**Hybrid Host-Based Intrusion Detection System Using System Call Analysis  
(Aryan Vats, Nitesh Jha, Shikhar Sharma)  
(Mentored by: Dr. Anand Kumar Mishra)**

# Concept Note

\*\*Background of Research:\*\*

With increasing sophistication in cyber-attacks, traditional security solutions such as firewalls and antivirus software are insufficient. Host-Based Intrusion Detection Systems (HIDS) that analyze system calls are an emerging solution. However, they suffer from challenges including high false positives, inefficiency in real-time detection, and poor handling of stealth attacks.

\*\*Research Statement / Problem:\*\*

Existing HIDS models often fail to detect zero-day and stealth attacks, struggle with dataset imbalance, and lack computational efficiency for practical deployment. A scalable, adaptive system is required.

\*\*Objectives:\*\*

- Design a hybrid HIDS using classical machine learning (ML) and deep learning (DL) approaches.

- Improve detection accuracy, recall, and computational efficiency.

- Ensure adaptability across Linux and Windows operating systems.

\*\*Methodology Adopted:\*\*

Three datasets were used: ADFA-LD (Linux), ADFA-WD (Windows), and ADFA-WD:SAA (stealth attacks).

- For ADFA-LD: TF-IDF (1-3 grams) feature extraction with XGBoost classifier.

- For ADFA-WD: TF-IDF 5-gram features from initial system calls, stacking ensemble learning (KNN, RF, DT as base models; XGBoost, AdaBoost, LightGBM as meta models).

- For ADFA-WD:SAA: CNN, LSTM, and CNN+LSTM architectures combined with TF-IDF and SVD features.

\*\*Data Set Used:\*\*

- ADFA-LD Dataset (Linux syscall traces)

- ADFA-WD Dataset (Windows DLL call traces)

- ADFA-WD:SAA Dataset (Windows stealth attack traces)

\*\*Key Findings:\*\*

- ADFA-LD: 97.2% accuracy with TF-IDF + XGBoost.

- ADFA-WD: 90.05% accuracy, 94.78% recall using stacking ensemble.

- ADFA-WD:SAA: 93.5% accuracy with CNN+LSTM model.

\*\*Significance / Implications:\*\*

The proposed HIDS framework is lightweight, highly accurate, scalable, and suitable for real-time deployment across diverse platforms including enterprise, government, and cloud infrastructures.

# Graphical Abstract

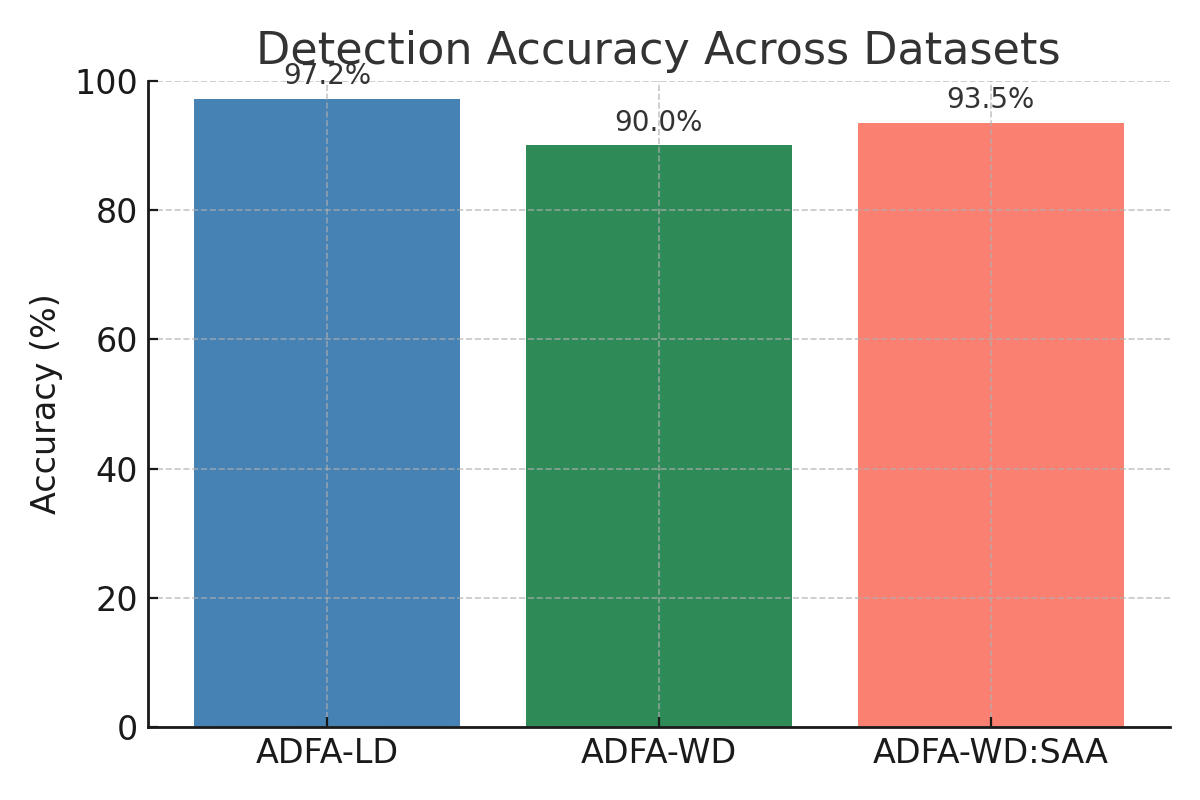


Fig 1: Detection Accuracy Across Datasets

|  |  |  |
| --- | --- | --- |
| Dataset | Best Model | Accuracy (%) |
| ADFA-LD | XGBoost + TF-IDF | 97.2 |
| DFA-WD | Stacking Ensemble | 90.05 |
| ADFA-WD:SAA | CNN+LSTM | 93.5 |

**Fig 2: Summary of Best Model Performance Across Datasets**