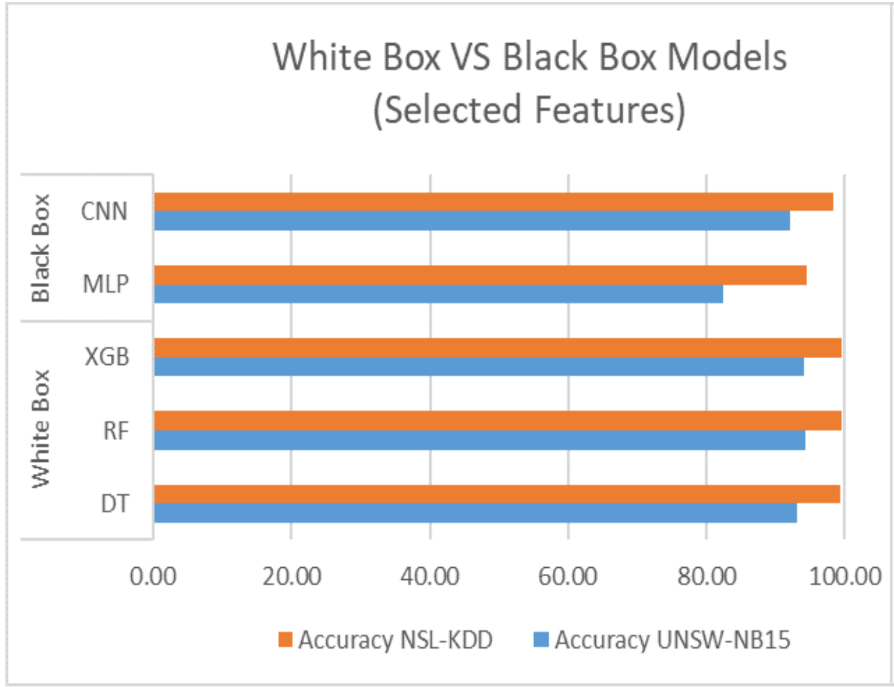
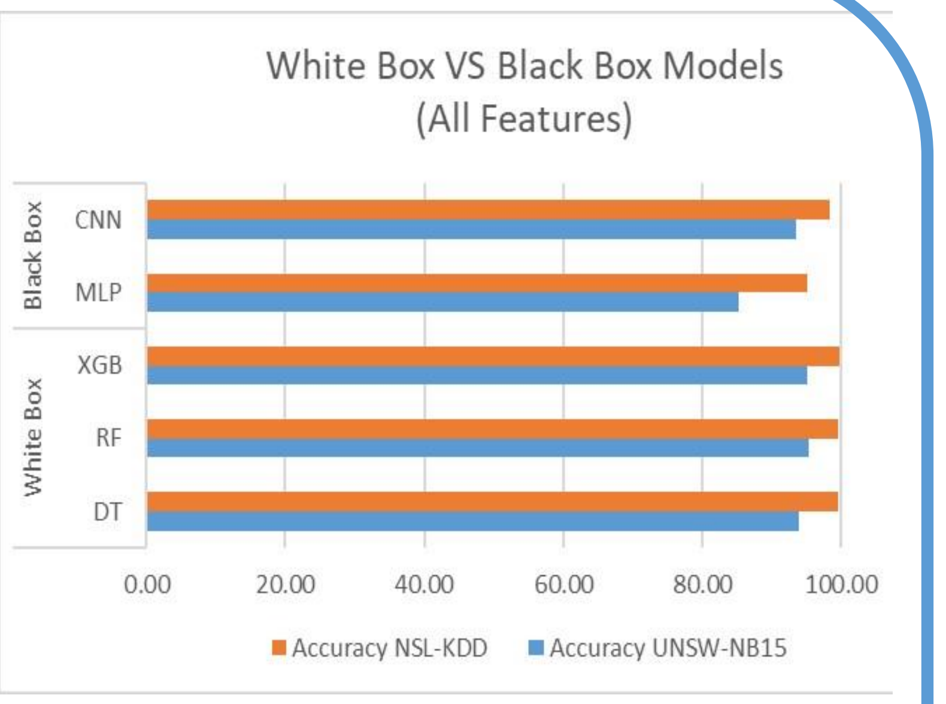
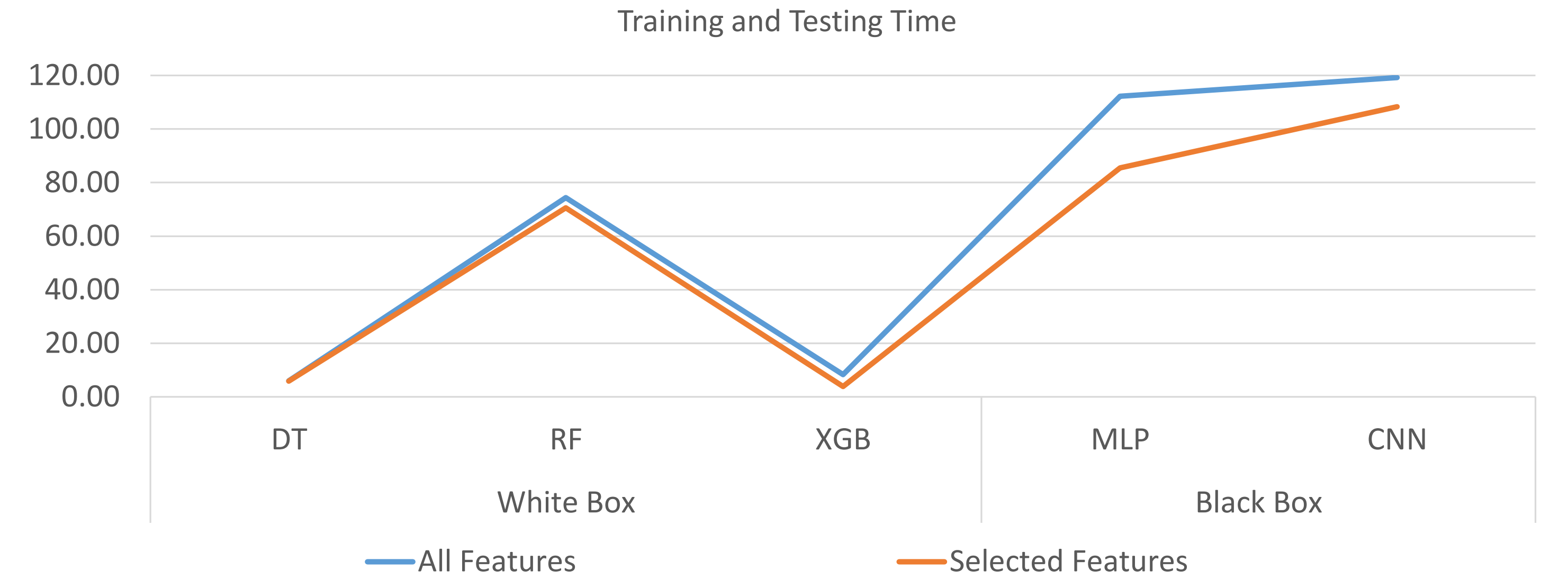

1. Data Collection
NSL-KDD
UNSW-NB15
2. Splitting Dataset
Stratified Method (80% Training Dataset, 20% Testing)
Regular Machine Learning
Deep Learning
3. Feature Extraction Method
Correlation Matrix + Univariate Analysis
All Features
All Features
Correlation Matrix + Univariate Analysis
Selected Features
Selected Features
4. Optimising and Training Models
Optimised ML Models using Grid Search
Optimised DL using Keras Tuner
Regular Machine Learning Classifiers
Decision Tree
Random Forest
XGBoost
Deep Learning Learning Algorithms
CNN
MLP
XAI Models
LIME
SHAP
5. Model Evaluation
Accuracy
Precision
Recall
F1 - Score
True Positive Rate
True Negative Rate
False Positive Rate

Fig 4: The workflow of Machine learning and Deep learning Models.

Results







- **Feature selection** shows improvement in computational efficiency though with a slight drop in performance**.**
- **White box models** outperforms **black box models**, with **Random Forest** achieving the highest performance.

Conclusion

- We developed an **IDS** with **XAI** approach.
- Phases includes **pre-modelling, modelling, and post-modelling,** focusing on data quality, model performance, and interpretability.
- We showed that cybersecurity capabilities can be enhanced by providing **actionable insights to stakeholders** for **informed** decision-making and **threat mitigation**.

References

- Gunning, D. & Aha, D. (2019). DARPA's explainable AI (XAI) program. AI Mag, 40(2), 44-58.
- Wang, M. et al. (2020). An explainable ML framework for intrusion detection systems. IEEE Access, 8, 73127-73141.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Google DeepMind

SOL PLAATJE UNIVERSITY