ABSTRAK

Peningkatan insiden serangan eksfiltrasi DNS mengharuskan penerapan sistem deteksi intrusi yang tidak hanya akurat tetapi juga transparan dalam operasinya, sehingga memastikan pemahaman keputusan model oleh pengguna manusia. Untuk mengatasi tantangan ini, studi ini mengusulkan penerapan dan perbandingan berbagai metode Kecerdasan Buatan yang Dapat Dijelaskan (XAI) untuk menjelaskan keputusan model Jaringan Neural Dalam (DNN) dalam konteks deteksi serangan eksfiltrasi DNS. Metodologi inti yang digunakan meliputi pembangunan model Jaringan Saraf Dalam (DNN) dengan arsitektur piramida dan penggunaan enam metode XAI. Global SHAP, PFI, dan ALE merupakan pendekatan yang digunakan untuk penjelasan global, sedangkan local SHAP, LIME, dan Anchor digunakan untuk penjelasan lokal. Data set CIC-Bell-DNS-EXF2021 digunakan untuk evaluasi. Hasil menunjukkan bahwa SHAP memberikan interpretasi paling komprehensif, baik secara global maupun lokal, meskipun memerlukan sumber daya pemrosesan yang signifikan, dengan 38,73 detik dan 36,8 byte untuk penjelasan global, serta 5,22 detik dan 31,01 byte untuk penjelasan lokal. Di sisi lain, PFI dan LIME menggunakan sumber daya yang lebih sedikit, yaitu 4,45 byte dalam 28,22 detik dan 27,69 byte dalam 2,58 detik, tetapi memberikan informasi yang kurang komprehensif. ALE mengonsumsi 10,43 byte dalam 51,02 detik, sedangkan Anchor mengonsumsi sumber daya terbanyak dengan 784,59 byte dalam 9,24 detik. Studi ini memberikan kontribusi yang signifikan dengan secara sistematis menganalisis manfaat, kelemahan, dan kemampuan beberapa pendekatan XAI. Temuan ini menekankan pentingnya integrasi XAI untuk meningkatkan transparansi sistem deteksi berbasis kecerdasan buatan. Studi ini menggunakan pendekatan XAI pasca-pengolahan untuk menjelaskan keputusan model DNN dalam mendeteksi serangan DNS Exfiltration.