Final Project in NT230 course– Malware's Modus Operandi

Windows Malware Detection: A research of evasion and detection techniques

Thành viên nhóm: Nguyễn Ngọc Diệu Duyên (23520401), Đoàn Việt Khải(23520673), Nguyễn Hoàng Bảo Minh (23520938)

Mã nhóm: G04, Mã đề tài: S17 Học kì: I Năm học: 2025 – 2026

Introduction

Over the past decades, the volume of malware has increased rapidly, accompanied by growing complexity and sophistication. According to the 2023 SonicWall report, 172,146 previously unseen samples were discovered in a single year, averaging over 956 new variants per day. Traditional static analysis methods are easily bypassed by techniques such as Code Obfuscation, Packing, and Metamorphism. Consequently, recent dynamic analysis research has primarily focused on API names and their invocation frequencies.

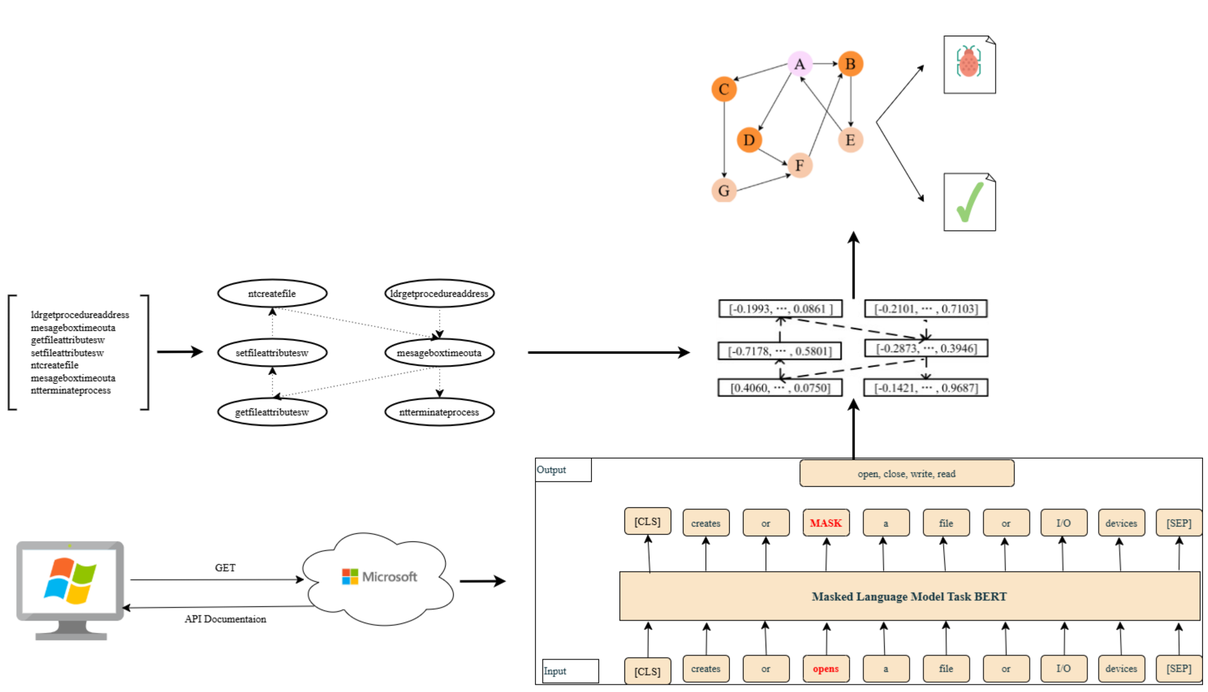

Fig.1 *Our method in project*

Methodology

Statement 1. API Documentation

Used BeautifulSoup & Selenium WebDriver to crawl API documentation and extracted names/descriptions.

Statement 2. API Embedding

Pre-trained on Masked Language Modeling (MLM) to capture API semantics.

Generated API Embedding from the Last Hidden Layer of BERT.

Statement 3. API Graph Construction

Represented unique APIs as nodes and consecutive call sequences as edges, mapping the API Embedding to each node.

Statement 4. GAT Classifier

Extracted structural features using Multi-head Attention and weighted message-passing to accurately classify programs as malicious or benign.

Experiments and Results

Result 1: Execution Flow

- **Description**: This setup reproduced the main execution flow, focusing on performance comparison between $BERT_{base}$ and $BERT_{small}$ across the MalBehavD-V1, PE_APICALLS, and APIMDS datasets.
- **Results**: Upgrading to $BERT_{base}$ improved overall metrics by 2.5% - 4.7% on MalBehavD-V1. The model achieved 100% precision on PE_APICALLS and reached 100% across all metrics on APIMDS.

Result 2: Comparison with baseline and hybrid models

- **Description**: A analysis was conducted between DawnGNN ($BERT_{base}$ + GAT), GNN algorithms (GCN, GIN), and hybrid models such as Word2Vec + GAT and $BERT_{base}$ + LSTM on the MalBehavD-V1.
- **Results**: DawnGNN outperformed all other architectures, with 93.39% Accuracy and 93.63% F1-Score.

Result 3: Efficiency Validation on new Dataset

- **Description**: The model was evaluated on the Malware Analysis Datasets: API Call Sequences to test stability.
- **Results**: Training on full dataset sustained between 95% - 98%, though the True Negative Rate (TNR) fluctuated 56% - 78%. In contrast, with 5% of the data, metrics ranged from 89% - 96%, but the TNR remained more stable at 63% - 85%.

Model	Precision	Recall	F1 - Score	Accuracy	Dataset
BERTsmall + GAT	0.9066	0.9000	0.9000	0.9000	MalBehavD-V1
BERTbase + GAT	0.9470	0.9259	0.9363	0.9339	MalBehavD-V1
BERTbase + GAT	0.9545	1.0000	0.9767	0.9677	PE_APICALLS
BERTbase + GAT	1.0000	1.0000	1.0000	1.0000	APIMDS

Tab.1 Our model achieves highly reliable classification result.

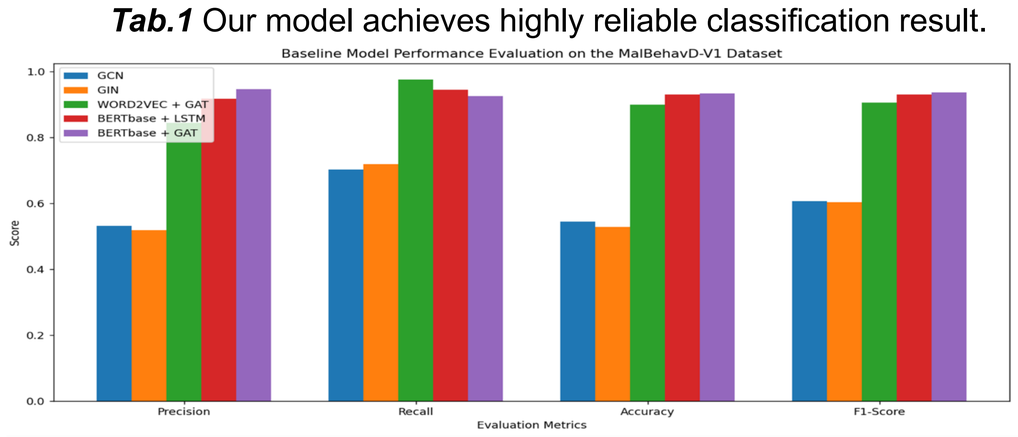

Fig.1 The proposed model outperforms other baselines across almost evaluation metrics.

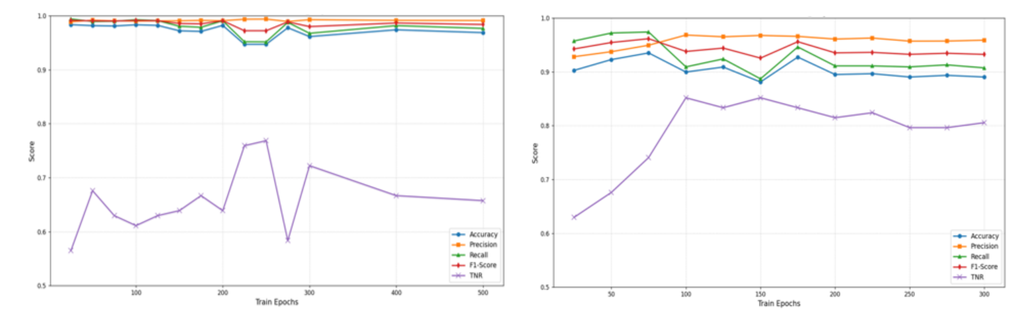

Fig.2 *Training* full-dataset maintains higher overall metrics, while using a subset provides a more stable True Negative Rate.