Application of various Machine Learning models on real cybersecurity data (BETH dataset)

Jamel Belgacem & Papa Moryba Kouate

GitHub

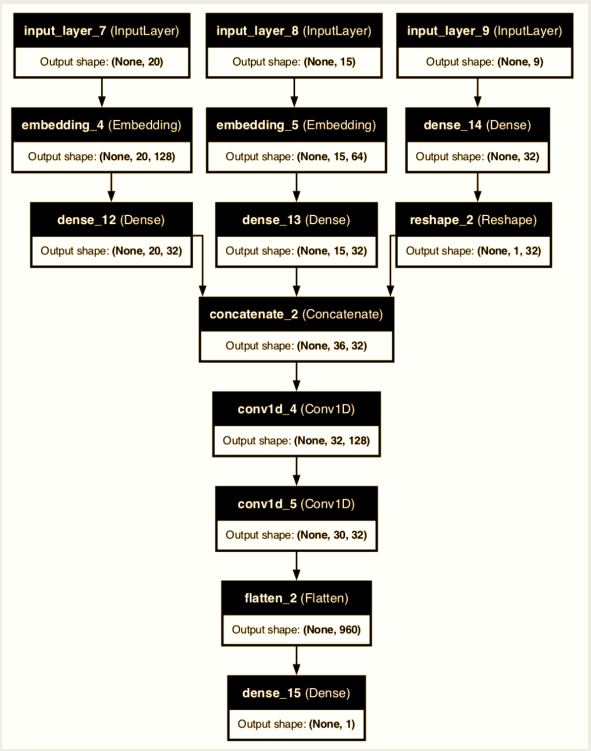
Abstract

The BETH dataset addresses a critical need in cybersecurity research: the availability of real-world, labelled data for anomaly detection. Unlike synthetic datasets, BETH captures genuine host activity and attacks, making it a valuable resource for developing robust machine learning models.

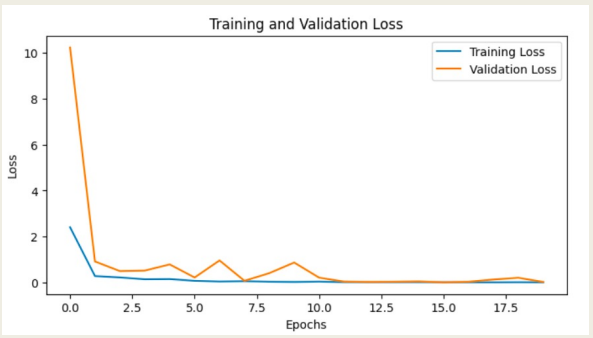
The scale, diversity, and structured heterogeneity of BETH dataset makes it an invaluable resource for advancing anomaly detection techniques and enhancing the robustness of machine learning models in the cybersecurity domain.

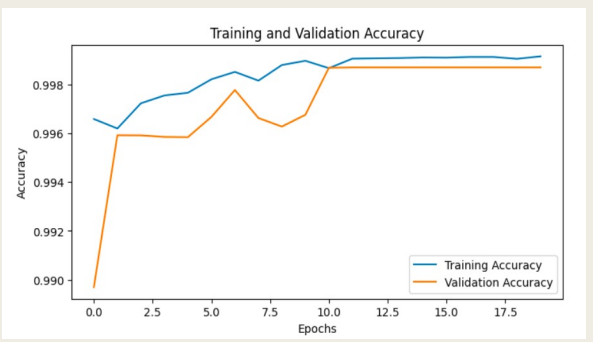
Purpose and modelisation

As Artificial Intelligence advances, the cybersecurity sector is placing greater emphasis on comprehending the potential effects and applications of AI on its methodologies and infrastructure. Therefore, the primary objective of this project is to investigate and evaluate how various machine learning models can be leveraged to effectively tackle the issue of anomaly detection, which is crucial for identifying unusual patterns or behaviours that may indicate security threats within cybersecurity systems

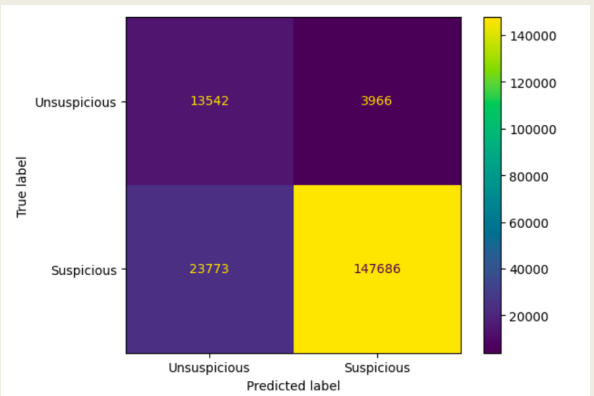


Convolutional neural network architecture





Training plots



Prediction

Discussion

Our experiments on the BETH dataset reveal that while most machine learning models, including CNNs, RNNs, and transformers, achieved high accuracy in detecting anomalies, the imbalance in the dataset posed significant challenges.

The transformer model, enhanced by positional encoding and multi-layer self-attention, demonstrated superior performance in identifying complex patterns.

However, techniques like SMOTE for data balancing and Shapelet Discovery for feature extraction showed mixed results, highlighting the need for more advanced methods.

Future improvements could involve integrating advanced data augmentation, feature engineering, model hybridization, and more efficient algorithms to enhance sensitivity to suspicious behavior.

Model	Accuracy	Precision avg	Recall avg	ROC score
Dense model	0.09	0.05	0.50	0.50
Dense model + embeddings	0.91	0.75	0.95	0.95
CNN model	0.11	0.53	0.51	0.51
CNN model + embeddings	0.95	0.82	0.97	0.97
RNN model	0.09	0.05	0.50	0.50
RNN model + embeddings.	0.95	0.82	0.97	0.97
Transformer	0.95	0.82	0.97	0.97

Shapelet Discovery method with 230.000 data points

Model	Accuracy	Precision avg	Recall avg	ROC score
Decision tree classifier	0.97	0.98	0.97	0.97
LSTM model	0.45	0.35	0.41	0.41
Dense Model	0.96	0.97	0.95	0.95

References

BETH Dataset: Real Cybersecurity Data for Anomaly Detection Research
Kate Highnam, Kai Arulkumaran, Zachary Hanif, Nicholas R. Jennings

Smote: Synthetic Minority Over-sampling Technique
Authors: Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer

Time Series Shapelets: A New Primitive for Data Mining.
Lexiang Ye, Eamonn Keogh