# Applying Ensemble Tree-Based Models and Explainable AI for IoT Botnet Detection in Heterogeneous Device

**M Ilham Yushronni[1], Parman Sukarno[2], Aulia Arif Wardana[3]**

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]hamyush@students.telkomuniversity.ac.id, [2]psukarno@telkomuniversity.ac.id,
[3]auliawardan@telkomuniversity.ac.id,

## Abstrak

**Serangan botnet menimbulkan risiko keamanan yang signifikan, membuat Internet of Things (IoT) semakin rentan. Keamanan sistem IoT merupakan aspek penting dalam mendeteksi serangan botnet. Kemanjuran keamanan sistem IoT dalam mendeteksi serangan botnet sangatlah penting. Meskipun pendekatan Machine Learning (ML) telah menunjukkan kemanjuran dalam hal ini, keterbatasan yang berkaitan dengan interpretabilitas dan transparansi model tetap menjadi tantangan yang cukup besar. Akibatnya, penelitian saat ini dilakukan dengan tujuan untuk meningkatkan interpretabilitas model ML, sehingga memfasilitasi pembuatan penjelasan yang lebih berdasar melalui pemanfaatan kecerdasan buatan yang dapat dijelaskan (XAI). Kumpulan data N BaIoT akan digunakan untuk melatih model ensemble untuk mendeteksi serangan botnet, dengan model mempelajari pola dari sembilan perangkat yang berbeda. Gradient Boost dengan Early Stopping, Catboost, dan Histogram Gradient Boost telah dipilih untuk mendeteksi serangan botnet, dengan model yang dirancang untuk menangani kumpulan data yang besar dan tidak seimbang secara efisien. Teknik Early Stopping digunakan untuk meminimalkan risiko overfitting. Untuk meningkatkan transparansi dan interpretabilitas, metode XAI seperti SHapley Additive Explanations (SHAP) dan Local Interpretable Model Agnostic Explanations (LIME) digunakan untuk mengidentifikasi fitur yang paling berpengaruh pada prediksi dan dengan demikian memperkuat keandalan model dalam deteksi serangan botnet. Temuan penelitian menunjukkan bahwa Device 7 mencapai akurasi tertinggi di antara ketiga model ensemble yang digunakan, dengan model Gradient Boosting dengan Early Stopping mencapai akurasi 99,94%, Catboost mencapai 99,98% dan Histogram Gradient Boosting juga mencapai 99,98%. Selain itu, SHAP dan LIME telah berhasil mengidentifikasi fitur utama yang memengaruhi prediksi kelas botnet, sementara juga mengungkap korelasi antara fitur yang berkontribusi untuk meningkatkan transparansi model. Ini membuktikan penggunaan model ensemble yang didukung oleh pendekatan XAI memiliki potensi besar dalam memahami dan mempercayai model Machine Learning.**

**Kata kunci: Botnet, Catboost, Early Stopping, Explainable Artificial Intelligence, Gradient Boosting, Internet of Things (IoT), Local Interpretable Model-Agnostic Explanations, Machine Learning, SHapley Additive exPlanations**

## Abstract

**Botnet attacks pose significant security risks, making the Internet of Things (IoT) increasingly vulnerable. IoT system security is a crucial aspect in detecting botnet attacks. The efficacy of IoT system security in the detection of botnet attacks is of paramount importance. While Machine Learning (ML) approaches have demonstrated efficacy in this regard, limitations pertaining to model interpretability and transparency persist as considerable challenges. Consequently, the present research was undertaken with the objective of enhancing the interpretability of ML models, thus facilitating the generation of more substantiated explanations through the utilisation of explainable artificial intelligence (XAI). The N BaIoT dataset will be used to train an ensemble model to detect botnet attacks, with the model learning patterns from nine different devices. Gradient Boost with Early Stopping, Catboost and Histogram Gradient Boost have been selected to detect botnet attacks, with the model designed to handle large and unbalanced datasets efficiently. The Early Stopping technique is used to minimise the risk of overfitting. To enhance trans parency and interpretability, XAI methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME) are employed to identify the most influential features on predictions and thereby strengthen the reliability of the model in botnet attack detection. The research findings indicate that Device 7 achieves the highest accuracy among the three ensemble models utilised, with the Gradient Boosting with Early Stopping model achieving an accuracy of 99.94%, Catboost achieving 99.98% and Histogram Gradient Boosting also reaching 99.98%. Additionally, SHAP and LIME have been successful in identifying key features that influence botnet class predictions, while also unveiling correlations between features that contribute to enhancing model**

**transparency. This proves the use of ensemble models supported by XAI approaches has great potential in under standing and trusting Machine Learning model.**

**Keywords: Botnet, Catboost, Early Stopping, Explainable Artificial Intelligence, Gradient Boosting, Internet of Things (IoT), Local Interpretable Model-Agnostic Explanations, Machine Learning, SHapley Additive exPlanations**