

# Cybersecurity *project* *documentation*

Daniele Russo

November 2025

## Contents

<b>1</b>	<b>Introduzione</b>	<b>3</b>
1.1	Contesto e motivazione . . . . .	3
1.2	Obiettivi del progetto . . . . .	3
1.3	Ipotesi di lavoro: resilienza dei campioni avvelenati alla misclas- sificazione . . . . .	3
<b>2</b>	<b>Stato dell'arte</b>	<b>3</b>
2.1	Data poisoning e backdoor attacks nei modelli ML . . . . .	3
2.1.1	Attacchi clean-label e backdoor feature-based . . . . .	3
2.1.2	Rumore intenzionale e inquinamento del dataset . . . . .	3
2.2	Tecniche di rilevazione e mitigazione . . . . .	3
2.3	Sparsità e robustezza dei modelli neurali . . . . .	3
<b>3</b>	<b>Metodologia</b>	<b>3</b>
3.1	Costruzione o scelta del dataset . . . . .	3
3.1.1	Dataset di malware avvelenato (es. MalwareBackdoors) . . . . .	3
3.1.2	Preprocessing e feature extraction . . . . .	3
3.2	Architettura del classificatore di malware . . . . .	3
3.2.1	Reti neurali utilizzate . . . . .	3
3.2.2	Metriche di valutazione . . . . .	3
3.3	Introduzione di rumore e sparsificazione . . . . .	3
3.3.1	Perturbazione dei pesi interni . . . . .	3
3.3.2	Applicazione di tecniche di pruning (Lottery Ticket Hy- pothesis) . . . . .	3
3.4	Ipotesi sperimentale . . . . .	3
3.4.1	Campioni avvelenati come outlier resilienti al rumore . . . . .	3
<b>4</b>	<b>Esperimenti</b>	<b>3</b>
4.1	Setup sperimentale . . . . .	3
4.1.1	Parametri di addestramento . . . . .	3
4.1.2	Procedure di test . . . . .	3
4.2	Valutazione con rete non perturbata . . . . .	3
4.3	Valutazione con rete perturbata/sparsificata . . . . .	3

4.4	Analisi della correlazione tra resilienza e avvelenamento . . . . .	3
<b>5</b>	<b>Risultati</b>	<b>3</b>
5.1	Prestazioni del modello . . . . .	3
5.2	Effetti del rumore sui campioni avvelenati . . . . .	3
5.3	Evidenze di separabilità o resistenza ai guasti . . . . .	3
5.4	Visualizzazione e interpretazione . . . . .	3
<b>6</b>	<b>Discussione</b>	<b>3</b>
6.1	Confronto con la letteratura . . . . .	3
6.2	Limiti dell'approccio . . . . .	3
6.3	Potenzialità future . . . . .	3
<b>7</b>	<b>Conclusioni</b>	<b>3</b>
7.1	Sintesi dei risultati . . . . .	3
7.2	Prospettive di ricerca . . . . .	3

# 1 Introduzione

## 1.1 Contesto e motivazione

## 1.2 Obiettivi del progetto

## 1.3 Ipotesi di lavoro: resilienza dei campioni avvelenati alla misclassificazione

# 2 Stato dell'arte

## 2.1 Data poisoning e backdoor attacks nei modelli ML

### 2.1.1 Attacchi clean-label e backdoor feature-based

### 2.1.2 Rumore intenzionale e inquinamento del dataset

## 2.2 Tecniche di rilevazione e mitigazione

## 2.3 Sparsità e robustezza dei modelli neurali

# 3 Metodologia

## 3.1 Costruzione o scelta del dataset

### 3.1.1 Dataset di malware avvelenato (es. MalwareBackdoors)

### 3.1.2 Preprocessing e feature extraction

## 3.2 Architettura del classificatore di malware

### 3.2.1 Reti neurali utilizzate

### 3.2.2 Metriche di valutazione

## 3.3 Introduzione di rumore e sparsificazione

### 3.3.1 Perturbazione dei pesi interni

### 3.3.2 Applicazione di tecniche di pruning (Lottery Ticket Hypothesis)

## 3.4 Ipotesi sperimentale

### 3.4.1 Campioni avvelenati come outlier resilienti al rumore

# 4 Esperimenti

## 4.1 Setup sperimentale

### 4.1.1 Parametri di addestramento

### 4.1.2 Procedure di test

## 4.2 Valutazione con rete non perturbata

## 4.3 Valutazione con rete perturbata/sparsificata

## 4.4 Analisi della correlazione tra resilienza e avvelenamento

# 5 Risultati

## 5.1 Prestazioni del modello