This document reviews three key research works that are closely related to my study on defending machine-learning-based malware detection systems against adversarial attacks. It summarises what each paper or thesis investigated, the main problems they identified, and the goals they aimed to achieve. It then explains how each of these works helped shape and refine my own research topic, problem statement, objectives, and research questions. Overall, the document shows how existing studies influenced my focus on evaluating simple and practical defenses against known adversarial attacks in malware detection systems.

STAR PAPER 1

1**. Source, Title and Reference**

Mostafa Jafari & Alireza Shameli-Sendi (2025)
Evaluating the Robustness of Adversarial Defenses in Malware Detection Systems
Published in Computers & Electrical Engineering, Volume 130 (Elsevier).

2**. Summary of the Problem/Gap Addressed**

This paper looks at how well current defense methods for malware detection stand up against strong evasion attacks. It highlights a gap: existing evaluations do not properly measure defense performance in domains where features are binary (as they often are in malware datasets). It introduces new ways to test defenses and shows many defenses fail badly when attacks are strong.

3. Aim and Objectives

- Main aim: Evaluate the strength and limitations of current defenses for machine-learning-based malware detectors.
- Objectives:

1. Introduce techniques to convert continuous attacks into binary feature space.
2. Design a new adversarial attack (called **sigma-binary**) for binary domains.
3. Run experiments to test different defenses and measure their robustness.

4. Main Questions Asked

- Do current defenses really protect malware detectors?
- How vulnerable are models in binary-constrained feature spaces?
- Which attacks can break these defenses even when they seem robust?

<u>Lessons Learned for My Research</u>

a) Tentative Research Topic
This paper confirms that existing defenses are often weak in real malware settings and need better evaluation. This strengthened my topic's focus on practical defense evaluation rather than proposing new attacks.

b) Problem Statement
It showed that many defenses thought to be robust fail badly under stronger attacks, but most studies don't

evaluate them thoroughly. That helped me shape my problem to emphasize the lack of systematic, realistic evaluations of simple defenses.

c) Objectives
From this paper, I saw the importance of evaluating defenses across attacks, not just one. My objectives now include comparing several defenses and measuring trade-offs like accuracy vs robustness.

d) Research Questions
Their questions confirmed I should not only compare attacks but focus on defense effectiveness, including real measures of performance under attack.

STAR PAPER 2

1. Source, Title and Reference

Adnan et al. (2025)
Evaluating Realistic Adversarial Attacks against Machine Learning Models for Windows PE Malware Detection
Published in Future Internet (MDPI).

2. Summary of the Problem/Gap Addressed

This paper investigates how machine-learning models that detect Windows PE malware can be tricked by carefully constructed adversarial inputs. It highlights that while these models are effective on clean data, attackers can modify malware so the models misclassify it as benign. It also argues that evaluation of adversarial attacks needs to be done on larger datasets and across different models.

3. Aim and Objectives

- Main aim: Evaluate several realistic evasion attacks on malware detection models.

1. Other objectives:
    a. Work with larger datasets than prior studies.
    b. Explain how attacks modify malware samples to evade models.
    c. Test whether adversarial training helps defend against such attacks.

4. Main Questions Asked

- How vulnerable are Windows PE malware detectors to different adversarial attacks?
- What changes are actually done to malware examples to fool models?
- Does adversarial training improve detection performance under attack?

Lessons Learned for My Research

a) Tentative Research Topic
This paper showed real attacks can be effective and highlighted the need for stronger evaluations. It helped refine my topic toward defense evaluation against real attack scenarios, not just theoretical attacks.

b) Problem Statement
It made me clearly see that the threat side (attack behavior and vulnerability) must be understood to design or evaluate defenses. I adapted my problem statement to include realistic threat models rather than abstract attacks.

c) Objectives
From this work, I learned that evaluation must consider:

- Larger and representative datasets
- Multiple models
- Realistic attack patterns
  So I refined my objectives to include these aspects.

d) Research Questions
It helped sharpen my questions to focus specifically on:

- Which simple defenses help in realistic attacks?
- How much does performance change under attack?

STAR PAPER 3

1. Source, Title and Reference

Aqib Rashid (2023)
Exploring Defenses Against Adversarial Attacks in Machine Learning-Based Malware Detection
Doctoral thesis, King's College London.

2. Summary of the Problem/Gap Addressed

This thesis identifies that most research on adversarial defenses has focused on other domains like image recognition, leaving malware detection less studied. It highlights limitations in existing malware defenses, such as ineffectiveness against new attacks and reduced accuracy on normal inputs.

3. Aim and Objectives

- Main aim: Understand and evaluate defense methods for malware detection.

- Other objectives:

  1. Introduce moving target defenses (MTDs) and stateful defenses.
  2. Compare their performance to prior techniques.
  3. Identify weaknesses in these approaches and provide recommendations.

4. Main Questions Asked

- Which defense methods are promising for adversarial malware detection?
- How do moving target defenses perform?

- Can stateful defenses reduce attack success rates?

Lessons Learned for My Research

a) Tentative Research Topic

This thesis confirmed that the field lacks systematic evaluation of defenses, while much research focuses on attacks. It strengthened my choice to focus on evaluating defenses rather than inventing them.

b) Problem Statement

It highlighted that some defenses are tested only in limited conditions and fail under stronger scenarios, reinforcing my emphasis on realistic evaluation across multiple defense*s*.

c) Objectives

From this work I learned:

- It is useful to compare multiple defense approaches
- It is important to test under varied attack intensities

  Thus, my objectives now include *evaluating several defenses under a common threat model*.

d) Research Questions

This thesis helped me refine my questions to ensure they include:

- Defense performance under a range of threat levels
- Differences between defense types

  instead of focusing only on one defense strategy.

Overall Reflection

All three star works helped shape my research by:

- Confirming that adversarial vulnerability is a real problem in malware detection.
- Showing that many defenses either fail under strong attacks or are resource expensive reinforcing the need for evaluation of lightweight defenses.
- Highlighting gaps in systematic, practical defense evaluation, not just attack design.
- Helping me refine my objectives and questions to focus on practical defense comparison rather than proposing new, complex methods.

Conclusion

All three articles agree that machine-learning malware detectors can be fooled by adversarial attacks and that many defenses are either weak or poorly tested. They differ in focus: one tests defenses under strong attacks, another studies real-world malware attacks, and the third explores various defense strategies. My research builds on these by evaluating lightweight, practical defenses, both individually and in combination, to see which actually improve robustness. This positions my work as a focused, practical study that bridges the gap between understanding attacks and applying deployable, resource-friendly defenses in real-world malware detection systems.