Rethinking Reproducibility: A Study on ML-Based IoT Malware Detection

Aditya Vikram Reddy V.[1], Josiah Hugo[2], Siddharth Kakumanu[3], Josiah Snyder[2], Dr. Dipakkumar Pravin[1]

***"If you're not able to reproduce the results, then it's not a scientific result."* – Richard Feynman**

UNT 1. Department of Information and Decision Sciences 2. Department of Computer Science and Engineering 3. Texas Academy of Mathematics and Science, University of North Texas

Abstract

The Internet of Things (IoT) refers to a network of interconnected "smart" devices that connect and exchange data over the internet. As this technology rapidly evolves, IoT devices have become increasingly vulnerable to malware attacks. In this paper, we replicate several research papers focused on detecting IoT malware using a variety of artificial intelligence and machine learning techniques. Our objective is to validate the findings of these studies and assess their reliability. To achieve this, we carefully analyze each paper, contact the original authors for datasets and additional information, and reproduce the experiments step by step. By comparing our results with the original findings, we aim to confirm the reproducibility of these approaches. Ultimately, this work serves as both a validation effort and a practical guide for other researchers interested in conducting malware detection studies.

Findings

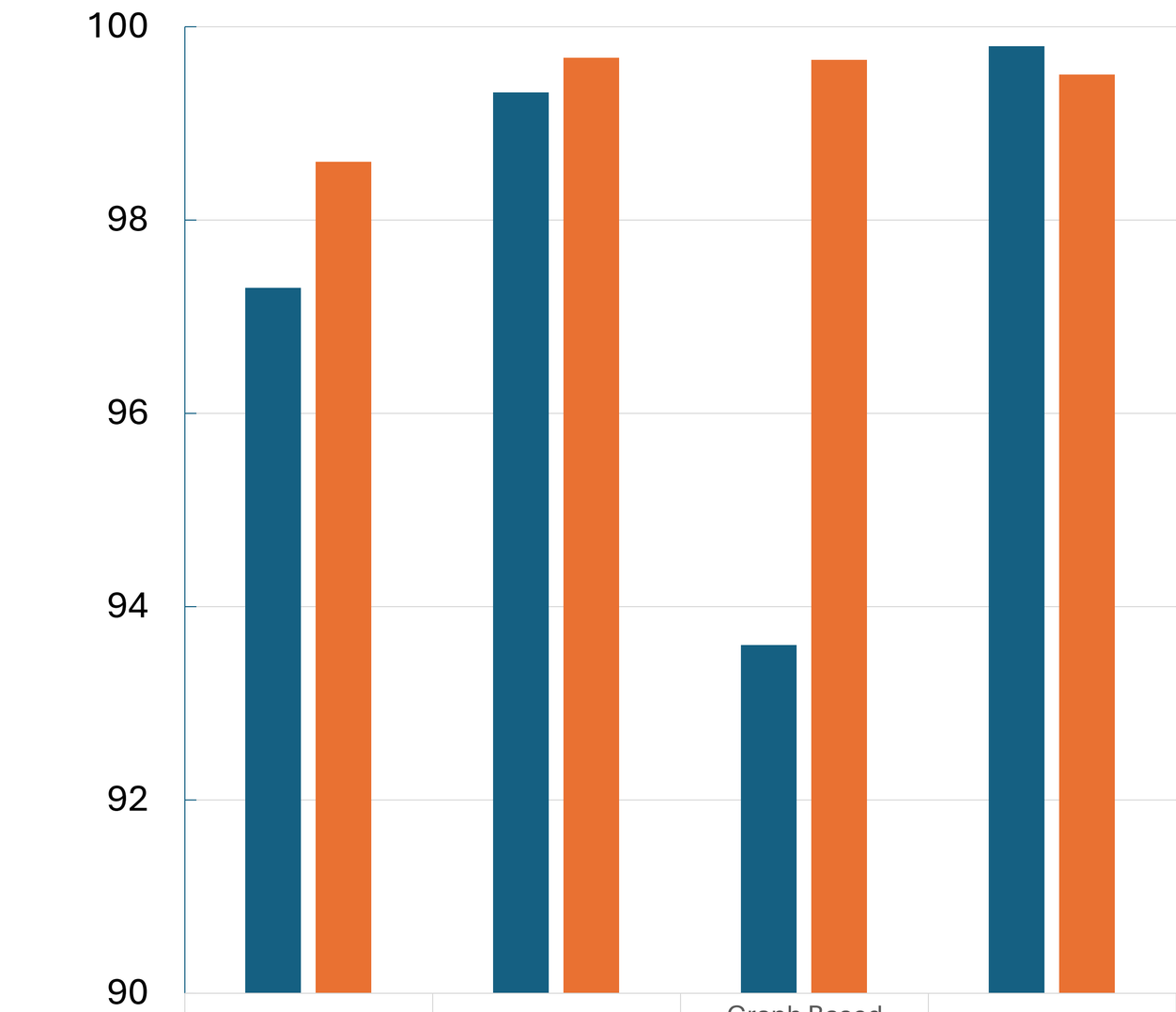
Research papers that are well written and sufficiently detailed are far more likely to be successfully replicated by other researchers. Based on our experience reproducing four different publications, we identified the following **key characteristics that make a paper properly replicable**:

- **Accessible datasets**: Ideally publicly available or clearly described
- **Step-by-step methodology**: Detailed enough to avoid assumptions
- **Provide source code**: Or pseudocode/detailed algorithmic descriptions
- **Clear evaluation metrics**: Have definitions and thresholds for performance measurement
- **Justifications for methodological choices:** Explain why approaches were taken
- **Author contact information**: Allowing clarification of ambiguities or missing details

Replication

Paper – Title | Authors | Who Replicated	AI/ML Technique Used; Technology Used	Replication Successful (Yes/ Somewhat/ No)	Paper vs Replication Accuracy
A Novel Detection and Multi-Classification Approach for IoT-Malware Using Random Forest Voting of Fine-Tuning Convolutional Neural Networks | Safa Ben Atitallah, et al.	• Transfer Learning • Ensemble Learning – Hard Voting, Soft Voting, Random-Forests Voting • Multiclass Image Classification	Yes	**Accuracy:** 98.6% vs. 97.3%
Robust Malware Detection or Internet of Battlefield Things Devices Using Deep Eigenspace Learning | Amin Azmoodeh, et al.	• Deep Eigenspace Learning • Class-Wise Information Gain (CIG) • Deep learning • Convolutional Network • Cross-validation	Yes	**Accuracy**: 99.68% vs. 99.32% **Precision**: 98.59% vs. 90.77% **Recall**: 98.37% vs. 96.84% **F-Measure**: 98.48% vs. 95.64%
Analyzing and Detecting Emerging Internet of Things Malware A Graph-Based Approach | Hisham Alasmary, Aminollah Khormali, et al.	• Deep learning • Convolutional Neural Network • Control Flow Graphs • K-Fold Cross Validation (k=10)	Somewhat	**Accuracy:** 99.66% vs. 93.6% **FNR**: .33% vs. 19.96% **FPR**: .33% vs. 2.27% **FOR**: .33% vs. 7.58% **FDR**: .33% vs. 8.66%
MADP-IIME: Malware Attack Detection Protocol in IoT-enabled Industrial Multimedia Environment using Machine Learning Approach | Sumit Pundir, . M.S. Odediat, et al.	• PCAP – Wireshark Conversion to CSVs • Naive Bayes • Logistic Regression • Artificial Neural Networks (ANN; n=50) • Random Forest • K-5-fold Cross – Validation • Classification Reports	Yes	**Original Accuracy:** 99.5% **Ours:** Best Model: Random Forest **Accuracies** (80/20 split): Train: 0.9974; Test: 0.9978; ~ **99.8%** **Accuracies** (70/20/10 split): Train: 0.9974; Test: 0.9978; ~ **99.8%**

Accuracies Overview



	Novel Detection	Robust Malware	Graph Based Approach	MADP
Replication	97.3	99.32	93.6	99.8
Papers	98.6	99.68	99.66	99.5

Replicated Paper Summaries

Research Paper - Atitallah Novel Detection

Our Experience:

We were able to easily replicate the described methods and get near-identical accuracy. The paper mentioned, in enough detail, the steps taken to achieve the described outcome, making replication a simple process. This paper was well written and had almost all the data needed to replicate its findings properly.

Strengths:

- Provided the dataset source and preprocessing steps in detail
- Provided hyperparameters in a concise and easily readable table
- Provided justification behind each algorithm and method that was used
- Used diagrams to explain the dataset and the models in an understandable manner.

Critiques:

- Did not provide or reference any source code
- Authors did not respond to attempts of contact

Research Paper - Azmoodeh Deep Eigenspace Learning

Overall Experience:

Replicating the paper proved to be a success as we were almost able to get the same accuracy as the paper. There were some troubles replicating the paper as some of the methods (specifically the deep learning section) were not clearly explained in detail, but the paper was overall well written and provided just enough information for replication.

Strengths:

- Provided dataset
- Provided formulas/equations and algorithms used
- Great explanation on methodologies and their justifications

Critiques:

- Go over results they might have gotten
- Explain the specifics of their CNN model
- Difficulties in finding author's contacts

Research Paper - Alasmary Graph

Our Experience:

Overall, we were close in replicating the model. The paper was well written with plenty of information on how to replicate it, but the original dataset can no longer be found which hurts the reproducibility of the paper.

Strengths:

- Had significant detail about algorithms and feature selection
- Provided formulas and equations used
- Gave explanation on methodologies and their justifications

Critiques:

- Provided dataset no longer accessible
- Did not specify learning rate or optimizer for CNN

Research Paper – Pundir MADP-IIME

Our Experience:

Replication was challenging slow PCAP to CSV conversion, limited source code, and missing feature/dataset details. Despite contacting authors, online resources and Gen-AI were crucial for accuracy and identifying weak points. While successful **(0.3%** performance increase), author guidance and complete source code are vital for future reproducibility.

Strengths:

- Addresses critical IoT security need.
- Uses diverse ML algorithms effectively.
- Reports high detection accuracy

Critiques:

- Poor reproducibility due to gaps .
- Ambiguous data/model specifics and limited architectural clarity (ANN)

***"Single occurrences that cannot be reproduced are of no significance to science."* – Karl Popper**

Contacts:
AdityaVikramReddyVennapuala@my.unt.edu | JosiahHugo@my.unt.edu | SiddharthKakumanu@my.unt.edu | JosiahSnyder@my.unt.edu
Dipakkumar.Pravin@unt.edu

Acknowledgements:
Special thanks to Dr. Mark V. Albert and the UNT AI/CS Summer Research Program.