Abstract

Malware is a malicious program that executes destructive functions to destroy the resources in a computer system, gain some financial benefits, steal the privacy and confidentiality of data, and use computing resources to make a service unavailable in a computer system. One of the ways to prevent malware attacks is by detecting Portable Executable (PE) malware files using machine learning. However, not all machine learning algorithms have optimal performance in detecting a malware PE File because some have several weaknesses that result in low performance in detecting a malware PE File. However, these shortcomings can be reduced by combining two or more two different individual algorithms into one hybrid machine learning algorithm, so the advantages of some individual algorithms can cover the shortcomings of other individual algorithms. Therefore, this research proposes research on the performance of the hybrid machine learning algorithms in detecting malware PE File. The hybrid machine learning algorithms use the voting classifier method and LightGBM, XGBoost, and Logistic Regression as their base model. This research proves that the hybrid machine learning algorithm produces a higher recall value than the ensemble algorithm LightGBM. The hybrid machine learning algorithm produces the highest recall value with a recall value of 99.5026%, while the LightGBM algorithm only produces a recall value of 99.4480%. Furthermore, the recall value of another base model is 99.5004% for the XGBoost algorithm and 98.0539% for the Logistic Regression algorithm.

Keywords: malware detection, hybrid machine learning, voting classifier, ensemble, PE file