BAB 1 PENDAHULUAN

1.1 Latar Belakang

Malware (malicious software) adalah perangkat lunak yang dibuat untuk merusak atau menjalankan proses yang dapat membahayakan sistem komputer, contohnya *virus*, *worms*, *trojan*, dan *spyware* [1]. Dengan berkembangnya internet, banyak variasi *malware* baru bermunculan yang seringkali digunakan untuk melakukan penyerangan dan tindakan kejahatan. Sebagai contoh, menurut laporan dari AV Test, dalam periode 2011 – November 2020, melaporkan sebanyak total jumlah *malware* yang terdeteksi sebanyak 1113.73 juta dan sebanyak 267.23 juta varian *malware* baru muncul dalam kurun 2 tahun terakhir [2]. Oleh karena itu, diperlukan sistem deteksi dan klasifikasi *malware* yang efektif untuk mencegah dan mengurangi terjadinya serangan *malware* pada sistem komputer.

Pada saat ini, pendekatan berbasiskan visualisasi citra *grayscale malware* banyak dikembangkan untuk melakukan klasifikasi *malware* [1] [3] [4] [5]. Citra *grayscale malware* dinilai lebih efisien dalam masalah klasifikasi *malware* karena keseluruhan struktur antar *malware* dapat diukur melalui kesamaan dan perbedaan fitur dan tekstur secara visual pada citra *grayscale*. Pada penelitian [3] menerapkan algoritme transformasi *wavelet* yang digunakan untuk metode ekstraksi fitur, sedangkan model *machine learning* yang dikembangkan adalah K-Nearest Neighbor (KNN) dan Support Vector Machine (SVM) untuk klasifikasi *malware*. Dalam penelitian ini, *dataset* yang digunakan adalah Malware Image (Malimg) khususnya *malware* yang berjenis *Trojan*. Akurasi terbaik yang dihasilkan model KNN adalah sebesar 89.11%, serta model SVM menghasilkan akurasi sebesar 73.55%. Kelemahan pada penelitian [3] adalah klasifikasi *malware* yang dilakukan belum memakai keseluruhan data di dalam *dataset* Malimg.

Pada penelitian [4], model klasifikasi *machine learning* yang dikembangkan untuk klasifikasi *malware* adalah Random Forest dengan akurasi 97.47%, K-Nearest Neighbor (KNN) dengan akurasi 96.23%, dan Support Vector Machine (SVM) dengan akurasi 95.23%. Metode *machine learning* konvensional, seperti Random Forest, KNN, dan SVM memiliki kelemahan, yaitu peningkatan performa akan melambat seiring berkembangnya jumlah data masukan sehingga kurang cocok untuk pelatihan dengan volume data yang banyak.

Metode berbasiskan *deep learning* juga banyak dimanfaatkan untuk pengolahan citra *grayscale* untuk melakukan deteksi dan klasifikasi *malware* [1]

[5]. Penelitian [1] menerapkan metode Convolution Neural Network (CNN) untuk klasifikasi *malware* terhadap citra *grayscale malware* menghasilkan akurasi 98.52% pada *dataset* Malimg dan 98.99% pada *dataset* Microsoft. Selain CNN, penelitian [5] mengembangkan metode *deep learning* lainnya, yaitu Extreme Learning Machine (ELM) untuk klasifikasi *malware* dengan akurasi 93.9%. Berdasarkan penelitian [1], [3], [4], [5] yang sudah dibahas, metode CNN dinilai mampu menganalisis pola tersembunyi pada data citra *grayscale* untuk klasifikasi *malware* sehingga dapat menghasilkan akurasi yang lebih tinggi dibandingkan metode lainnya.

Penelitian [6] melakukan eksperimen klasifikasi *malware* dengan model CNN menggunakan citra RGB yang didapat dari *file malware* pada *platform* Android. Klasifikasi *malware* menggunakan citra RGB dinilai efektif dengan akurasi sebesar 98.77%.

Oleh karena itu, penelitian ini akan menerapkan metode CNN untuk melakukan klasifikasi *malware* menggunakan visualisasi citra *malware* dengan menggunakan teknik *pseudocolor*, serta Principal Component Analysis (PCA) untuk ekstraksi fitur. Berdasarkan penelitian [7], PCA bertujuan untuk melakukan *dimension reduction* atau mengurangi dimensi pada masukan untuk mempercepat proses pembelajaran dan meningkatkan performa klasifikasi dengan menyeleksi dan mengekstrak fitur-fitur penting pada data. Kemudian, pada penelitian ini teknik *pseudocolor* dilakukan pada citra *grayscale malware* dari *dataset* untuk menghasilkan citra berwarna RGB, tujuannya adalah citra RGB memuat 3 *channel* warna, yaitu merah (*red*), hijau (*green*), dan biru (*blue*) sehingga memiliki informasi fitur yang lebih banyak dibandingkan citra *grayscale* yang hanya berisi 1 *channel* warna saja. Penelitian dibuat untuk membandingkan besar nilai akurasi pada klasifikasi terhadap citra *grayscale* dan citra RGB. Pengujian kinerja model klasifikasi CNN menggunakan *confusion matrix* untuk mengukur nilai akurasi, *recall*, dan *F-measure*.

1.2 Rumusan Masalah

Berikut ini adalah rumusan masalah yang menyangkut ide pokok dari setiap masalah yang akan dibahas dan dipecahkan melalui penelitian ini:

1. Berapa nilai akurasi dan *recall* dari metode Convolutional Neural Network untuk klasifikasi *malware* terhadap data citra *grayscale* maupun citra RGB ?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, berikut adalah tujuan dari penelitian tugas akhir ini:

1. Mengimplementasikan metode CNN dengan menerapkan PCA untuk melakukan klasifikasi *malware* menggunakan data visualisasi citra *malware*.
2. Mengimplementasikan dan mencari nilai *hyperparameter* terbaik untuk jumlah *convolution layer*, jumlah *filter*, *epoch*, dan *learning rate* terhadap model CNN untuk klasifikasi *malware*.
3. Melihat pengaruh citra RGB dibandingkan citra *grayscale* terhadap performa akurasi dan *recall* pada metode CNN untuk klasifikasi *malware*

1.4 Batasan Masalah

Agar penelitian ini lebih fokus dan terarah, maka peneliti akan memberikan beberapa batasan masalah sebagai berikut:

1. *Dataset* yang digunakan adalah Malware Image (Malimg) Dataset, berisi data citra *grayscale* yang terdiri dari 9339 citra dan 25 *malware family*.

1.5 Kontribusi Penelitian

Terdapat penelitian tugas akhir yang pernah dikembangkan terkait klasifikasi *malware* dengan mengimplementasikan metode Random Forest, tugas akhir tersebut berjudul ”Penerapan Metode Random Forest Untuk Klasifikasi *Malware* Dengan Sumber Data Berbasis *Gray-Scale Image*”. Sedangkan, pada penelitian ini akan mengambil pendekatan dengan metode yang berbeda untuk sistem klasifikasi *malware*, yaitu mengimplementasikan metode CNN dan PCA untuk ekstraksi fitur, serta membandingkan akurasi antara klasifikasi citra *grayscale* dan citra RGB. Berikut ini merupakan kontribusi penelitian dalam penelitian ini:

1. Menerapkan metode CNN sebagai metode klasifikasi dengan menggunakan teknik *pseudocolor* dan PCA untuk melakukan klasifikasi *malware* berdasarkan data visualisasi citra *malware*.
2. Membandingkan dan melakukan analisis pengaruh antara citra *grayscale* dan citra RGB terhadap performa akurasi dan *recall* pada sistem klasifikasi *malware* dengan metode CNN.

1.6 Metodologi Penelitian

Berikut ini merupakan metodologi penelitian yang dilakukan dalam penelitian ini:

1. Studi Literatur
   Penulisan penelitian diawali dengan melakukan studi kepustakaan yang bersumber dari jurnal penelitian terkait topik yang sudah ada.
2. Data Sampling
   Data sampling yang digunakan untuk penelitian berupa citra *grayscale malware*.
3. Analisis Masalah
   Pada tahap ini, dilakukan analisis masalah berdasarkan batasan masalah yang ada.
4. Perancangan dan Implementasi
   Pada tahap ini dilakukan pembangunan model pembelajaran mesin dengan metode CNN, serta pelatihan dengan kedua jenis data citra *grayscale* dan RGB *malware*.
5. Pengujian
   Pada tahap ini dilakukan pengujian terhadap performa metode CNN dan mengukur keakuratan dalam melakukan klasifikasi *malware*.
6. Dokumentasi
   Di tahap ini akan dilakukan dokumentasi hasil analisis dan implementasi secara tertulis dalam bentuk laporan metode penelitian.

1.7 Sistematika Pembahasan

Penelitian ini disusun berdasarkan sistematika penulisan sebagai berikut:

Bab I Pendahuluan

Bab ini menjelaskan latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, kontribusi penelitian, metodologi penelitian serta sistematika pembahasan.

Bab II Landasan Teori

Bab ini menjelaskan teori mengenai tinjauan pustaka, tinjauan studi, dan tinjauan objek yang mendukung implementasi penelitian ini.

Bab III Analisis dan Perancangan

Bab ini menjelaskan analisis terhadap masalah dan perancangan metode yang akan digunakan.

Bab IV Implementasi dan Pengujian

Bab ini menjelaskan implementasi dan pengujian pada sistem yang dikembangkan serta pengukuran evaluasi sistem berdasarkan nilai akurasi.

Bab V Kesimpulan dan Saran

Bab ini menjelaskan kesimpulan berdasarkan hasil sistem klasifikasi yang dibuat dan saran untuk pengembangan lebih lanjut di masa mendatang.