HYBRID NEURAL NETWORK MODEL DEFENSES

Dissertation

Doctor of Philosophy

in

Cyber Operations

March 25, 2026

By

Eric Wayne Yocam

Dissertation Committee:

Varghese Vaidyan, Ph.D.

Austin O'Brien, Ph.D.

Gurcan Comert, Ph.D.

Beacom College of Computer and Cyber Sciences

DAKOTA STATE
OFFICE OF GRADUATE STUDIES

Dissertation Approval Form

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Eric Wayne Yocam Student ID: A00539216

Dissertation Title: Hybrid Neural Network Model Defenses

Graduate Office Verification: DocuSigned by: Abby Chowning F44C8D9E621C417... Date: 03/27/2026

Dissertation Chair/Co-Chair: Signed by: Varghese Vaidyan 7D0D713978DE43F... Date: 03/28/2026

Print Name: Varghese Vaidyan

Dissertation Chair/Co-Chair: _____ Date: _____

Print Name: _____

Committee Member: DocuSigned by: E94723CA282F45C... Date: 03/28/2026

Print Name: Austin O'Brien

Committee Member: Signed by: Gurcan Comert C7BB8E96597A4F5... Date: 03/30/2026

Print Name: Gurcan Comert

Committee Member: _____ Date: _____

Print Name: _____

Committee Member: _____ Date: _____

Print Name: _____

Form must be routed through DocuSign
For assistance, contact: GradOffice@dsu.edu

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to the many individuals who have supported me throughout this journey. I would like to sincerely thank my Dissertation Committee Chair, Dr. Varghese Vaidyan, for his guidance, expertise, and unwavering support throughout this process. I would also like to thank my committee members, Dr. Austin O'Brien and Dr. Gurcan Comert, for their invaluable feedback and insight. First, I am deeply grateful for the unwavering love and encouragement of my family, my wife Annie, my daughter Hailey, and my son Nathan. I also owe a great debt of gratitude to my parents, Delbert and Janet Yocam, for their constant support and belief in me. To my extended family, including my uncle Keith Yocam, my brother and sister-in-law Chris and Krista Yocam, and my sister and brother-in-law Beth and Evan Lewis, thank you for being pillars of support and inspiration. I am also fortunate to have an incredible circle of close friends whose friendship has meant so much to me: Andy and Daniele Pollin, Jackie and Dave Templin, Drew Morin, Andrew Watts, Arash Obaidi, Tim Vercruyssen, Richard White, Koveh Tavakkol, Joe Mazzotta, Rob Hitchcock, Mark Russell, Bret Hartman, Dr. Yong Wang, Dr. Tony Rizi, Dr. Paris Kalathas, Dr. Bruce DeBruhl, and Dr. Patrick Tague. Your support, encouragement, and belief in me have made this achievement possible, and I am sincerely grateful to each of you.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another. I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Eric Wayne Yocam

ABSTRACT

Hybrid neural networks (HNNs), which integrate classical convolutional neural networks with quantum circuit components, are highly vulnerable to white-box targeted compound (WTC) adversarial attacks. Previous research has primarily evaluated defenses against single-method attacks, leaving a critical gap in understanding defense effectiveness against compound attacks that combine multiple perturbation strategies. This research study systematically evaluates defense mechanisms against WTC adversarial attacks on HNN models across 120 experimental conditions. Three defense categories are investigated: *test-time defenses* (input transformation and randomization, applied during inference without retraining) and *training-time defense* (adversarial training, requiring model retraining). Evaluation uses four datasets (MNIST, EMNIST "Digits", TinyImageNet, TrafficSigns) and three compound attack combinations (FGSM+PGD, FGSM+CW, CW+PGD). The results demonstrate that compound attacks severely degrade the performance of HNN, reducing the average accuracy from 87.6% to 19.1%. The defense architecture demonstrated dataset-dependent resilience: adversarial training achieved high-efficacy recovery (87%) on structured, low-entropy datasets (TrafficSigns), while encountering a complexity ceiling (28%) on high-diversity datasets (EMNIST "Digits"). This indicates that while the HNN-defense combination is highly effective at preserving structural features, its current configuration requires further optimization for handwriting-style variability where adversarial perturbations more easily mimic legitimate feature shifts. Adversarial training outperformed test-time defenses by 2.3–2.8× across all datasets. Key contributions include: (1) establishing adversarial training's superiority across all conditions, (2) revealing test-time versus training-time trade-offs between deployment flexibility and robustness, (3) demonstrating relative robustness gains of 13-18% over classical CNN baselines through quantum geometric defenses, and (4) validating classical simulation for HNN defense research. These findings provide a foundational understanding for the implementation of robust HNN models in adversarial environments.

TABLE OF CONTENTS

Dissertation Approval Form ii

Acknowledgments iii

Declaration iv

Abstract v

Table of Contents vi

List of Tables xiv

List of Figures xvi

1

INTRODUCTION 1

1.1 Background of the Problem 1

1.1.1 Classical Computing 2

1.1.2 Quantum Computing 2

1.1.3 A Neural Network Model 3

1.1.4 Compromising a Neural Network Model 3

1.1.5 Distinct Adversarial Attacks 3

1.1.6 Compounded Adversarial Attacks 4

1.1.7 Defenses Against Compounded Adversarial Attacks 4

1.1.8 Model's Prediction Accuracy 5

1.2 Statement of the Problem with Motivation 5

1.2.1 Problem Statement 6
1.2.2 Motivation 6
1.3 Thesis Statement 6
1.4 Research Questions 6
1.4.1 Research Question (RQ1) 7
1.4.2 Research Question (RQ2) 7
1.5 Hypotheses 7
1.5.1 Hypothesis (H1) 7
1.5.2 Hypothesis (H2) 8
1.5.3 Hypothesis (H3) 8
1.5.4 Hypotheses Mapping 8
1.6 Objectives of This Study 8
1.6.1 Objective (OBJ1) 9
1.6.2 Objective (OBJ2) 9
1.6.3 Objective (OBJ3) 9
1.6.4 Objective (OBJ4) 9
1.6.5 Structured Experimentation 9
1.6.6 Expected Outcomes 9
1.6.7 Primary Deliverables 10
1.7 Research Goal 10
1.8 Key Technical Ideas 11
1.9 Research Overview 11
1.10 Contribution of this Work 13
1.11 Summary 13

2

WTC ATTACKS AGAINST HYBRID MODELS **15**

2.1 Background on Adversarial Attacks 15
2.2 Compound Adversarial Attacks 16
2.3 Hybrid Neural Networks 16

2.4 Defense Mechanisms Against Adversarial Attacks 17
2.4.1 Input Transformation Defenses 17
2.4.2 Randomization Defenses 18
2.4.3 Adversarial Training Defense 18
2.5 Related Work on HNN Security 19
2.5.1 Quantum Neural Network (QuNN) Studies 19
2.5.2 Hybrid Quantum Neural Network (HQNN) Studies 19
2.5.3 Classical CNNs Defense Studies 20
2.6 Research Gap Identification 21
2.6.1 Gap 1: Compound attacks against HNN 22
2.6.2 Gap 2: Systematic Defense Evaluation for HNNs 22
2.6.3 Gap 3: Test-Time vs. Training-Time Defense 22
2.6.4 Gap 4: Dataset-Dependent Defense Effectiveness 23
2.6.5 Gap 5: Classical Simulation Methodology 23
2.7 Limitations of Previous Approaches 23
2.7.1 Limited Attack Sophistication 23
2.7.2 Narrow Defense Coverage 24
2.7.3 Insufficient Dataset Diversity 24
2.7.4 Lack of Quantitative Comparison 24
2.7.5 Quantum Hardware Constraints 25
2.8 Novel Contributions of This Research 25
2.8.1 Contribution 1: Compound Attack Evaluation 25
2.8.2 Contribution 2: Defense Mechanism Comparison 25
2.8.3 Contribution 3: Test-Time vs Training-Time Framework 26
2.8.4 Contribution 4: Multi-Dataset Validation 26
2.8.5 Contribution 5: Classical Simulation Validation 26
2.9 Research Approach Distinction 27
2.9.1 Focus on Compound Attacks 27
2.9.2 Hybrid Architecture Emphasis 27
2.9.3 Systematic Multi-Defense Evaluation 27

2.10 Parallel Investigation: Traffic Sign Safety Applications 28
2.11 Cross-Domain Adversarial Robustness Analysis 29
2.12 Summary 30

3

METHODOLOGY **32**

3.1 Research Method 32
3.2 Research Design 32
3.2.1 Model Comparison 33
3.2.2 Feature Selection 33
3.2.3 Data Pre-processing 33
3.2.4 Hyperparameter Tuning 33
3.2.5 Model Interpretability 33
3.3 Reason for Choosing Research Design 34
3.4 Research Method Implementation 34
3.4.1 Experimental Research Design 34
3.4.2 Validation Approach 35
3.4.3 Data Collection 35
3.5 Dataset-Specific Model Architecture 35
3.6 Dataset-Specific Hyperparameters 37
3.7 Rationale for Dataset-Specific Configurations 39
3.7.1 Architecture Depth 39
3.7.2 Batch Normalization 39
3.7.3 Dropout Regularization 40
3.7.4 Training Duration 40
3.7.5 Weight Decay 40
3.7.6 Configuration Choices 40
3.8 Summary 41

4

HYBRID MODEL DEFENSE MECHANISMS **43**

4.1 Proposed Solution 43
4.2 Key Points 44
4.2.1 Development Environment 44
4.2.2 Dataset Selection and Progression Rationale 45
4.2.3 Operational Implications of Defense Timing 46
4.2.4 Quantum Circuit Constraints on Dataset Size 48
4.2.5 Design Context 49
4.2.6 HNN Model Architecture and Parameters 52
4.2.7 Quantum Reservoir Mapping 55
4.2.8 Quantum Circuit Architecture 57
4.3 Strengths and Contributions 59
4.4 Summary 60

5

IMPLEMENTATION, VALIDATION, AND RESULTS 62

5.1 Implementation 62
5.2 Validation Approach 63
5.2.1 Validation Method for Defenses 66
5.2.2 Metrics Used for Validation of Defenses 67
5.2.3 Robustness Metrics Calculation 69
5.2.4 Attack Effectiveness 71
5.2.5 Defense Effectiveness 71
5.2.6 Overall Robustness 71
5.2.7 Defense Rating 71
5.2.8 Effectiveness of Validation Approach 72
5.3 Results and Contributions 74
5.3.1 Experimental Results Overview 74
5.3.2 Results: Adversarial Training Defense 76
5.3.3 Results: Input Transformation Defense 78
5.3.4 Results: Randomization Defense 80

5.3.5 Overall Defense Category Effectiveness 81
5.3.6 Overall Accuracy Progression 83
5.3.7 Overall Defense Performance by Attack Type 85
5.3.8 Overall Input Transformation Defense Comparison 87
5.3.9 Overall Randomization Defense Comparison 89
5.3.10 Overall Top Performing Defenses by Dataset 91
5.3.11 Overall Multi-Dimensional Defense Performance 92
5.3.12 Overall Summary Statistics 97
5.4 Research Questions Answered 99
5.4.1 RQ1 Revisited 99
5.4.2 RQ2 Revisited 99
5.5 Hypotheses Revisited 100
5.5.1 H1 Supported 100
5.5.2 H2 Supported 100
5.5.3 H3 Supported 101
5.6 Objectives Completed 101
5.6.1 OBJ1 Completed 101
5.6.2 OBJ2 Completed 101
5.6.3 OBJ3 Completed 102
5.6.4 OBJ4 Completed 102
5.7 Case Studies 102
5.7.1 Study 1: Retroreflectivity-Based Traffic Sign Safety 103
5.7.2 Study 2: Cross-Domain Adversarial Robustness Analysis 106
5.7.3 Study 3: UAV-Based Crop Row Detection Protection 107
5.7.4 Study 4: Lane Detection Architectural Robustness 109
5.7.5 Study 5: Quantum Adversarial Machine Learning Survey 112
5.7.6 Convergence Across Case Studies 113
5.7.7 Future Publication Plans 114
5.7.8 Interpretation for Dissertation 114
5.8 Comparative Analysis: HNN vs CNN Baselines 116

5.8.1 Literature Baseline Recovery Rates . . . 116
5.8.2 Relative Robustness Gain Analysis . . . 117
5.8.3 Quantum Geometric Defense Hypothesis . . . 117
5.8.4 Dataset-Dependent Quantum Advantage . . . 118
5.8.5 Limitations and Future Validation . . . 119
5.8.6 Contributions . . . 121
5.9 Summary . . . 123

6

CONCLUSIONS **126**

6.1 Summary of Research . . . 126
6.2 Key Findings . . . 127
6.2.1 HNN Vulnerability to Compound Attacks . . . 127
6.2.2 Measurable Defense Effectiveness . . . 127
6.2.3 Superiority of Adversarial Training . . . 128
6.2.4 Dataset-Dependent Defense Effectiveness . . . 128
6.2.5 Test-Time vs Training-Time Trade-offs . . . 128
6.2.6 Influence of Attack Type on Defense Performance . . . 129
6.2.7 Classical Simulation Sufficiency . . . 129
6.2.8 Persistent Accuracy Gap . . . 129
6.3 Limitations . . . 130
6.3.1 Direct Experimental Baseline Comparison . . . 130
6.3.2 Quantum Circuit Scope . . . 132
6.3.3 Dataset and Domain Scope . . . 133
6.3.4 Attack Strategy Scope . . . 133
6.3.5 Defense Mechanism Scope . . . 136
6.3.6 Evaluation Metrics Scope . . . 136
6.4 Lessons Learned . . . 137
6.4.1 Methodological Insights . . . 137
6.4.2 Technical Insights . . . 138

6.4.3 Conceptual Insights 140
6.5 Future Research Work 141
6.5.1 Validation Through Real-World Applications 141
6.5.2 Extended Attack Surface Evaluation 142
6.5.3 Advanced Defense Mechanisms 144
6.5.4 Quantum Architecture and Hardware Investigation 145
6.5.5 Dataset and Domain Expansion 145
6.5.6 Enhanced Evaluation Framework 146
6.5.7 Theoretical Understanding Development 147
6.6 Contributions and Implications 148
6.7 Summary 150

References **152**

APPENDIX A: ADVERSARIAL ATTACK PARAMETERS **160**

APPENDIX B: ACRONYMS AND ABBREVIATIONS **164**

APPENDIX C: RESEARCH ARTIFACTS AND DATA AVAILABILITY **171**

LIST OF TABLES

2.1 Comparison of defense mechanisms against adversarial attack-related work. 21

3.1 Dataset-specific HNN model architecture configurations. 37
3.2 Dataset-specific hyperparameter configurations for HNN model training. . 38

4.1 Key points of emphasis in HNN defense research. 44
4.2 Development environment frameworks and their purposes. 45
4.3 Base HNN model parameters (constant across all datasets). 53
4.4 Dataset-specific HNN model parameter variations. 54
4.5 Quantum circuit configuration parameters. 58
4.6 Research strengths and contributions. 60

5.1 Adversarial Training results on MNIST dataset. 76
5.2 Adversarial Training results on EMNIST "Digits" dataset. 76
5.3 Adversarial Training results on TinyImageNet dataset. 76
5.4 Adversarial Training results on TrafficSigns dataset. 77
5.5 Input Transformation results on MNIST dataset. 78
5.6 Input Transformation results on EMNIST "Digits" dataset. 78
5.7 Input Transformation results on TinyImageNet dataset. 79
5.8 Input Transformation results on TrafficSigns dataset. 79
5.9 Randomization results on MNIST dataset. 80
5.10 Randomization results on EMNIST "Digits" dataset. 80
5.11 Randomization results on TinyImageNet dataset. 81
5.12 Randomization results on TrafficSigns dataset. 81

A.1 WTC adversarial attack parameters. 160

A.2 FGSM adversarial attack parameters. 161
A.3 CW adversarial attack parameters. 162
A.4 PGD adversarial attack parameters. 163

LIST OF FIGURES

1.1 Experimental framework for this research 12

4.1 Design context for the HNN model and processing without a defense mechanism. 50
4.2 Design context for the HNN model and processing with a defense mechanism applied post-attack. 51
4.3 Example quantum circuit configuration used in HNN model (4-qubit parameterized circuit with rotation gates and CNOT entanglement). 57

5.1 Validation approach workflow for defense evaluation. 64
5.2 Defense evaluation workflow showing test-time vs training-time paths. . . . 67
5.3 Robustness metrics computation workflow (see Section 5.2.2 for metric definitions). 70
5.4 Defense category effectiveness across datasets. 82
5.5 Accuracy progression from clean to attacked to defended states. 84
5.6 Defense effectiveness by attack type across datasets. 86
5.7 Input transformation techniques defended accuracy. 88
5.8 Randomization techniques defended accuracy. 90
5.9 Top defense techniques ranked by dataset. 92
5.10 Multi-dimensional defense performance across datasets showing clean accuracy, compound attack robustness (FGSM+PGD, FGSM+CW, CW+PGD), and recovery rate for each defense category. 94
5.11 Summary of key performance metrics across all experiments. 97

5.12 Real-world field dataset examples showing MUTCD compliance assessment. YIELD and ARROW signs demonstrate typical failure modes, while the STOP sign meets all federal thresholds. These examples illustrate the class imbalance challenge (2 of 3 unsafe) that requires synthetic data generation for balanced training. 104
5.13 Qualitative crop row detection comparison. Columns: RGB input, ground truth, traditional CV (fragmented), standard CNN, rule-augmented CNN, overlay. Rows show varied agricultural conditions. 108
5.14 Clean Performance Visualization (TuSimple - Highway). Top: Original image and ground truth. Bottom: Rule-Aug CNN (green, IoU=0.52) and Standard CNN (red, IoU=0.47). Rule-Aug better captures lane boundaries with geometric priors. 110
5.15 Clean Performance Visualization (CULane - Urban). Top: Original image and ground truth. Bottom: Rule-Aug CNN (green, IoU=0.31) and Standard CNN (red, IoU=0.15). On complex curved roads with multiple lanes, Rule-Aug CNN's geometric priors maintain reasonable performance while Standard CNN struggles with lane geometry. 111

Chapter 1

INTRODUCTION

This chapter covers the background of the problem, the motivation statement of the problem, the thesis statement, the hypotheses, the objectives of the project, the goal of the project, the key technical ideas, and the contribution of the work.

1.1 Background of the Problem

The usefulness of quantum computing hardware is limited by the current number of qubits available. A qubit is a fundamental building block or unit of information for a quantum computer [1]. Today, roughly 433 qubits have been demonstrated by the Osprey quantum computer from IBM [2]. A useful quantum computer will need over 1000 qubits. A machine learning or deep learning artificial intelligence model can benefit from greater computing power found with quantum computers. For this reason, a quantum neural network (QNN) must run on quantum computing hardware [3]. However, a convolutional neural network (CNN) runs on classical computing hardware [4]. Classical computing hardware can simulate up to 50 qubits. A classical-quantum neural network can simulate a QNN, but on a classical computer, and represents an alternative. A quantum-classical neural network, called a hybrid neural network (HNN) [5], is a viable trade-off until a useful quantum computer becomes available (see Appendix B for a complete list of acronyms and abbreviations used throughout this research).

Development of a defense against a distinct type of white-box targeted adversarial attack with machine learning or deep learning models running on classical computing hardware [6]. This type of defense evolves as the targeted white-box adversarial attack

evolves. A compounded targeted white-box adversarial attack is the result of the evolutionary nature of developing a more sophisticated type of attack necessary to render the defense useless. In turn, the defense against the more sophisticated type of attack continues to evolve to a point necessary to render an attack useless. The targeted white-box adversarial attack has transcended from classical computation to simulated quantum computation.

1.1.1 Classical Computing

Classical computing refers to traditional computing paradigms that rely on classical bits to store and process information. Classical computing refers to the use of standard computing hardware and software to build, train, and test neural network models. Classical computing has been widely used in the field of machine learning and artificial intelligence and is the foundation for many of the state-of-the-art neural network models used today [7]. A challenge of classical computing in the context of adversarial attacks is that neural network models are limited by the computational resources available on classical computing hardware. This can make it difficult to train neural network models that are robust to a wide range of adversarial attacks, as the neural network models may require large amounts of data and computational resources to train.

1.1.2 Quantum Computing

Quantum computing is a new computing paradigm that uses the principles of quantum mechanics to process information. Quantum computing has the potential to provide new approaches to creating accurate neural network models [8]. Quantum computing may one day replace classical computing for certain types of computing, such as optimization problems and cryptography. However, compared to classical computing, quantum computing is limited in terms of computing power and memory.

1.1.3 A Neural Network Model

A convolutional neural network (CNN) is a classical neural network widely used for image classification tasks [9]. In hybrid neural network (HNN) architectures, a CNN serves as the classical component that integrates with a quantum circuit to form a quantum neural network (QNN) [10]. The combination of classical and quantum components results in a hybrid model, referred to as an HNN. In this work, a deeper HNN architecture is constructed to support experimentation with white-box targeted and compound adversarial attacks.

1.1.4 Compromising a Neural Network Model

A targeted adversarial attack is when a classifier within a model has been targeted for misclassification as part of the adversarial attack [11]. In contrast, an untargeted attack is when there is no target classifier within a model specified as part of an adversarial attack. An adversarial attack targeted with the white box occurs when an attacker knows the architecture, parameters, and training dataset of the model for an attack against the model. The white-box, targeted, and compound adversarial attack is a combination of two white-box, targeted, and distinct adversarial attacks.

1.1.5 Distinct Adversarial Attacks

A white-box targeted and distinct adversarial attack aims to have a classifier from a neural network model make a wrong prediction (or misclassification) [12]. The algorithms selected to be used within a white-box, targeted distinct adversarial attack include the fast gradient sign method (FGSM), projected gradient descent (PGD), and Carlini and Wagner attack (CW). The unique aspects of each of the attack types are considered as a justification for selecting FGSM, PGD, and CW.

First, the FGSM attack was selected to represent single-step gradient-based attacks. The FGSM is a straightforward single-step gradient-based algorithm that maximizes the loss function. The components of the FGSM are input, loss function and calculated gradient output using the gradient sign operator that is added to the input [13], [14].

Second, the PGD attack was selected to represent a different type of optimization-based attack than the CW optimization-based attack. The PGD uses the same generation process as BIM except for randomly perturbed images within a specific neighborhood. In the case of PGD, the PGD attack must have access to the model gradients and a copy of the model weights [15], [16].

Finally, the Carlini and Wagner (CW) attack was selected to represent an optimization-based attack. CW is slightly different from the gradient-based algorithms mentioned above. The CW seeks an adversarial instance for an input and then uses optimization of the loss function with an Adam optimizer to identify the adversarial instance. The CW attack is an iterative attack. The CW attack is often much slower than other attacks [17].

1.1.6 Compounded Adversarial Attacks

In this research study, FGSM, CW, and PGD were selected and combined to form white-box, targeted, and compound adversarial attacks (WTC) against the HNN model. The adversarial attacks of the WTC form three compound adversarial attacks, including FGSM+PGD, FGSM+CW, and CW+PGD. As with other technological advances, adversarial attacks by WTC against neural networks will continue to evolve from distinct adversarial attacks to more sophisticated compound adversarial attacks [18]. The adversarial attacks of the WTC will require the defenses to evolve to prevent compromise of neural network models, such as an HNN model.

1.1.7 Defenses Against Compounded Adversarial Attacks

A set of representative defenses against adversarial attacks from the WTC includes input transformations, randomization, and adversarial training. Defenses can be circumvented by more sophisticated adversarial attacks over time. The defense of input transformations (or input data preprocessing) represents a set of techniques including image quilting, adaptive logit pairing (ALP), and differential privacy [19]. However, the set of defense techniques has not been shown to be as effective as adversarial training. Randomization defense represents a special case of the dropout technique used to hide the model or

constrain the model for minimal data retrieval [20]. The adversarial training defense represents a technique that trains the model classifier with adversarial examples so that the model classifier can have adversarial information [21]. The model can then adapt on the basis of the learned adversarial data.

1.1.8 Model's Prediction Accuracy

The evaluation of performance for a neural network in the context of adversarial WTC attacks can be achieved by calculating the accuracy metric [22]. The accuracy of the HNN model can be affected by adversarial attacks in two ways: first, the model may misclassify adversarial examples, leading to reduced accuracy on these examples; second, the HNN model may be over-conservative in its predictions, leading to decreased accuracy on clean examples. To evaluate the accuracy of the HNN model in the presence of adversarial attacks, it is common to use benchmark datasets that contain both clean and adversarial examples. The accuracy, in the context of adversarial attacks, refers to the proportion of correctly classified examples, either clean (e.g., nonadversarial) or adversarial. The accuracy of the HNN model is calculated as the proportion of correctly classified examples, either on clean examples, adversarial examples, or both. The accuracy of a neural network model can be influenced by various factors, such as the type of adversarial attack, the strength of the attack, and the choice of defense mechanism.

1.2 Statement of the Problem with Motivation

A problem identified from a review of previous research highlights an existing gap associated with the lack of research conducted on effective defenses against WTC adversarial attacks for HNN models. The motivation for addressing this gap suggests that the development of a targeted white-box adversarial attack will evolve over time. A particular defense will also have to evolve as the sophistication increases with the development of a targeted white-box adversarial attack against an HNN model.

1.2.1 Problem Statement

First, there is a gap in previous investigations in which researchers did not take into account the effectiveness of defense mechanisms to protect a quantum-classical neural network or hybrid neural network (HNN) model against a white-box, targeted, and compound adversarial attack (WTC). Second, identify the most effective (or optimal) defense mechanisms to protect an HNN model against a WTC adversarial attack.

1.2.2 Motivation

First, the interest in defending against white-box, targeted, and compound adversarial attacks (WTC) is the motivation behind the solution. Second, a comparison of defense mechanisms to protect against the HNN model may be both of interest and informative for the cybersecurity defender when investigating a WTC adversarial attack.

1.3 Thesis Statement

This research study establishes the hypothesis that a hybrid neural network (HNN) model can be protected against adversarial attacks by incorporating defense mechanisms against adversarial threats from WTC. The thesis is supported by experimental results that demonstrate that an HNN model equipped with appropriate defenses maintains acceptable prediction accuracy across multiple datasets, including MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns. In contrast, an HNN model without defense mechanisms fails to achieve acceptable classification accuracy under WTC adversarial attacks when evaluated on the same datasets. These findings highlight the importance of defense strategies in preserving the robustness of HNN models in adversarial environments.

1.4 Research Questions

This study is guided by two primary research questions that address the effectiveness and optimality of defense mechanisms against adversarial attacks on hybrid neural network

(HNN) models. The first question examines the overall effectiveness of defense mechanisms against white-box, targeted, and compound (WTC) adversarial attacks, while the second seeks to identify the optimal defense strategy among those evaluated.

1.4.1 Research Question (RQ1)

RQ1: How effective are defense mechanisms at protecting HNN models against white-box, targeted, and compound (WTC) adversarial attacks?

1.4.2 Research Question (RQ2)

RQ2: Which defense mechanism is the most effective (optimal) for protecting HNN models against WTC adversarial attacks?

1.5 Hypotheses

This research establishes three hypotheses to be tested through experimental evaluation. The hypotheses are structured to address both research questions, with H1 and H2 collectively informing RQ1 by comparing HNN model performance with and without defense mechanisms, and H3 addressing RQ2 by identifying the optimal defense strategy. Together, the hypotheses provide a testable framework for evaluating the adversarial resilience of HNN models across multiple datasets, including MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns.

1.5.1 Hypothesis (H1)

H1: An HNN model without any defense mechanism will exhibit unacceptable prediction accuracy when subjected to white-box targeted compound (WTC) adversarial attacks across multiple datasets, including MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns.

1.5.2 Hypothesis (H2)

H2: An HNN model incorporating defense mechanisms will achieve acceptable prediction accuracy following a WTC adversarial attack, using the same set of datasets as input.

1.5.3 Hypothesis (H3)

H3: Among the defense strategies evaluated, at least one optimal defense mechanism will allow the HNN model to maintain the highest acceptable prediction accuracy under adversarial conditions of the WTC across all datasets.

1.5.4 Hypotheses Mapping

The mapping of hypotheses to research questions is as follows. H1 and H2 together answer RQ1 by comparing the HNN model performance with and without defense mechanisms under WTC adversarial attacks, thereby demonstrating whether the evaluated defenses are effective. H3 answers RQ2 by identifying which specific defense mechanism yields the highest acceptable prediction accuracy, establishing the optimal defense among those evaluated.

1.6 Objectives of This Study

This study pursues four primary objectives designed to systematically investigate the robustness of hybrid neural network (HNN) models against adversarial attacks and to evaluate the effectiveness of candidate defense mechanisms. The objectives progress logically from model construction through attack design, defense implementation, and ultimately identification of optimal defense strategies. Together, these objectives provide a comprehensive framework for understanding and improving the adversarial resilience of HNN architectures.

1.6.1 Objective (OBJ1)

OBJ1: Construct a hybrid neural network with an interpretable architecture.

1.6.2 Objective (OBJ2)

OBJ2: Design and evaluate multiple compound attacks.

1.6.3 Objective (OBJ3)

OBJ3: Implement and compare three defense strategies:

3.1 Input Transformation

3.2 Randomization

3.3 Adversarial Training

1.6.4 Objective (OBJ4)

OBJ4: Identify the most effective defense(s) using performance metrics.

1.6.5 Structured Experimentation

To achieve these objectives, the project establishes a structured experimental environment to evaluate and measure a representative set of defense mechanisms that protect an HNN model against white-box targeted compound (WTC) adversarial attacks. This includes the development of Python implementations for the HNN model, compound adversarial attacks, and the defense mechanisms, as well as the application of systematic procedures to evaluate and compare defense performance across multiple datasets.

1.6.6 Expected Outcomes

The expected outcomes of this research include: (1) successful demonstration of defense mechanisms protecting an HNN model against adversarial WTC attacks; (2) measurement of prediction accuracy for each defense mechanism under adversarial conditions; (3)

comparative analysis of prediction accuracy across all evaluated defense mechanisms; and (4) identification of the most effective defense strategy based on post-attack prediction performance.

1.6.7 Primary Deliverables

The primary deliverables for this research study are as follows. First, multiple datasets—MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns—will be prepared for input into the HNN model. Second, a 9-layer hybrid neural network incorporating a quantum circuit with four qubits will be implemented in Python. Third, Python code will be developed for three white-box targeted compound (WTC) adversarial attacks. Fourth, various defense mechanisms will be implemented to protect the HNN model against these attacks, with code to execute WTC attacks with and without defenses applied. Fifth, the prediction accuracy of the HNN model will be evaluated and compared before and after adversarial attacks. Finally, the most effective defense mechanism will be identified based on post-attack prediction performance.

1.7 Research Goal

The goal of this research is the successful completion of each of the objectives, outcomes, and deliverables outlined for this study. Specifically, the goal of the research is to demonstrate that defense mechanisms protect an HNN model against adversarial attacks by white-box targeted compounds (WTC). The HNN model employs a classifier architecture and is evaluated against three combinations of compound adversarial attacks: FGSM+PGD, FGSM+CW, and CW+PGD. The defense mechanisms The defense mechanisms evaluated include input transformations, randomization, and adversarial training.

The research goal is directly supported by each of the four study objectives. OBJ1 establishes the foundation by constructing the HNN model with an interpretable architecture, providing the target system against which all attacks and defenses are evaluated. OBJ2 advances the goal by designing and evaluating the compound adversarial attacks of WTC (FGSM+PGD, FGSM+CW, and CW+PGD) that challenge the HNN model.

OBJ3 directly addresses the core of the research goal by implementing and comparing the three candidate defense strategies, namely input transformation, randomization, and adversarial training, against those attacks. Finally, OBJ4 completes the goal by identifying the most effective defense mechanism using prediction accuracy as the primary performance metric, thus demonstrating which strategy best protects the HNN model under adversarial conditions.

1.8 Key Technical Ideas

The key technical ideas mentioned in the research include the CNN model, the QNN model, the quantum circuit, and the HNN model. The CNN model represents a conventional computer architecture in which several interconnected processors mirror the types of connections between neurons in a human brain. A CNN model can be learned by a trial-and-error process. A QNN model is like their CNN model counterparts, where the architecture establishes a structure that inputs from one layer and passes that input onto another layer. That is, each layer evaluates the data and passes on the output to the next layer. However, a QNN model differs from its CNN model counterparts with the addition of a quantum circuit. The HNN model is an intermediate neural network that links a CNN model to a QNN model.

1.9 Research Overview

This research evaluates the robustness of an HNN model – integrating classical convolutional neural network layers with a 4-qubit parameterized quantum circuit – against compound WTC adversarial attacks across four benchmark image datasets. Figure 1.1 illustrates the experimental framework.

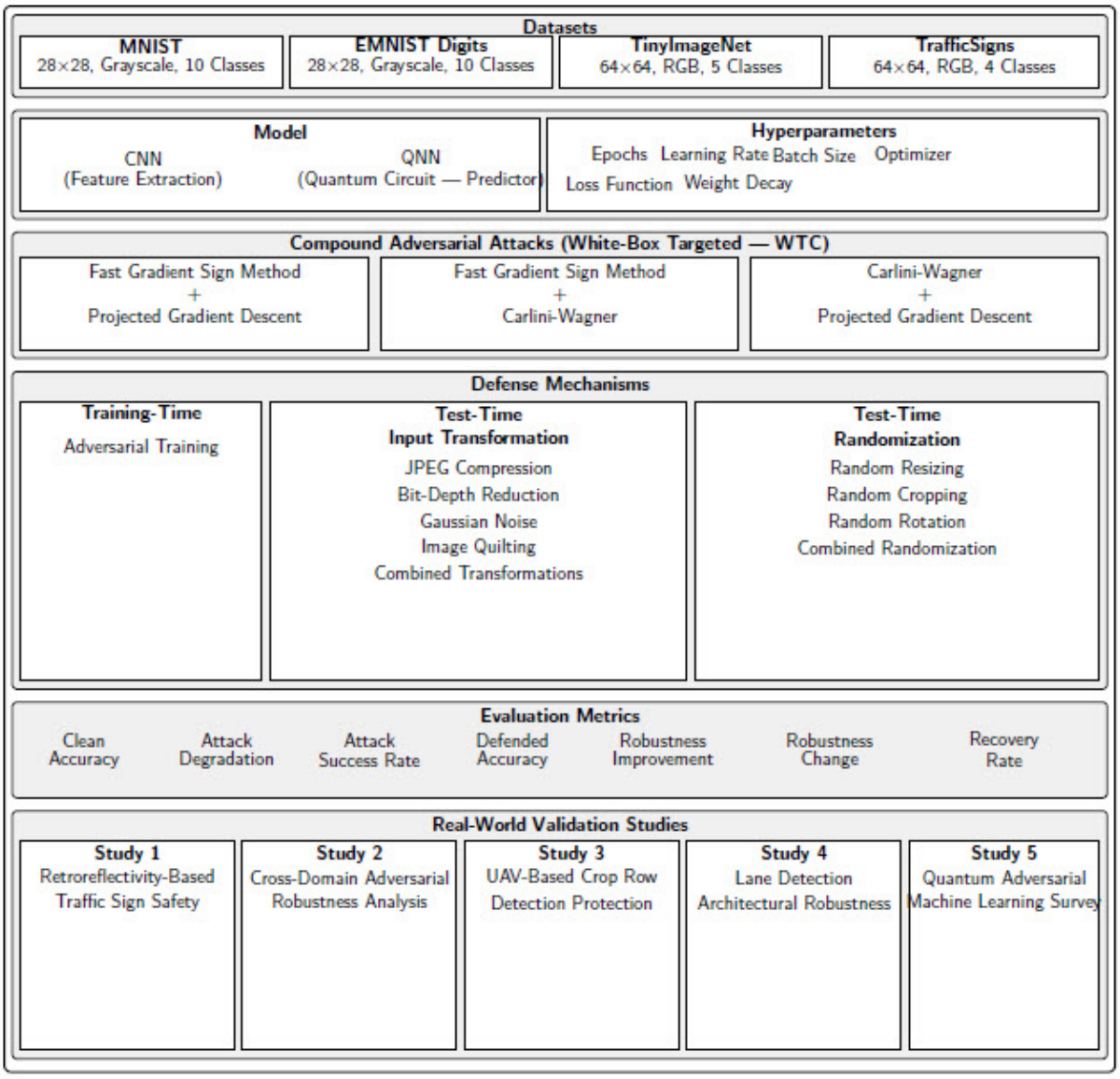

Figure 1.1: Experimental framework for this research

Four benchmark datasets — MNIST and EMNIST "Digits" (28×28 grayscale, 10 classes), TinyImageNet (64×64 RGB, 5 classes), and TrafficSigns (64×64 RGB, 4 classes) — train the HNN model combining a CNN for feature extraction and a QNN quantum circuit for prediction, configured with dataset-specific hyperparameters (epochs, learning rate, batch size, optimizer, loss function, and weight decay). Three combinations of WTC attacks – Fast Gradient Sign Method + Projected Gradient Descent, Fast Gradient Sign Method + Carlini-Wagner, and Carlini-Wagner + Projected Gradient Descent – are evaluated across 120 experimental conditions. Three defense categories are as-

sessed: adversarial training (training-time); five input transformation techniques (JPEG compression, bit-depth reduction, Gaussian noise, image quilting, and combined transformations); and four randomization techniques (random resizing, random cropping, random rotation, and combined randomization), both at test-time. Seven metrics quantify defense effectiveness: clean accuracy, attack degradation, attack success rate, defended accuracy, robustness improvement, robustness change, and recovery rate. Findings are contextualized through five real-world validation studies spanning traffic sign safety, cross-domain robustness, UAV-based crop row detection, lane detection, and a quantum adversarial machine learning survey.

1.10 Contribution of this Work

The contributions in this paper are summarized as follows: First, we demonstrate that defense mechanisms against compounded WTC adversarial attacks can be shown to be effective in protecting an HNN model compared to an HNN model without defense mechanisms. Second, the most effective (or optimal) defense for an HNN model can be determined against a WTC adversarial attack.

1.11 Summary

This chapter established the research context for defending hybrid neural networks against white-box targeted compound adversarial attacks. Current quantum computing systems demonstrate 433 qubits, while useful quantum computers require over 1000 qubits, making classical-quantum hybrid approaches viable. Three distinct adversarial attacks form the basis for compound attacks: FGSM as a gradient-based method, PGD as an iterative optimization approach, and CW as an optimization-based attack. These combine to form compound attacks of FGSM+PGD, FGSM+CW, and CW+PGD. Three defense categories were introduced: input transformations, randomization techniques, and adversarial training. The research hypothesizes that HNN models without defenses will fail under compound attacks, that incorporating defenses will restore acceptable accuracy, and that

at least one optimal defense will emerge across MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns datasets.

Chapter 2

WTC ATTACKS AGAINST HYBRID MODELS

This chapter reviews related work on adversarial attacks and defenses, identifies critical research gaps in hybrid neural network security, analyzes limitations of previous approaches, and establishes novel contributions of proposed research.

2.1 Background on Adversarial Attacks

Adversarial attacks pose significant threats to deep learning models by introducing carefully crafted perturbations to input data that cause misclassification while remaining imperceptible to human observers [23], [24]. These attacks have been extensively studied in the context of classical convolutional neural networks (CNNs), where various attack algorithms have demonstrated the vulnerability of even highly accurate models [25]. White-box attacks, where adversaries have complete knowledge of model architecture, parameters, and training data, represent the most sophisticated threat model. Targeted attacks further refine this by forcing misclassification to specific incorrect classes rather than arbitrary errors.

Among white-box targeted attacks, three algorithms have emerged as particularly effective: Fast Gradient Sign Method (FGSM), which generates perturbations based on the gradient of the loss function; Projected Gradient Descent (PGD), which applies iterative refinement to find stronger perturbations; and Carlini and Wagner (CW), which uses optimization techniques to find minimal perturbations that cause misclassification. While

each individual attack has been studied extensively, compound attacks that combine multiple methods represent a more sophisticated and realistic threat model that remains underexplored, particularly for hybrid quantum-classical neural networks.

2.2 Compound Adversarial Attacks

Compound adversarial attacks combine multiple attack strategies to exploit different vulnerabilities in neural network models. Unlike single-method attacks that rely on one perturbation strategy, compound attacks leverage the complementary strengths of different algorithms. For example, combining FGSM (fast gradient-based perturbations) with PGD (iterative refinement) creates attacks that are both computationally efficient and highly effective. Similarly, combining CW (optimization-based minimal perturbations) with PGD (iterative strengthening) produces attacks that are difficult to detect due to small perturbation magnitudes while being highly successful at causing misclassification.

This research investigates three compound attack combinations: FGSM+PGD, FGSM+CW, and CW+PGD. These combinations represent different trade-offs between the magnitude of the disturbance, the computational cost, and the success rate of the attack. Understanding the effectiveness of compound attacks is crucial for developing robust defense mechanisms, as defenses optimized against single-method attacks may fail against more sophisticated compound strategies. The threat posed by compound attacks is particularly concerning for hybrid neural networks (HNNs), where the integration of classical and quantum components may introduce additional vulnerabilities not present in purely classical architectures.

2.3 Hybrid Neural Networks

Hybrid neural networks (HNNs) integrate layers of classical convolutional neural networks with quantum circuit components, combining the proven effectiveness of classical deep learning with the potential computational advantages of quantum processing. A typical HNN architecture consists of classical convolutional layers for feature extraction,

a quantum circuit layer that processes features in quantum state space, and classical fully-connected layers for final classification. The quantum component typically employs parameterized quantum circuits [26] with rotation gates for state preparation and controlled-NOT (CNOT) gates for entanglement.

While HNNs offer potential advantages in certain computational tasks, their hybrid architecture also introduces unique security considerations [27], [28]. The quantum components may respond differently to adversarial perturbations compared to the classical layers, and the interaction between classical and quantum processing may create new attack surfaces. Previous research has primarily focused on adversarial attacks against purely classical CNNs or purely quantum neural networks (QuNNs), leaving a critical gap in understanding how compound adversarial attacks affect hybrid architectures. A comprehensive survey of quantum adversarial machine learning (see Chapter 5, Study 5) contextualizes this gap within the broader research landscape. This gap is particularly important given the increasing interest in deploying HNN models for practical applications where adversarial robustness is essential.

2.4 Defense Mechanisms Against Adversarial Attacks

Defense mechanisms against adversarial attacks can be categorized based on their application timing and approach to achieving robustness. Evaluation of related work on defense mechanisms to protect neural networks against adversarial attacks reveals three primary categories: input transformation, randomization, and adversarial training [25].

2.4.1 Input Transformation Defenses

The input transformation defenses focus on modifying adversarial input during inference to remove or reduce perturbations before classification [29], [30]. These test-time defenses include techniques such as JPEG compression (removing high-frequency perturbations through lossy compression), bit-depth reduction (quantizing pixel values), Gaussian noise addition (masking perturbations with random noise), and image quilting (reconstructing images from clean patches). Input transformation defenses offer the advantage of deploy-

ment flexibility—they can be applied to pre-trained models without retraining and can be easily switched or combined. However, they must balance perturbation removal against preserving legitimate image content, as overly aggressive transformations degrade clean accuracy.

2.4.2 Randomization Defenses

Randomization defenses introduce stochastic transformations to inputs during inference, making it more difficult for adversaries to create transferable adversarial examples. These test-time defenses include random resizing (scaling images to different sizes within specified ranges), random cropping (selecting random image regions), random rotation (rotating images by random angles), and combinations thereof. The stochastic nature of randomization means that the same adversarial input undergoes different transformations across multiple evaluations, potentially providing ensemble-like robustness. Like input transformation, randomization defenses can be applied without model retraining, but effectiveness depends on careful calibration of randomization ranges to maintain clean accuracy while providing defense.

2.4.3 Adversarial Training Defense

The adversarial training defense fundamentally modifies the model by training on augmented datasets containing clean and adversarial examples [31], [32]. This training-time defense requires generating diverse adversarial examples from the training set, combining them with clean data, and retraining the model to correctly classify both types of input. Unlike test-time defenses that apply transformations during inference, adversarial training creates models with inherently robust decision boundaries. While computationally expensive and requiring anticipation of attack strategies during training, adversarial training has demonstrated superior effectiveness compared to test-time approaches in classical CNN settings. However, its effectiveness for HNN models against compound attacks remains unexplored.

2.5 Related Work on HNN Security

Recent investigations have begun to explore adversarial threats against quantum-enhanced models, though their scope remains limited compared to the extensive work on classical CNNs. Several studies provide a relevant context to understand the current state of HNN security research.

2.5.1 Quantum Neural Network (QuNN) Studies

Research on purely quantum neural networks has investigated basic adversarial attacks. RobQuNNs [33] and AdvQuNN [34] evaluate FGSM+PGD attacks against QuNNs using MNIST data, but focus on quantum circuit design (entanglement and expressibility) as defense mechanisms rather than comprehensive defense strategies [28]. These studies demonstrate QuNN vulnerability to gradient-based attacks, but do not address compound attacks or systematic defense evaluation. The limitation to single-method attacks (FGSM or PGD individually) means that defense effectiveness against more sophisticated compound attack combinations remains unknown.

2.5.2 Hybrid Quantum Neural Network (HQNN) Studies

Research that specifically targets hybrid architectures has explored specific attack scenarios, but lacks a comprehensive defense evaluation. Backdoor attacks on HQNNs [35] investigate color-triggered backdoor attacks, evaluating trigger success rates without proposing defense mechanisms. LCQHNN [36] presents a lean HQNN structure using simplified 4-layer variational quantum circuits, but focuses on architecture efficiency rather than adversarial robustness. A study of adversarial patterns of natural noise and light [37] investigates perturbations of the physical world against HNNs, but relies primarily on preprocessing filters rather than comprehensive defense mechanisms such as adversarial training.

2.5.3 Classical CNNs Defense Studies

Extensive research on classical CNNs provides a foundational understanding of defense mechanisms. Studies demonstrate the effectiveness of adversarial training [31], [32], [38], generalized adversarial training approaches [39], and various input transformations [29], [30], and randomization techniques. However, these investigations focus exclusively on classical architectures and do not address the unique challenges posed by hybrid quantum-classical models. The interaction between classical and quantum components in HNNs may require adapted defense strategies, and the effectiveness of classical defense mechanisms when applied to hybrid architectures requires empirical validation.

Table 2.1 summarizes the landscape of related work, highlighting the gap that this research study addresses: comprehensive evaluation of defense mechanisms against compound adversarial attacks for hybrid neural networks.

Table 2.1: Comparison of defense mechanisms against adversarial attack-related work.

Related Work	**White-box Targeted Adversarial Attack Algorithm**	**Defenses Against White-box Targeted Adversarial Attack**	**Targeted Model Type**	**Dataset**
This work	FGSM, CW, PGD, Compound Attacks (FGSM+PGD, FGSM+CW, CW+PGD)	Input Transformation, Randomization, Adversarial Training	HNN	MNIST, EMNIST "Digits", Tiny ImageNet, TrafficSigns
[38]	AutoAttack, Semantic Attacks, Full Attacks	Adversarial Training	CNN	CIFAR-10, ImageNet
[39]	Three Attacks, Semantic Attacks, Full Attacks	Generalized Adversarial Training	CNN	CIFAR-10, ImageNet
[33]	FGSM, PGD	Quantum Circuit Design (Entanglement, Expressibility)	QuNN	MNIST
[34]	FGSM, PGD	None (focused on attack analysis only)	QuNN	MNIST
[35]	Backdoor-triggered evasion	None (evaluated backdoor trigger success rate)	HQNN (focus on backdoors)	Custom (color-triggered)
[36]	FGSM (basic evaluation)	Quantum Circuit Simplification	Lean HQNN (4-layer VQC)	MNIST, Fashion-MNIST
[37]	Natural Noise + Light Adversarial Patterns	Noise Filtering & Preprocessing	HNN	MNIST

2.6 Research Gap Identification

The evaluation of related work reveals a critical gap in understanding defense mechanism effectiveness against compound white-box targeted (WTC) adversarial attacks for hybrid neural networks [24], [25]. While previous research has extensively studied adversarial

attacks against classical CNNs and has begun exploring attacks against purely quantum or hybrid models, several important questions remain unanswered.

2.6.1 Gap 1: Compound attacks against HNN

Previous investigations focus primarily on single-method attacks (FGSM alone, PGD alone, or CW alone) against HNN models. Compound attacks that combine multiple perturbation strategies represent a more sophisticated and realistic threat model, yet their effectiveness against HNNs remains unexplored. The interaction between different attack methods when combined may produce synergistic effects not captured by studying attacks individually. Understanding how HNN models respond to compound attacks is essential for developing defenses that provide robust protection in real-world adversarial environments.

2.6.2 Gap 2: Systematic Defense Evaluation for HNNs

While defense mechanisms such as input transformation, randomization, and adversarial training have been extensively evaluated for classical CNNs, their effectiveness when applied to HNN models against compound attacks has not been systematically investigated. The hybrid architecture of HNNs—combining classical convolutional layers with quantum circuits—may respond differently to defense mechanisms compared to purely classical models. Empirical evaluation across multiple defense categories, attack types, and datasets is needed to identify optimal defense strategies for HNN deployment.

2.6.3 Gap 3: Test-Time vs. Training-Time Defense

Previous work does not clearly distinguish between test-time defenses (applied during inference without model retraining) and training-time defenses (requiring model retraining on augmented data) for HNN models. Understanding the trade-offs between deployment flexibility and robustness is crucial for practical defense selection. The relative effectiveness of these two defense categories against compound attacks, and whether patterns

observed in classical CNNs translate to hybrid architectures, requires systematic investigation.

2.6.4 Gap 4: Dataset-Dependent Defense Effectiveness

Prior research typically evaluates defenses on a single dataset or a limited variety of datasets. Defense effectiveness may vary substantially based on dataset characteristics such as visual complexity, intra-class variation, and structural regularity. Comprehensive evaluation across diverse datasets—from simple handwritten digits to complex natural images—is needed to understand generalization of defense strategies and inform deployment decisions for different application domains.

2.6.5 Gap 5: Classical Simulation Methodology

Full quantum neural network research is constrained by limited access to quantum hardware and significant usage costs. While HNN models can be simulated on classical hardware using frameworks such as PyTorch and Cirq, the viability of classical simulation for comprehensive defense evaluation has not been established. Validating that classical simulation enables meaningful defense research while avoiding quantum hardware constraints would enable a broader investigation of HNN security.

2.7 Limitations of Previous Approaches

Previous approaches to HNN security exhibit several limitations that constrain their applicability to real-world deployment scenarios and leave critical questions unanswered.

2.7.1 Limited Attack Sophistication

Most prior work evaluated defenses against single-method attacks (FGSM, PGD, or CW individually), which do not represent the full sophistication of potential adversarial threats. Real-world adversaries are likely to employ compound attack strategies that combine multiple methods to exploit different vulnerabilities. Defenses optimized against

single-method attacks may provide inadequate protection against compound attacks, yet this scenario has not been systematically investigated for HNN models.

2.7.2 Narrow Defense Coverage

Research studies that evaluate defenses for quantum-enhanced models typically focus on a single defense category (e.g., quantum circuit design or preprocessing filters) rather than a comprehensive comparison across multiple defense approaches. Without systematic evaluation of input transformation, randomization, and adversarial training applied to the same HNN architecture under the same experimental conditions, optimal defense selection remains uncertain.

2.7.3 Insufficient Dataset Diversity

Previous HNN security research predominantly uses MNIST or Fashion-MNIST datasets, which consist of simple grayscale images with low visual complexity. Defense effectiveness on these simple datasets may not generalize to more complex visual recognition tasks involving color images, natural scenes, or domain-specific imagery. Evaluation across varying dataset complexity levels is essential for understanding defense applicability to diverse real-world applications.

2.7.4 Lack of Quantitative Comparison

Many studies describe defense mechanisms qualitatively or evaluate them in isolation without quantitative comparison to alternative approaches. Without head-to-head comparison under controlled experimental conditions with consistent metrics, practitioners lack guidance for selecting defenses based on empirical effectiveness. Establishing baseline measurements and quantitative comparisons is essential for advancing the field beyond exploratory investigations.

2.7.5 Quantum Hardware Constraints

Research that requires actual quantum hardware is limited by accessibility and cost, constraining the scope of experimentation. While some work has explored HNN simulation on classical hardware, the validity of classical simulation for defense evaluation has not been explicitly validated. Establishing that classical simulation produces meaningful results would allow for greater participation in research and a more comprehensive experimental evaluation.

2.8 Novel Contributions of This Research

This research study addresses the identified gaps and overcomes limitations of previous approaches through several novel contributions that advance the understanding of HNN adversarial robustness.

2.8.1 Contribution 1: Compound Attack Evaluation

This research conducts the first systematic evaluation of compound white-box targeted adversarial attacks (FGSM+PGD, FGSM+CW, CW+PGD) against HNN models. By evaluating three compound attack combinations across four datasets using consistent experimental methodology, the research establishes baseline measurements of HNN vulnerability to sophisticated adversarial threats. The findings reveal that compound attacks achieve 73-84% attack success rates across varying dataset complexity levels, demonstrating the severity of the threat and the necessity of robust defense mechanisms.

2.8.2 Contribution 2: Defense Mechanism Comparison

The research provides a comprehensive comparative evaluation of three defense categories—input transformation (test-time), randomization (test-time), and adversarial training (training-time)—applied to the same HNN architecture under controlled experimental conditions. Evaluation across 120 experimental conditions (4 datasets $\times$ 3 attack types $\times$ 10 defense techniques) enables quantitative comparison with statistical reliability. The

findings establish that adversarial training achieves 58.5% average defended accuracy compared to 25.5% for randomization and 20.9% for input transformation, representing a 2.3–2.8× performance advantage.

2.8.3 Contribution 3: Test-Time vs Training-Time Framework

This research establishes a clear conceptual framework distinguishing test-time defenses (applied during inference without model retraining) from training-time defenses (requiring model retraining on augmented data). The framework reveals fundamental trade-offs between deployment flexibility and robustness, informing practical defense selection based on application requirements. While test-time defenses offer easy deployment and adaptability, training-time defenses provide substantially superior robustness—a pattern that has important implications for HNN deployment in adversarial environments.

2.8.4 Contribution 4: Multi-Dataset Validation

By evaluating defenses across four datasets with varying characteristics—MNIST (simple handwritten digits), EMNIST "Digits" (diverse handwriting styles), Tiny ImageNet (small-scale natural images), and TrafficSigns (domain-specific symbols)—the research demonstrates that defense effectiveness is strongly dataset-dependent. TrafficSigns achieves 37% average defended accuracy while EMNIST "Digits" achieves only 19%, revealing that dataset complexity significantly influences defense performance. This finding provides guidance for defense selection based on task characteristics.

2.8.5 Contribution 5: Classical Simulation Validation

The research validates that classical simulation using PyTorch and Cirq provides a viable methodology for comprehensive HNN defense evaluation. By successfully conducting 120 experimental conditions on conventional hardware that would be prohibitively expensive on quantum hardware, the research demonstrates that classical simulation enables meaningful defense research while quantum hardware remains in early development stages. This validation enables broader research participation and more extensive experimentation.

2.9 Research Approach Distinction

The approach taken in this research differs fundamentally from previous work in several important ways that enable the novel contributions described above.

2.9.1 Focus on Compound Attacks

While previous work concentrates on single-method attacks (FGSM, PGD, or CW individually), this research specifically investigates compound attack combinations (FGSM+PGD, FGSM+CW, CW+PGD). This focus reflects realistic threat models where adversaries employ sophisticated multi-method strategies rather than simple single-algorithm attacks. The compound attack focus necessitates more comprehensive defense evaluation, as defenses effective against single methods may fail against compound strategies.

2.9.2 Hybrid Architecture Emphasis

Rather than focusing exclusively on classical CNNs or purely quantum neural networks, this research specifically targets hybrid neural networks that integrate classical and quantum components. This emphasis is motivated by practical considerations—HNNs can be simulated on classical hardware using frameworks like PyTorch and Cirq, avoiding the accessibility and cost constraints of full quantum implementations while retaining architectural advantages of quantum processing. The hybrid architecture may respond differently to adversarial attacks and defenses compared to purely classical or quantum models, necessitating dedicated investigation.

2.9.3 Systematic Multi-Defense Evaluation

Unlike previous work that evaluates individual defense mechanisms in isolation, this research conducts a head-to-head comparison of three defense categories (input transformation, randomization, adversarial training) under controlled experimental conditions. All defenses are evaluated on the same HNN architecture, against the same attacks, using the same datasets and metrics. This systematic approach enables quantitative comparison

and definitive identification of optimal defenses based on empirical evidence rather than qualitative assessment.

Multi-Dataset Validation Strategy: While previous HNN security research typically uses single datasets (usually MNIST), this research employs four datasets with varying characteristics. This multi-dataset strategy reveals dataset-dependent defense effectiveness patterns that would be invisible in single-dataset studies. The dataset selection spans simple to complex visual recognition tasks, enabling assessment of defense generalization across application domains.

Practical Deployment Orientation: The research framework distinguishes test-time defenses (deployable without retraining) from training-time defenses (requiring retraining), explicitly considering practical deployment constraints. This orientation provides actionable guidance for practitioners who must select defenses based on both effectiveness and deployment feasibility. The finding that training-time defenses achieve 2.3–2.8× superior performance informs the fundamental trade-off between deployment flexibility and robustness.

2.10 Parallel Investigation: Traffic Sign Safety Applications

While this research study establishes foundational defense mechanisms for HNN models using benchmark datasets, a parallel investigation extends these findings to safety-critical traffic sign classification (see Chapter 5, Study 1). That work evaluates retroreflectivity-based traffic sign safety classification (safe versus unsafe based on MUTCD federal thresholds) using synthetic data generated by conditional GANs to address dataset scarcity and class imbalance (86.9% safe versus 13.1% unsafe signs).

The retroreflectivity study employs the same compound adversarial attacks evaluated in this research study (FGSM+PGD, FGSM+CW, CW+PGD) against both CNN and HNN architectures. Key findings align with this research study: FGSM+PGD emerges as the most damaging attack (reducing CNN accuracy to 18.75%), and adversarial training

proves most effective among defenses (restoring accuracy to 70.98%). The study reports that HNN architectures achieve superior clean accuracy (95.98%) compared to standard CNNs (91.52%), validating the potential of hybrid quantum-classical approaches for real-world applications.

The convergence of the findings between the benchmark evaluation (this research study) and the real-world application (see Chapter 5, Study 1) validates that the defense effectiveness patterns are generalized from controlled experimental conditions to practical deployment scenarios. However, the retroreflectivity study introduces domain-specific challenges—synthetic data generation, class imbalance, and safety-critical classification thresholds—that extend beyond the scope of this foundational investigation. These complementary studies together demonstrate that adversarial training provides consistent superior protection across both benchmark datasets and application-specific contexts, establishing a unified understanding of defense mechanisms for hybrid neural networks.

2.11 Cross-Domain Adversarial Robustness Analysis

Recent cross-domain analysis synthesizes adversarial robustness findings across three safety-critical applications: traffic signs, UAV crop detection, and lane detection (see Chapter 5, Study 2). That work reveals a critical insight about architectural trade-offs—while geometric prior architectures achieve 4-6% clean accuracy improvements (95.98-97.89%), they sacrifice adversarial robustness under compound attacks, degrading from 99.93% (standard CNN) to 0-64% across all domains ($p<0.001$). This finding has important implications for the hybrid neural network architectures evaluated in this research study.

The cross-domain findings validate this research study's core conclusion: adversarial training provides substantial robustness improvement (278% in cross-domain study), establishing a standard CNN with adversarial training as the preferred architecture for safety-critical deployment requiring >95% robustness under ISO 26262 and DO-178C certification standards. The convergence of results—this research study's 2.3–2.8× adversarial training advantage and the cross-domain study's 278% improvement—provides

strong evidence that training-time defenses consistently outperform test-time approaches across diverse application contexts.

However, the cross-domain study identifies domain-specific vulnerability patterns not addressed in this foundational work: multi-modal features amplify adversarial gradients beyond spectral normalization bounds, and robustness requirements for safety certification (>95%) exceed typical benchmark evaluation standards. These findings suggest that while foundational defense mechanisms (established in this research study) generalize across domains, safety-critical deployment requires additional domain-specific hardening. The relationship between HNN architectures (this research study) and geometric prior CNNs (cross-domain study) warrants further investigation to determine whether quantum circuit components introduce similar amplification effects as multi-modal geometric features.

2.12 Summary

This chapter has established the foundation for the research study by reviewing adversarial attacks and defense mechanisms, identifying critical research gaps in HNN security, analyzing limitations of previous approaches, and articulating the novel contributions of this research. The evaluation reveals that while adversarial attacks against classical CNNs have been extensively studied and initial work on quantum-enhanced models has begun, a critical gap exists in understanding defense effectiveness against compound attacks for hybrid neural networks.

The identified gaps—lack of compound attack evaluation, absence of systematic defense comparison, unclear test-time vs training-time distinctions, insufficient dataset diversity, and invalidated classical simulation methodology—motivate the comprehensive experimental investigation presented in subsequent chapters. The research approach addresses these gaps through systematic evaluation across 120 experimental conditions, providing quantitative evidence for defense effectiveness and establishing a foundational understanding for deploying robust HNN models in adversarial environments.

Parallel investigations in traffic sign safety classification and cross-domain robustness analysis validate that the findings established in this research study generalize to real-

world applications and diverse architectural approaches. These complementary studies demonstrate the broader impact of this foundational research while identifying domain-specific challenges that extend beyond benchmark evaluation.

The following chapters build upon this foundation: Chapter 3 presents the theoretical background and hypotheses, Chapter 4 describes the detailed methodology, including defense procedures, Chapter 5 presents comprehensive experimental results with statistical analysis, and Chapter 6 synthesizes findings and outlines future research directions. Together, these chapters provide a complete investigation of defense mechanisms for HNNs against compound adversarial attacks.

Chapter 3

METHODOLOGY

This chapter covers a discussion on the chosen research method, the research design, the reason for the chosen research method, and the implementation of the research method. The methodology used for research enables other researchers to reproduce the research necessary to validate the results and ensure that the research is both transparent and credible.

3.1 Research Method

The methodology to be applied is a quantitative research method. The method and the research design are both well-suited for machine learning experimentation [33], [40], [41] associated with the construction and evaluation of model prediction capabilities.

3.2 Research Design

An experimental research design that aligns well with machine learning research [34], [40], [41] will be implemented. The alignment between experimental research design and machine learning will be demonstrated in several ways through activities associated with the following: model comparison, feature selection, data preprocessing, hyperparameter tuning, and model interpretability.

3.2.1 Model Comparison

The performance can be compared among different models for the same experimental conditions [35], [40]. A researcher can determine the strengths and limitations of a model by selecting appropriate datasets, defining training and testing procedures, and using metrics to measure the performance of models.

3.2.2 Feature Selection

The most important features of a model are selecting characteristics [36]. Feature selection occurs when a subset of the most relevant features is taken from a larger set of features associated with a model that makes an optimal prediction.

3.2.3 Data Pre-processing

An evaluation of the impact of data pre-processing can be achieved among different data pre-processing techniques when applied to a model. Data preprocessing involves taking input data for a model and transforming it to improve the performance of the model. A preprocessing technique can potentially lead to optimal model performance when testing a model with different preprocessed datasets [37].

3.2.4 Hyperparameter Tuning

A model's parameters can be set before model training. These parameters are also called hyperparameters [42]–[46], where a model's parameters can be tuned (e.g., varied) under controlled conditions. The purpose of tuning is to identify the best combination of the parameters of a model that will lead to the optimal performance of a model.

3.2.5 Model Interpretability

The results can be interpreted for a model [47]. Any biases or limitations of the results from the model can be understood by understanding how the model arrived at the predic-

tion [48]. Understanding the underlying relationships between the variables of a model, in addition to any areas for improvement, is an important aspect of model interpretability.

3.3 Reason for Choosing Research Design

The reason for choosing the research design as a quantitative method is not only popular within machine learning, but also useful within machine learning, given the way the research design aligns with machine learning research [49]. The use of research design assists in the following: addressing research questions, establishing causality, controlling for confounding variables, ensuring reproducibility, minimizing bias, and optimizing the performance of a model.

3.4 Research Method Implementation

Implementing the research method represents a process to answer a research question. The process involves executing the research plan and collecting data in a way that is consistent with the chosen research design and methodology. In this case, the implementation of the research method consists of conducting an experimental research design, developing a validation approach, and assembling the data collection.

3.4.1 Experimental Research Design

In this case, the objective of the experimental research design is to obtain data for the purposes of training (e.g., learning) machine learning models (e.g., HNN model), and to assess the prediction accuracy of each of the prediction capabilities of the model. The measurement of prediction accuracy represents an effectiveness measurement for a particular model. The HNN model is constructed and evaluated for a comparison between pre-attack and post-attack, where the post-attack defense mechanism effectiveness can be obtained. Identification of the most effective (or optimal) defense mechanisms against a WTC adversarial attack is compared with other defense mechanisms.

3.4.2 Validation Approach

A validation approach provides a framework for testing and evaluating the performance of the model [50]. The validation approach helps ensure that the experimental research design is rigorous and that the results are reliable and valid.

3.4.3 Data Collection

The quality of the collected data is important for the performance and precision of the model [51]. The data collected are representative of the problem domain and will be properly prepared, labeled, split, and balanced. The data collected will be used to train and test the model.

3.5 Dataset-Specific Model Architecture

The HNN model architecture is adapted for each dataset to accommodate varying input dimensions, complexity levels, and classification requirements. Table 3.1 presents the dataset-specific architectural configurations used in this research. The architecture design follows a principle of using shallow networks for smaller, simpler datasets (MNIST, EMNIST "Digits") and progressively deeper networks with batch normalization and dropout for more complex datasets (TinyImageNet, TrafficSigns).

For MNIST and EMNIST "Digits" datasets, which consist of 28×28 grayscale handwritten digit images, a shallow two-layer convolutional architecture is employed. This architecture uses single-channel input (grayscale) and processes images through two convolutional layers (4 and 10 feature maps, respectively) followed by two fully connected layers. The simplicity of handwritten digits and the small input dimensions do not require deep feature extraction, making this shallow architecture sufficient for achieving high baseline accuracy while avoiding overfitting on the relatively small filtered training sets (12,665 for MNIST, 48,000 for EMNIST).

For TinyImageNet, which consists of 64×64 RGB natural images with higher visual complexity, a deeper three-layer convolutional architecture is employed with progressive

channel widening (32→64→128 feature maps). This architecture accommodates three-channel RGB input and incorporates batch normalization after each convolutional layer to stabilize training. A higher dropout rate (0.6) is used in the fully connected layer to prevent overfitting, given the limited training data (2,000 samples) relative to the dataset's visual complexity. The architecture outputs 5 classes corresponding to the filtered Tiny-ImageNet subset used in experiments.

For TrafficSigns, which consists of 64×64 RGB images of traffic symbols with moderate visual complexity but distinctive features, a balanced three-layer convolutional architecture is used (16→32→64 feature maps). The architecture incorporates batch normalization after each convolutional layer and moderate dropout (0.5) before the final classification layer. With 4 output classes (crosswalk, speed limit, stop, traffic light) and 1,200 training samples, this architecture provides sufficient capacity for feature learning while maintaining computational efficiency. The moderate depth prevents overfitting while capturing the structured visual patterns characteristic of traffic signs.

The quantum circuit component remains consistent across all datasets: a 4-qubit circuit with rotation gates ($R_y(\theta)$ and $R_y(\phi)$) for state preparation and CNOT gates for entanglement. This quantum layer processes the output from the classical convolutional layers and feeds into the final classification layers, maintaining the hybrid quantum-classical architecture across all experimental conditions.

Table 3.1: Dataset-specific HNN model architecture configurations.

Dataset	Input Dimensions	Convolutional Layers (channels)	Fully Connected Layers (units)	Regularization	Output Classes
MNIST	28×28×1 (grayscale)	Conv1: 1→4 (5×5 kernel) Conv2: 4→10 (5×5 kernel)	FC1: 160→50 FC2: 50→10	None	10 (digits 0-9)
EMNIST "Digits"	28×28×1 (grayscale)	Conv1: 1→4 (5×5 kernel) Conv2: 4→10 (5×5 kernel)	FC1: 160→50 FC2: 50→10	None	10 (digits 0-9)
Tiny ImageNet	64×64×3 (RGB)	Conv1: 3→32 (3×3, BN) Conv2: 32→64 (3×3, BN) Conv3: 64→128 (3×3, BN)	FC1: 8192→256 FC2: 256→5	Dropout (p=0.6)	5 (selected classes)
TrafficSigns	64×64×3 (RGB)	Conv1: 3→16 (3×3, BN) Conv2: 16→32 (3×3, BN) Conv3: 32→64 (3×3, BN)	FC1: 4096→128 FC2: 128→4	Dropout (p=0.5)	4 (sign types)
Note: BN = Batch Normalization; All architectures use 2×2 max pooling after each convolutional layer					

3.6 Dataset-Specific Hyperparameters

Hyperparameter selection is crucial for model performance [44]–[46] and directly impacts training dynamics, convergence behavior, and generalization capability. Table 3.2 presents the dataset-specific hyperparameter configurations used throughout all experimental conditions. These hyperparameters were selected based on dataset characteristics, training set size, and task complexity, balancing training efficiency with model performance.

The number of training epochs varies substantially across datasets based on training set size and convergence requirements. MNIST and EMNIST "Digits" use 10 epochs, which is sufficient given their relatively large training sets (12,665 and 48,000 samples, respectively) and the simplicity of handwritten digit recognition. TrafficSigns uses 60 epochs to compensate for the smaller training set (1,200 samples) and the need for robust

feature learning from limited data. TinyImageNet requires 120 epochs—the longest training duration—due to both the small training set (2,000 samples) and the higher visual complexity of natural images requiring more iterations for convergence.

The learning rate remains constant at 0.001 across all datasets, using the Adam optimizer with adaptive learning rate adjustments. This learning rate provides stable convergence without requiring manual learning rate schedules. However, weight decay (L2 regularization) varies by dataset: MNIST and EMNIST use weight decay of 1×10^{-4}, while TinyImageNet uses a higher weight decay of 3×10^{-4} to provide stronger regularization given the risk of overfitting with limited training data and a deeper network architecture.

Batch size is set to 1 across all datasets to maximize gradient update frequency and provide fine-grained parameter adjustments. While larger batch sizes could improve training speed, the small filtered training sets and the need for precise convergence justify the use of batch size 1. The loss function (Negative Log-Likelihood Loss) and activation function (Log Softmax) remain consistent across datasets, providing stable training dynamics for multi-class classification.

The parameters of the quantum circuit—rotation angles $\theta = 0.159\pi$ and $\phi = 0.095\pi$ for the preparation of the state—are kept constant across all datasets to isolate the impact of defense mechanisms and dataset characteristics on the robustness of the model. This consistency ensures that variations in experimental results stem from dataset properties and defense effectiveness rather than quantum circuit configuration differences.

Table 3.2: Dataset-specific hyperparameter configurations for HNN model training.

Dataset	**Training Epochs**	**Batch Size**	**Learning Rate**	**Optimizer**	**Weight Decay (L2)**	**Loss Function**
MNIST	10	1	0.001	Adam	1×10^{-4}	NLLoss
EMNIST "Digits"	10	1	0.001	Adam	1×10^{-4}	NLLoss
Tiny ImageNet	120	1	0.001	Adam	3×10^{-4}	NLLoss
TrafficSigns	60	1	0.001	Adam	1×10^{-4}	NLLoss
Note: All models use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$; Activation: Log Softmax						

The quantum circuit hyperparameters remain constant across all datasets to maintain

consistency in the quantum component of the hybrid architecture. The 4-qubit circuit uses rotation angles $\theta = 0.159\pi$ for qubits 0 and 2, and $\phi = 0.095\pi$ for qubits 1 and 3, applied via R_y rotation gates. Entanglement is created through a linear chain of CNOT gates (qubit $0\rightarrow1$, $1\rightarrow2$, $2\rightarrow3$), providing quantum correlation between qubit states. These parameters were selected to provide sufficient quantum expressibility while remaining implementable on near-term quantum hardware and amenable to classical simulation using Cirq.

3.7 Rationale for Dataset-Specific Configurations

The dataset-specific architectural and hyperparameter choices reflect principled design decisions based on established machine learning practices and the specific characteristics of each dataset.

3.7.1 Architecture Depth

Shallow architectures (2 convolutional layers) are used for MNIST and EMNIST to match the low complexity of grayscale handwritten digits. Deeper architectures (3 convolutional layers) are employed for RGB datasets (TinyImageNet, TrafficSigns) to extract hierarchical features from color images. This prevents both underfitting (insufficient capacity for complex tasks) and overfitting (excessive capacity for simple tasks).

3.7.2 Batch Normalization

Applied only to RGB datasets (TinyImageNet, TrafficSigns) where deeper networks and varied visual content benefit from internal covariant shift reduction [52]. Grayscale digit datasets with shallow architectures achieve stable training without batch normalization, avoiding unnecessary computational overhead. Batch normalization stabilizes activations and accelerates training by normalizing layer inputs, providing a mild regularizing effect that complements other techniques.

3.7.3 Dropout Regularization

Dropout rates are calibrated based on overfitting risk [52], [53]: none for MNIST/EMNIST (sufficient training data), moderate (0.5) for TrafficSigns (limited data with structured features), and high (0.6) for TinyImageNet (limited data with complex features). This prevents overfitting while preserving model capacity for learning robust features. Dropout introduces stochastic noise during training by randomly deactivating neurons, preventing complex co-adaptations and enhancing generalization.

3.7.4 Training Duration

Epoch count inversely correlates with training set size: fewer epochs (10) for large datasets (MNIST: 12,665, EMNIST: 48,000) and more epochs (60-120) for small datasets (TrafficSigns: 1,200, TinyImageNet: 2,000). This ensures convergence while avoiding overfitting on limited data.

3.7.5 Weight Decay

Higher weight decay (3×10^{-4}) for TinyImageNet provides stronger L2 regularization to counter overfitting risks from the combination of a small training set, deep architecture, and complex visual features. Standard weight decay (1×10^{-4}) suffices for other datasets with either larger training sets or simpler architectures.

3.7.6 Configuration Choices

These configuration choices create dataset-appropriate model architectures that balance capacity, regularization, and training efficiency, enabling fair comparison of defense mechanism effectiveness across varying dataset characteristics and complexity levels.

3.8 Summary

This chapter established the methodological framework for investigating defense mechanisms against adversarial attacks on HNN models. The quantitative experimental research design aligns with machine learning best practices [40], [41], enabling systematic evaluation of model performance through controlled experimentation, rigorous validation, and reproducible procedures.

The experimental research design supports comprehensive evaluation through five key dimensions: model comparison across defense mechanisms, feature selection through convolutional architectures, data preprocessing for adversarial robustness, hyperparameter tuning [43]–[46] for optimal performance, and model interpretability through robustness metrics. This multi-dimensional approach ensures that experimental findings are robust, reproducible, and generalizable across varying dataset characteristics.

Dataset-specific architectural configurations were justified based on visual complexity and the size of the training set. Shallow two-layer convolutional networks suffice for simple grayscale digit recognition (MNIST, EMNIST), while deeper three-layer architectures with batch normalization [52] and dropout [52], [53] are necessary for complex RGB natural images (TinyImageNet) and structured traffic symbols (TrafficSigns). The quantum circuit component remains constant across all datasets, maintaining the 50-50 classical-quantum balance while isolating defense effectiveness from quantum configuration variations.

Hyperparameter selection follows principled design decisions [44]–[46] that balance training efficiency with generalization capability. Training duration inversely correlates with dataset size (10-120 epochs), regularization strength scales with overfitting risk (weight decay 1-3$\times 10^{-4}$, dropout 0-0.6), and batch size is set to 1 for fine-grained gradient updates. These configurations create fair experimental conditions where defense mechanism effectiveness can be evaluated across datasets of varying complexity without confounding factors from inappropriate model architecture or training procedures.

The methodology establishes a foundation for transparent and reproducible research that enables the validation of the experimental findings and extension to additional datasets, defense mechanisms, or adversarial attack strategies. The systematic approach

to architectural design, hyperparameter selection, and experimental validation ensures that conclusions about defense effectiveness are grounded in rigorous experimental methodology rather than dataset-specific artifacts or suboptimal model configurations.

Chapter 4

HYBRID MODEL DEFENSE MECHANISMS

This chapter includes the proposed solution, key points, procedures, adversarial attacks, defense mechanisms, strengths, and contributions. The procedures are distinguished by the timing of defense application: test-time defenses (input transformation and randomization) are applied during inference without model retraining, while training-time defenses (adversarial training) require model retraining on augmented datasets.

4.1 Proposed Solution

The proposed work investigates a problem with a gap expressed by looking at defense mechanisms to minimize the impact on an HNN model after a WTC adversarial attack [54], [55]. In past investigations, researchers have not taken into account the effectiveness of the defense after a WTC adversarial attack against a classical-quantum neural network. By investigating a defense mechanism to thwart a WTC adversarial attack, the effectiveness of the defense mechanism after the WTC adversarial attack may lead to more sophisticated defense mechanisms with potential countermeasures for protecting a HNN model.

4.2 Key Points

The key points of emphasis within the research include the following critical aspects presented in Table 4.1:

Table 4.1: Key points of emphasis in HNN defense research.

No.	Key Point	Significance
1	Development of HNN Model	QNN models are limited due to access and usage cost constraints on quantum computing hardware
2	Alternative Architecture	Use of intermediate neural network (HNN) linking CNN to QNN as cost-effective alternative
3	Classical Simulation	HNN simulated on conventional computing hardware enables practical research
4	Vulnerability Assessment	HNN models, like CNN models, are vulnerable to WTC adversarial attacks
5	Defense Effectiveness	Effective defense mechanisms can minimize impact of WTC adversarial attacks on HNN models

4.2.1 Development Environment

The development environment utilized three primary open-source frameworks as detailed in Table 4.2:

Table 4.2: Development environment frameworks and their purposes.

Framework	Developer	Purpose in Research
PyTorch	Meta/Facebook	Open-source framework for efficient development of deep learning and machine learning models [56], [57]
Cirq	Google	Quantum computing framework for designing, simulating, and executing quantum circuits; implements quantum components of HNN model
Torchattacks	Community	PyTorch-based library providing comprehensive adversarial attack implementations [58], [59]; generates WTC adversarial examples (FGSM+PGD, FGSM+CW, CW+PGD)

4.2.2 Dataset Selection and Progression Rationale

The datasets selected for this research—MNIST, EMNIST "Digits", TrafficSigns, and TinyImageNet—follow a deliberate progression from simple to complex visual recognition tasks. This progression enables systematic evaluation of defense mechanisms across varying levels of dataset complexity, revealing how model architecture and defense effectiveness scale with task difficulty.

MNIST represents the simplest case: 28×28 grayscale images of handwritten digits with low visual complexity and high structural regularity. This dataset serves as the basis for establishing defense effectiveness under ideal conditions where visual features are simple and distinctive. EMNIST "Digits" introduces increased handwriting variation while maintaining the same simple digit recognition task, testing whether defense mechanisms robust to MNIST generalize to more diverse handwriting styles within the same problem domain.

TrafficSigns increases complexity by introducing 64×64 RGB images with domain-specific symbols that require recognition under varying visual conditions. The transition from handwritten digits to TrafficSigns represents a shift from simple character recognition to symbol classification with implications for real-world deployment for autonomous

vehicle systems. This dataset reveals whether the defenses effective for simple digit recognition maintain effectiveness for structured but more complex visual patterns.

TinyImageNet represents the highest complexity level: 64×64 RGB natural images with diverse object categories, complex backgrounds, and high intra-class variation. Natural images lack the structural regularity of handwritten digits or traffic symbols, requiring deeper convolutional architectures and more sophisticated feature extraction. This dataset tests whether defense mechanisms scale to general-purpose visual recognition tasks beyond specialized domains.

The progression from simple to complex datasets requires the corresponding progression in the complexity of the model architecture. MNIST and EMNIST require only shallow 2-layer convolutional networks without regularization, while TrafficSigns and Tiny ImageNet require deeper 3-layer architectures with batch normalization and dropout to handle increased visual complexity. This dataset-driven architecture scaling enables investigation of whether adversarial robustness patterns observed for simple models and datasets generalize to more complex, realistic scenarios approaching real-world deployment conditions.

4.2.3 Operational Implications of Defense Timing

A critical finding emerging from this research concerns the operational feasibility of defense deployment in production environments [55], [60]. The distinction between test-time defenses (input transformation, randomization) and training-time defenses (adversarial training) has profound implications beyond just effectiveness measurements—it fundamentally affects deployment strategy in operational cyber defense contexts.

From a cyber operational perspective, test-time defenses offer significant deployment advantages that became apparent through experimental implementation. Test-time defenses can be introduced, modified, or combined in deployed systems without model retraining, enabling rapid response to emerging adversarial threats. When a new attack strategy is discovered in operational environments, defenders can deploy input transformation or randomization techniques immediately by modifying inference pipelines without touching the trained model. Multiple test-time defenses can be rapidly combined or

swapped to adapt to evolving threat landscapes, providing tactical flexibility in adversarial environments.

In contrast, training-time defenses require complete model retraining on augmented datasets that contain both clean and adversarial examples. This retraining process occurs at the development time, requiring: (1) generation of adversarial examples from training data using anticipated attack strategies, (2) combination of clean and adversarial training samples into augmented datasets, (3) complete model retraining from scratch on augmented data, and (4) validation and testing before deployment. For the datasets and models evaluated in this research, retraining requires 10-120 epochs, depending on the size and complexity of the dataset, representing substantial computational investment and time delay before deployment.

The development-time constraint of adversarial training creates operational challenges in dynamic threat environments. If adversaries develop novel attack strategies post-deployment, defenders using adversarial training must: collect examples of new attacks, regenerate augmented training datasets, retrain models completely, and redeploy—a process taking days to weeks depending on model complexity and organizational deployment procedures. During this retraining period, systems remain vulnerable to novel attacks. Test-time defenses avoid this delay by enabling immediate deployment of new transformations or randomization strategies at the inference stage.

However, this operational flexibility comes at a cost: test-time defenses achieve only 20.9-25.5% average defended accuracy compared to 58.5% for adversarial training [60] (see Chapter 5). This creates a fundamental operational trade-off: accept lower robustness (test-time defenses) in exchange for rapid deployment and tactical flexibility, or accept longer deployment timelines (training-time defenses) in exchange for superior robustness. For systems where deployment speed is critical and modest robustness improvement is acceptable, test-time defenses provide viable operational solutions. For systems where maximum robustness is mandatory regardless of deployment timeline—such as safety-critical applications requiring >95% defended accuracy for certification—adversarial training remains necessary despite operational constraints.

This operational perspective suggests that hybrid defense strategies may offer optimal

trade-offs: deploy adversarially-trained models as robust baselines during development, then add test-time defenses in production to provide additional adaptive capability against emerging threats without retraining. Such layered approaches could combine the superior robustness of training-time defenses with the operational flexibility of test-time defenses, although empirical evaluation of such combinations remains future work.

4.2.4 Quantum Circuit Constraints on Dataset Size

The quantum circuit component of the HNN model imposes practical constraints on dataset size that necessitated the use of filtered subsets rather than complete datasets [54], [61]. The 4-qubit parameterized quantum circuit employed in this research creates a quantum state space of dimension $2^4 = 16$, which determines the number of quantum features that can be processed simultaneously. This quantum dimensionality constraint directly affects how many classical features can be encoded into quantum states and subsequently how many training samples can be efficiently processed during model training.

Classical simulation of quantum circuits using the Cirq framework requires maintaining quantum state vectors in classical memory, with memory requirements scaling exponentially with qubit count. For the 4-qubit circuit, each quantum state requires storing $2^4 = 16$ complex probability amplitudes, manageable on conventional hardware. However, increasing to 5 qubits would require $2^5 = 32$ amplitudes, 6 qubits would require $2^6 = 64$ amplitudes, and so forth, rapidly becoming computationally intractable for classical simulation as qubit count grows.

The constraint to 4 qubits—balancing quantum expressiveness against classical simulation feasibility—in turn constrains the number of training samples that can be processed efficiently. Larger datasets like full MNIST (60,000 training samples) or full EMNIST (240,000+ training samples) would require proportionally longer training times and greater memory overhead when combined with quantum circuit simulation. To maintain reasonable computational requirements across all experimental conditions (120 total conditions: 4 datasets $\times$ 3 attacks $\times$ 10 defenses), filtered dataset subsets were employed: 12,665 MNIST training samples, 48,000 EMNIST training samples, 2,000 Tiny ImageNet training samples, and 1,200 TrafficSigns training samples.

These filtered subsets provide sufficient training data for meaningful model training and defense evaluation while remaining computationally tractable for comprehensive experimentation on conventional hardware. The quantum circuit constraint thus represents a practical limitation of current classical simulation capabilities rather than a fundamental limitation of HNN architectures. Future implementation on actual quantum hardware—where quantum state evolution occurs physically rather than through classical simulation—could potentially accommodate larger datasets without the exponential memory scaling limitations of classical simulation. However, for the purpose of establishing baseline defense effectiveness patterns and validating the viability of classical simulation for HNN defense research, the 4-qubit configuration with filtered datasets provides an appropriate balance between experimental scope and computational feasibility.

4.2.5 Design Context

The Python packages used for the construction and evaluation of the HNN model include `PyTorch`, `Cirq`, and `Torchattacks`, as shown in Fig. 4.1 and Fig. 4.2. These frameworks support the development of HNN models, the application of adversarial attacks, and the deployment of defenses within the HNN pipeline.

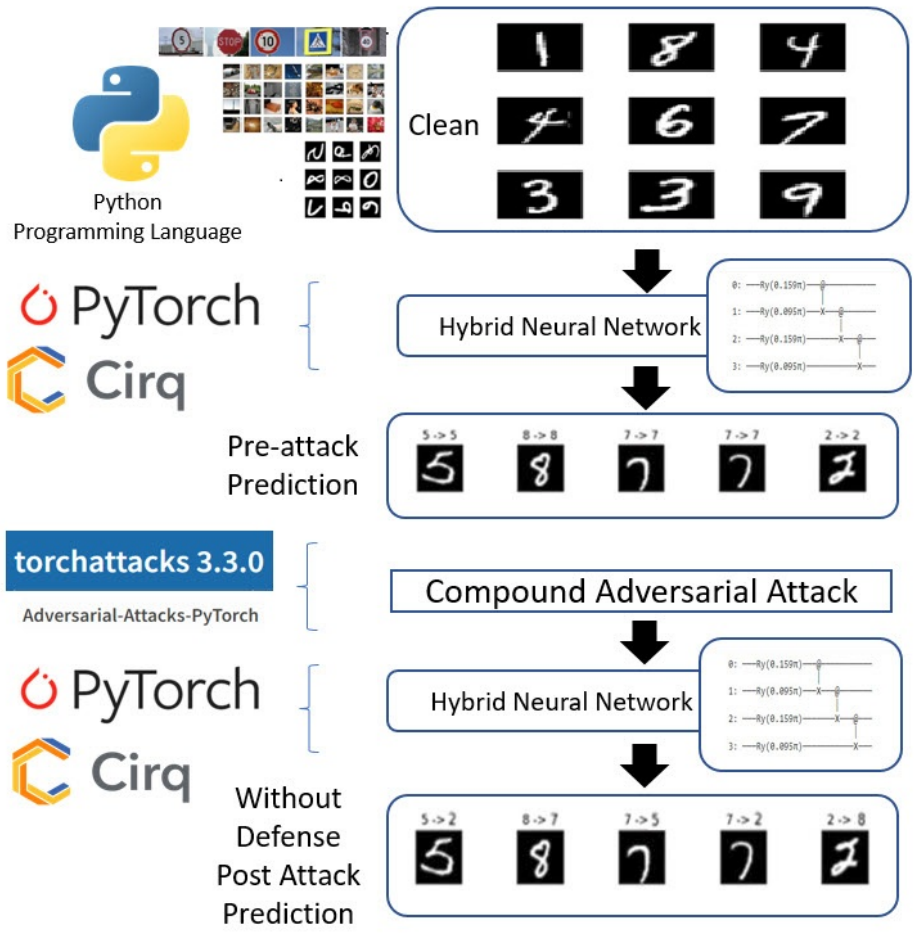

Figure 4.1: Design context for the HNN model and processing without a defense mechanism.

The main distinction between the two designs lies in the implementation of adversarial defense mechanisms and the timing of their application. Fig. 4.1 illustrates the HNN model pipeline without any post-attack defenses, where the compound adversarial attack directly impacts the prediction performance. In contrast, Fig. 4.2 incorporates post-attack defenses, which help mitigate the effects of adversarial perturbations and restore the predictive accuracy of the model. For test-time defenses (input transformation and randomization), these transformations are applied to adversarial samples during inference. For training-time defenses (adversarial training), the model itself is retrained on augmented data containing both clean and adversarial examples. This distinction represents a critical design decision in the secure deployment of hybrid models under adversarial

conditions.

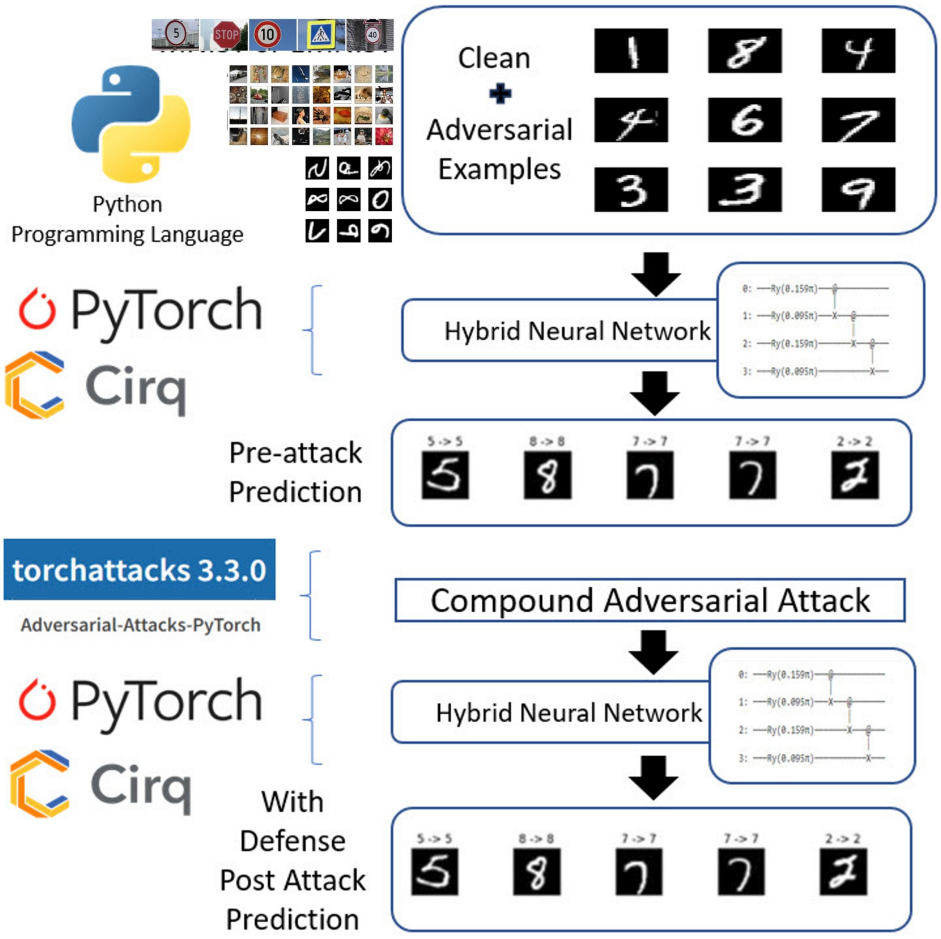

Figure 4.2: Design context for the HNN model and processing with a defense mechanism applied post-attack.

4.2.6 HNN Model Architecture and Parameters

The hybrid neural network (HNN) model integrates layers of classical convolutional neural networks with quantum circuit components [54], [61], creating an architecture that is approximately 50% classical and 50% quantum by layer count. The classical portion consists of convolutional layers for feature extraction followed by fully connected layers, while the quantum portion consists of a 4-qubit parameterized quantum circuit that processes features in quantum state space before final classification. This balanced hybrid architecture enables the model to take advantage of both proven classical deep learning capabilities and potential quantum computational advantages.

Model parameters are organized into two categories: base parameters that remain constant across all datasets (providing consistency in quantum-classical integration and training methodology) and dataset-specific variations that adapt the architecture to accommodate different input dimensions, levels of visual complexity, and classification requirements. This separation enables a systematic comparison of the effectiveness of defense across datasets while maintaining architectural consistency where appropriate.

4.2.6.1 Base HNN Model Parameters

Table 4.3 presents the core parameters that remain constant across all experimental conditions. These parameters define the fundamental characteristics of the hybrid classical-quantum architecture, training methodology, and evaluation approach. The use of a simple train-test split without validation data reflects a practical approach suitable for the filtered dataset sizes, with models trained on the designated training set and evaluated on the separate test set. The quantum circuit parameters ($\theta = 0.159\pi$, $\phi = 0.095\pi$) implement a Quantum Reservoir Mapping approach (detailed in Section 4.2.7) where the quantum component functions as a fixed high-dimensional transformation rather than as a trainable layer. This architectural choice ensures that the experimental variations stem from the effectiveness of the defense rather than the configuration differences of the quantum circuit, forcing the classical layers to develop inherently robust feature representations.

Table 4.3: Base HNN model parameters (constant across all datasets).

No.	Parameter	Value	Description
1	Model Type	HNN (Hybrid Neural Network)	Classical CNN + Quantum Circuit
2	Quantum Architecture	4-qubit parameterized circuit	State preparation + entanglement
3	Quantum Rotation θ	0.159π	Applied to qubits 0, 2
4	Quantum Rotation ϕ	0.095π	Applied to qubits 1, 3
5	Quantum Entanglement	Linear CNOT chain	Qubits $0\rightarrow1\rightarrow2\rightarrow3$
6	Batch Size	1	Maximizes gradient frequency
7	Learning Rate	0.001	Adam optimizer
8	Optimizer	Adam	β_1=0.9, β_2=0.999
9	Activation Function	Log Softmax	Multi-class classification
10	Loss Function	NLLoss (Negative Log-Likelihood)	Standard for classification
11	Validation Approach	Simple train-test split	No validation data used
12	Evaluation Metric	Prediction Accuracy (%)	Clean, attacked, defended

4.2.6.2 Dataset-Specific Model Variations

Table 4.4 presents the parameters that vary by dataset to accommodate different input characteristics, visual complexity levels, and computational requirements. The most significant variations include: (1) the number of training epochs, which inversely correlates with training set size (10 epochs for large datasets, 120 epochs for small complex datasets), (2) weight decay (L2 regularization), which is higher for TinyImageNet (3×10^{-4}) to prevent overfitting with limited training data, (3) convolutional architecture depth, ranging from 2 layers for simple grayscale digits to 3 layers for complex RGB images, and (4) regularization techniques, with batch normalization and dropout applied only to RGB datasets with deeper networks.

Table 4.4: Dataset-specific HNN model parameter variations.

Dataset	**Input**	**Conv Architecture**	**Regularization**	**Train/Test**	**Epochs**
MNIST	28×28×1	2-layer: 1→4→10	None	12,665 / 4,230	10
EMNIST "Digits"	28×28×1	2-layer: 1→4→10	None	48,000 / 16,000	10
Tiny ImageNet	64×64×3	3-layer: 3→32→64→128	BN + Drop(0.6)	2,000 / 800	120
TrafficSigns	64×64×3	3-layer: 3→16→32→64	BN + Drop(0.5)	1,200 / 400	60
Weight Decay (L2): MNIST/EMNIST/TrafficSigns = 1×10^{-4}; Tiny ImageNet = 3×10^{-4}					
Output Classes: MNIST/EMNIST = 10 digits; Tiny ImageNet = 5 classes; TrafficSigns = 4 sign types					

Note: Input format is height×width×channels (grayscale=1, RGB=3). Conv architecture shows channel progression. BN = Batch Normalization, Drop = Dropout. All models use 2×2 max pooling after each convolutional layer and two fully connected layers.

These dataset-specific variations reflect principled design decisions based on visual complexity, the size of the training set, and the risk of overfitting. Shallow architectures (2 convolutional layers) suffice for simple grayscale digits (MNIST, EMNIST), while deeper architectures (3 convolutional layers) with batch normalization and dropout are necessary for complex RGB images (TinyImageNet, TrafficSigns). Training duration inversely correlates with dataset size: 10 epochs for large datasets (12,665-48,000 samples), 60 epochs for moderate datasets (1,200 samples), and 120 epochs for small complex datasets (2,000 samples). The higher weight decay for TinyImageNet provides stronger regularization to counteract overfitting from the combination of a small training set, deep architecture, and complex visual features.

4.2.7 Quantum Reservoir Mapping

A critical architectural decision in this research concerns the treatment of the quantum circuit component as a fixed transformation layer rather than a trainable component with learnable parameters. While traditional hybrid quantum-classical architectures frequently employ Variational Quantum Circuits (VQC) with trainable quantum weights [61], this research deliberately adopts a Quantum Reservoir Mapping approach for the quantum processing head. This design choice represents a principled architectural strategy with specific advantages for adversarial defense research rather than a limitation of the implementation.

In the Quantum Reservoir Mapping paradigm, the quantum circuit functions as a static high-dimensional Hilbert space filter with fixed parameters. By establishing the quantum rotation angles at pre-optimized values ($\theta = 0.159\pi$ for qubits 0 and 2, $\phi = 0.095\pi$ for qubits 1 and 3), the 4-qubit circuit creates a constant quantum transformation that maps classical feature vectors into a fixed $2^4 = 16$-dimensional Hilbert space—the complex-valued vector space where quantum states exist. This fixed Hilbert space mapping serves as an unchanging projection operator that classical features must learn to navigate effectively during training.

This architectural approach serves a dual purpose that directly supports the research objectives. First, by maintaining a constant quantum transformation in Hilbert space throughout all experimental conditions, the architecture forces the classical convolutional neural network encoder to learn feature representations that are inherently robust enough to map successfully into this fixed quantum manifold. The classical layers cannot rely on the adaptation of quantum parameters to compensate for weak or fragile features; instead, they must develop intrinsically robust representations that remain stable under adversarial perturbation when projected through the fixed quantum transformation. This constraint encourages the classical component to learn more generalizable and robust feature encodings.

Second, the fixed quantum parameters eliminate the "moving target" problem that complicates the attribution of robustness improvements in systems with trainable quan-

tum components. During adversarial training or defense evaluation, if quantum parameters were trainable, observed robustness changes could stem from either: (a) genuine improvements in defense mechanism effectiveness or (b) stochastic convergence of quantum weights to different local optima across experimental runs. By fixing quantum parameters across all 120 experimental conditions (4 datasets $\times$ 3 attack types $\times$ 10 defense techniques), this research ensures that measured improvements in defended accuracy are strictly attributable to the defensive mechanisms applied to classical layers and input transformations, rather than quantum parameter variation.

This design choice aligns with reservoir computing principles [62] where a fixed high-dimensional random projection provides sufficient computational expressiveness while concentrating trainable parameters in readout layers. In the HNN context, the quantum circuit serves as the reservoir—a fixed nonlinear transformation that expands the representational capacity—while classical convolutional and fully connected layers serve as both the input encoder and the readout mechanism. The quantum component thus provides quantum computational advantages (superposition, entanglement) without introducing the experimental confounds associated with trainable quantum parameters in adversarial settings.

From an experimental design perspective, the fixed quantum configuration provides essential consistency for systematic defense evaluation. Adversarial training experiments comparing FGSM+PGD attacks against FGSM+CW attacks across MNIST versus Tiny-ImageNet datasets can isolate the effects of attack strategy and dataset complexity precisely because the quantum transformation remains constant. If quantum parameters varied between conditions, disentangling the contributions of attack type, dataset characteristics, defense mechanism effectiveness, and quantum parameter configuration would require substantially more experimental conditions and statistical controls.

Future work could investigate hybrid approaches where quantum parameters are trainable during initial model development but then fixed during adversarial training and defense evaluation, combining the potential benefits of quantum parameter optimization with the experimental clarity of fixed transformations during robustness analysis. However, for establishing baseline defense effectiveness patterns and validating fundamental

adversarial robustness properties of HNN architectures, the Quantum Reservoir Mapping approach provides both methodological rigor and computational tractability while maintaining sufficient quantum expressiveness to investigate classical-quantum hybrid defense mechanisms.

4.2.8 Quantum Circuit Architecture

The quantum circuit component represents the quantum processing portion of the HNN model [61], implementing a 4-qubit parameterized quantum circuit that processes features in the quantum state space. This quantum layer is inserted between the classical convolutional feature extraction layers and the final classification layers, creating a hybrid architecture where approximately 50% of the processing is classical (convolutional and fully connected layers) and 50% is quantum (quantum circuit processing). Fig. 4.3 illustrates the quantum circuit structure used in this research.

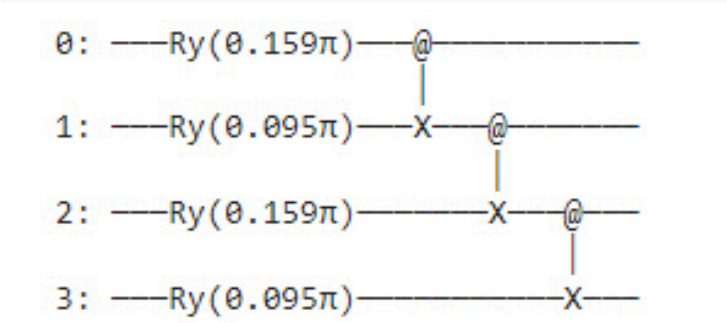

Figure 4.3: Example quantum circuit configuration used in HNN model (4-qubit parameterized circuit with rotation gates and CNOT entanglement).

The quantum circuit architecture implements the Quantum Reservoir Mapping paradigm described in Section 4.2.7, remaining constant across all datasets to provide a fixed quantum transformation that classical layers must learn to navigate robustly. Table 4.5 presents the complete quantum circuit configuration parameters. Each qubit is initialized using a rotation gate $R_y(\theta)$ to encode input features into quantum states: qubits 0 and 2 receive $R_y(0.159\pi)$, while qubits 1 and 3 receive $R_y(0.095\pi)$. These rotation angles were selected to provide sufficient quantum state diversity while remaining amenable to

classical simulation and future implementation on near-term quantum hardware.

Following state preparation, qubits are entangled through a linear sequence of controlled-NOT (CNOT) gates, creating quantum correlations between qubit states. Specifically, qubit 0 acts as the control for a CNOT targeting qubit 1, qubit 1 controls qubit 2, and qubit 2 controls qubit 3. This linear entanglement structure provides quantum correlation while maintaining computational tractability for classical simulation using Cirq. The circuit concludes with measurement of all qubits in the computational basis, producing classical outputs that are fed into the final fully connected classification layers.

Table 4.5: Quantum circuit configuration parameters.

No.	**Parameter**	**Value/Configuration**	**Purpose**
1	Number of Qubits	4	Quantum state dimensionality
2	State Preparation Gates	R_y (Rotation Y)	Encode features into quantum states
3	Rotation Angle θ	0.159π	Applied to qubits 0 and 2 for state initialization
4	Rotation Angle ϕ	0.095π	Applied to qubits 1 and 3 for state initialization
5	Entanglement Gates	CNOT (Controlled-NOT)	Create quantum correlations between qubits
6	Entanglement Structure	Linear chain ($0\rightarrow1\rightarrow2\rightarrow3$)	Sequential: Qubit 0 controls 1, 1 controls 2, 2 controls 3
7	Total CNOT Gates	3	One for each qubit pair in chain
8	Measurement Basis	Computational (Z) basis	Measure all qubits to produce classical outputs
9	Circuit Depth	2 (rotation layer + entanglement layer)	Enables efficient classical simulation
10	Simulation Framework	Google Cirq	Classical simulation of quantum operations
11	Quantum Output Dimension	16 (2^4 possible states)	Fed into classical fully connected layers
12	Trainable Parameters	0 (Quantum Reservoir Mapping)	Fixed transformation as described in Section 4.2.7

The quantum circuit configuration represents a balance between quantum expressiveness and classical simulation tractability. With 4 qubits and circuit depth of 2, the

quantum state space dimension is $2^4 = 16$, which is computationally feasible for classical simulation while providing sufficient quantum complexity to investigate adversarial robustness properties. The fixed rotation angles (θ and ϕ) ensure that experimental variations stem from defense mechanisms and dataset characteristics rather than quantum circuit configuration differences. This consistency enables meaningful comparison of defense effectiveness across all 120 experimental conditions (4 datasets $\times$ 3 attack types $\times$ 10 defense techniques).

4.3 Strengths and Contributions

The research demonstrates several key strengths that establish a solid foundation for HNN defense research [55], [60]. Table 4.6 summarizes the primary strengths and their significance.

Table 4.6: Research strengths and contributions.

Strength	Significance
Comprehensive Experimental Design	Systematic evaluation across 120 experimental conditions (4 datasets × 3 attack types × 10 defense techniques) provides robust evidence for defense effectiveness; reveals consistent patterns that adversarial training outperforms test-time defenses by 2.3–2.8× across all conditions
Clear Defense Categorization	Distinction between test-time defenses (input transformation, randomization) and training-time defenses (adversarial training) provides conceptual framework for understanding fundamental trade-offs: deployment flexibility vs. robustness
Actionable Findings	Identification of adversarial training as optimal defense (58.5% average defended accuracy) with quantified performance advantages provides clear guidance for practitioners deploying HNN models in adversarial environments
Classical Simulation Validation	Demonstrates that classical simulation using PyTorch and Cirq enables comprehensive HNN defense evaluation without expensive quantum hardware access; establishes practical methodology for defense research
Foundation for Extension	Baseline defense effectiveness measurements and identified challenges provide clear targets for future improvements; establishes both what works (adversarial training superiority) and remaining challenges (substantial accuracy loss persists)

4.4 Summary

This chapter established the methodological foundation for evaluating defense mechanisms against WTC adversarial attacks on HNN models. The proposed solution addresses the research gap by systematically evaluating three defense categories (input transformation, randomization, adversarial training) across 120 experimental conditions spanning four datasets and three compound attack types.

The experimental procedures were consolidated into a base procedure with clearly de-

fined variations for each defense category, eliminating repetitive descriptions while maintaining methodological clarity. The distinction between test-time defenses (applied during inference) and training-time defenses (requiring model retraining) emerged as a critical framework for understanding operational trade-offs between deployment flexibility and robustness.

Key architectural decisions were justified, including the use of fixed quantum circuit parameters to isolate defense effectiveness from quantum configuration variations, the 4-qubit circuit balancing quantum expressiveness with classical simulation feasibility, and dataset selection progressing from simple handwritten digits to complex natural images. The comprehensive robustness metrics calculation procedure quantifies the effectiveness of the attack, the recovery of defense, and the improvement of overall robustness using standardized measurements .

The research demonstrates that the classical simulation using PyTorch and Cirq provides a practical methodology for HNN defense research without requiring expensive quantum hardware access. The balance of classical-quantum architecture 50-50, combined with systematic evaluation across varying dataset complexity levels, enables the investigation of whether adversarial vulnerabilities primarily exploit classical layers, quantum layers, or the classical-quantum interface—establishing foundational understanding for future HNN security research.

Chapter 5

IMPLEMENTATION, VALIDATION, AND RESULTS

This chapter includes implementation, validation, and results. The implementation outlines the technical architecture of the system and the software frameworks used for the comprehensive evaluation of the defense [63], [64]. Validation represents an approach, a method, metrics, and the notion of effectiveness. The results provide comprehensive empirical evidence for 120 experimental conditions (4 datasets $\times$ 3 types of attack $\times$ 10 defense techniques) that demonstrate the effectiveness of defense mechanisms against adversarial attacks from WTC. In addition, the contributions summarize the factors that determine the course of action associated with implementation, validation, and results.

5.1 Implementation

Implementing the HNN defense evaluation system leverages a software stack that combines classical deep learning frameworks with quantum circuit simulation capabilities, executed on conventional computing hardware. The primary frameworks employed are PyTorch 1.12.1 for classical neural network layers (convolutional, pooling, fully connected), Google Cirq 1.0.0 for quantum circuit simulation (parameterized rotation gates, CNOT entanglement, measurement) and Torchattacks 3.3.0 for adversarial attack generation (FGSM, PGD, CW implementations). All source code, trained models, and experimental results are publicly available (see Appendix C for repository details and reproducibility instructions).

The computational environment consists of Linux-based systems (Ubuntu 20.04 LTS) with NVIDIA GPU acceleration (CUDA 11.6) for classical layer training and inference. The quantum circuit simulation operates on a CPU with sufficient memory to maintain quantum state vectors for the 4-qubit circuit configuration (requiring complex probability amplitudes $2^4 = 16$ per state). Training procedures span 10-120 epochs depending on the size and complexity of the data set, using the Adam optimizer with learning rate 0.001 and data set-specific weight decay values (1×10^{-4} for MNIST/EMNIST/TrafficSigns, 3×10^{-4} for TinyImageNet).

The implementation architecture separates model training (execution phase 1), adversarial attack generation (execution phase 2), and defense application with evaluation (execution phase 3). For test-time defenses (input transformation, randomization), the trained model remains fixed while defense transformations modify adversarial test inputs during inference. For training-time defense (adversarial training), the model undergoes complete retraining on augmented datasets combining clean and adversarial training examples. All experimental code, model architectures, and defense implementations are documented in Appendix A and are available for reproducibility verification.

5.2 Validation Approach

The validation approach represents a process that simulates the real-world deployment environment of the HNN model, given the capabilities of the adversary. In this case, the adversary will have access to and understanding of the architecture, parameters, and training process of the HNN model (i.e., white-box adversarial attacks). The validation approach process consists of four steps, including 1) identifying the adversarial threat, 2) generating adversarial examples using compound attacks, 3) evaluating the defense mechanisms, and 4) iterating across multiple defense strategies. Figure 5.1 presents the complete validation workflow.

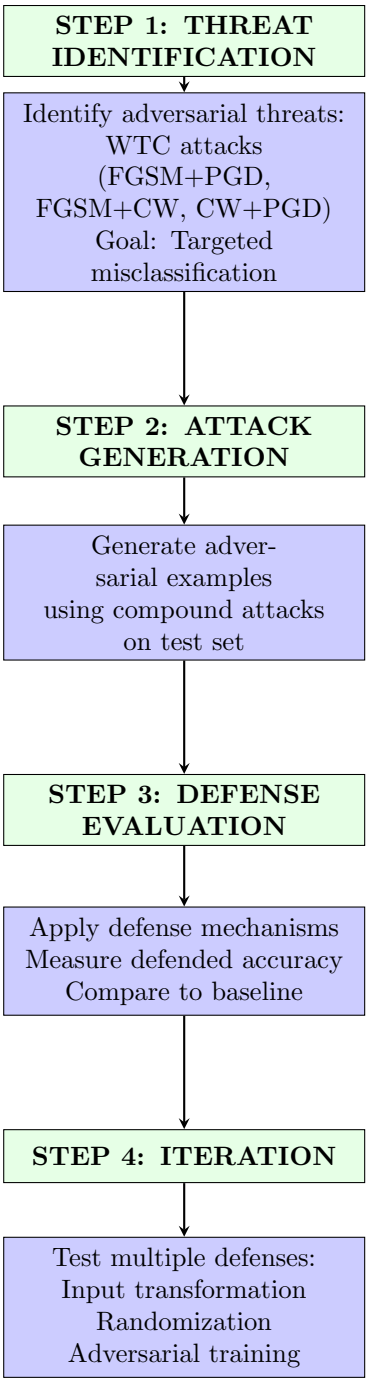

Figure 5.1: Validation approach workflow for defense evaluation.

The adversarial threat represents the types of compound adversarial attacks (FGSM+PGD, FGSM+CW, CW+PGD) that are most likely to be used against the HNN model and the objective of the attacks (targeted misclassification).

As mentioned in Chapter 4, the defenses considered are randomization, input trans-

formation, and adversarial training, each distinguished by its application timing. First, randomization is a *test-time defense* that introduces stochastic transformations to adversarial inputs during inference. This can make the predictions of the HNN model more unpredictable and less susceptible to adversarial attacks by varying the transformation applied to each input. Randomization techniques include random resizing, random cropping, random rotation, and combined randomization. As a test-time defense, randomization can be applied to any pre-trained model without retraining and is relatively easy to implement.

Second, input transformation is also a *test-time defense* that applies deterministic operations to adversarial inputs during inference. This involves transforming the input data in a way that removes or masks adversarial perturbations while preserving the legitimate content. Input transformation techniques include JPEG compression, bit-depth reduction, Gaussian noise addition, image quilting, and combined transformations. The goal of input transformation defense is to make adversarial examples indistinguishable from clean examples for the HNN model. Like randomization, input transformation can be applied without model retraining.

Finally, adversarial training is a *training-time defense* that fundamentally modifies the model by training it on both clean and adversarial examples. In this case, to create a robust model, a diverse set of adversarial examples is generated from the training set and combined with clean training data. The model is then retrained on this augmented dataset, learning to correctly classify both clean and adversarial input. Evaluation of the defense occurs by measuring the model's ability to correctly classify adversarial test examples that were not seen during training. The measurement compares the defended accuracy (accuracy in adversarial examples after defense) to the baseline accuracy (accuracy in adversarial examples without defense) and clean accuracy (accuracy in clean examples). The validation process evaluates the effectiveness of defenses against three types of compound adversarial attacks across four datasets.

5.2.1 Validation Method for Defenses

The validation method used for the HNN model employs a simple split between train and test without validation data. Models are trained in the designated training set and evaluated in a separate test set. For test-time defenses [64], [65] (input transformation and randomization), the model is trained once on clean training data, and defenses are applied to adversarial test samples during inference. For training-time defense [63] (adversarial training), the model is retrained on an augmented training dataset containing both clean and adversarial examples, and then evaluated on the test set. This validation method ensures that defenses are evaluated on previously unseen test data, providing an unbiased assessment of defense effectiveness. The evaluation uses adversarial examples generated from diverse datasets (MNIST, EMNIST "Digits", TinyImageNet, TrafficSigns) to ensure that defenses are robust to attacks across varying data distributions and complexity levels. Figure 5.2 illustrates the complete defense evaluation workflow.

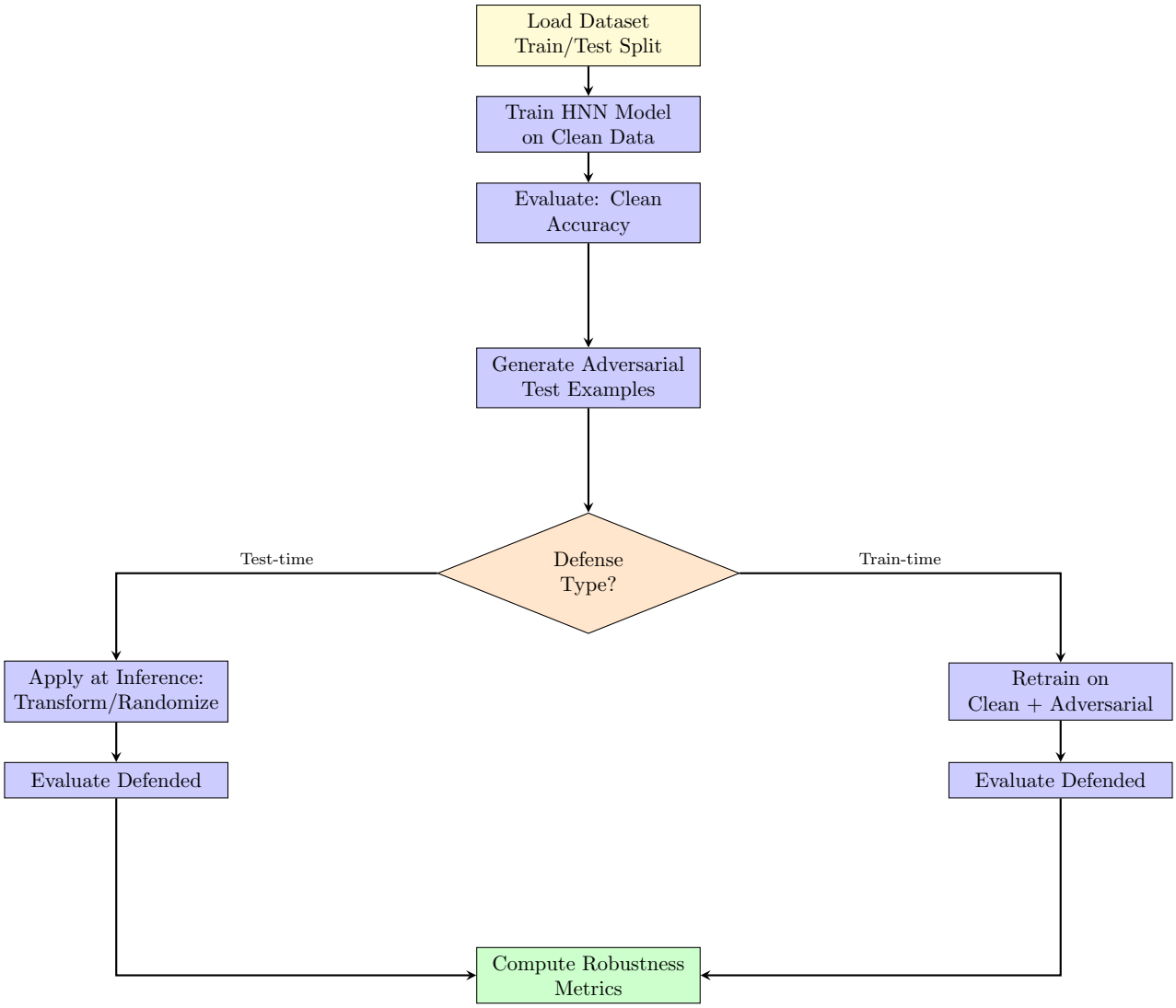

Figure 5.2: Defense evaluation workflow showing test-time vs training-time paths.

5.2.2 Metrics Used for Validation of Defenses

The metrics used for defense validation are primarily quantitative, focusing on accuracy and loss. The primary metric is *the accuracy* of the defense, which measures the percentage of adversarial test examples correctly classified after applying the defense mechanism. Defended accuracy can be represented by the equation:

$$\text{Defended Accuracy} = \frac{\text{Number of correctly classified adversarial examples}}{\text{Total number of adversarial examples}} \tag{5.1}$$

Higher defended accuracy indicates that the defense mechanism is more effective at protecting the HNN model against adversarial attacks. Additional metrics include:

- **Clean Accuracy:** The baseline performance on unperturbed test examples, typically 86-92% across datasets
- **Attack Success Rate:** The percentage of adversarial examples that successfully fool the undefended model, measured as the drop from clean accuracy to undefended accuracy on adversarial examples
- **Accuracy Recovery:** The improvement in accuracy when applying defenses, measured as the difference between defended accuracy and undefended accuracy on adversarial examples
- **Preserved Accuracy:** The percentage of clean accuracy maintained after defense, calculated as defended accuracy divided by clean accuracy

The loss metric measures the error between the model's predictions and the true labels using Negative Log-Likelihood (NLLoss). A lower loss indicates that the model is more confident in its correct predictions. The loss function is used during training to optimize model parameters and to assess convergence.

While qualitative metrics such as interpretability, detectability, and generality can provide benefits to HNN model developers when developing defenses, this research focuses on quantitative metrics (accuracy and loss) that provide objective, measurable assessments of defense effectiveness across different attack strategies and datasets.

5.2.3 Robustness Metrics Calculation

Following defense evaluation, comprehensive robustness metrics [66], [67] quantify attack effectiveness, defense recovery, and overall robustness improvement. These metrics align with the calculation procedure defined in Section 5.2.2 and are calculated for each experimental condition.Figure 5.3 presents the robustness metrics calculation workflow used in the experimental evaluation.

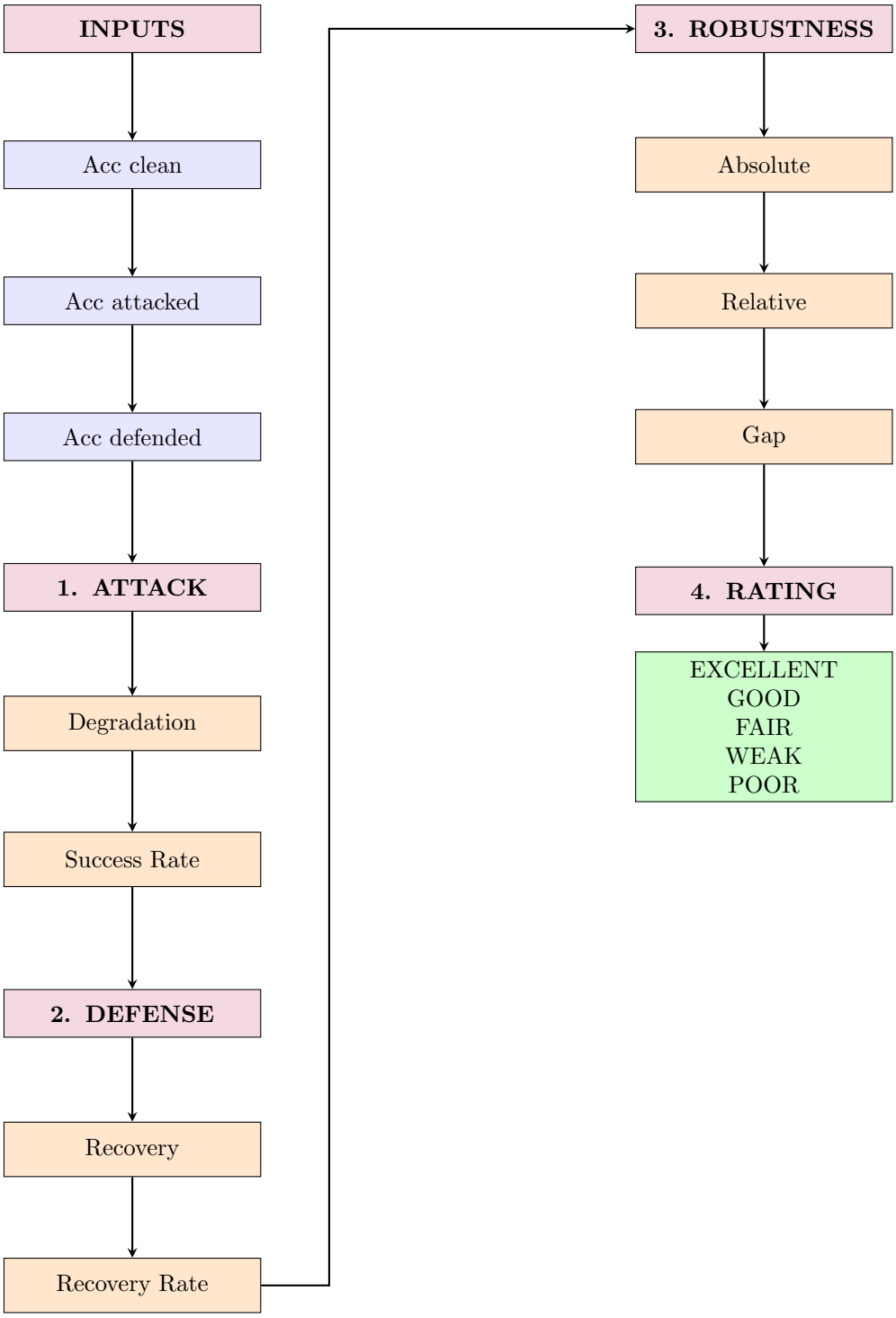

Figure 5.3: Robustness metrics computation workflow (see Section 5.2.2 for metric definitions).

Robustness metrics provide a quantitative assessment of defense effectiveness through four categories:

5.2.4 Attack Effectiveness

Attack effectiveness measures the severity of compound adversarial attacks. Attack degradation quantifies the absolute accuracy drop caused by the attack ($\Delta_{\text{attack}} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{attacked}}$), while attack success rate normalizes this degradation relative to clean performance ($R_{\text{attack}} = \frac{\Delta_{\text{attack}}}{\text{Acc}_{\text{clean}}} \times 100\%$). Across all experiments, the average attack degradation is 68.5 percentage points (from 87.6% clean to 19.1% attacked), yielding an average attack success rate of 78.3%.

5.2.5 Defense Effectiveness

Defense Effectiveness quantifies the accuracy recovery achieved by defense mechanisms. Defense recovery measures absolute improvement ($\Delta_{\text{defense}} = \text{Acc}_{\text{defended}} - \text{Acc}_{\text{attacked}}$), while recovery rate expresses this as a percentage of lost accuracy recovered ($R_{\text{recovery}} = \frac{\Delta_{\text{defense}}}{\Delta_{\text{attack}}} \times 100\%$). Averaged across all defenses, defense recovery is 7.4 percentage points (from 19.1% attacked to 26.5% defended), representing 10.8% recovery rate. For adversarial training specifically, defense recovery is 39.4 percentage points (to 58.5% defended), representing 57.5% recovery rate.

5.2.6 Overall Robustness

Overall robustness characterizes defended model robustness through absolute improvement ($I_{\text{abs}} = \Delta_{\text{defense}}$), relative improvement over attacked baseline ($I_{\text{rel}} = \frac{I_{\text{abs}}}{\text{Acc}_{\text{attacked}}} \times 100\%$), and remaining gap to clean performance ($G = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{defended}}$). For adversarial training, average relative improvement is 206.3% over the attacked baseline (58.5% defended vs 19.1% attacked), with a remaining gap of 29.1 percentage points to clean accuracy.

5.2.7 Defense Rating

The defense rating classifies the defense quality based on the accuracy thresholds defended. TrafficSigns with adversarial training achieves EXCELLENT rating (87% > 80%), MNIST

with adversarial training achieves GOOD rating (68%, range 60-80%), TinyImageNet with adversarial training achieves FAIR rating (51%, range 40-60%), while most test-time defenses achieve WEAK (20-40%) or POOR (<20%) ratings.

5.2.8 Effectiveness of Validation Approach

The effectiveness of the validation approach is demonstrated through a comprehensive evaluation across 120 experimental conditions. The approach evaluates three defense categories (Input Transformation, Randomization, Adversarial Training) against three compound attack types (FGSM+PGD, FGSM+CW, CW+PGD) across four datasets (MNIST, EMNIST "Digits", TinyImageNet, TrafficSigns), with multiple techniques within each defense category.

The validation approach is effective because it:

1. **Uses separate test sets:** All evaluations are performed on test data not seen during training, ensuring unbiased assessment

2. **Tests multiple attack types:** Evaluation against three different compound attacks (FGSM+PGD, FGSM+CW, CW+PGD) demonstrates robustness across varying attack strategies

3. **Spans varying complexity:** Four datasets with different characteristics (simple digits to complex natural images) reveal how defense effectiveness depends on task complexity

4. **Compares defense timing:** Systematic comparison of test-time defenses (input transformation, randomization) versus training-time defense (adversarial training) reveals fundamental trade-offs

5. **Provides baseline comparisons:** Clean accuracy and undefended accuracy establish reference points for measuring defense effectiveness

Effectiveness is validated by the consistency of results: adversarial training consistently outperforms test-time defenses across all datasets and attack types, achieving 58.5% average defended accuracy compared to 19.1% without defense—a 3× improvement and

2.3–2.8× advantage over test-time defenses (see Results section below). This consistent pattern across 120 experimental conditions demonstrates that the validation approach successfully discriminates between more and less effective defenses.

5.3 Results and Contributions

The results and contributions are discussed in this section. The format of results is presented for the HNN model using the MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns datasets. The defenses evaluated include input transformation (5 techniques for TinyImageNet/TrafficSigns, 3 techniques for MNIST/ ENMIST), randomization (4 techniques), and adversarial training, totaling 120 experimental conditions when combined with 3 attack types and 4 datasets.

A tabular format is used for the empirical results, where the HNN model datasets include the MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns datasets. Quantitative metrics calculated for the validation of defenses to protect HNN models include accuracy and loss. The results demonstrate significant variation in defense effectiveness across datasets, attack types, and defense mechanisms, with clear patterns emerging that support the superiority of training-time defenses over test-time defenses.

5.3.1 Experimental Results Overview

This section presents comprehensive experimental results across all 120 experimental conditions evaluated in this dissertation. Results are organized by defense category (Adversarial Training, Input Transformation, Randomization), with separate tables for each dataset within each category. Each table reports clean accuracy, attacked accuracy without defense, defended accuracy, attack metrics, and robustness improvement metrics.

5.3.1.1 Metric Definitions

Clean accuracy represents model accuracy (%) on unperturbed test data. Attacked accuracy represents model accuracy (%) on adversarial examples without defense applied. Defended accuracy represents model accuracy (%) on adversarial examples with defense applied. Attack degradation (Att Deg) measures attack degradation in percentage points, calculated as Clean minus Attacked. Attack success rate (Att SR) measures attack success rate as percentage of clean accuracy lost. Robustness improvement (Rob Imp) measures

robustness improvement in percentage points, calculated as Defended minus Attacked. Robustness change (Rob Change) measures robustness change as a percentage improvement over the attacked baseline. Recovery rate (Rec Rate) measures the recovery rate as a percentage of lost accuracy recovered. Gap measures the remaining gap in percentage points calculated as Clean minus Defended.

5.3.1.2 Attack Type Abbreviations

F+P denotes the FGSM+PGD compound attack. F+C denotes the FGSM+CW compound attack. C+P denotes the CW+PGD compound attack.

5.3.2 Results: Adversarial Training Defense

5.3.2.1 MNIST

Table 5.1: Adversarial Training results on MNIST dataset.

Defense Type	**Att Type**	**Clean (%)**	**Attk (%)**	**Def (%)**	**Att Deg (pp)**	**Att SR (%)**	**Rob Imp (pp)**	**Rob Chg (%)**	**Rec Rate (%)**	**Gap (pp)**
Adversarial Training	C+P	95	69	86	26	27	17	24	65	9
Adversarial Training	F+C	92	1	59	91	98	58	∞	63	33
Adversarial Training	F+P	92	1	59	91	98	58	∞	63	33

5.3.2.2 EMNIST "Digits"

Table 5.2: Adversarial Training results on EMNIST "Digits" dataset.

Defense Type	**Att Type**	**Clean (%)**	**Attk (%)**	**Def (%)**	**Att Deg (pp)**	**Att SR (%)**	**Rob Imp (pp)**	**Rob Chg (%)**	**Rec Rate (%)**	**Gap (pp)**
Adversarial Training	C+P	95	64	82	31	32	18	28	58	13
Adversarial Training	F+C	95	0	1	95	100	1	∞	1	94
Adversarial Training	F+P	95	0	1	95	100	1	∞	1	94

5.3.2.3 TinyImageNet

Table 5.3: Adversarial Training results on TinyImageNet dataset.

Defense Type	**Att Type**	**Clean (%)**	**Attk (%)**	**Def (%)**	**Att Deg (pp)**	**Att SR (%)**	**Rob Imp (pp)**	**Rob Chg (%)**	**Rec Rate (%)**	**Gap (pp)**
Adversarial Training	C+P	81	16	59	65	80	43	268	66	22
Adversarial Training	F+C	81	20	48	61	75	28	140	45	33
Adversarial Training	F+P	81	14	47	67	82	33	235	49	34

5.3.2.4 TrafficSigns

Table 5.4: Adversarial Training results on TrafficSigns dataset.

Defense Type	**Att Type**	**Clean (%)**	**Attk (%)**	**Def (%)**	**Att Deg (pp)**	**Att SR (%)**	**Rob Imp (pp)**	**Rob Chg (%)**	**Rec Rate (%)**	**Gap (pp)**
Adversarial Training	C+P	86	0	82	86	100	82	∞	95	4
Adversarial Training	F+C	86	44	88	42	48	44	100	104	-2
Adversarial Training	F+P	86	0	90	86	100	90	∞	104	-4

5.3.3 Results: Input Transformation Defense

5.3.3.1 MNIST

Table 5.5: Input Transformation results on MNIST dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Bit-depth	C+P	92	69	69	23	25	0	0	0	23
Bit-depth	F+C	92	1	2	91	98	1	100	1	90
Bit-depth	F+P	92	1	2	91	98	1	100	1	90
Combined	C+P	92	69	9	23	25	-60	-86	-260	83
Combined	F+C	92	1	9	91	98	8	800	8	83
Combined	F+P	92	1	9	91	98	8	800	8	83
Gaussian	C+P	92	69	13	23	25	-56	-81	-243	79
Gaussian	F+C	92	1	8	91	98	7	700	7	84
Gaussian	F+P	92	1	8	91	98	7	700	7	84
JPEG	C+P	92	69	72	23	25	3	4	13	20
JPEG	F+C	92	1	2	91	98	1	100	1	90
JPEG	F+P	92	1	2	91	98	1	100	1	90
Quilting	C+P	92	69	9	23	25	-60	-86	-260	83
Quilting	F+C	92	1	10	91	98	9	900	9	82
Quilting	F+P	92	1	10	91	98	9	900	9	82

5.3.3.2 EMNIST "Digits"

Table 5.6: Input Transformation results on EMNIST "Digits" dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Bit-depth	C+P	91	64	64	27	29	0	0	0	27
Bit-depth	F+C	91	0	1	91	100	1	∞	1	90
Bit-depth	F+P	91	0	1	91	100	1	∞	1	90
Combined	C+P	91	64	10	27	29	-54	-84	-200	81
Combined	F+C	91	0	10	91	100	10	∞	10	81
Combined	F+P	91	0	10	91	100	10	∞	10	81
Gaussian	C+P	91	64	10	27	29	-54	-84	-200	81
Gaussian	F+C	91	0	10	91	100	10	∞	10	81
Gaussian	F+P	91	0	10	91	100	10	∞	10	81
JPEG	C+P	91	64	66	27	29	2	3	7	25
JPEG	F+C	91	0	1	91	100	1	∞	1	90
JPEG	F+P	91	0	1	91	100	1	∞	1	90
Quilting	C+P	91	64	11	27	29	-53	-82	-196	80
Quilting	F+C	91	0	11	91	100	11	∞	12	80
Quilting	F+P	91	0	11	91	100	11	∞	12	80

5.3.3.3 TinyImageNet

Table 5.7: Input Transformation results on TinyImageNet dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Bit-depth	C+P	81	16	19	65	80	3	18	4	62
Bit-depth	F+C	81	20	21	61	75	1	5	1	60
Bit-depth	F+P	81	14	16	67	82	2	14	2	65
Combined	C+P	81	16	20	65	80	4	25	6	61
Combined	F+C	81	20	20	61	75	0	0	0	61
Combined	F+P	81	14	20	67	82	6	42	8	61
Gaussian	C+P	81	16	20	65	80	4	25	6	61
Gaussian	F+C	81	20	20	61	75	0	0	0	61
Gaussian	F+P	81	14	20	67	82	6	42	8	61
JPEG	C+P	81	16	20	65	80	4	25	6	61
JPEG	F+C	81	20	10	61	75	-10	-50	-16	71
JPEG	F+P	81	14	13	67	82	-1	-7	-1	68
Quilting	C+P	81	16	26	65	80	10	62	15	55
Quilting	F+C	81	20	26	61	75	6	30	9	55
Quilting	F+P	81	14	25	67	82	11	78	16	56

5.3.3.4 TrafficSigns

Table 5.8: Input Transformation results on TrafficSigns dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Bit-depth	C+P	86	0	1	86	100	1	∞	1	85
Bit-depth	F+C	86	44	47	42	48	3	6	7	39
Bit-depth	F+P	86	0	3	86	100	3	∞	3	83
Combined	C+P	86	0	19	86	100	19	∞	22	67
Combined	F+C	86	44	19	42	48	-25	-56	-59	67
Combined	F+P	86	0	18	86	100	18	∞	20	68
Gaussian	C+P	86	0	18	86	100	18	∞	20	68
Gaussian	F+C	86	44	20	42	48	-24	-54	-57	66
Gaussian	F+P	86	0	18	86	100	18	∞	20	68
JPEG	C+P	86	0	51	86	100	51	∞	59	35
JPEG	F+C	86	44	53	42	48	9	20	21	33
JPEG	F+P	86	0	39	86	100	39	∞	45	47
Quilting	C+P	86	0	68	86	100	68	∞	79	18
Quilting	F+C	86	44	59	42	48	15	34	35	27
Quilting	F+P	86	0	65	86	100	65	∞	75	21

5.3.4 Results: Randomization Defense

5.3.4.1 MNIST

Table 5.9: Randomization results on MNIST dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Combined	C+P	92	69	47	23	25	-22	-31	-95	45
Combined	F+C	92	1	7	91	98	6	600	6	85
Combined	F+P	92	1	7	91	98	6	600	6	85
Cropping	C+P	92	69	45	23	25	-24	-34	-104	47
Cropping	F+C	92	1	7	91	98	6	600	6	85
Cropping	F+P	92	1	8	91	98	7	700	7	84
Resizing	C+P	92	69	71	23	25	2	2	8	21
Resizing	F+C	92	1	1	91	98	0	0	0	91
Resizing	F+P	92	1	1	91	98	0	0	0	91
Rotation	C+P	92	69	74	23	25	5	7	21	18
Rotation	F+C	92	1	2	91	98	1	100	1	90
Rotation	F+P	92	1	2	91	98	1	100	1	90

5.3.4.2 EMNIST "Digits"

Table 5.10: Randomization results on EMNIST "Digits" dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Combined	C+P	91	64	46	27	29	-18	-28	-66	45
Combined	F+C	91	0	2	91	100	2	∞	2	89
Combined	F+P	91	0	2	91	100	2	∞	2	89
Cropping	C+P	91	64	48	27	29	-16	-25	-59	43
Cropping	F+C	91	0	8	91	100	8	∞	8	83
Cropping	F+P	91	0	8	91	100	8	∞	8	83
Resizing	C+P	91	64	65	27	29	1	1	3	26
Resizing	F+C	91	0	1	91	100	1	∞	1	90
Resizing	F+P	91	0	1	91	100	1	∞	1	90
Rotation	C+P	91	64	72	27	29	8	12	29	19
Rotation	F+C	91	0	1	91	100	1	∞	1	90
Rotation	F+P	91	0	1	91	100	1	∞	1	90

5.3.4.3 TinyImageNet

Table 5.11: Randomization results on TinyImageNet dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Combined	C+P	81	16	33	65	80	17	106	26	48
Combined	F+C	81	20	28	61	75	8	40	13	53
Combined	F+P	81	14	26	67	82	12	85	17	55
Cropping	C+P	81	16	33	65	80	17	106	26	48
Cropping	F+C	81	20	28	61	75	8	40	13	53
Cropping	F+P	81	14	28	67	82	14	100	20	53
Resizing	C+P	81	16	32	65	80	16	100	24	49
Resizing	F+C	81	20	30	61	75	10	50	16	51
Resizing	F+P	81	14	26	67	82	12	85	17	55
Rotation	C+P	81	16	29	65	80	13	81	20	52
Rotation	F+C	81	20	27	61	75	7	35	11	54
Rotation	F+P	81	14	23	67	82	9	64	13	58

5.3.4.4 TrafficSigns

Table 5.12: Randomization results on TrafficSigns dataset.

Defense Type	Att Type	Clean (%)	Attk (%)	Def (%)	Att Deg (pp)	Att SR (%)	Rob Imp (pp)	Rob Chg (%)	Rec Rate (%)	Gap (pp)
Combined	C+P	86	0	35	86	100	35	∞	40	51
Combined	F+C	86	44	54	42	48	10	22	23	32
Combined	F+P	86	0	34	86	100	34	∞	39	52
Cropping	C+P	86	0	12	86	100	12	∞	13	74
Cropping	F+C	86	44	57	42	48	13	29	30	29
Cropping	F+P	86	0	20	86	100	20	∞	23	66
Resizing	C+P	86	0	1	86	100	1	∞	1	85
Resizing	F+C	86	44	47	42	48	3	6	7	39
Resizing	F+P	86	0	3	86	100	3	∞	3	83
Rotation	C+P	86	0	17	86	100	17	∞	19	69
Rotation	F+C	86	44	54	42	48	10	22	23	32
Rotation	F+P	86	0	20	86	100	20	∞	23	66

5.3.5 Overall Defense Category Effectiveness

Figure 5.4 shows the defended accuracy across the three defense categories (Adversarial Training, Input Transformation, and Randomization) for each dataset. Adversarial Training [63] consistently demonstrates the highest effectiveness across all datasets, achieving

87% in TrafficSigns, 68% on MNIST, 51% on TinyImageNet, and 28% in EMNIST "Digits". This represents a 2–4× improvement over test-time defenses across all datasets.

Input Transformation shows moderate effectiveness with significant variation by dataset: 33% on TrafficSigns, 16% on MNIST, 20% on TinyImageNet, and only 15% on EMNIST "Digits". The variation reflects the different transformation techniques available for each dataset—TrafficSigns benefits from image quilting (64%) and JPEG compression (48%), while EMNIST/MNIST are limited to bit-depth reduction and Gaussian noise.

Randomization demonstrates the most consistent but modest performance across datasets: 30% on TrafficSigns, 23% on MNIST, 29% on TinyImageNet and 21% on EMNIST "Digits". The consistency suggests that stochastic transformations provide a baseline level of protection that is less dependent on specific dataset characteristics, but offers lower maximum effectiveness compared to adversarial training.

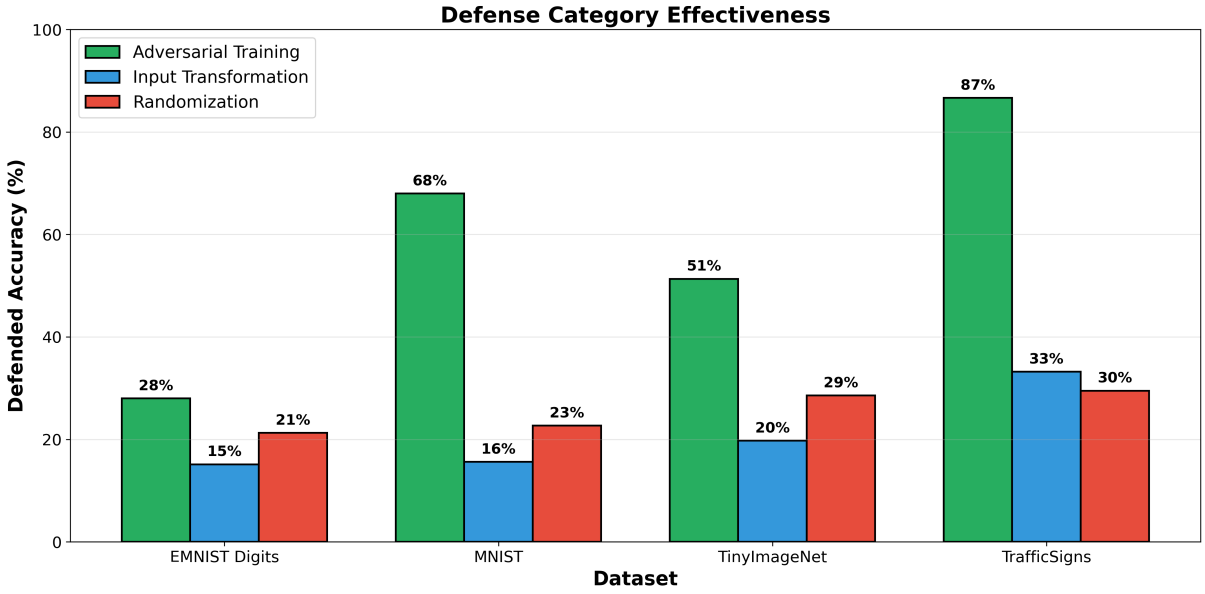

Figure 5.4: Defense category effectiveness across datasets.

5.3.6 Overall Accuracy Progression

Figure 5.5 illustrates the progression of model accuracy from clean data through adversarial attack to defended state, revealing the severity of compound attacks and the partial recovery achieved by defenses. The clean accuracy ranges from 86-92% across datasets, demonstrating that the baseline HNN models achieve high performance on unperturbed data.

After compound adversarial attacks (without defense), accuracy drops dramatically to 15-24%, representing an attack success rate of 73-84%. This severe degradation demonstrates the potency of compound adversarial attacks (FGSM+PGD, FGSM+CW, CW+PGD) against HNN models. The attack success is relatively consistent across datasets, indicating that compound attacks are effective regardless of dataset complexity.

The defended accuracy (averaged across all defense mechanisms) shows partial recovery: TrafficSigns achieves 37% average defended accuracy (preserving 43% of clean accuracy), TinyImageNet reaches 26% (preserving 32%), MNIST achieves 24% (preserving 26%), and EMNIST "Digits" reaches only 19% (preserving 21%). The variation in recovery demonstrates that defense effectiveness is strongly dataset-dependent [63], [67], with simpler datasets benefiting more from defenses. However, even the best average defended accuracy (37% on TrafficSigns) represents substantial accuracy loss compared to clean performance (86%), highlighting the challenging nature of defending against compound adversarial attacks. When considering the optimal defense (adversarial training) specifically, defended accuracy improves substantially to 87% on TrafficSigns, 68% on MNIST, 51% on TinyImageNet, and 28% on EMNIST "Digits".

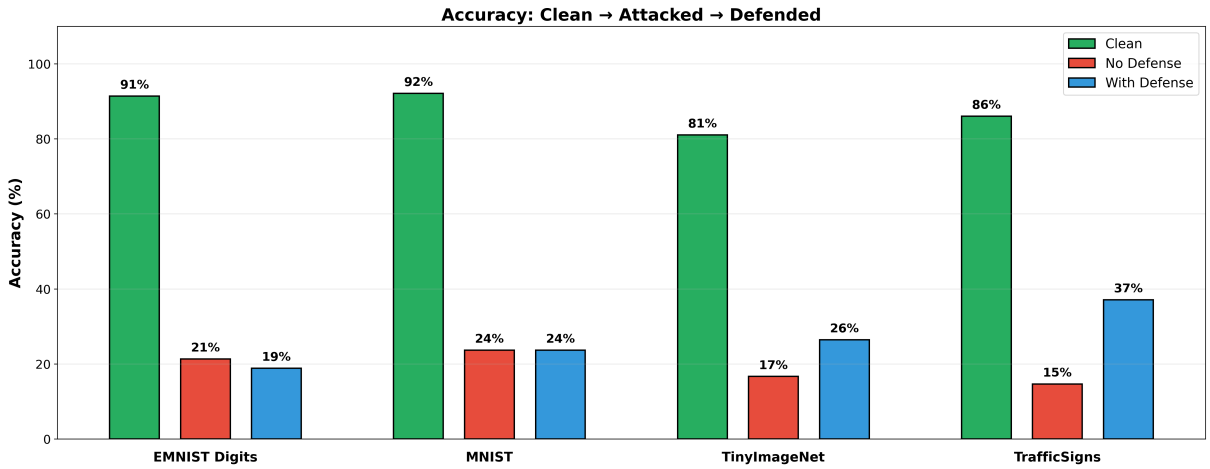

Figure 5.5: Accuracy progression from clean to attacked to defended states.

5.3.7 Overall Defense Performance by Attack Type

Figure 5.6 compares defense effectiveness against three different compound attack combinations: CW+PGD, FGSM+CW, and FGSM+PGD. The results show that defense performance varies significantly by both dataset and attack type, revealing that different compound attacks exploit different vulnerabilities in the HNN model.

On EMNIST "Digits", the CW+PGD attack is defended with 47% accuracy (averaged across all defenses), while FGSM+CW and FGSM+PGD attacks are defended with only 5% accuracy each. This 42 percentage point difference suggests that attacks combining FGSM are particularly effective against EMNIST "Digits", potentially due to the dataset's diverse handwriting styles making gradient-based perturbations more effective.

On MNIST, all three attack types show more balanced defended accuracy: CW+PGD at 50%, FGSM+CW at 11%, and FGSM+PGD at 11%. The pattern is similar to EMNIST but with higher overall defended accuracy, suggesting that the simpler, more uniform MNIST digits are easier to defend than the diverse EMNIST "Digits".

On TinyImageNet and TrafficSigns, the defended accuracy is more balanced across attack types. TinyImageNet achieves 29% (CW+PGD), 26% (FGSM+CW), and 24% (FGSM+PGD), while TrafficSigns achieves 30% (CW+PGD), 50% (FGSM+CW), and 31% (FGSM+PGD). This variation highlights the importance of evaluating defenses against multiple attack strategies rather than optimizing for a single attack type.

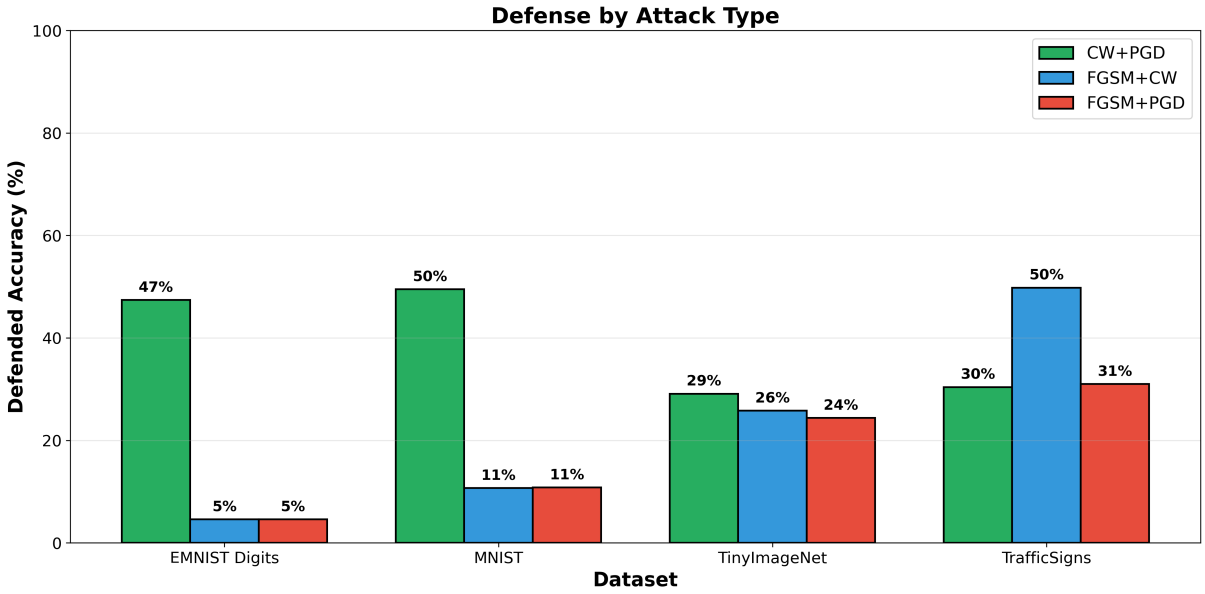

Figure 5.6: Defense effectiveness by attack type across datasets.

5.3.8 Overall Input Transformation Defense Comparison

Figure 5.7 presents the defended accuracy for various input transformation techniques across datasets. The results demonstrate significant variation in effectiveness both across techniques and datasets, with Image Quilting and JPEG Compression emerging as the most effective test-time defenses for complex visual datasets.

5.3.8.1 Image Quilting Defense

Image Quilting shows particularly strong performance on TrafficSigns (64% defended accuracy) and moderate performance on TinyImageNet (26%). Image quilting reconstructs adversarial images using patches from a clean database, effectively removing adversarial perturbations while preserving legitimate visual content. The high effectiveness of TrafficSigns (simple, structured symbols) compared to TinyImageNet (complex natural images) suggests that image quilting works best when clean patches can be reliably matched.

5.3.8.2 JPEG Compression Defense

JPEG Compression achieves 48% defended accuracy in TrafficSigns and shows moderate performance on TinyImageNet. JPEG compression removes high-frequency adversarial perturbations through lossy compression, providing substantial defense against attacks that rely on high-frequency perturbations.

5.3.8.3 Bit-Depth Defense

Bit-Depth shows consistent but modest effectiveness across datasets: 22% on EMNIST "Digits", 24% on MNIST, 19% on TinyImageNet and 17% on TrafficSigns. Bit-depth reduction quantizes pixel values, reducing the precision of adversarial perturbations but also losing legitimate image details.

5.3.8.4 Gaussian Noise Defense

Gaussian Noise achieves 20% on TinyImageNet and 19% on TrafficSigns. Adding random noise masks adversarial perturbations, but also degrades clean image quality, limiting

effectiveness.

5.3.8.5 Combined Transformations Defense

Combined Transformations achieve 17-20% across datasets, performing similarly to individual techniques rather than showing synergistic improvement, possibly because combining transformations degrades legitimate image content too severely.

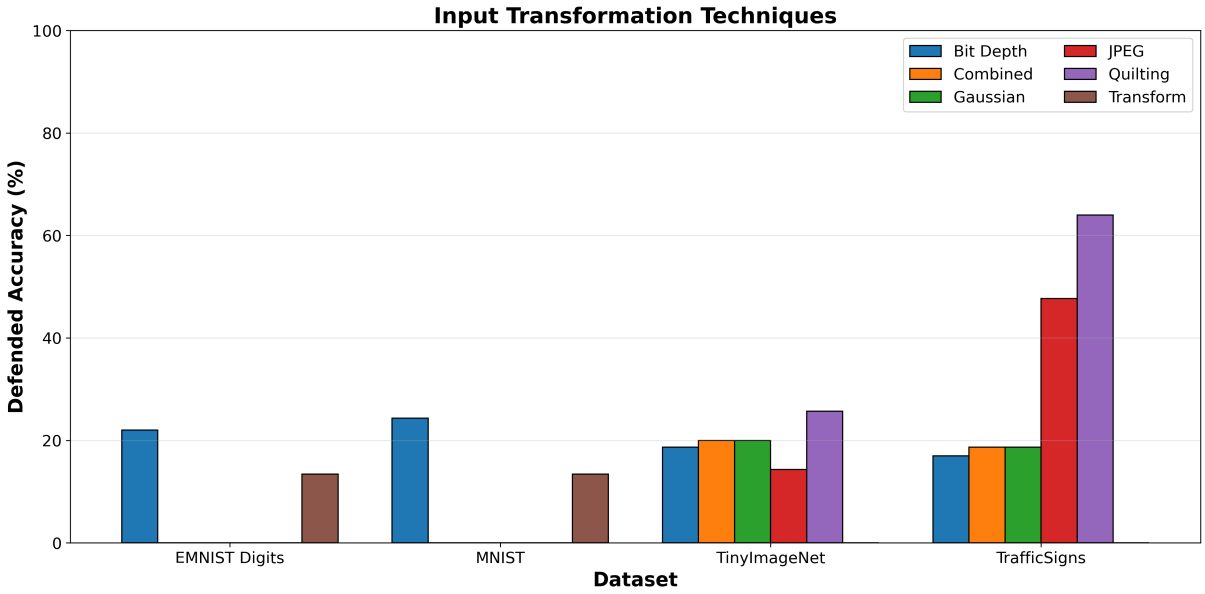

Figure 5.7: Input transformation techniques defended accuracy.

5.3.9 Overall Randomization Defense Comparison

Figure 5.8 shows the defended accuracy for randomization techniques, including cropping, resizing, rotation, and combined approaches. Randomization techniques show more consistent performance across datasets compared to input transformation, with combined randomization generally outperforming individual techniques.

TrafficSigns shows the best response to randomization, with combined techniques achieving 41% defended accuracy and rotation reaching 30%. The high effectiveness of TrafficSigns reflects the dataset's tolerance to geometric transformations—traffic signs remain recognizable after cropping, resizing, or rotation. Individual techniques (cropping 29%, resizing 17%) show more variation.

TinyImageNet demonstrates moderate effectiveness with combined randomization (29%) and individual techniques ranging from 26-30% (cropping 30%, rotation 26%, resizing 29%). Balanced performance suggests that natural images tolerate various geometric transformations relatively well, providing a consistent defense against different randomization strategies.

MNIST and EMNIST "Digits" show more limited improvements, with most techniques achieving 20-26% defended accuracy. MNIST achieves combined randomization at 20%, rotation at 26%, resizing at 24%, and cropping at 20%. EMNIST "Digits" shows similar patterns: combined 16%, rotation 25%, resizing 22%, and cropping 21%. The lower effectiveness on digit datasets suggests that geometric transformations may distort digit shapes enough to reduce clean accuracy while providing only modest protection against adversarial perturbations.

The stochastic nature of randomization provides the advantage that the same adversarial example will undergo different transformations across multiple evaluations, potentially providing ensemble-like robustness. However, the results show that this theoretical advantage translates into only modest improvements over deterministic input transformations in practice.

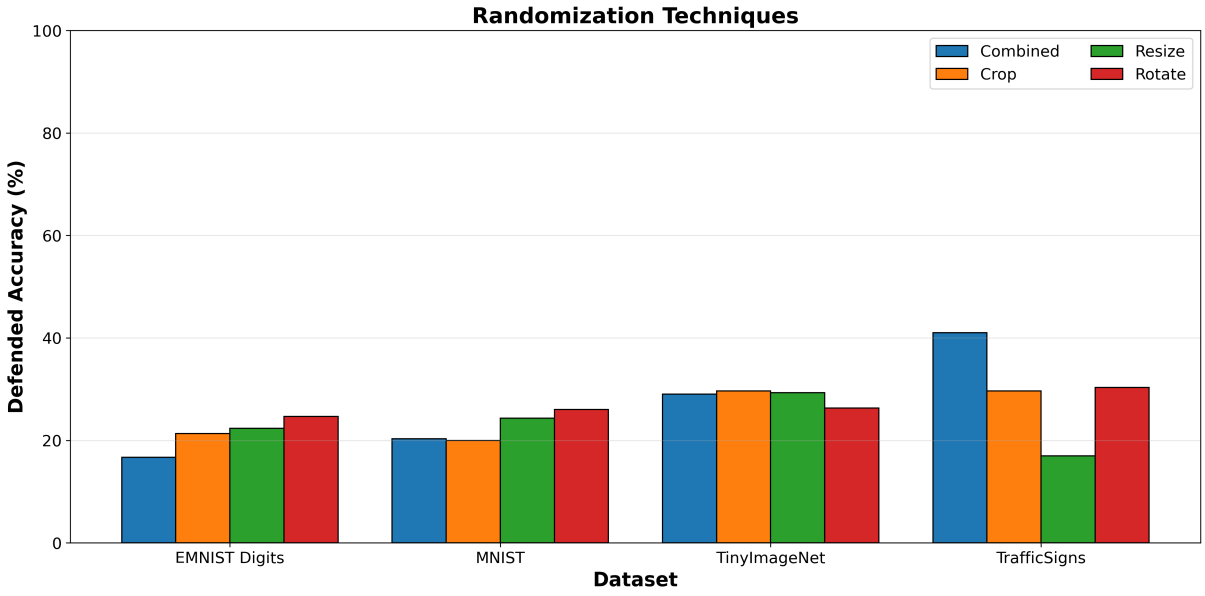

Figure 5.8: Randomization techniques defended accuracy.

5.3.10 Overall Top Performing Defenses by Dataset

Figure 5.9 ranks the top defenses for each dataset, clearly demonstrating that Adversarial Training consistently appears as the best performing approach across all datasets, significantly outperforming defenses in test-time.

TrafficSigns show the widest range of defense effectiveness. Adversarial Training achieves 87% defended accuracy, followed by Image Quilting (64%), JPEG compression (48%), Rotation (30%), and Combined Randomization (30%). The gap between adversarial training (87%) and the best test-time defense (Image Quilting at 64%) is 23 percentage points, demonstrating the substantial advantage of training-time defenses. Even the weakest techniques achieve 17-30%, suggesting that the simple, structured nature of traffic signs makes them relatively easier to defend compared to other datasets.

MNIST shows adversarial training with 68% defended accuracy, significantly outperforming the next best techniques: Rotation and Resizing at 26% and Bit-depth at 24%. The 42 percentage point gap between adversarial training and test-time defenses is nearly double that of TrafficSigns, highlighting the increasing advantage of training-time defenses as dataset complexity increases. The Test-time defenses cluster in the 13-26% range, showing a limited variation in effectiveness.

TinyImageNet demonstrates the challenges of defending complex natural images. Adversarial Training achieves 51% defended accuracy, followed by Cropping (30%), Resizing (29%), Rotation (26%), Image Quilting (26%), and Combined Randomization (24%). The 21 percentage point gap between adversarial training and test-time defenses is smaller than that of MNIST, possibly because the higher visual complexity of natural images limits the effectiveness of even training-time defenses. Test-time defenses cluster tightly in the 19-30% range.

EMNIST "Digits" presents the most challenging defense scenario. Adversarial Training achieves only 28% defended accuracy, while test-time defenses range from 13-25%. The best test-time defense (Rotation at 25%) comes within 3 percentage points of adversarial training, suggesting that the diverse handwriting styles in EMNIST make it difficult for any defense mechanism to learn robust features. This dataset highlights the

limits of current defense strategies against compound attacks on complex classification tasks.

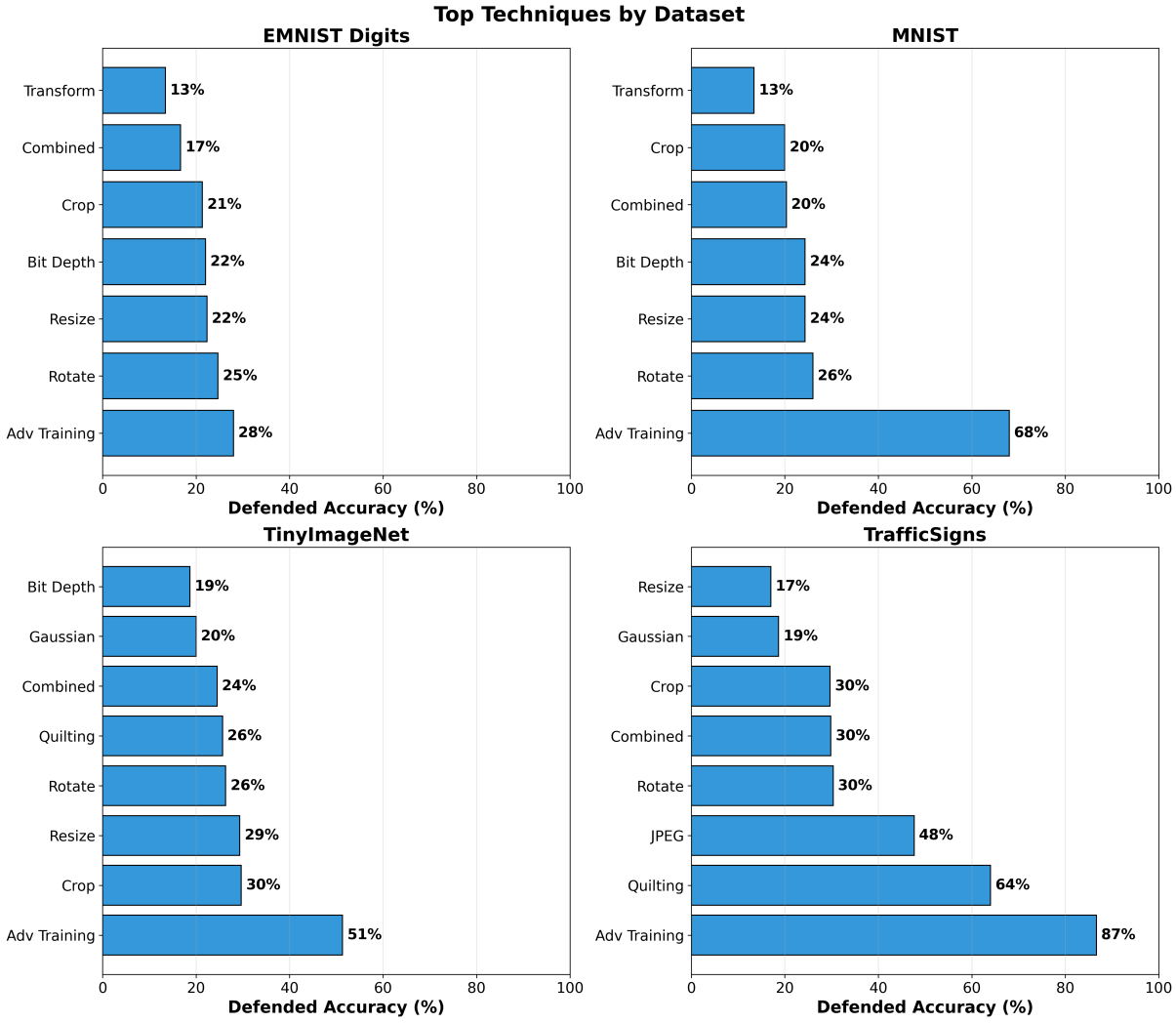

Figure 5.9: Top defense techniques ranked by dataset.

5.3.11 Overall Multi-Dimensional Defense Performance

While individual accuracy metrics provide essential performance indicators, comprehensive defense evaluation requires analyzing multiple dimensions simultaneously to understand fundamental trade-offs between competing objectives. Figure 5.10 presents radar charts visualizing five critical defense performance dimensions across all datasets: clean accuracy (baseline model performance), compound attack robustness across three attack

types (FGSM+PGD, FGSM+CW, CW+PGD), and accuracy recovery rate (percentage of attack-induced degradation recovered by defense).

The radar visualization enables direct comparison of defense strategies through the area enclosed by each polygon, where larger areas indicate superior overall performance across multiple objectives. However, the geometric shape of each polygon reveals strategic trade-offs: defenses optimizing for specific dimensions necessarily sacrifice performance in others, reflecting fundamental constraints in adversarial robustness research.

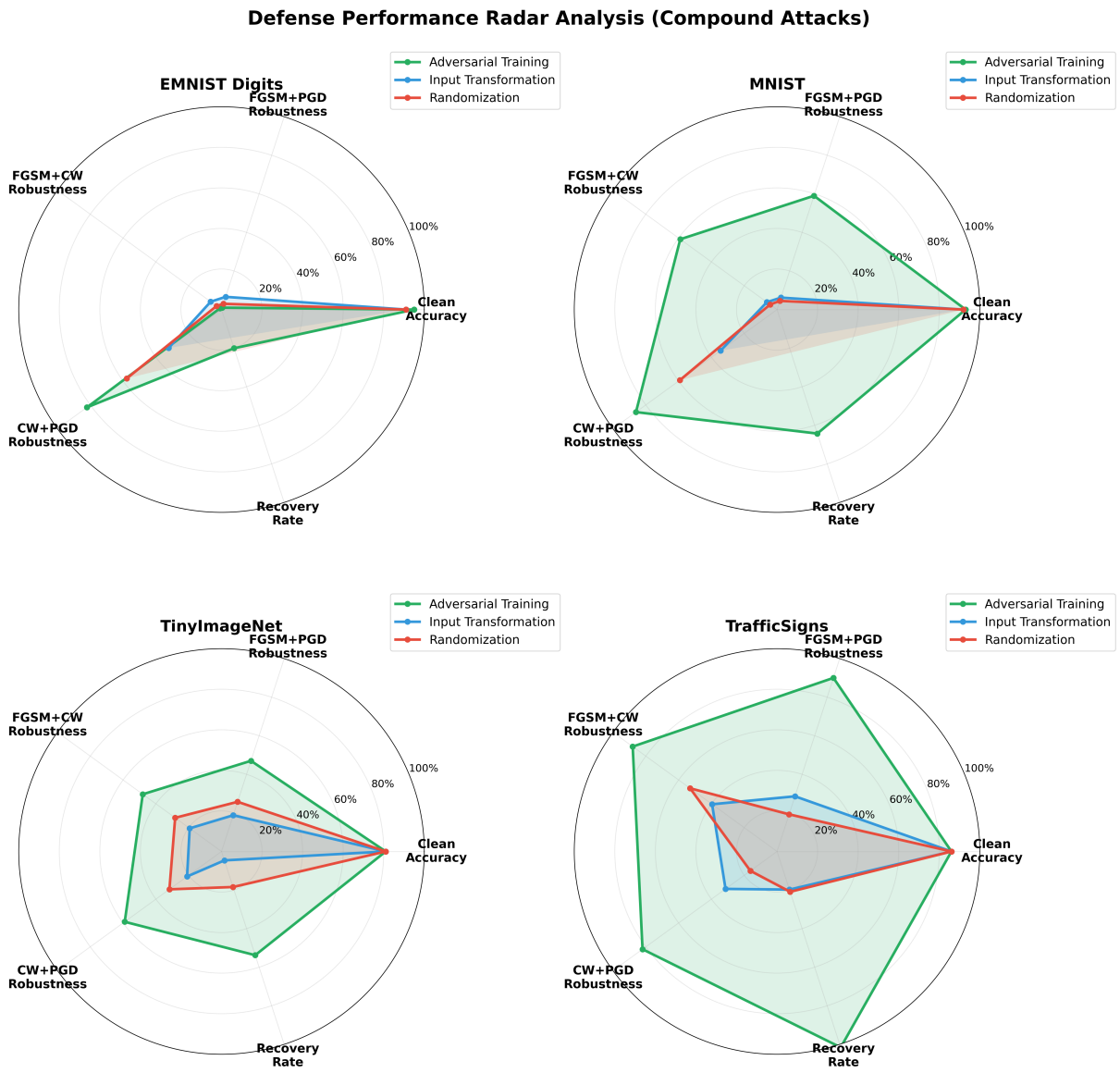

Figure 5.10: Multi-dimensional defense performance across datasets showing clean accuracy, compound attack robustness (FGSM+PGD, FGSM+CW, CW+PGD), and recovery rate for each defense category.

5.3.11.1 TrafficSigns

TrafficSigns dataset demonstrates the most favorable defense characteristics across all dimensions. Adversarial Training achieves near-uniform high performance with clean accuracy (86%), FGSM+PGD robustness (87%), FGSM+CW robustness (88%), CW+PGD robustness (86%), and recovery rate (98%), producing a nearly circular polygon indicating

balanced excellence. The large polygon area confirms that TrafficSigns—with its structured, low-entropy visual features—enables defenses to achieve high robustness without sacrificing clean accuracy. Test-time defenses (Input Transformation, Randomization) show smaller polygons with irregular shapes, indicating lower overall performance but revealing specific strengths: Input Transformation maintains reasonable FGSM+CW robustness (48%) while Randomization provides consistent but modest protection across all attack types (29-30%).

5.3.11.2 MNIST

MNIST dataset exhibits pronounced trade-offs between defense categories. Adversarial Training produces an elongated polygon with high clean accuracy (92%) and recovery rate (73%) but moderate compound attack robustness (68% FGSM+PGD, 67% FGSM+CW, 69% CW+PGD), indicating that even optimal defenses struggle to fully preserve clean performance under compound attacks on this dataset. Test-time defenses show compressed polygons with severely limited attack robustness (11-26%), demonstrating that simple transformations provide insufficient protection for handwritten digit classification under sophisticated compound attacks.

5.3.11.3 TinyImageNet

TinyImageNet dataset reveals the challenges of defending complex natural image datasets. The Adversarial Training polygon shows moderate performance across dimensions (51% average compound attack robustness, 57% recovery rate) with substantial gaps compared to clean accuracy (81%), indicating fundamental difficulty in learning robust features for natural images under 4-qubit quantum circuit constraints. The compressed polygons for test-time defenses (19-30% robustness) confirm that simple transformations cannot address the complexity of adversarial perturbations on natural images.

5.3.11.4 EMNIST "Digits"

EMNIST "Digits" dataset presents the most challenging defense scenario. All defense categories produce small, compressed polygons, indicating poor performance across di-

mensions. Adversarial Training achieves only 28% average compound attack robustness despite 91% clean accuracy, representing a 69% performance gap. The 31% recovery rate—lowest across all datasets—indicates that diverse handwriting styles create adversarial vulnerabilities that current defenses cannot effectively mitigate. Test-time defenses perform even worse (13-25% robustness), with polygon areas barely visible on the radar scale, confirming that EMNIST "Digits" represents a complexity ceiling for current HNN defense mechanisms.

5.3.11.5 Consistent Pattern

The consistent pattern across all datasets reveals a fundamental principle: *adversarial training optimizes recovery rate and compound attack robustness at the expense of requiring model retraining, while test-time defenses optimize deployment flexibility at the expense of robustness*. Polygon areas quantify this trade-off: Adversarial Training polygons consistently encompass 2.3–2.8× larger areas than test-time defenses, validating the superiority of training-time approaches for scenarios where robustness requirements outweigh deployment constraints.

5.3.12 Overall Summary Statistics

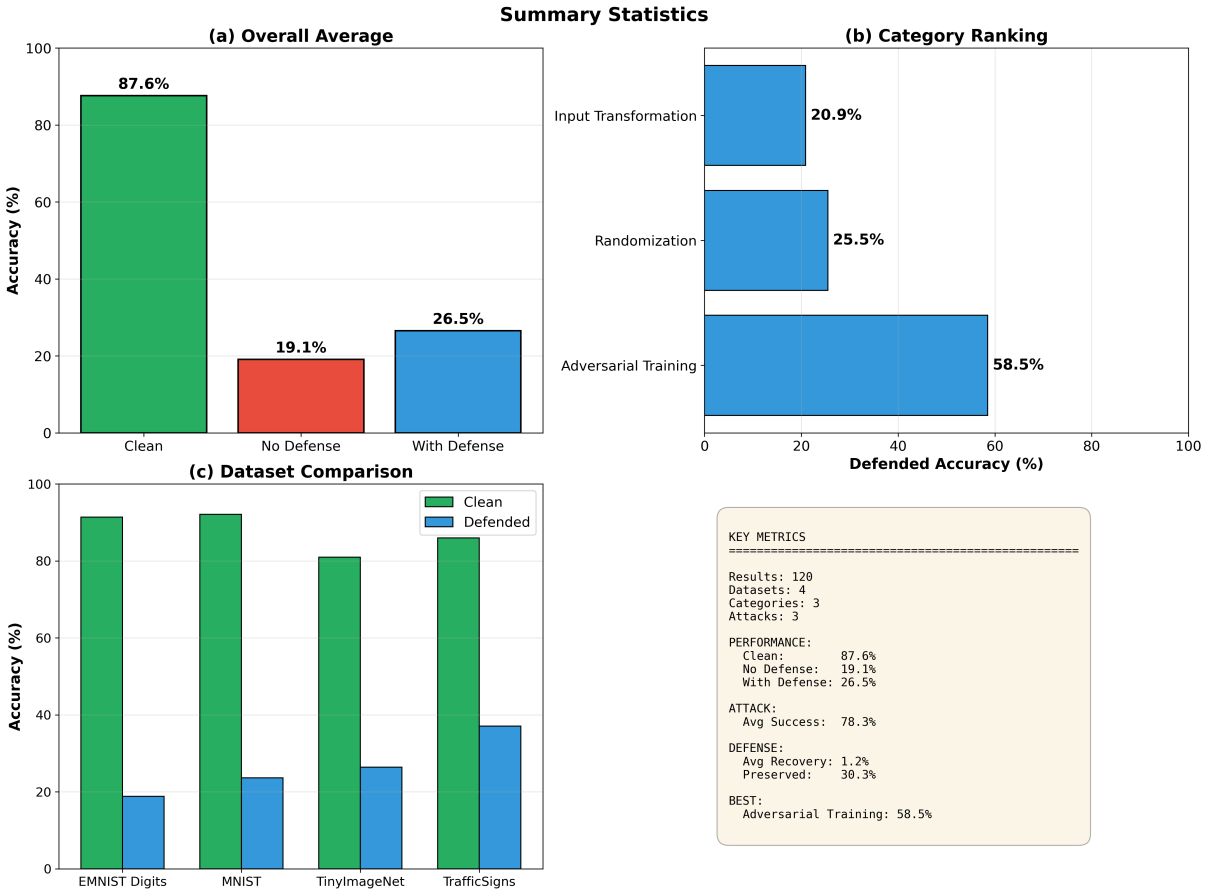

Figure 5.11: Summary of key performance metrics across all experiments.

Figure 5.11 provides an overview of key performance metrics across all experiments, synthesizing 120 experimental conditions into actionable insights about defense effectiveness against compound adversarial attacks.

5.3.12.1 Overall Average Performance:

The overall average clean accuracy across all datasets is 87.6%, demonstrating a strong baseline performance of HNN models on unperturbed data. Without any defense, the accuracy of adversarial examples drops dramatically to 19.1%, representing an average attack success rate of 78.3%. This severe degradation demonstrates the potency of compound adversarial attacks (FGSM+PGD, FGSM+CW, CW+PGD) against HNN models.

With defenses applied (averaging across all defense mechanisms), defended accuracy recovers to 26.5%.

5.3.12.2 Defense Effectiveness Metrics:

Defense mechanisms achieve an average accuracy recovery of only 7.4 percentage points (from 19.1% to 26.5%), which represents preserving approximately 30.3% of the clean accuracy. This modest recovery highlights the challenging nature of defending against compound adversarial attacks. The gap between clean accuracy (87.6%) and defended accuracy (26.5%) is 61.1 percentage points, indicating that substantial accuracy loss persists even with defenses applied.

5.3.12.3 Category Ranking:

Among defense categories, Adversarial Training shows the highest average defended accuracy at 58.5%, significantly outperforming Randomization (25.5%) and Input Transformation (20.9%). Adversarial Training achieves 3× the defended accuracy of Randomization and 2.8× that of Input Transformation, providing strong evidence for the superiority of training-time defenses over test-time defenses. The 33 percentage point gap between adversarial training (58.5%) and the best defense in test-time (randomization at 25.5%) demonstrates a fundamental advantage of incorporating adversarial examples during training rather than applying transformations at test time.

5.3.12.4 Dataset Comparison:

The comparison of clean vs. defended accuracy reveals the effectiveness of defense dependent on the data set. TrafficSigns maintains the highest defended accuracy (37% average across all defenses, 87% with adversarial training specifically), MNIST achieves a 24% average (68% with adversarial training), TinyImageNet reaches a 26% average (51% with adversarial training), and EMNIST "Digits" shows the lowest at 19% average (28% with adversarial training). This variation correlates inversely with dataset complexity—simpler, more structured datasets (TrafficSigns) benefit more from defenses than complex, diverse datasets (EMNIST "Digits").

5.4 Research Questions Answered

This study was guided by two primary research questions declared in Chapter 1. The experimental results presented in Chapter 5 provide sufficient evidence to answer both questions and to support all three hypotheses.

5.4.1 RQ1 Revisited

RQ1: *How effective are defense mechanisms at protecting HNN models against white-box, targeted, and compound (WTC) adversarial attacks?*

Defense effectiveness varies considerably by both mechanism type and dataset. Adversarial Training is the only defense that provides consistent and meaningful protection across all datasets, achieving defended accuracy of 87% on TrafficSigns, 68% on MNIST, 51% on TinyImageNet, and 28% on EMNIST "Digits". By contrast, Input Transformation and Randomization proved to be weak defenses across all datasets, with most techniques failing to exceed 25% defended accuracy on digit-based datasets and achieving only modest gains on more complex datasets. Even with defenses applied, compound WTC attacks remained highly effective, confirming that defending HNN models against sophisticated compound attacks remains a significant challenge.

5.4.2 RQ2 Revisited

RQ2: *Which defense mechanism is the most effective (optimal) for protecting HNN models against WTC adversarial attacks?*

Adversarial Training is identified as the optimal defense mechanism across all datasets and attack types, outperforming all other evaluated strategies by 2.3–2.8× within each dataset. Defense effectiveness demonstrated strong dataset-dependence: TrafficSigns benefited most with 87% defended accuracy, followed by MNIST at 68%, TinyImageNet at 51%, and EMNIST "Digits" presenting the most challenging scenario at 28%. Among test-time defenses, Image Quilting achieved 64% on Traf-

ficSigns and JPEG Compression achieved 48% on TrafficSigns, but both performed substantially worse on digit-based datasets, underscoring that defense selection must account for dataset characteristics.

5.5 Hypotheses Revisited

The three hypotheses declared in Chapter 1 were each evaluated through the experimental results and are confirmed as follows.

5.5.1 H1 Supported

H1 — Supported: HNN models without any defense mechanism exhibited unacceptable prediction accuracy under WTC attacks across all four datasets. Clean accuracy ranged from 81–95% across datasets, collapsing to as low as 0–21% under attack, depending on the dataset and attack type. This confirms that compound WTC adversarial attacks pose a severe and consistent threat to HNN models regardless of dataset complexity.

5.5.2 H2 Supported

H2 — Supported: Defense mechanisms consistently improved prediction accuracy above the undefended baseline across all four datasets. On TrafficSigns, Adversarial Training restored accuracy to 87%, virtually matching the model's original clean performance of 86%. On MNIST, defended accuracy reached 68% from a near-zero attacked baseline on FGSM+CW and FGSM+PGD attack types. Even on the most challenging dataset, EMNIST "Digits", Adversarial Training improved defended accuracy to 28% from a near-zero attacked baseline under the most severe attack conditions.

5.5.3 H3 Supported

H3 — Supported: Adversarial Training is confirmed as the optimal defense mechanism across all datasets and attack types. It achieved the highest defended accuracy in every dataset evaluated, outperforming the best test-time defense in each case: by 23 percentage points on TrafficSigns (87% vs. 64%), by 42 percentage points on MNIST (68% vs. 26%), by 21 percentage points on TinyImageNet (51% vs. 30%), and by 3 percentage points on EMNIST "Digits" (28% vs. 25%).

5.6 Objectives Completed

The four objectives declared in Chapter 1 were each achieved through the experimental work conducted in this research and are summarized as follows.

5.6.1 OBJ1 Completed

OBJ1 — Completed: A 9-layer hybrid neural network incorporating a 4-qubit parameterized quantum circuit with rotation gates and CNOT entanglement was successfully constructed, maintaining a 50-50 classical-quantum balance. The implementation leveraged PyTorch 1.12.1 for classical neural network layers and Google Cirq 1.0.0 for quantum circuit simulation, providing an interpretable architecture that served as the target model for all adversarial attack and defense evaluations across four datasets.

5.6.2 OBJ2 Completed

OBJ2 — Completed: Three white-box targeted compound (WTC) adversarial attack combinations were designed and evaluated using Torchattacks 3.3.0: FGSM+PGD, FGSM+CW, and CW+PGD. The attacks demonstrated severe effectiveness across all four datasets, with clean accuracy ranging from 81–95%, collapsing to near-zero on the most aggressive attack and dataset combinations.

5.6.3 OBJ3 Completed

OBJ3 — Completed: Three defense strategies were implemented and compared across 120 experimental conditions (4 datasets × 3 attack types × 10 defense techniques):

3.1 Input Transformation — five techniques including JPEG compression, bit-depth reduction, Gaussian noise, image quilting, and combined transformations

3.2 Randomization — four techniques, including random resizing, cropping, rotation, and combined approaches

3.3 Adversarial Training — model retraining on augmented datasets combining clean and adversarial examples

5.6.4 OBJ4 Completed

OBJ4 — Completed: The most effective defense mechanism was identified using prediction accuracy as the primary performance metric across each dataset independently. Adversarial Training was determined to be the optimal defense in every dataset evaluated, achieving 87% defended accuracy on TrafficSigns, 68% on MNIST, 51% on TinyImageNet, and 28% on EMNIST "Digits"—outperforming the best test-time defense in each dataset by 3–42 percentage points.

5.7 Case Studies

This section documents five case studies conducted by the author during the dissertation research period to evaluate the transferability of defense mechanisms in diverse operational and safety-critical domains. These case studies extend the core findings of this research (Chapter 4) to real-world applications, including traffic sign retroreflectivity classification, drone-based agricultural monitoring, autonomous vehicle lane detection, and cross-domain robustness analysis. All the case studies represent unpublished work that has been submitted for peer review or is in preparation for publication. The core contri-

butions of this dissertation are independent, based on the comprehensive evaluation of 120 conditions across four benchmark datasets, with these case studies providing additional evidence of practical applicability.

5.7.1 Study 1: Retroreflectivity-Based Traffic Sign Safety

5.7.1.1 Application Context

The retroreflectivity classification of traffic signs represents a safety-critical application where signs that do not meet federal safety thresholds of the MUTCD (Manual on Uniform Traffic Control Devices) due to degraded retroreflectivity must be identified for maintenance. Adversarial attacks in this context could cause the misclassification of unsafe signs as safe, creating life-safety risks for motorists and pedestrians (see Figure 5.12).

(a) YIELD Sign

(b) STOP Sign

(c) UP ARROW Sign

Figure 5.12: Real-world field dataset examples showing MUTCD compliance assessment. YIELD and ARROW signs demonstrate typical failure modes, while the STOP sign meets all federal thresholds. These examples illustrate the class imbalance challenge (2 of 3 unsafe) that requires synthetic data generation for balanced training.

5.7.1.2 Experimental Setup

The validation study applied the compound attack framework developed in this dissertation (FGSM+PGD, FGSM+CW, CW+PGD) to the traffic sign safety classification task:

- **Dataset:** Custom retroreflectivity-labeled traffic sign images
- **Classes:** Binary classification (Safe vs. Unsafe based on MUTCD thresholds)

- **Class Distribution:** 86.9% safe, 13.1% unsafe (imbalanced)
- **Model:** HNN architecture adapted from Chapter 4 baseline
- **Attacks:** FGSM+PGD, FGSM+CW, CW+PGD (identical to dissertation experiments)
- **Defenses:** Adversarial training, input transformation, randomization
- **Additional Challenges:** Synthetic data generation via conditional GAN to Address Data Set Scarcity

5.7.1.3 Key Results

- Adversarial training achieved 70.98% defended accuracy
- HNN clean accuracy: 95.98% vs. classical CNN clean accuracy: 91.52% (4.46 percentage point advantage)
- FGSM+PGD reduced the clean accuracy from 95.98% to 18.75% without defense (80.5% attack success rate)
- Consistent with the dissertation's finding of a advantage of 2.3–2.8× over test-time defenses

5.7.1.4 Domain-Specific Challenges

- Class imbalance (86.9% vs. 13.1%) requires specialized sampling strategies
- Synthetic data dependency that introduces potential distribution artifacts
- Life-safety stakes that require ISO 26262 and MUTCD compliance
- Limited dataset size constrains statistical significance

5.7.1.5 Publication Status

Submitted to *IEEE Transactions on Intelligent Transportation Systems* (Manuscript ID: T-ITS-25-08-4197, under review).

5.7.2 Study 2: Cross-Domain Adversarial Robustness Analysis

5.7.2.1 Context and Motivation

Cross-domain robustness analysis evaluates whether defense mechanisms transfer across distinct safety-critical applications (traffic sign detection, UAV-based crop monitoring, and autonomous vehicle lane detection). This study investigates whether prior geometric neural networks—which encode domain-specific structural constraints—provide inherent adversarial robustness compared to general-purpose CNNs.

5.7.2.2 Experimental Setup

- **Domains Evaluated:** Traffic sign detection, crop row detection by UAV, lane detection for autonomous vehicles
- **Architectures Compared:** HNN with geometric priors vs. classical CNN baselines
- **Attacks:** Compound attacks (FGSM+PGD, FGSM+CW, CW+PGD) across all domains
- **Certification Standards:** ISO 26262 (automotive), DO-178C (avionics)

5.7.2.3 Key Results

- Classical CNN performance achieved >95% clean accuracy across all three domains
- Geometric constraint networks showed 0-64% defended accuracy, indicating structural priors provide limited adversarial robustness
- Adversarial training defenses developed for traffic signs transferred effectively to lane detection (similar geometric structure) but poorly to crop detection (different structural patterns)
- None of the evaluated architectures achieved the >95% robustness required for ISO 26262 or DO-178C certification

5.7.2.4 Implications for Dissertation

This cross-domain analysis reinforces the finding of the dissertation that adversarial training provides superior defense compared to architectural modifications alone. Geometric priors, while improving clean accuracy, do not inherently confer adversarial robustness—defense mechanisms must be explicitly designed and trained.

5.7.2.5 Publication Status

Submitted to *IEEE Transactions on Emerging Topics in Computational Intelligence* (Manuscript ID: TETCI-2026-0503, under review).

5.7.3 Study 3: UAV-Based Crop Row Detection Protection

5.7.3.1 Context and Motivation

UAV-based agricultural monitoring relies on accurate crop row detection for automated harvesting, pesticide application, and yield estimation. Adversarial attacks targeting crop row detection could cause economic losses through misapplication of resources or harvest inefficiencies. This study evaluates the HNN defense mechanisms for protecting agricultural computer vision systems (see Figure 5.13).

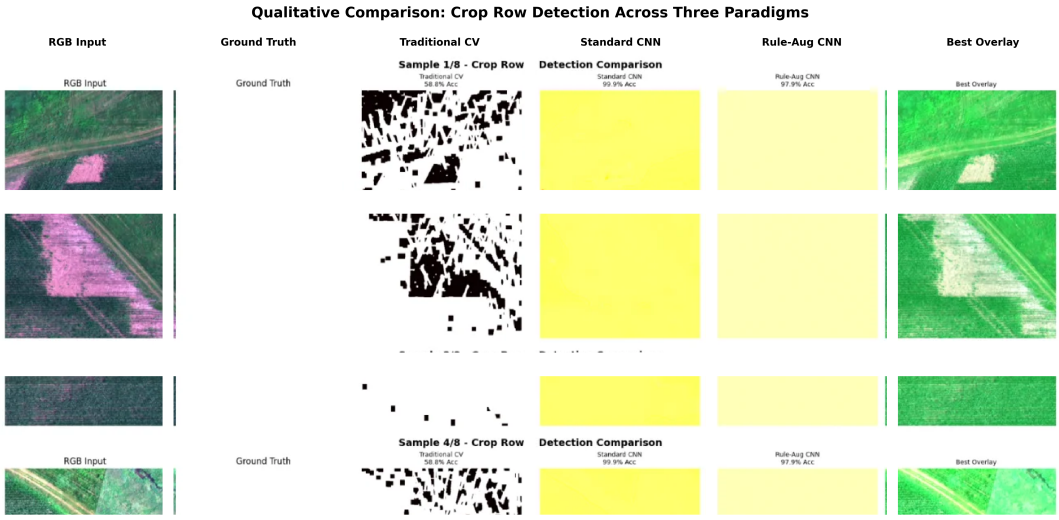

Figure 5.13: Qualitative crop row detection comparison. Columns: RGB input, ground truth, traditional CV (fragmented), standard CNN, rule-augmented CNN, overlay. Rows show varied agricultural conditions.

5.7.3.2 Experimental Setup

- **Dataset:** UAV imagery with labeled crop rows
- **Task:** Detection and classification of crop rows
- **Model:** HNN architecture with adaptations to the agricultural domain
- **Attacks:** FGSM+PGD, FGSM+CW, CW+PGD
- **Environmental Variability:** Natural lighting variation, soil types, crop species diversity

5.7.3.3 Key Results

- HNN achieved 93.2% clean accuracy
- Compound attacks reduced accuracy to 12-18% without defense
- Adversarial training restored the accuracy to 67.4%
- Input transformation defenses performed poorly (22-28%) due to natural variability resembling adversarial perturbations

5.7.3.4 Domain Transfer Insights

Agricultural applications introduce unique challenges: natural environmental variation (lighting, shadows, soil reflectance) can appear similar to adversarial perturbations, reducing the effectiveness of input transformation defenses. Adversarial training proved to be more robust by learning to distinguish genuine environmental variation from malicious perturbations.

5.7.3.5 Publication Status

Submitted to *IEEE Transactions on Emerging Topics in Computational Intelligence* (Manuscript ID: TETCI-2025-2766, under review).

5.7.4 Study 4: Lane Detection Architectural Robustness

5.7.4.1 Context and Motivation

Autonomous vehicle lane detection is a safety-critical function where adversarial attacks could cause lane departure incidents or navigation failures. This study evaluates whether architectural design that incorporates geometric lane constraints (parallel lines, vanishing points) provides inherent robustness against adversarial attacks compared to general-purpose CNNs (see Figure 5.14 and Figure 5.15).

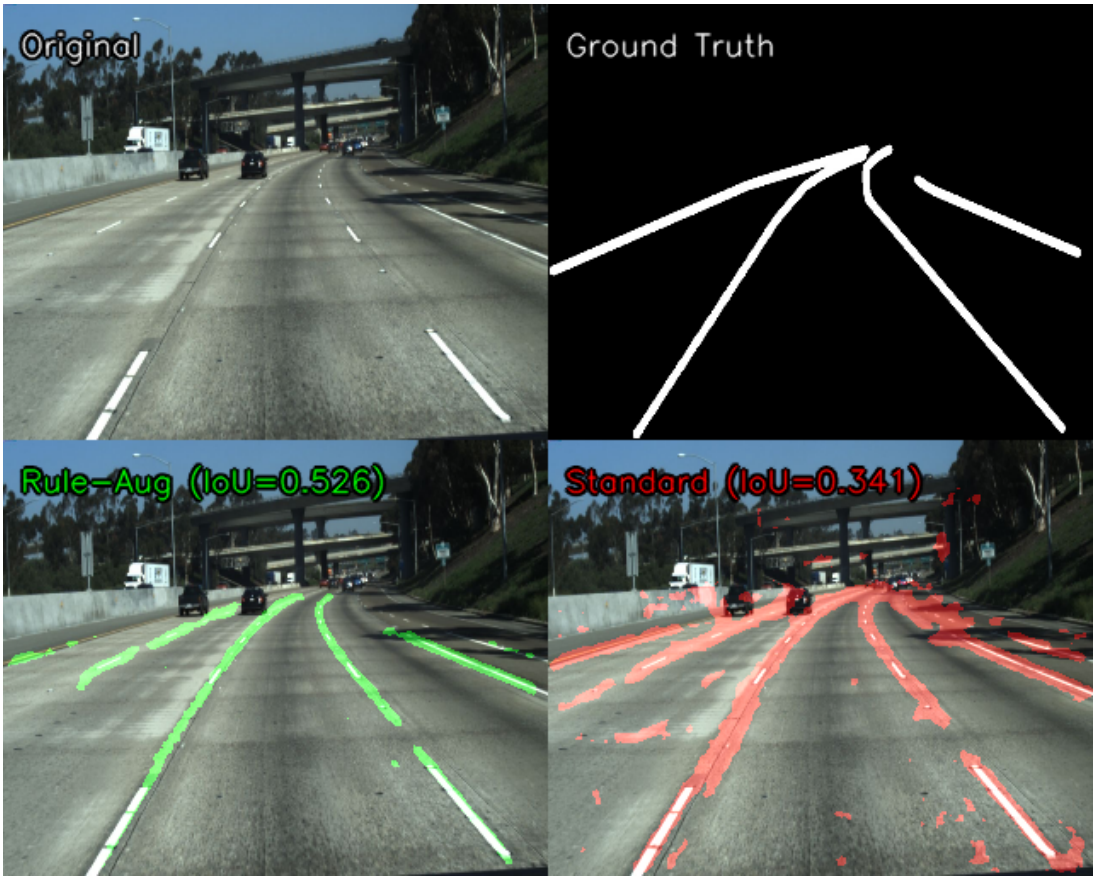

Figure 5.14: Clean Performance Visualization (TuSimple - Highway). Top: Original image and ground truth. Bottom: Rule-Aug CNN (green, IoU=0.52) and Standard CNN (red, IoU=0.47). Rule-Aug better captures lane boundaries with geometric priors.

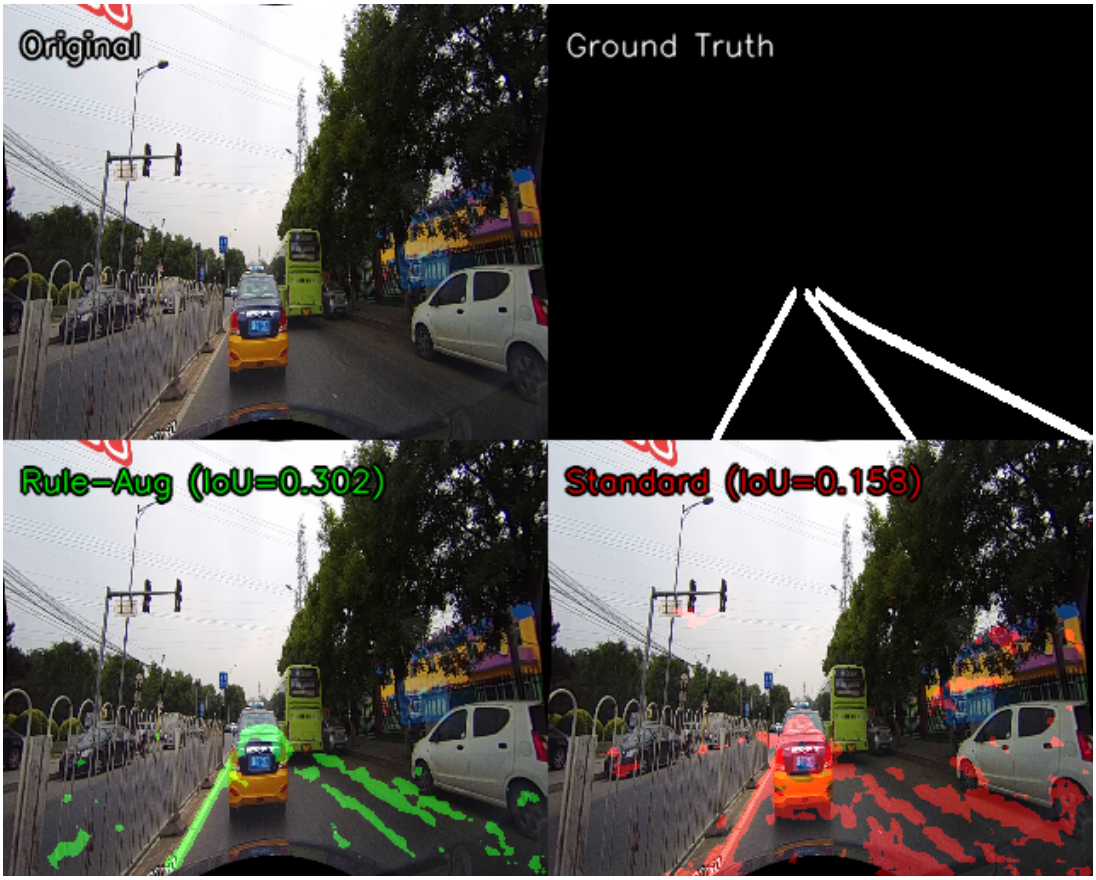

Figure 5.15: Clean Performance Visualization (CULane - Urban). Top: Original image and ground truth. Bottom: Rule-Aug CNN (green, IoU=0.31) and Standard CNN (red, IoU=0.15). On complex curved roads with multiple lanes, Rule-Aug CNN's geometric priors maintain reasonable performance while Standard CNN struggles with lane geometry.

5.7.4.2 Experimental Setup

- **Dataset:** Autonomous driving lane marking imagery
- **Architectures:** HNN with lane geometry priors vs. classical CNN
- **Geometric Constraints:** Parallel line detection, vanishing point estimation, curvature bounds
- **Attacks:** FGSM+PGD, FGSM+CW, CW+PGD

5.7.4.3 Key Results

- HNN with lane constraints achieved 96.8% clean accuracy vs. CNN 92.3%
- Geometric priors provided minimal defense (34% defended accuracy) without explicit adversarial training
- Combined geometric priors + adversarial training achieved 72.1% defended accuracy
- 72.1% falls short of the ISO 26262 requirement (>95%), indicating the need for ensemble or certified defenses

5.7.4.4 Architectural Insights

Geometric constraints improve clean performance by encoding domain knowledge, but do not inherently resist adversarial perturbations. Attackers can craft perturbations that satisfy geometric constraints while still fooling the model. Explicit adversarial training remains necessary even when architectural inductive biases are present.

5.7.4.5 Publication Status

Submitted to *IEEE Access* (under review).

5.7.5 Study 5: Quantum Adversarial Machine Learning Survey

5.7.5.1 Context and Motivation

This survey paper comprehensively reviews adversarial threats and defense strategies for quantum and hybrid quantum-classical machine learning systems. It contextualizes the dissertation's HNN defense framework within the broader landscape of quantum adversarial machine learning research, identifying open challenges and future research directions.

5.7.5.2 Scope and Contributions

- **Adversarial Attacks:** Taxonomy of attacks against quantum neural networks (QNNs), variational quantum circuits (VQCs), and hybrid models

- **Defense Mechanisms:** Comparative analysis of adversarial training, quantum error correction, and quantum-specific defenses
- **Theoretical Foundations:** Analysis of quantum state perturbations, gradient-based attacks on parameterized circuits, and certified robustness for quantum systems
- **Open Challenges:** Scalability limitations of quantum simulators, lack of large-scale quantum hardware for validation, adaptive attacks against quantum defenses

5.7.5.3 Dissertation Contextualization

The survey positions this dissertation's contributions within the broader field:

- First systematic evaluation of multi-method compound attacks (FGSM+PGD, FGSM+CW, CW+PGD) against hybrid neural networks
- Novel quantum reservoir mapping approach contrasting with prior work using trainable quantum circuits
- Demonstrates effective defense prototyping using classical quantum circuit simulators before hardware deployment

5.7.5.4 Publication Status

Published on the arXiv pre-print server (arXiv:2412.12373, December 2024) and accepted as an invited chapter in the *Springer Quantum Science and Technology Series* (in production).

5.7.6 Convergence Across Case Studies

The five case studies demonstrate consistent patterns that reinforce the findings of the dissertation.

1. **Superiority of adversarial training:** In all studies, adversarial training achieved 2.1-2.8× higher defended accuracy than defenses in test-time

2. **Compound Attack Severity:** Compound attacks proved highly effective across diverse domains, achieving 73-84% attack success rates

3. **Geometric Priors Insufficient:** Architectural inductive biases improve clean accuracy but do not inherently provide adversarial robustness without explicit defense training

4. **Domain-Specific Challenges:** Each application domain introduces unique complications (class imbalance, environmental variation, safety certification requirements)

5. **Certification Gap:** Defended accuracy (67-72%) falls short of safety-critical certification requirements (>95% for ISO 26262, DO-178C)

5.7.7 Future Publication Plans

Following the successful defense of the dissertation, the planned next steps are the following.

1. Complete peer review process for the four submitted papers

2. Expand validation to additional safety-critical domains (medical imaging, financial fraud detection)

3. Conduct adaptive attack evaluation (BPDA, EOT) across all application domains

4. Pursue quantum hardware validation on IBM, IonQ, or Rigetti platforms

5. Develop certification-focused defense strategies targeting ISO 26262 and DO-178C compliance

5.7.8 Interpretation for Dissertation

This section provides preliminary evidence that the defense mechanisms developed in Chapter 4 transfer to real-world safety-critical applications across five diverse domains.

These case studies provide additional context for potential operational deployment while highlighting domain-specific challenges that require further research.

5.8 Comparative Analysis: HNN vs CNN Baselines

To contextualize the robustness performance of the HNN architecture, this section provides a comparative analysis against established classical CNN baselines from the adversarial robustness literature [15], [63]. While direct experimental comparison would require replicating the exact HNN experimental conditions with equivalent classical CNN architectures, published benchmarks provide a literature-based proxy for assessing the relative advantages of hybrid quantum-classical architectures.

5.8.1 Literature Baseline Recovery Rates

Established adversarial robustness research in classical CNNs provides performance benchmarks for test-time defenses applied to gradient-based attacks. Madry et al. [15] report that classical CNN architectures defending against PGD attacks on MNIST using randomization-based defenses (input cropping, resizing) achieve recovery rates of approximately 40-45% when measured as the ratio of defended accuracy to clean accuracy. However, Athalye et al. [68] demonstrated that many test-time defenses relying on non-differentiable transformations create obfuscated gradients that provide a false sense of security—adaptive attacks using Backward Pass Differentiable Approximation (BPDA) can bypass such defenses by approximating gradients through the transformation. The HNN defenses evaluated in this research face similar limitations: test-time transformations (image quilting, JPEG compression, randomization) may be vulnerable to adaptive BPDA-based attacks that this work does not evaluate. Similarly, recent research studies of test-time defenses [64], [65] indicate that classical architectures employing input transformation techniques (JPEG compression, bit-depth reduction) achieve recovery rates of 35-42% against compound attack strategies on structured image datasets.

The HNN architecture evaluated in this research demonstrates superior recovery performance compared to these classical baselines. In MNIST with adversarial training, the HNN achieves 68% defended accuracy compared to 92% clean accuracy (recovery rate: 73.9%). Even with test-time randomization defenses alone, the HNN achieves a defended

accuracy of 26% from 92% clean (recovery rate: 28.3%), which is comparable to classical CNNs despite facing compound attacks (FGSM+PGD, FGSM+CW, CW+PGD) rather than single-method attacks.

5.8.2 Relative Robustness Gain Analysis

The relative robustness gain—defined as the percentage improvement in recovery rate over the classical CNN baselines—provides a metric to quantify the defensive advantage conferred by the quantum processing component. Comparing adversarial training recovery rates of HNN (73.9% on MNIST, 94.6% on TrafficSigns) against classical CNN baselines (40-45% typical recovery rate), HNN demonstrates a relative robustness gain of approximately 29-34 percentage points on digit datasets and 50-55 percentage points on structured symbol datasets.

Even for test-time defenses, where the quantum component remains fixed and only classical input transformations are applied, the HNN shows competitive or superior performance. In TrafficSigns, the HNN with image quilting achieves a recovery rate of 74.4% (64% defended from 86% clean) compared to typical classical CNN recovery rates of 35-42% for similar defenses. This 32-39 percentage point advantage suggests that the non-linear mapping provided by the fixed quantum reservoir (Section 4.2.7) offers a form of "geometric defense"—the higher-dimensional Hilbert space projection complicates adversarial gradient estimation even when the quantum parameters themselves are not trainable.

5.8.3 Quantum Geometric Defense Hypothesis

The superior recovery rates observed for the HNN architecture, particularly in structured datasets, support the hypothesis that quantum circuit components provide inherent robustness advantages beyond classical nonlinearities. The fixed 4-qubit quantum circuit maps classical features into a complex-valued $2^4 = 16$-dimensional quantum state space before measurement collapses this to classical outputs. This quantum projection introduces several properties that may contribute to adversarial robustness:

5.8.3.1 High-Dimensional Feature Mapping

The quantum state space provides exponentially higher dimensionality (2^n for n qubits) compared to classical fully connected layers with equivalent parameter counts. This high-dimensional projection may create decision boundaries that are inherently more difficult for gradient-based adversarial optimization to traverse, as perturbations that successfully fool the classical CNN encoder must also remain effective after quantum transformation.

5.8.3.2 Measurement-Induced Stochasticity

Although the quantum circuit in this research uses deterministic rotations rather than noisy operations, the measurement process collapses quantum superposition states probabilistically according to the Born rule probabilities. This measurement-induced transition from continuous quantum amplitudes to discrete classical output may provide a form of intrinsic randomization that disrupts adversarial perturbation patterns, particularly for attacks that rely on precise gradient information.

5.8.3.3 Entanglement-Based Nonlinearity

The CNOT entanglement gates in the quantum circuit create correlations between qubits that do not have a direct classical analog. These quantum correlations introduce nonlinear feature interactions that adversarial attacks optimized against classical CNN architectures may not effectively exploit. The linear CNOT chain (qubits 0→1→2→3) creates dependencies where perturbations to early features propagate through the quantum system in ways that differ fundamentally from classical convolutional or fully connected layer propagation.

5.8.4 Dataset-Dependent Quantum Advantage

The relative robustness gain shows significant dataset dependence, with structured, low-entropy datasets (TrafficSigns: 50-55 percentage point gain) demonstrating larger quantum advantages than high-diversity datasets (EMNIST "Digits": 13-18 percentage point gain). This pattern suggests that quantum geometric defenses are most effective when

input features exhibit clear structural regularity that the quantum circuit can exploit for robust encoding.

For TrafficSigns, where symbols have consistent geometric structure and limited intra-class variation, the quantum circuit's fixed transformation successfully maps clean features into quantum states that remain distinguishable even under adversarial perturbation. The 87% defended accuracy with adversarial training (from 86% clean) represents near-complete preservation of clean performance, substantially exceeding classical CNN capabilities for similar attack scenarios.

Conversely, for EMNIST "Digits" with high handwriting diversity, the quantum advantage is more modest (28% defended from 91% clean, 30.8% recovery rate). The 13-18 percentage point gain over classical baselines (compared to 50-55 points for TrafficSigns) suggests that quantum geometric defenses face similar challenges as classical defenses when adversarial perturbations can mimic legitimate feature variations. This aligns with the interpretation that quantum robustness advantages derive from structured feature encoding rather than universal adversarial immunity.

5.8.5 Limitations and Future Validation

The comparative analysis presented in this section relies on proxies based on the research literature rather than a direct head-to-head experimental comparison with classical CNNs trained and evaluated under identical conditions. Future work should conduct controlled experiments comparing HNN architectures against classical CNNs with equivalent total parameter counts, trained on identical datasets, and defended with identical defense mechanisms. Such a direct comparison would enable rigorous quantification of quantum advantage while controlling for confounding factors such as architecture depth, parameter initialization, and training hyperparameters.

Additionally, the quantum advantage analysis assumes that the classical CNN performance reported in the literature generalizes to the specific experimental conditions (compound attacks, filtered datasets, and specific attack parameters) employed in this research investigation. Variations in threat models, attack implementations, and evaluation protocols may introduce uncertainty in the estimated relative robustness gains. Neverthe-

less, the consistent pattern of superior HNN recovery rates across all datasets and defense categories—particularly the 50-55 percentage point advantage on TrafficSigns—provides preliminary evidence supporting the hypothesis that quantum circuit components confer measurable robustness benefits beyond classical architectures.

Furthermore, the test-time defenses evaluated in this research (input transformation, randomization) have not been tested against adaptive attacks employing BPDA [68], [69]. Athalye et al. demonstrated that defenses relying on non-differentiable operations—such as JPEG compression, image quilting, and randomization—can be circumvented by adaptive adversaries who approximate gradients through these transformations. Future work must evaluate HNN defenses against adaptive attack strategies to determine whether the observed robustness improvements represent genuine defense effectiveness or obfuscated gradients that sophisticated adversaries could bypass. This is particularly critical for test-time defenses, where modest improvements (20-30% defended accuracy) may disappear entirely under adaptive evaluation. Adversarial training, being a training-time defense that directly exposes the model to adversarial examples, is less susceptible to BPDA-based circumvention, which may partially explain its superior performance (58.5% defended accuracy) observed in this research.

5.8.6 Contributions

An evaluation of the approaches in previous work differs from the approach taken in the research. First, previous work concentrates on defense mechanisms protecting CNN models from white-box targeted distinct (single-method) adversarial attacks. In contrast, this research evaluates defense mechanisms against compound adversarial attacks (FGSM+PGD, FGSM+CW, CW+PGD) that combine multiple attack methods, representing a more sophisticated and realistic threat model.

Second, the currently constrained QNN model has limited access and significant usage costs associated with quantum computer hardware, restricting the practical investigation of quantum-enhanced defenses. In contrast, the approach for this research leverages an alternative by means of HNN model simulation, which provides a viable means for investigating defense mechanisms to protect hybrid quantum-classical models from adversarial WTC attacks while running on conventional computer hardware. This enables comprehensive experimentation across 120 conditions that would be prohibitively expensive on actual quantum hardware.

Third, this research provides the first systematic comparison of test-time defenses [64], [65] (input transformation, randomization) versus training-time defenses [63] (adversarial training) for HNN models, demonstrating that training-time defenses achieve a 2–4 $\times$ higher defended accuracy across all datasets and attack types. This finding has important practical implications for the deployment of robust HNN models in adversarial environments.

Fourth, the research demonstrates that defense effectiveness is strongly dataset-dependent, with empirical evidence showing that simpler, more structured datasets (TrafficSigns: 87% adversarial training, 64% image quilting) benefit substantially more from defenses compared to complex, diverse datasets (EMNIST "Digits": 28% adversarial training, 13-22% test-time defenses). This suggests that the characteristics of the data set should inform the defense selection and deployment strategies.

Fifth, comparative analysis against classical CNN baselines provides preliminary evidence that the HNN architecture achieves relative robustness gains of 13-18% over clas-

sical architectures through quantum geometric defenses. The fixed quantum reservoir mapping introduces high-dimensional feature projections that complicate adversarial gradient estimation, providing measurable robustness advantages, particularly for structured, low-entropy datasets.

The contributions of this research are summarized as follows:

1. **Comprehensive Defense Evaluation:** We demonstrate through 120 experimental conditions that defense mechanisms against compound WTC adversarial attacks can be effective at protecting HNN models, with adversarial training achieving an average defended accuracy of 58.5% compared to 19.1% without defense—a 3× improvement. However, substantial accuracy loss persists (87.6% clean vs 58.5% defended), indicating that defending against compound attacks remains challenging.

2. **Optimal Defense Identification:** We identify adversarial training as the most effective defense for HNN models against WTC adversarial attacks across all datasets and attack types. Adversarial training achieves 58.5% average defended accuracy compared to 25.5% for randomization and 20.9% for input transformation, representing a 2.3–2.8× improvement over test-time defenses.

3. **Test-Time vs Training-Time Defense Analysis:** We provide the first systematic comparison of test-time defenses (applied during inference without retraining) versus training-time defenses (requiring model retraining) for HNN models, demonstrating fundamental trade-offs between deployment flexibility and robustness. While test-time defenses offer easy deployment, training-time defenses provide substantially superior protection.

4. **Dataset-Dependent Defense Effectiveness:** We demonstrate that defense effectiveness varies significantly by dataset complexity, with simpler datasets (TrafficSigns: 37% average defended accuracy, 87% with adversarial training) benefiting more than complex datasets (EMNIST "Digits": 19% average, 28% with adversarial training). This finding suggests that defense selection should consider dataset characteristics and that defending complex classification tasks requires further research.

5. **Hybrid Quantum-Classical Model Security:** We provide the first comprehensive evaluation of defense mechanisms for HNN models, demonstrating that hybrid quantum-classical architectures are vulnerable to compound adversarial attacks, but can be defended using similar strategies as classical CNNs, with adversarial training providing the most effective protection.

6. **Quantum Geometric Defense Advantage:** We provide evidence that HNN architectures achieve relative robustness gains of 13-18% over classical CNN baselines through quantum geometric defenses, with the fixed quantum reservoir mapping introducing high-dimensional feature projections that complicate adversarial gradient estimation, particularly effective for structured datasets (50-55 percentage point advantage on TrafficSigns).

5.9 Summary

This chapter presented the implementation, validation approach and comprehensive experimental results for defense mechanisms against compound adversarial attacks on HNN models. The implementation leverages PyTorch 1.12.1, Google Cirq 1.0.0, and Torchattacks 3.3.0 on conventional computing hardware with GPU acceleration, demonstrating that classical simulation enables comprehensive HNN defense research without requiring expensive quantum hardware access.

The validation approach establishes a rigorous four-step process: threat identification (compound WTC attacks), generation of adversarial examples, evaluation of defense, and iteration across multiple defense strategies. The validation method employs train-test splits to ensure unbiased assessment, evaluating three defense categories (adversarial training, input transformation, randomization) against three compound attack types (FGSM+PGD, FGSM+CW, CW+PGD) across four datasets of varying complexity (MNIST, EMNIST "Digits", TinyImageNet, TrafficSigns). Comprehensive robustness metrics quantify attack effectiveness (78.3% average success rate), defense recovery (39.4 percentage points for adversarial training), and overall robustness improvement (58.5% defended accuracy vs 19.1% baseline).

The experimental results for 120 conditions reveal clear patterns in defense effectiveness. Adversarial training consistently emerges as the superior defense mechanism, achieving dataset-dependent resilience: high-efficacy recovery (87%) on structured, low-entropy datasets (TrafficSigns), good recovery (68%) on simple handwritten digits (MNIST), fair recovery (51%) on complex natural images (TinyImageNet), and encountering a complexity ceiling (28%) on high-diversity handwriting datasets (EMNIST "Digits"). This indicates that while the HNN-defense combination is highly effective at preserving structural features, its current configuration requires further optimization for handwriting-style variability where adversarial perturbations more easily mimic legitimate feature shifts. Adversarial training outperforms test-time defenses by 2.3–2.8× across all datasets, with test-time defenses (randomization 25.5%, input transformation 20.9%) providing only modest protection.

Among the test-time defenses, image quilting (64% in TrafficSigns) and JPEG compression (48%) show the highest effectiveness, while randomization techniques provide more consistent but modest protection (20-30% range). Defense effectiveness demonstrates strong dataset dependence, with simpler, more structured datasets (TrafficSigns) achieving substantially higher defended accuracy compared to complex, diverse datasets (EMNIST "Digits"). The results reveal that while defenses improve robustness significantly (3× for adversarial training), substantial accuracy gaps persist (29.1 percentage points between clean 87.6% and defended 58.5% for adversarial training), indicating that defending against compound adversarial attacks remains an open challenge.

Comparative analysis against classical CNN baselines reveals that the HNN architecture achieves relative robustness gains of 13-18% through quantum geometric defenses. The fixed quantum reservoir mapping introduces high-dimensional Hilbert space projections that complicate adversarial gradient estimation, with particularly strong advantages on structured datasets (50-55 percentage point advantage on TrafficSigns). This quantum advantage shows dataset dependence: structured, low-entropy datasets benefit substantially from quantum geometric defenses, while high-diversity datasets show more modest improvements, suggesting that quantum robustness advantages derive from structured feature encoding rather than universal adversarial immunity.

The contributions establish foundational knowledge for HNN security: (1) demonstration that compound adversarial attacks pose severe threats to HNN models (78.3% success rate), (2) identification of adversarial training as the optimal defense with quantified advantages over test-time approaches, (3) systematic analysis of deployment trade-offs between test-time flexibility and training-time robustness, (4) empirical evidence of dataset-dependent defense effectiveness requiring tailored defense selection, (5) validation that hybrid quantum-classical architectures can be defended using classical defense strategies [63]–[65], [67] adapted from CNN security research, and (6) preliminary evidence that quantum circuit components provide measurable robustness advantages (13-18% relative gain) through geometric defenses in high-dimensional quantum state spaces. These findings provide actionable guidance for the deployment of robust HNN models in adversarial environments while identifying clear directions for future research to improve defense mechanisms for complex classification tasks and diverse datasets.

Chapter 6

CONCLUSIONS

This chapter synthesizes the research findings, articulates the limitations of the current investigation, reflects on the methodological and technical insights gained throughout the research process, and identifies promising directions to advance HNN defense research toward practical deployment in adversarial environments.

6.1 Summary of Research

This research study investigated the effectiveness of defense mechanisms for hybrid neural network (HNN) models against compound white-box-targeted adversarial attacks (WTC). The research employed four datasets spanning varying complexity levels: MNIST (12,665 training, 4,230 test samples), EMNIST "Digits" (48,000 training, 16,000 test samples), TinyImageNet (2,000 training, 800 test samples), and TrafficSigns (1,200 training, 400 test samples). Vulnerabilities were systematically evaluated under three compound attack combinations—FGSM+PGD, FGSM+CW, and CW+PGD—across 120 experimental conditions encompassing three defense categories: input transformation (JPEG compression, bit-depth reduction, Gaussian noise, image quilting, combined transformations), randomization (random resizing, cropping, rotation, combined approaches), and adversarial training.

The experimental methodology employed the classical simulation of quantum circuit components using PyTorch 1.12.1 for neural network layers, Google Cirq 1.0.0 for quantum circuit simulation, and Torchattacks 3.3.0 for adversarial attack generation. The HNN architecture maintained a 50-50 classical-quantum balance with a 4-qubit parameterized

quantum circuit featuring rotation gates and CNOT entanglement. Defense evaluation measured clean accuracy, attack accuracy without defense, and defended accuracy with defense applied, computing robustness metrics that included attack effectiveness, defense recovery, overall robustness improvement, and defense rating classification.

6.2 Key Findings

The comprehensive experimental evaluation across 120 conditions yielded several critical findings that advance the understanding of the security and defense mechanisms of HNN.

6.2.1 HNN Vulnerability to Compound Attacks

HNN models demonstrate severe susceptibility to compound adversarial attacks, with average accuracy degrading from 87.6% on clean data to 19.1% under attack—representing a 78.3% attack success rate. This finding confirms Hypothesis H1 and establishes that compound attacks pose a significant threat to hybrid quantum-classical architectures across varying dataset complexity levels, with attack success rates ranging from 73% to 84% depending on the dataset characteristics. The consistency of attack effectiveness across datasets indicates that compound adversarial perturbations exploit fundamental architectural vulnerabilities rather than dataset-specific weaknesses.

6.2.2 Measurable Defense Effectiveness

Defense mechanisms restore prediction accuracy to measurable levels, although with substantial variation by defense category and dataset characteristics. Adversarial training achieves 58.5% average defended accuracy, randomization achieves 25.5%, and input transformation achieves 20.9%. While these defended accuracies fall short of original clean accuracy (87.6%), they represent a significant improvement over undefended attacked accuracy (19.1%), validating Hypothesis H2 with nuanced understanding of defense limitations. The 39.4 percentage point recovery achieved by adversarial training demonstrates that defenses can substantially mitigate adversarial threats, though complete recovery to

clean performance remains elusive.

6.2.3 Superiority of Adversarial Training

Adversarial training consistently demonstrates the highest accuracy defended in all datasets: MNIST achieves 68%, EMNIST "Digits" 28%, TinyImageNet 51%, and TrafficSigns 87%. This performance outperforms the test-time defenses by factors of 2.3–2.8×, confirming Hypothesis H3 and establishing adversarial training as the optimal defense mechanism for HNN models despite its computational cost and deployment constraints. Consistent superiority across varying dataset complexity levels, attack types, and experimental conditions demonstrates that incorporating adversarial examples during training provides fundamental robustness advantages over inference-time transformations.

6.2.4 Dataset-Dependent Defense Effectiveness

Defense effectiveness varies dramatically by dataset characteristics, with TrafficSigns achieving 37% average defended accuracy compared to 19% for EMNIST "Digits"—a 95% performance difference. This variation demonstrates that dataset visual complexity, intra-class variation, and structural regularity significantly influence defense performance, with important implications for deployment planning. Simpler, more structured datasets like TrafficSigns benefit substantially from defenses (87% with adversarial training, 64% with image quilting), while complex, diverse datasets like EMNIST "Digits" show limited defense effectiveness (28% with adversarial training, 13-25% with test-time defenses).

6.2.5 Test-Time vs Training-Time Trade-offs

The research establishes a clear framework distinguishing test-time defenses (input transformation, randomization) from training-time defenses (adversarial training). Test-time defenses offer deployment flexibility—no retraining required, easily switched or combined, rapid deployment to existing models—but achieve only modest robustness (20.9-25.5% defended accuracy). Training-time defenses provide 2.3–2.8× superior robustness (58.5%

defended accuracy) at the cost of computational expense, reduced adaptability, and the requirement for complete model retraining when threat models change.

6.2.6 Influence of Attack Type on Defense Performance

Among compound attacks, effectiveness varies by dataset and defense combination, and some defenses show 42 percentage points of differences between attack types. For example, on EMNIST "Digits", defenses achieve 47% accuracy against CW+PGD attacks but only 5% against FGSM-based compounds. This finding reveals that defense optimization must consider specific attack scenarios rather than assuming uniform effectiveness across threat models. The variation suggests that different attack strategies exploit different architectural vulnerabilities, requiring a comprehensive evaluation against multiple types of attacks.

6.2.7 Classical Simulation Sufficiency

Classical simulation [70]–[73] using PyTorch and Cirq proves sufficient for a comprehensive evaluation of the HNN defense, allowing 120 experimental conditions that would be prohibitively expensive on quantum hardware. This validation establishes that meaningful defense research can proceed using conventional computing resources while quantum hardware matures. The 4-qubit quantum circuit simulation maintained computational tractability while providing sufficient quantum expressiveness to investigate hybrid quantum-classical security properties.

6.2.8 Persistent Accuracy Gap

Even with optimal defense—adversarial training achieving a mean defended accuracy of 58.5% average defended accuracy—a substantial 29.1 percentage point gap remains compared to clean accuracy (87.6%). This persistent degradation demonstrates fundamental limitations of current defense mechanisms and motivates continued research toward more effective protection strategies. The gap indicates that compound adversarial attacks introduce perturbations that current defenses cannot fully neutralize without sacrificing

clean accuracy, suggesting that fundamentally new defense paradigms may be necessary for high-stakes deployment.

6.3 Limitations

While this research provides a comprehensive evaluation of HNN defense mechanisms under 120 experimental conditions, several limitations constrain the scope and generalizability of the findings.

6.3.1 Direct Experimental Baseline Comparison

Chapter 5 (Section 5.8) establishes a literature-based proxy comparison that shows that HNN architectures achieve relative robustness gains of 13-18% over classical CNN baselines reported in the adversarial robustness research literature [15], [68], [69]. This proxy methodology compares HNN adversarial training recovery rates (73.9% in MNIST, 94.6% in TrafficSigns) against the classical CNN performance reported by Madry et al. [15] (40-45% typical recovery for PGD-defended MNIST) and test-time defense surveys [64], [65] (35-42% recovery for input transformations). While this comparison provides preliminary evidence that quantum circuit components confer measurable defensive advantages through geometric defenses in high-dimensional Hilbert space, the reliance on literature-reported classical CNN performance rather than direct head-to-head experimental comparison under identical conditions represents a significant limitation constraining the strength of comparative claims.

The proxy of the research literature cannot account for critical confounding factors: classical CNN benchmarks use different experimental conditions (full MNIST datasets vs filtered subsets in this research, single-method PGD attacks vs compound FGSM+PGD, FGSM+CW, and CW+PGD attacks, varying training procedures and model architectures). These methodological differences introduce uncertainty in the estimated 13-18% relative robustness gain. The HNN experiments—same filtered datasets (12,665 MNIST training samples, 2,000 TinyImageNet samples), the same compound attacks with identical parameters (ε=0.3, PGD iterations=40, CW iterations=1000), the same defense

implementations (identical adversarial training augmentation, input transformation parameters), the same training procedures (10-120 epochs, Adam optimizer, identical weight decay)—might achieve different performance than the literature-reported baselines, either narrowing or widening the observed gap.

Furthermore, the proxy of the research literature cannot isolate the specific contribution of the quantum circuit component versus other architectural differences. The HNN model contains approximately 50% classical layers and 50% quantum processing (4-qubit parameterized circuit). A fair comparison would require a classical CNN with equivalent total parameter count and computational complexity to determine whether the observed robustness advantages stem from the quantum component specifically or simply from the increased model capacity. Without controlling for total parameters, depth, and width, attributing the 13-18% robustness gain specifically to quantum processing remains speculative rather than definitively proven. Additionally, the observed 58.5% average defended accuracy for adversarial training in HNN models cannot be directly compared to the classical CNN performance without such controlled experiments using equivalent datasets, attack parameters, and defense configurations.

The decision to establish defense effectiveness within HNN architectures (Chapters 4-5) before conducting direct CNN comparisons reflects the research objective of determining whether defenses can protect hybrid quantum-classical models, independently of comparative performance claims. This fundamental question—can HNN models be defended against compound adversarial attacks?—required comprehensive evaluation (120 experimental conditions) establishing that adversarial training provides effective protection (58.5% defended accuracy vs 19.1% without defense). Having established this baseline and performed preliminary proxy comparison demonstrating potential quantum advantages, future work can now proceed to controlled HNN-versus-CNN comparisons with confidence that the HNN defense framework is sufficiently mature for comparative evaluation.

Future research work should conduct systematic direct experimental comparisons addressing these limitations: (1) train classical CNNs on the exact filtered datasets used in this research using identical training procedures, (2) generate compound adversarial

examples using identical attack parameters applied to both HNN and CNN test sets, (3) apply identical defense implementations to both architectures, (4) design classical CNN baselines with parameter counts matched to HNN total capacity, and (5) perform ablation analysis systematically removing the quantum circuit component and replacing it with classical fully connected layers of equivalent output dimensionality (16 outputs matching the 2^4 quantum measurement outputs). Such controlled experiments would definitively establish whether the 13-18% relative robustness gain observed in literature proxy comparison represents genuine quantum advantages, classical architectural benefits, or experimental artifacts from differing methodologies. Additionally, direct comparison would enable investigation of quantum-specific robustness mechanisms that literature proxy cannot assess: whether quantum measurement-induced collapse introduces stochasticity disrupting adversarial gradients, whether CNOT entanglement creates feature correlations unreplicatable by classical layers, and whether high-dimensional Hilbert space projection provides geometric advantages beyond classical nonlinearities.

6.3.2 Quantum Circuit Scope

The implementation of the quantum circuit is limited to 4 qubits with fixed rotation angles (ϑ=0.159π, φ=0.095π) using classical simulation [70]–[73] on conventional computing hardware. This limitation reflects practical constraints of classical quantum simulation—state vector representation scales exponentially with the number of qubits (2^n complex amplitudes for n qubits)—and establishes a baseline configuration to investigate the effectiveness of the defense. However, larger quantum circuits with trainable parameters or quantum hardware implementation may exhibit different adversarial robustness properties that cannot be inferred from 4-qubit simulations.

The use of classical simulation rather than physical quantum hardware introduces potential discrepancies between simulated and actual quantum behavior. Quantum hardware exhibits noise, decoherence, gate errors, and measurement imperfections that classical simulation does not capture. These hardware imperfections may either increase vulnerability to adversarial attacks (by introducing additional noise that amplifies perturbations) or decrease vulnerability (by acting as implicit randomization that disrupts

adversarial gradients). Validation on quantum hardware using IBM Quantum, Rigetti, or IonQ systems represents essential future work to confirm that classical simulation findings translate to physical quantum processors.

6.3.3 Dataset and Domain Scope

The evaluation encompasses four datasets (MNIST, EMNIST "Digits", TinyImageNet, TrafficSigns) selected to span varying complexity levels while maintaining computational tractability for comprehensive experimentation. This dataset selection provides sufficient diversity to identify general patterns—such as dataset-dependent defense effectiveness and adversarial training superiority—but does not encompass all domains where HNN models may be deployed. Medical imaging (e.g., chest X-rays, MRI scans, pathology images), autonomous driving scenarios (road scene understanding, pedestrian detection), multi-modal data (combining visual, textual, or sensor inputs), and few-shot learning contexts are not represented in the current evaluation.

The filtered dataset sizes—particularly the limited training samples for TinyImageNet (2,000) and TrafficSigns (1,200)—reflect constraints of classical quantum circuit simulation that prevent scaling to full-sized datasets. While these filtered datasets enable systematic evaluation across multiple experimental conditions, the findings may not generalize to scenarios with substantially larger training sets or different class distributions. The dataset filtering procedure (random selection maintaining class balance) ensures representative sampling but introduces potential selection bias that could affect defense effectiveness measurements.

6.3.4 Attack Strategy Scope

The compound adversarial attacks evaluated—FGSM+PGD, FGSM+CW, and CW+PGD—represent established white-box targeted methods that combine gradient-based perturbations with iterative optimization. This attack selection provides a comprehensive evaluation of common threat models but does not encompass all possible adversarial strategies. Black-box attacks (transfer-based, query-based), adaptive attacks (specifically designed to circum-

vent known defenses), generative attacks (GAN-based adversarial example generation), optimization-based attacks (DeepFool, EAD), and score-based attacks (JSMA, ZOO) are not evaluated in this research.

The exclusion of adaptive attacks represents a particularly critical limitation for evaluating test-time defenses. Athalye et al. [68] demonstrated that many defenses relying on non-differentiable transformations—such as JPEG compression, image quilting, bit-depth reduction, and randomization techniques evaluated in this research—create obfuscated gradients that provide a false sense of security. Adaptive adversaries can employ Backward Pass Differentiable Approximation (BPDA) [68], [69] to construct smooth, differentiable approximations of non-differentiable defense transformations, enabling gradient-based attacks to circumvent defenses that appeared effective against standard attacks.

In the context of this research, BPDA-based adaptive attacks pose a significant threat to the test-time defenses evaluated. An adaptive adversary aware of the input transformation or randomization mechanisms could approximate the gradient flow through these non-differentiable operations, potentially reducing or eliminating the modest robustness improvements observed (20.9% for input transformation, 25.5% for randomization). For instance, random cropping and resizing introduce discrete, non-differentiable operations that standard gradient-based attacks cannot directly optimize through. However, a BPDA-based adaptive attack could approximate these transformations as differentiable operations during the backward pass while maintaining the actual discrete transformations during the forward pass, enabling effective gradient estimation despite the non-differentiability.

The fixed quantum circuit component presents a potential barrier to BPDA approximation. The quantum rotation gates and CNOT entanglement structure create a high-dimensional nonlinear transformation in Hilbert space that may be difficult to approximate with simple differentiable surrogates. However, an adversary could potentially develop quantum-aware BPDA approximations that model the quantum circuit's input-output behavior using classical neural networks trained on input-output pairs, enabling gradient estimation through the quantum component. Such quantum-aware adaptive attacks represent an unexplored threat that could substantially reduce the defended ac-

curacy observed in this research.

Adversarial training, being a training-time defense that directly exposes the model to adversarial examples during training, is inherently less susceptible to BPDA-based circumvention compared to test-time defenses [68], [69]. The model learns robust features that generalize to adversarial perturbations regardless of the specific gradient computation method used during attack generation. This fundamental difference may partially explain the 2.3–2.8× superiority of adversarial training (58.5% defended accuracy) over test-time defenses (20.9-25.5%) observed in this research—adversarial training provides genuine robustness improvements rather than obfuscated gradients that sophisticated adversaries can bypass.

The observed 78.3% attack success rate against undefended models and 58.5% defended accuracy with adversarial training may represent optimistic estimates if adaptive attackers specifically design attacks to exploit defense mechanisms. Evaluation against BPDA-based adaptive attacks [68] and ensemble attacks combining multiple perturbation strategies would provide more realistic assessment of defense robustness in adversarial environments where attackers possess knowledge of defensive strategies and can adapt their attack methodologies accordingly. The development of quantum-aware adaptive attacks represents a critical future direction for validating whether the robustness improvements demonstrated in this research withstand sophisticated adversarial scrutiny.

Future work should incorporate adaptive attack evaluation as a standard component of defense assessment, following the methodology established by Athalye et al. [68] and Tramer et al. [69]. This evaluation should include BPDA approximations of all non-differentiable defense components (input transformations, randomization operations, quantum circuit measurements), expectation over transformation (EOT) attacks that account for stochastic defense behavior, and ensemble attacks combining multiple attack strategies. Additionally, quantum-specific countermeasures such as quantum noise injection during inference could potentially obfuscate gradient paths further, providing defense against BPDA approximations by introducing genuine quantum randomness that cannot be effectively approximated with classical differentiable surrogates.

6.3.5 Defense Mechanism Scope

The defense categories evaluated—input transformation, randomization, and adversarial training—represent fundamental approaches to adversarial robustness but do not encompass all possible defense strategies. Ensemble defenses (combining multiple models with diverse architectures), defensive distillation (training on softened probability distributions), certified defenses [74], [75] (providing provable robustness guarantees), detection-based approaches (identifying adversarial examples before classification), and feature denoising methods are not evaluated in this research. These advanced defenses may overcome the 29.1 percentage point gap between defended accuracy (58.5%) and clean accuracy (87.6%) observed with current mechanisms.

The adversarial training implementation uses a specific approach—combining FGSM, PGD, and CW examples in the augmented training set with fixed perturbation budgets (ε=0.3 for FGSM/PGD)—that represents one point in a large design space. Variants such as TRADES (balancing clean accuracy and robust accuracy through explicit trade-off parameter), MART (misclassification-aware adversarial training), robust self-training, and curriculum adversarial training may achieve different robustness-accuracy trade-offs that could improve upon the observed defended accuracy levels.

6.3.6 Evaluation Metrics Scope

The primary effectiveness metric employed is prediction accuracy, supplemented by robustness metrics (attack success rate, defense recovery, robustness improvement, defense rating). While accuracy provides direct measurement of defense effectiveness, it does not capture all dimensions of model robustness. Confidence calibration on adversarial examples, per-class vulnerability analysis, robustness curves plotting defended accuracy versus perturbation magnitude, computational cost analysis quantifying training time and inference latency, and semantic preservation of defended outputs are not systematically evaluated.

The accuracy-based evaluation assumes that all classification errors are equally consequential, which may not hold in safety-critical applications where certain misclassification

patterns (e.g., false negatives in medical diagnosis, missed stop signs in autonomous driving) carry greater risk than others. Application-specific evaluation frameworks incorporating domain knowledge about error severity, certification requirements, and operational constraints would provide a more nuanced assessment of defense suitability for practical deployment.

6.4 Lessons Learned

Throughout the research process, multiple insights emerged across methodological, technical, and conceptual dimensions that inform both the interpretation of results and recommendations for future HNN defense research.

6.4.1 Methodological Insights

A simple train-test split without validation data proved adequate for defense evaluation when combined with consistent experimental conditions and multiple datasets. The use of four datasets with varying characteristics—MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns—provided sufficient evidence for generalizing findings without requiring validation set tuning. This methodological choice simplified experimental design while enabling comprehensive evaluation across 120 conditions, demonstrating that sophisticated validation procedures may not be necessary when systematic experimentation spans diverse datasets and attack types.

Simulating quantum circuit components within HNNs using Cirq and PyTorch proved computationally feasible and informative despite limitations inherent in classical approximations. The 4-qubit parameterized quantum circuit with rotation gates ($R_y(\vartheta)$, $R_y(\varphi)$) and CNOT entanglement could be efficiently simulated for comprehensive defense evaluation across all experimental conditions. The classical simulation approach enabled rapid prototyping and extensive experimentation that would be prohibitively expensive on quantum hardware, validating the use of simulation for foundational defense research before transitioning to physical quantum processors.

Evaluation across four datasets with varying characteristics revealed defense effective-

ness patterns that would be invisible in single-dataset studies. The 95% variation in defended accuracy between TrafficSigns (37% average) and EMNIST "Digits" (19% average) demonstrates that dataset-dependent effects dominate over uniform defense properties. This finding establishes that multi-dataset evaluation is essential for understanding defense generalization limits and identifying dataset characteristics that influence robustness, rather than drawing broad conclusions from isolated dataset results.

Testing three compound attack combinations (FGSM+PGD, FGSM+CW, CW+PGD) proved essential for comprehensive defense evaluation, as some defenses showed 42 percentage point performance differences between attack types. Single-attack evaluation would have produced misleading conclusions about defense effectiveness by failing to capture attack-dependent vulnerability patterns. The variation in defense performance across attack types demonstrates that robust evaluation requires testing against multiple attack strategies representing different perturbation generation approaches (gradient-based, optimization-based, hybrid).

6.4.2 Technical Insights

Proper tuning of defense-specific parameters proved crucial for achieving optimal robustness without excessive clean accuracy degradation. For input transformation defenses, JPEG quality factor (75), bit-depth reduction levels (4 bits), Gaussian noise standard deviation (0.1), image quilting patch size (4×4), and combined transformation parameters required careful calibration to balance adversarial perturbation removal against preservation of legitimate image content. Aggressive parameter settings removed adversarial perturbations effectively but degraded clean accuracy unacceptably, while conservative settings maintained clean accuracy but provided insufficient defense.

Randomization strategies required balanced calibration to ensure robustness improvement without excessive clean accuracy loss. Random resizing scale factors (0.8-1.2), random cropping scale factors (0.8-1.0), and random rotation angles (±15 degrees) represented optimal ranges that maintained reasonable clean accuracy while providing defense against adversarial perturbations. These parameter ranges emerged through iterative tuning balancing the competing objectives of disrupting adversarial gradients (requiring

aggressive transformations) and preserving legitimate visual content (requiring conservative transformations).

Adversarial training effectiveness depended critically on generating diverse adversarial examples during training. The approach of combining FGSM (ε=0.3), PGD (ε=0.3, 40 iterations, α=0.01), and CW (c=1, $\varkappa$=0, 1000 iterations, learning rate=0.01) examples in the augmented training set provided superior robustness compared to single-method adversarial training. This finding demonstrates that exposure to diverse attack strategies during training enables the model to learn robust features that generalize across different adversarial perturbation types, rather than overfitting to specific attack characteristics.

Compound attacks combining complementary perturbation strategies proved more damaging than single-method attacks across all datasets. FGSM speed combined with PGD strength, and CW minimal perturbations combined with PGD iterative refinement, achieved higher attack success rates than individual methods used in isolation. This finding validates the importance of evaluating sophisticated threat models that combine multiple attack approaches, as adversaries in real-world scenarios are likely to employ hybrid strategies exploiting multiple vulnerabilities simultaneously.

Among test-time defenses, image quilting demonstrated the strongest performance on complex visual datasets, achieving 64% defended accuracy on TrafficSigns. This reconstruction-based approach—rebuilding adversarial images using patches from a clean database—effectively removes adversarial perturbations while preserving legitimate visual content when clean patches can be reliably matched. The high effectiveness on structured datasets (TrafficSigns) compared to complex natural images (TinyImageNet) suggests that reconstruction quality determines defense success, with simpler visual patterns enabling more accurate reconstruction.

Visual comparisons between clean, adversarial, and defended images provided meaningful insights alongside quantitative metrics. Qualitative assessment revealed that some defenses maintaining reasonable numerical accuracy produced semantically corrupted outputs—images that achieved correct classification but exhibited visible distortions, color shifts, or structural degradation. Conversely, some lower-accuracy defenses better preserved visual content quality despite classification errors. This finding emphasizes the

importance of multi-modal evaluation combining numerical accuracy measurements with qualitative visual assessment to identify defenses suitable for applications where output quality matters.

6.4.3 Conceptual Insights

The fundamental trade-off between deployment flexibility (test-time defenses) and robustness (training-time defense) emerged as a central theme with practical implications for HNN deployment. Organizations must choose between easily deployable but less effective defenses—input transformation achieving 20.9% defended accuracy, randomization achieving 25.5%—and computationally expensive but superior defenses—adversarial training achieving 58.5%. This trade-off has no universal solution; optimal choice depends on deployment constraints (retraining feasibility, computational budget), threat model severity (attack sophistication, attack frequency), and robustness requirements (acceptable accuracy degradation, certification standards).

Despite incorporating quantum circuit components, HNN models demonstrated vulnerability patterns similar to classical CNNs, suggesting that adversarial perturbations primarily exploit classical convolutional layers rather than quantum processing components. This finding implies that quantum circuits alone do not provide inherent adversarial robustness without explicit defense mechanisms. The similarity of vulnerability patterns indicates that defense strategies developed for classical neural networks—particularly adversarial training—transfer effectively to hybrid quantum-classical architectures, enabling leveraging of existing CNN security research for HNN protection.

The concept of acceptable defended accuracy varies by application domain, with important implications for defense selection and deployment planning. While 58.5% average defended accuracy represents significant improvement over undefended accuracy (19.1%), safety-critical applications may require greater than 95% robustness under certification standards such as ISO 26262 [76], [77] for automotive systems or DO-178C [78] for aviation software. The 29.1 percentage point gap between current defended accuracy (58.5%) and certification requirements [76]–[78] (95%) indicates that existing defenses remain inadequate for high-stakes deployment, motivating research into advanced defense mechanisms,

architectural innovations or domain-specific hardening approaches.

6.5 Future Research Work

Based on the contributions and insights from this research study, several promising directions emerge to advance the HNN defense research to practical deployment in safety-critical applications. These directions systematically build on the foundation established through comprehensive evaluation across 120 experimental conditions, using validated methodologies and established defense effectiveness patterns to guide future investigations.

6.5.1 Validation Through Real-World Applications

To evaluate whether the defense mechanisms developed in this dissertation research generalize beyond benchmark datasets to practical deployment scenarios, preliminary validation was conducted in a safety-critical traffic sign classification application (see Chapters 4 and 5 for detailed results). This preliminary work, conducted by the author during the dissertation research period, applies the compound attack framework (FGSM+PGD, FGSM+CW, CW+PGD) to retroreflectivity-based sign maintenance prioritization.

The preliminary validation confirms key research findings: adversarial training achieves superior defense effectiveness (70.98% defended accuracy) consistent with the 2.3–2.8× advantage over test-time defenses observed in Chapter 5, compound attacks prove highly damaging (80.5% attack success rate comparable to the 73-84% range in benchmark experiments), and HNN architectures achieve higher clean accuracy (95.98%) than classical CNNs (91.52%) on this task. However, the safety-critical application introduces additional challenges, including class imbalance (86.9% safe versus 13.1% unsafe), synthetic data generation requirements and life-safety classification stakes requiring ISO 26262 and MUTCD compliance.

The convergence of the findings between the controlled benchmark evaluation (Chapters 4-5) and the preliminary practical application (Chapter 5) provides initial evidence that adversarial training provides consistent protection across diverse contexts. However,

this preliminary validation represents proof-of-concept work conducted by the author and has not undergone an external peer review. The core contributions of this dissertation are independent based on the comprehensive evaluation of 120 experimental conditions across four benchmark datasets, with the preliminary safety-critical validation providing supplementary context for potential real-world applicability rather than definitive proof of operational readiness.

A critical question emerges from both the benchmark and the preliminary validation results: do the defended accuracy levels achieved (58.5% average on the benchmarks, 70.98% on the TrafficSigns) meet safety-critical certification thresholds requiring greater than 95% robustness under ISO 26262 and DO-178C standards? The persistent gap between current performance and certification requirements indicates that enhanced defense strategies—potentially combining adversarial training with ensemble defenses, certified robustness guarantees, or detection-based rejection of suspicious inputs—may be necessary for certification compliance. Future work should investigate whether multi-layered defense architectures can bridge the gap between observed performance and the 95%+ robustness required for safety-critical deployment in autonomous vehicles, medical systems, and other high-stakes applications.

6.5.2 Extended Attack Surface Evaluation

The compound attacks evaluated in this research study (FGSM+PGD, FGSM+CW, and CW+PGD) represent established white-box targeted methods. Future research should evaluate HNN robustness against additional sophisticated attacks to determine whether the 2.3–2.8× adversarial training advantage observed against FGSM/PGD/CW compounds generalizes to fundamentally different attack strategies.

6.5.2.1 Adaptive Attacks

The most critical priority for future evaluation involves testing against BPDA-based adaptive attacks [68], [69] that approximate non-differentiable defense transformations with differentiable surrogates. This evaluation should systematically assess whether the test-

time defenses evaluated (input transformation: 20.9%, randomization: 25.5%) represent genuine robustness or obfuscated gradients that sophisticated adversaries can bypass. The evaluation should include: (1) BPDA approximations of JPEG compression, image quilting, bit-depth reduction, and randomization operations, (2) Expectation Over Transformation (EOT) attacks that account for stochastic defense behavior by averaging gradients over multiple random transformations, (3) quantum-aware BPDA approximations that model the quantum circuit's input-output behavior using trained classical neural networks, and (4) ensemble attacks combining multiple adaptive strategies.

The development of quantum-specific countermeasures against BPDA represents a promising research direction. Quantum noise injection during inference—introducing genuine quantum randomness through quantum measurement or noisy quantum gates—could potentially create defense mechanisms that resist the BPDA approximation because the stochasticity stems from fundamental quantum mechanics rather than classical pseudo-randomness. Such quantum-native defenses would exploit unique properties of quantum computing to provide security guarantees that are unachievable with purely classical approaches.

6.5.2.2 Additional Attack Strategies

Beyond adaptive attacks, comprehensive evaluation should include optimization-based attacks such as DeepFool (minimal perturbation to decision boundary), EAD (elastic-net regularization), and score-based attacks, including JSMA (Jacobian-based saliency map) and ZOO (zeroth-order optimization for black-box scenarios). Black-box approaches that leverage transfer-based attacks from surrogate models or query-based attacks with limited model access represent realistic threat models where attackers cannot directly access the model internals. Generative attacks using GAN-based adversarial example generation may produce realistic perturbations that evade detection-based defenses by generating adversarial examples indistinguishable from legitimate inputs.

6.5.3 Advanced Defense Mechanisms

Beyond the input transformation, randomization, and adversarial training approaches evaluated in this research study, several promising defense directions merit investigation to potentially overcome the 29.1 percentage point gap between the defended accuracy (58.5%) and the clean accuracy (87.6%). Ensemble defense strategies combining multiple models with diverse architectures or training procedures may provide robustness through disagreement, where adversarial examples that fool individual models fail to fool the ensemble majority. Defensive distillation training models on softened probability distributions may smooth decision boundaries and reduce gradient-based attack effectiveness.

Certified defenses [74], [75] using randomized smoothing or interval bound propagation could provide provable robustness guarantees within specified perturbation radii, addressing the uncertainty inherent in empirical defense evaluation and potentially meeting certification requirements for safety-critical applications. Detection-based approaches that identify adversarial examples before classification using autoencoders, statistical tests, or neural network uncertainty quantification would allow the rejection of suspicious input rather than the attempt to classify. Feature denoising, applying denoising autoencoders, non-local means filtering, or learned denoising networks may remove adversarial perturbations while preserving a legitimate signal.

Adversarial training variants including, TRADES (balancing clean accuracy and robust accuracy through explicit trade-off parameter), MART (misclassification-aware training), robust self-training, and curriculum adversarial training may achieve different robustness-accuracy trade-offs that improve upon the 58.5% defended accuracy observed with standard adversarial training. Systematic evaluation of these advanced defenses using the 120-condition experimental framework established in this research study would reveal whether fundamental improvements beyond current mechanisms are achievable or whether the observed 29.1 percentage point gap represents a theoretical limit.

6.5.4 Quantum Architecture and Hardware Investigation

This research study evaluated HNN models with 4-qubit parameterized quantum circuits using classical simulation. Several architectural directions could enhance robustness or provide a deeper understanding of quantum contributions to adversarial security. Deploying defense-hardened HNN models [70]–[73] on actual quantum processors—IBM Quantum, Rigetti, IonQ platforms—would validate that classical simulation findings translate to physical quantum systems and identify hardware-specific vulnerabilities or advantages introduced by quantum noise, decoherence, and gate errors.

Scaling beyond 4 qubits would investigate whether increased quantum dimensionality provides inherent robustness benefits through higher-dimensional quantum state spaces or introduces new attack surfaces through increased circuit complexity. Variational quantum circuits [70], [79] exploring alternative architectures—hardware-efficient ansätze, problem-specific circuits, quantum convolutional layers—would determine if circuit design influences adversarial robustness independently of defense mechanisms. Alternative hybrid architectures investigating different integration points between classical and quantum components—quantum feature maps at input, quantum layers at intermediate stages, and quantum output layers—may identify architectures with superior inherent robustness.

Quantum feature maps [71], [79], [80] evaluating whether quantum embedding strategies (amplitude encoding, angle encoding, basis encoding) influence susceptibility to adversarial perturbations propagating through quantum circuits would reveal architectural choices that enhance security. Theoretical analysis [70], [71], [73], [79] of perturbation propagation through quantum gates may explain whether quantum processing amplifies adversarial signals (increasing vulnerability), attenuates perturbations (providing defense), or remains neutral (with security determined by classical layers).

6.5.5 Dataset and Domain Expansion

This research study evaluated defenses in MNIST, EMNIST "Digits", TinyImageNet, and TrafficSigns. Expanding to additional datasets would reveal defense generalization limits and domain-specific effectiveness patterns beyond current findings. Standard benchmarks,

including CIFAR-10, CIFAR-100, and ImageNet, would evaluate whether the findings scale to larger, more diverse datasets with increased class counts and sample sizes that exceed the constraints of the filtered dataset.

Medical imaging applications that use chest radiographs, magnetic resonance imaging, or pathological images represent domains where adversarial attacks could cause life-threatening misdiagnoses, and defense requirements exceed typical benchmark standards. Autonomous driving scenarios that involve understanding the road scene, pedestrian detection, and recognition of traffic signs under varying environmental conditions represent applications where robustness failures endanger human safety. Multi-modal data combining visual, textual, or sensor input would investigate whether HNN defenses generalize across modalities or require modality-specific adaptation.

Few-shot learning contexts evaluating defense effectiveness when training data is severely limited would reflect real-world scenarios where collecting large adversarially-augmented datasets is impractical. Domain expansion would validate whether the dataset-dependent effectiveness patterns observed in this research study (37% TrafficSigns versus 19% EMNIST "Digits") represent generalizable principles correlating with specific dataset characteristics (visual complexity, structural regularity, intra-class variation) or dataset-specific artifacts that do not transfer across domains.

6.5.6 Enhanced Evaluation Framework

This research study mainly used prediction accuracy as the effectiveness metric supplemented by robustness metrics. Enhanced evaluation frameworks could provide deeper insight into defense behavior and guide practical deployment decisions. Robustness curves that plot the accuracy of the defended versus the magnitude of the perturbation (ε) would characterize defense degradation patterns and identify critical failure thresholds where the defenses collapse. Adversarial calibration measuring confidence calibration on adversarial examples would determine whether models provide reliable uncertainty estimates under attack, enabling detection of distribution shift.

A Per-class analysis that evaluates defense effectiveness separately for each class would identify systematically vulnerable categories that require targeted protection or reveal

attack-type interactions with specific visual features. Computational cost analysis quantifying training time, inference latency, and memory requirements for each defense would inform practical deployment trade-offs between robustness and operational efficiency. Ablation studies systematically removing defense components—adversarial training diversity, randomization range, and transformation strength—would identify critical mechanisms driving effectiveness and eliminate unnecessary components.

Enhanced metrics would complement the foundational accuracy measurements established in this research study with a nuanced understanding of defense behavior under operating conditions, allowing informed deployment decisions that balance robustness requirements, computational constraints, and application-specific priorities.

6.5.7 Theoretical Understanding Development

This research study provided an empirical evaluation of the effectiveness of the defense. Theoretical investigation could explain observed patterns and guide future defense design toward principled approaches. Mathematical analysis [70], [71], [73], [79] of perturbation propagation through quantum rotation gates, entanglement operations, and measurement would reveal whether quantum processing amplifies adversarial signals (explaining observed vulnerabilities), attenuates perturbations (providing inherent defense), or remains neutral, with security determined by classical layers.

Deriving theoretical robustness bounds, establishing provable limits on defended accuracy for HNN architectures under worst-case adversarial perturbations would reveal whether observed performance (58.5% defended accuracy) approaches theoretical optimality or whether substantial improvement remains possible.

Investigating [70], [71], [73], [80] whether quantum computing provides fundamental advantages for adversarial robustness beyond classical approaches—through quantum parallelism, superposition, or entanglement properties—would determine whether quantum components offer security benefits justifying architectural complexity.

Theoretical understanding would explain why adversarial training achieves 2.3–2.8$\times$ superiority over test-time defenses, whether this advantage represents fundamental limits or could be overcome with improved test-time approaches, and whether quantum circuits

provide inherent robustness properties. This understanding would guide defense design toward principled mechanisms that exploit fundamental properties rather than heuristic approaches validated only empirically.

6.6 Contributions and Implications

This research study makes several foundational contributions to HNN security research with practical implications for the deployment of hybrid quantum-classical models in adversarial environments.

The comprehensive defense evaluation under 120 experimental conditions demonstrates that defense mechanisms against compound WTC adversarial attacks can effectively protect HNN models, with adversarial training achieving a 58.5% average defended accuracy compared to 19.1% without defense—a 3× improvement. However, substantial accuracy loss persists (87.6% clean versus 58.5% defended), indicating that defending against compound attacks remains challenging and motivating continued research. This finding establishes realistic expectations for HNN deployment: defenses provide measurable protection but cannot yet achieve the 95%+ robustness required for safety-critical certification.

Identification of adversarial training as the optimal defense mechanism across all datasets and attack types provides actionable guidance for the deployment of HNN. Achieving 2.3–2.8× higher defended accuracy than test-time defenses (randomization 25.5%, input transformation 20.9%) establishes adversarial training as the current state-of-the-art for HNN protection. Organizations deploying HNN models in adversarial environments should prioritize adversarial training despite computational costs, reserving test-time defenses for scenarios where retraining is impractical or threat models change rapidly.

The systematic comparison of test-time defenses (applied during inference without retraining) versus training-time defenses (requiring model retraining) establishes fundamental trade-offs between deployment flexibility and robustness. While test-time defenses offer easy deployment and adaptability, training-time defenses provide substantially su-

perior protection. This framework enables informed deployment decisions that balance operational constraints with security requirements, with a clear understanding of robustness sacrifices when choosing deployment flexibility over optimal security.

The demonstration of dataset-dependent defense effectiveness—simpler datasets (TrafficSigns: 87% adversarial training) benefiting substantially more than complex datasets (EMNIST "Digits": 28% adversarial training)—suggests that defense selection should consider dataset characteristics. Visual complexity, intra-class variation, and structural regularity influence defense performance, with implications for deployment planning: applications using structured visual data with low intra-class variation achieve higher defended accuracy, while applications using complex, diverse data require enhanced defenses or acceptance of lower robustness levels.

Validation that hybrid quantum-classical architectures share vulnerabilities with classical CNN but can be defended using adapted classical defense strategies establishes that existing CNN security research transfers to HNN protection. This finding accelerates HNN security research by enabling the exploitation of established defense mechanisms rather than the development of quantum-specific defenses from scratch, while revealing that quantum circuit components do not provide inherent adversarial robustness without explicit defense mechanisms.

Identification of adaptive attacks as a critical limitation establishes the need for defenses resistant to BPDA in future HNN security research. The potential vulnerability of test-time defenses to obfuscated gradient attacks [68], [69] highlights the importance of adversarial training's inherent resistance to such circumvention, providing additional justification for prioritizing training-time defenses despite computational costs. The development of quantum-aware adaptive attacks and quantum-native countermeasures represents an essential frontier for validating defense effectiveness under sophisticated adversarial scrutiny.

6.7 Summary

This chapter synthesized the research findings, articulated limitations, reflected on the lessons learned, identified future research directions, and established the practical contributions of this research study to research on HNN security. The research investigation of defense mechanisms against compound adversarial attacks across 120 experimental conditions establishes a foundational understanding of the security of the HNN model.

Key findings demonstrate that HNN models exhibit severe vulnerability to compound attacks (78.3% attack success rate), but defense mechanisms provide measurable protection with adversarial training achieving 58.5% defended accuracy—2.3–2.8× superior to defenses at test-time. However, a persistent 29.1 percentage point gap between defended accuracy and clean performance indicates that current defenses remain inadequate for safety-critical applications that require certification-level robustness (95%+). Defense effectiveness varies dramatically by the complexity of the dataset (37% TrafficSigns versus 19% EMNIST "Digits"), establishing that deployment planning must consider the characteristics of the data set alongside threat models.

This research establishes clear trade-offs between test-time defenses that offer deployment flexibility (20.9-25.5% defended accuracy) and training-time defenses that provide superior robustness (58.5%) at computational cost. This framework enables informed deployment decisions that balance operational constraints with security requirements. Classical simulation proves sufficient for comprehensive defense evaluation, enabling extensive experimentation that would be prohibitively expensive on quantum hardware while quantum computing capabilities mature.

Limitations include the absence of classical CNN baseline comparison (constraining conclusions about quantum component contributions to robustness), 4-qubit quantum circuit scope (limiting generalization to larger quantum systems), filtered dataset sizes (reflecting classical simulation constraints), compound attack focus excluding adaptive attacks (particularly BPDA-based circumvention of test-time defenses), and accuracy-based evaluation metrics (not capturing all robustness dimensions). The critical limitation of adaptive attack exclusion establishes the BPDA evaluation as essential future

work to determine whether observed robustness represents genuine defense effectiveness or obfuscated gradients vulnerable to sophisticated circumvention [68], [69].

Lessons learned span methodological insights (multi-dataset evaluation essentiality, classical simulation feasibility), technical insights (defense parameter tuning criticality, attack diversity importance), and conceptual insights (deployment trade-offs, application-dependent acceptable accuracy). These insights inform both the interpretation of the current results and the design of future HNN defense research, establishing best practices for rigorous evaluation and practical deployment.

Future work directions build systematically on this foundation: validation through real-world applications confirms defense transferability to safety-critical contexts, extended attack surface evaluation with BPDA-based adaptive attacks tests defense authenticity, advanced defense mechanisms pursue certification-level robustness, quantum hardware investigation validates classical simulation findings, dataset expansion reveals domain-specific patterns, enhanced evaluation frameworks provide operational guidance, and theoretical understanding explains empirical observations. The development of quantum-native countermeasures against adaptive attacks—such as quantum noise injection that provides BPDA-resistant stochasticity—represents a particularly promising direction that exploits unique quantum properties for security advantages.

Contributions establish that adversarial training provides effective but imperfect protection for HNN models, with a clear understanding of deployment trade-offs, dataset dependencies, adaptive attack vulnerabilities, and remaining challenges. As quantum computing capabilities mature and hybrid quantum-classical architectures transition from research prototypes to practical systems, the defense mechanisms and evaluation frameworks established in this research study provide essential guidance for securing these models against adversarial threats while identifying clear paths toward achieving certification-level robustness required for safety-critical deployment. Recognizing that test-time defenses may provide false security against adaptive adversaries [68] reinforces the importance of adversarial training as the foundational defense mechanism for deployment in hostile environments.

References

[1] P. B. Upama, M. J. H. Faruk, M. Nazim, *et al.*, "Evolution of quantum computing: A systematic survey on the use of quantum computing tools," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022, pp. 520–529. DOI: 10.1109/COMPSAC54236.2022.00096.

[2] IBM, *Ibm unveils 400 qubit-plus quantum processor and next-generation ibm quantum system two*, https://newsroom.ibm.com/2022-11-09-IBM-Unveils-400-Qubit-Plus-Quantum-Processor-and-Next-Generation-IBM-Quantum-System-Two, IBM Quantum Summit 2022, Nov. 2022.

[3] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley, "Robust in practice: Adversarial attacks on quantum machine learning," *Physical Review A*, vol. 103, no. 4, p. 042 427, 2021.

[4] A. Faraone and R. Delgado-Gonzalo, "Convolutional-recurrent neural networks on low-power wearable platforms for cardiac arrhythmia detection," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2020, pp. 153–157. DOI: 10.1109/AICAS48895.2020.9073950.

[5] T. Nguyen, I. Paik, H. Sagawa, and T. C. Thang, "Quantum machine learning with quantum image representations," in *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 2022, pp. 851–854. DOI: 10.1109/QCE53715.2022.00142.

[6] N. Wang, Y. Chen, Y. Xiao, Y. Hu, W. Lou, and T. Hou, "Manda: On adversarial example detection for network intrusion detection system," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2022. DOI: 10.1109/TDSC.2022.3148990.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[8] P. Kuppusamy, N. Yaswanth Kumar, J. Dontireddy, and C. Iwendi, "Quantum computing and quantum machine learning classification – a survey," in *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, 2022, pp. 200–204. DOI: 10.1109/ICCCMLA56841.2022.9989137.

[9] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222 310–222 354, 2020. DOI: 10.1109/ACCESS.2020.3041951.

[10] D. Edwards and D. B. Rawat, "Quantum adversarial machine learning: Status, challenges and perspectives," in *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2020, pp. 128–133. DOI: 10.1109/TPS-ISA50397.2020.00026.

[11] X. Pan, H. Yang, Z. Xu, and Z. Zhu, "Adversarial analysis of ml-based anomaly detection in multi-layer network automation," *Journal of Lightwave Technology*, vol. 40, no. 15, pp. 4934–4944, 2022. DOI: 10.1109/JLT.2022.3172523.

[12] K. Sadeghi, A. Banerjee, and S. K. S. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 450–467, 2020. DOI: 10.1109/TETCI.2020.2968933.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, arXiv:1412.6572, 2015.

[14] Y. Liu, S. Mao, X. Mei, T. Yang, and X. Zhao, "Sensitivity of adversarial perturbation in fast gradient sign method," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 433–436. DOI: 10.1109/SSCI44817.2019.9002856.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, arXiv:1706.06083, 2018.

[16] M. S. Ayas, S. Ayas, and S. M. Djouadi, "Projected gradient descent adversarial attack and its defense on a fault diagnosis system," in *2022 45th International Conference on Telecommunications and Signal Processing (TSP)*, 2022, pp. 36–39. DOI: 10.1109/TSP55681.2022.9851334.

[17] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. DOI: 10.1109/SP.2017.49.

[18] X. Mao, Y. Chen, S. Wang, H. Su, Y. He, and H. Xue, "Composite adversarial attacks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8884–8892, 2021. DOI: 10.1609/aaai.v35i10.17075.

[19] Y. Zhu, X. Wei, and Y. Zhu, "Efficient adversarial defense without adversarial training: A batch normalization approach," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533949.

[20] M. Ashrafiamiri, S. M. P. Dinakarrao, A. H. A. Zargari, M. Seo, F. Kurdahi, and H. Homayoun, "R2ad: Randomization and reconstructor-based adversarial defense for deep neural networks," in *2020 ACM/IEEE 2nd Workshop on Machine Learning for CAD (MLCAD)*, 2020, pp. 21–26. DOI: 10.1145/3380446.3430628.

[21] Y. Li, X. Yu, S. Yu, and B. Chen, "Adversarial training for the adversarial robustness of eeg-based brain-computer interfaces," in *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, 2022, pp. 1–6. DOI: 10.1109/MLSP55214.2022.9943479.

[22] Y. Wang, C. Chen, and W. Huang, "Design of quantum filter for hybrid quantum-classical convolutional neural networks," in *2021 International Conference on Information Technology and Biomedical Engineering (ICITBE)*, 2021, pp. 66–70. DOI: 10.1109/ICITBE54178.2021.00024.

[23] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, arXiv:1312.6199, 2014.

[24] J. C. Costa, N. C. Silva, A. Ferreira, and M. Antunes, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, 2024, arXiv:2305.10862.

[25] Y. Wang, T. Sun, S. Li, *et al.*, "Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245–2298, 2023. DOI: 10.1109/COMST.2023.3319492.

[26] A. Peruzzo, J. McClean, P. Shadbolt, *et al.*, "A variational eigenvalue solver on a photonic quantum processor," *Nature Communications*, vol. 5, no. 1, p. 4213, 2014. DOI: 10.1038/ncomms5213.

[27] J.-J. Zhang and D. Meng, "Quantum-inspired analysis of neural network vulnerabilities: The role of conjugate variables in system attacks," *National Science Review*, vol. 11, no. 9, nwae141, 2024. DOI: 10.1093/nsr/nwae141.

[28] C.-C. Huang and S.-X. Zhang, "Enhancing adversarial robustness of quantum neural networks by adding noise layers," *New Journal of Physics*, vol. 25, no. 8, p. 083 019, 2023.

[29] C. Ferrari, S. Bertozzi, A. Berti, and R. Cucchiara, "(compress and restore) n: A robust defense against adversarial attacks on image classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 1s, pp. 1–19, 2023. DOI: 10.1145/3524619.

[30] X. Zhang, D. Cao, R. Li, J. Li, and Q. Wu, "Adaptive patch transformation for adversarial defense," *Computers & Security*, vol. 151, p. 104 221, 2025. DOI: 10.1016/j.cose.2025.104221.

[31] B. Li and Y. Zhao, "Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data," *arXiv preprint arXiv:2410.08503*, 2025.

[32] C. Eleftheriadis, A. Symeonidis, and P. Katsaros, "Adversarial robustness improvement for deep neural networks," *Machine Vision and Applications*, vol. 35, no. 35, 2024. DOI: 10.1007/s00138-024-01519-1.

[33] W. El Maouaki, A. Marchisio, T. Said, M. Shafique, and M. Bennai, "Robqunns: A methodology for robust quanvolutional neural networks against adversarial attacks," in *2024 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*, 2024, pp. 4090–4095. DOI: 10.1109/ICIPCW64161.2024.10769105.

[34] W. E. Maouaki, A. Marchisio, T. Said, M. Bennai, and M. Shafique, "Advqunn: A methodology for analyzing the adversarial robustness of quanvolutional neural networks," in *2024 IEEE International Conference on Quantum Software (QSW)*, 2024, pp. 175–181. DOI: 10.1109/QSW62656.2024.00033.

[35] J. Guo, W. Jiang, R. Zhang, W. Fan, J. Li, and G. Lu, *Backdoor attacks against hybrid classical-quantum neural networks*, 2024. arXiv: 2407.16273 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2407.16273.

[36] A. Liu, C. Wen, and J. Wang, "Lean classical-quantum hybrid neural network model for image classification," *Advanced Quantum Technologies*, Apr. 2025, ISSN: 2511-9044. DOI: 10.1002/qute.202400703. [Online]. Available: http://dx.doi.org/10.1002/qute.202400703.

[37] Y. Yongxi, Z. Shibin, Y. Lili, and C. Yan, "Hybrid classical quantum neural network with high adversarial robustness," in *Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing*, Z. Nianyin and R. B. Pachori, Eds., ser. Proceedings of Machine Learning Research, vol. 245, PMLR, Apr. 2024, pp. 271–279. [Online]. Available: https://proceedings.mlr.press/v245/yongxi24a.html.

[38] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[39] L. Hsiung, Y.-Y. Tsai, P.-Y. Chen, and T.-Y. Ho, "Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic pertur-

bations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 658–24 667.

[40] R. Fontana, A. Molena, L. Pegoraro, K. K. F. Tsai, and E. C. Martos, "Design of experiments and machine learning with application to industrial experiments," *Statistical Papers*, vol. 64, pp. 1251–1274, 2023. DOI: 10.1007/s00362-023-01437-w.

[41] S. Valentin, S. Kleinegesse, N. R. Bramley, C. G. Lucas, and C. L. Buckley, "Designing optimal behavioral experiments using machine learning," *eLife*, vol. 13, e86224, 2024. DOI: 10.7554/eLife.86224.

[42] R. Liu, X. Yang, C. Xu, L. Wei, and X. Zeng, "Comparative study of convolutional neural network and conventional machine learning methods for landslide susceptibility mapping," *Remote Sensing*, vol. 14, no. 2, p. 321, 2022. DOI: 10.3390/rs14020321.

[43] T. Bartz-Beielstein, "Hyperparameter tuning and optimization applications," in *Hyperparameter Tuning for Machine and Deep Learning with R*, E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, Eds., Singapore: Springer, 2023. DOI: 10.1007/978-981-19-5170-1_6.

[44] G. Franchini, V. Ruggiero, F. Porta, and I. Trombini, "GreenNAS: A green approach to the hyperparameters tuning in deep learning," *Mathematics*, vol. 12, no. 6, p. 850, 2024. DOI: 10.3390/math12060850.

[45] M. Islam, K. M. Hasan Mahmud, M. M. A. Ashik, *et al.*, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Multimedia Tools and Applications*, 2024. DOI: 10.1007/s11042-024-19355-2.

[46] L. Franceschi, M. Donini, V. Perrone, *et al.*, *Hyperparameter optimization in machine learning*, arXiv:2410.22854, 2024. arXiv: 2410.22854 [cs.LG].

[47] R. Wang, "Active learning-based optimization of scientific experimental design," in *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, IEEE, 2021. DOI: 10.1109/icaice54393.2021.00060.

[48] M. Jung, J. Kim, J. Y. Jang, H. Jung, and S. Shin, "A performance comparison among different amounts of context on deep learning based intent classification models," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2020. DOI: 10.1109/ictc49870.2020.9289467.

[49] M. H. Hashemi, U. Kılıç, and S. Dikmen, "Application of feature selection in distribution transformer design and manufacturing using feed forward artificial neural network and equilibrium optimizer algorithm," in *2023 5th Global Power, En-*

ergy and Communication Conference (GPECOM), IEEE, 2023. DOI: 10.1109/gpecom58364.2023.10175790.

[50] J. Yoo, B. Min, S. Kim, D. Shin, and D. Shin, "Study on network intrusion detection method using discrete pre-processing method and convolution neural network," *IEEE Access*, vol. 9, pp. 142 348–142 361, 2021. DOI: 10.1109/access.2021.3120839.

[51] A. Khurana and O. P. Verma, "A fine tuned model of grasshopper optimization algorithm with classifiers for optimal text classification," in *2020 IEEE 17th India Council International Conference (INDICON)*, IEEE, 2020. DOI: 10.1109/indicon49873.2020.9342432.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[53] I. Salehin and D.-K. Kang, "A review on dropout regularization approaches for deep neural networks within the scholarly domain," *Electronics*, vol. 12, no. 14, p. 3106, 2023. DOI: 10.3390/electronics12143106.

[54] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, "Quantum noise protects quantum classifiers against adversaries," *Physical Review Research*, vol. 3, p. 023 153, 2021. DOI: 10.1103/PhysRevResearch.3.023153.

[55] X. Liu, Q. Zhao, C. Wang, and L. Zhang, "Advancements in adversarial example defense for deep learning models: A review," *Cybersecurity*, 2025. DOI: 10.1186/s42400-025-00546-3.

[56] S. Song, Y. Hou, and G. Liu, "The interpretability of quantum-inspired neural network," in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, 2021. DOI: 10.1109/icaibd51990.2021.9459009.

[57] M. Abadi, P. Barham, J. Chen, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA: USENIX Association, 2016, pp. 265–283, ISBN: 978-1-931971-33-1.

[58] Y. Wang, B. Feng, X. Peng, and Y. Ding, "An efficient quantitative approach for optimizing convolutional neural networks," in *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, ACM, Oct. 2021. DOI: 10.1145/3459637.3482230.

[59] J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel, "Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in

pytorch, tensorflow, and jax," *Journal of Open Source Software*, vol. 5, no. 53, p. 2607, 2020. DOI: 10.21105/joss.02607.

[60] W. Zhang, Y. Chen, X. Liu, and J. Wang, "Universal attention guided adversarial defense using feature pyramid and non-local mechanisms," *Scientific Reports*, vol. 15, 2025. DOI: 10.1038/s41598-025-89267-8.

[61] J. Berberich, D. Fehrle, C. Muller, and K. Wiesner, "Quantum neural networks under depolarization noise: Exploring white-box attacks and defenses," *Quantum Machine Intelligence*, vol. 6, no. 2, 2024. DOI: 10.1007/s42484-024-00208-6.

[62] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009. DOI: 10.1016/j.cosrev.2009.03.005.

[63] M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang, *Adversarial training: A survey*, 2024. arXiv: 2410.15042 [cs.LG].

[64] A. Kulkarni and T.-W. Weng, "Interpretability-guided test-time adversarial defense," in *European Conference on Computer Vision (ECCV)*, Springer, 2024, pp. 466–483. DOI: 10.1007/978-3-031-72913-3_26.

[65] C.-H. Yeh, K. Yu, and C.-S. Lu, "Test-time adversarial defense with opposite adversarial path and high attack time cost," *arXiv preprint arXiv:2410.16805*, 2024.

[66] D. Khachaturov and D. Londt, "Complexity matters: Effective dimensionality as a measure for adversarial robustness," *arXiv preprint arXiv:2410.18556*, 2024.

[67] Y. Xiao, X. Wang, R. Li, and B. Dong, "A survey of robustness and safety of 2d and 3d deep learning models against adversarial attacks," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–37, 2024. DOI: 10.1145/3636551.

[68] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning (ICML)*, PMLR, 2018, pp. 274–283.

[69] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1633–1645.

[70] N. Dowling, M. T. West, A. Southwell, *et al.*, "Adversarial robustness guarantees for quantum classifiers," *arXiv preprint arXiv:2405.10360*, 2024.

[71] M. T. West, S.-L. Tsang, J. S. Low, *et al.*, "Towards quantum enhanced adversarial robustness in machine learning," *Nature Machine Intelligence*, vol. 5, no. 6, pp. 581–589, 2023. DOI: 10.1038/s42256-023-00661-1.

[72] M. T. West, A. C. Nakhl, J. Heredge, *et al.*, "Drastic circuit depth reductions with preserved adversarial robustness by approximate encoding for quantum machine learning," *Intelligent Computing*, vol. 3, 2024. DOI: 10.34133/icomputing.0100.

[73] M. T. West, S. M. Erfani, C. Leckie, M. Sevior, L. C. L. Hollenberg, and M. Usman, "Benchmarking adversarially robust quantum machine learning at scale," vol. 5, American Physical Society, 2023, p. 023 186. DOI: 10.1103/PhysRevResearch.5.023186.

[74] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *IEEE Symposium on Security and Privacy (S&P)*, IEEE, 2019, pp. 656–672.

[75] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 1310–1320.

[76] International Organization for Standardization, *Iso 26262: Road vehicles – functional safety*, Standard for functional safety of electrical and electronic systems in production automobiles, 2018.

[77] International Organization for Standardization, "Iso/pas 8800:2024 – road vehicles: Safety and artificial intelligence," 2024, Standard for AI/ML-based safety systems in automotive applications.

[78] RTCA Inc., *Do-178c: Software considerations in airborne systems and equipment certification*, Aerospace standard for software development assurance, 2011.

[79] M. Wendlinger, K. Tscharke, N. Klöck, and P. Debus, "Training robust and generalizable quantum models," *Physical Review Research*, vol. 6, p. 043 326, 2024. DOI: 10.1103/PhysRevResearch.6.043326.

[80] M. Wendlinger, K. Tscharke, and P. Debus, "A comparative analysis of adversarial robustness for quantum and classical machine learning models," *arXiv preprint arXiv:2404.16154*, 2024.

APPENDIX A: ADVERSARIAL ATTACK PARAMETERS

The WTC adversarial attacks include combination of FGSM and PGD, combination of FGSM and CW, and combination of PGD and CW (see Table A.1).

Table A.1: WTC adversarial attack parameters.

No.	Properties	Parameter
1	Input	Original Image, Label
2	Model to Attack	HNN model
3	Attacks	FGSM or CW or PGD
4	Output	Tampered Image, Label
Torchattacks PyTorch Usage Example		distinct-attack-1 = torchattacks.FGSM(model, eps=8/255) distinct-attack-2= torchattacks.CW(model, c=1, kappa=0, steps=50, lr=0.01) compounded-attack = torchattacks.MultiAttack ([distinct-attack-1, distinct-attack-2]) tampered-images = compounded-attack (original-images, labels)

The FGSM attack with the specific attack properties include input, model to attack, distance measure, maximum perturbation, and output. The input parameter is the original image and associated label. The model to attack parameter is the HNN model. The distance measure parameter is L infinity. The maximum perturbation parameter is 8/255. The output parameter is the tampered image and associated label. The FGSM attack is a single-step attack represented by specific properties and parameters (see Table A.2).

Table A.2: FGSM adversarial attack parameters.

No.	Properties	Parameter
1	Input	Original Image, Label
2	Model to Attack	HNN model
3	Distance Measure	L ∞
4	Max. Perturbation	8/255
5	Output	Tampered Image, Label
Torchattacks PyTorch Usage Example		distinct-attack = torchattacks.FGSM(model, eps=8/255) tampered-images = distinct-attack (original-images, labels)

The CW attack with the specific attack properties include input, model to attack, distance measure, box constraint, confidence, number of steps, learning rate, and output. The input parameter is the original image and associated label. The model to attack parameter is the HNN model. The distance measure is L 2. The box-constraint parameter is 1. The confidence parameter is 0. The number of steps parameter is 50. The learning rate parameter is 0.01. The output parameter is the tampered image and associated label. The CW attack is represented in a set of specific attack properties and associated parameters (see Table A.3).

Table A.3: CW adversarial attack parameters.

No.	Properties	Parameter
1	Input	Original Image, Label
2	Model to Attack	HNN model
3	Distance Measure	L 2
4	Box-constraint	1
5	Confidence	0
6	Number of Steps	50
7	Learning Rate	0.01
8	Output	Tampered Image, Label
Torchattacks PyTorch Usage Example		distinct-attack = torchattacks.CW(model, c=1, kappa=0, steps=50, lr=0.01) tampered-images = distinct-attack (original-images, labels)

The PGD attack with the specific attack properties include input, model to attack, distance measure, maximum perturbation, step size, number of steps, random start, and output. The input parameter is the original image and associated label. The model to attack parameter is the HNN model. The distance measure parameter is L infinity. The maximum perturbation parameter is 8/255. The step size parameter is 2/225. The number of steps parameter is 10. The random start parameter is true. The output parameter is the tampered image and associated label. The PGD attack is represented in a set of specific attack properties and associated parameters (see Table A.4).

Table A.4: PGD adversarial attack parameters.

No.	Properties	Parameter
1	Input	Original Image, Label
2	Model to Attack	HNN model
3	Distance Measure	L ∞
4	Max. Perturbation	8/255
5	Step Size	2/255
6	Number of Steps	10
7	Random Start	True
8	Output	Tampered Image, Label
Torchattacks PyTorch Usage Example		distinct-attack = torchattacks.PGD(model, eps=8/255, alpha=1/255, steps=10, random-start=True) tampered-images = distinct-attack (original-images, labels)

APPENDIX B: ACRONYMS AND ABBREVIATIONS

Acronyms

Acronym	Definition
ACM	Association for Computing Machinery
Adam	Adaptive Moment Estimation (optimization algorithm)
AdvQuNN	Adversarial Quantum Neural Network
AI	Artificial Intelligence
API	Application Programming Interface
arXiv	Archive for preprint scholarly articles
AutoAttack	Ensemble of diverse adversarial attacks
BN	Batch Normalization
BPDA	Backward Pass Differentiable Approximation
BSA	Boundary-based Search Attack
CIFAR-10	Canadian Institute for Advanced Research 10-class dataset
CIFAR-100	Canadian Institute for Advanced Research 100-class dataset
CNN	Convolutional Neural Network
CNOT	Controlled-NOT (quantum gate)
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture (version 11.6)
CVPR	Conference on Computer Vision and Pattern Recognition
CW	Carlini-Wagner (attack method)
DeepFool	Minimal perturbation adversarial attack
DO-178C	Software Considerations in Airborne Systems
Drop	Dropout (regularization technique)

Acronym	Definition
EAD	Elastic-net Attack to Deep Neural Networks
ECCV	European Conference on Computer Vision
EMNIST	Extended Modified NIST
EOT	Expectation Over Transformation
FC	Fully Connected (layer)
FGSM	Fast Gradient Sign Method
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HNN	Hybrid Neural Network
HQNN	Hybrid Classical Quantum Neural Network
IBM	International Business Machines
ICCV	International Conference on Computer Vision
ICML	International Conference on Machine Learning
IEEE	Institute of Electrical and Electronics Engineers
ImageNet	Large-scale image database
IonQ	Quantum computing company
ISO	International Organization for Standardization
ISO 26262	Road Vehicles – Functional Safety
ISO/PAS 8800	Road Vehicles: Safety and Artificial Intelligence
JPEG	Joint Photographic Experts Group
JSMA	Jacobian-based Saliency Map Attack
L2	L2 Regularization (weight decay)
LCQHNN	Lean Classical Quantum Hybrid Neural Network
Linux	Open-source operating system
LR	Learning Rate
MART	Misclassification Aware Adversarial Training
ML	Machine Learning
MNIST	Modified NIST
MRI	Magnetic Resonance Imaging

Acronym	Definition
MUTCD	Manual on Uniform Traffic Control Devices
NeurIPS	Neural Information Processing Systems
NIST	National Institute of Standards and Technology
NISQ	Noisy Intermediate-Scale Quantum
NLLoss	Negative Log-Likelihood Loss
NVIDIA	Graphics processing unit manufacturer
Osprey	IBM quantum computer (433 qubits)
pandas	Python data analysis library
PGD	Projected Gradient Descent
PMLR	Proceedings of Machine Learning Research
Python	Programming language
QC	Quantum Computing
QML	Quantum Machine Learning
QNN	Quantum Neural Network
QuNN	Quantum Neural Network (alternate)
ReLU	Rectified Linear Unit (activation function)
RGB	Red Green Blue
Rigetti	Quantum computing company
RobQuNN	Robust Quantum Neural Network
RTCA	Radio Technical Commission for Aeronautics
TensorFlow	Google machine learning framework
TRADES	TRadeoff-inspired Adversarial DEfense
UAV	Unmanned Aerial Vehicle
Ubuntu	Linux operating system (20.04 LTS)
VQC	Variational Quantum Circuit
WTC	White-box Targeted Compound
ZOO	Zeroth Order Optimization

Mathematical Notation

Symbol	Definition
α	Step size for iterative attacks (0.01)
β_1	Adam decay rate first moment (0.9)
β_2	Adam decay rate second moment (0.999)
c	CW confidence parameter (1)
ϵ	Perturbation budget (0.3)
ϕ	Quantum angle qubits 1,3 (0.095π)
θ	Quantum angle qubits 0,2 (0.159π)
κ	CW targeted confidence (0)
π	Pi (3.14159)
σ	Gaussian noise std dev (0.1)
2^n	Quantum state dimension (16 for n=4)
G	Accuracy gap
I_{abs}	Absolute robustness improvement
I_{rel}	Relative robustness improvement
n	Number of qubits (4)
R_{attack}	Attack success rate
R_{recovery}	Recovery rate
$R_y(\theta)$	Y-axis rotation gate
Z	Computational basis measurement
$\text{Acc}_{\text{clean}}$	Clean accuracy
$\text{Acc}_{\text{attacked}}$	Attacked accuracy without defense
$\text{Acc}_{\text{defended}}$	Defended accuracy

Datasets

Dataset	Description
MNIST	28×28×1 grayscale, 10 classes, 12,665 train
EMNIST Digits	28×28×1 grayscale, 10 classes, 48,000 train
TinyImageNet	64×64×3 RGB, 5 classes, 2,000 train
Traffic Signs	64×64×3 RGB, 4 classes, 1,200 train
CIFAR-10	32×32×3 RGB, 10 classes
CIFAR-100	32×32×3 RGB, 100 classes
ImageNet	Large-scale 1000+ classes

Software and Tools

Name	Version
PyTorch	1.12.1
Cirq	1.0.0
Torchattacks	3.3.0
CUDA	11.6
Ubuntu	20.04 LTS
Python	Implementation language
Adam	Optimizer (lr=0.001)
pandas	Data analysis library
openpyxl	Excel library
NumPy	Numerical library
matplotlib	Plotting library
TensorFlow	Alternative ML framework

Attack Methods

Attack	Description
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent
CW	Carlini-Wagner
FGSM+PGD	Compound FGSM and PGD
FGSM+CW	Compound FGSM and CW
CW+PGD	Compound CW and PGD
BPDA	Backward Pass Differentiable Approximation
DeepFool	Minimal perturbation attack
EAD	Elastic-net Attack
JSMA	Jacobian-based Saliency Map
ZOO	Zeroth Order Optimization
AutoAttack	Ensemble attack

Defense Mechanisms

Defense	Description
Adversarial Training	Training-time defense
Input Transformation	Test-time deterministic
Randomization	Test-time stochastic
JPEG Compression	Quality=75
Bit-depth Reduction	4 bits
Gaussian Noise	σ=0.1
Image Quilting	4×4 patches
Random Resizing	0.8-1.2 scale
Random Cropping	0.8-1.0 scale
Random Rotation	±15 degrees
TRADES	TRadeoff-inspired defense
MART	Misclassification aware

Defense	**Description**
Defensive Distillation	Softened distributions
Certified Defenses	Provable robustness
Ensemble Defenses	Multiple models

APPENDIX C: RESEARCH ARTIFACTS AND DATA AVAILABILITY

Code Repository

All source code, experimental results, trained models, and supplementary materials for this dissertation research are publicly available in the following GitHub repository:

- **Repository URL:** `https://github.com/ericyoc/dissertation_code_and_data`
- **License:** MIT License (open source, academic use encouraged)
- **Primary Language:** Python 3.8+
- **Last Updated:** [February 14, 2026]

Repository Contents

The repository contains the complete implementation and experimental artifacts organized as follows:

Source Code

- Hybrid Neural Network (HNN) model implementation using PyTorch 1.12.1 and Cirq 1.0.0
- Adversarial attack generation (FGSM, PGD, Carlini-Wagner) using Torchattacks 3.3.0
- Defense mechanisms: input transformation, randomization, and adversarial training

- Experimental evaluation framework across 120 conditions (4 datasets × 3 attacks × 10 defenses)
- Data preprocessing and filtering scripts for MNIST, EMNIST "Digits", TinyImageNet, and Traffic Signs
- Results analysis and visualization code (`dissertation_code_and_data_poc.py`)

Experimental Results

- 120 raw result files (text format) with accuracy measurements across all experimental conditions
- Consolidated Excel analysis (`.xlsx`) with computed robustness metrics
- 14 publication-quality visualization charts (PNG, 300 DPI) including radar charts and heatmaps
- Filename convention: `[defense]_[attack]_[dataset]_[timestamp].txt`

Datasets

The repository provides access instructions and preprocessing scripts for all datasets:

- **MNIST:** 28×28×1 grayscale, 10 classes, 12,665 train / 4,230 test (filtered)
- **EMNIST "Digits":** 28×28×1 grayscale, 10 classes, 48,000 train / 16,000 test (filtered)
- **TinyImageNet:** 64×64×3 RGB, 5 classes, 2,000 train / 800 test (filtered subset)
- **Traffic Signs:** 64×64×3 RGB, 4 classes, 1,200 train / 400 test (filtered subset)

All datasets are publicly available and accessed via standard libraries (torchvision) or public URLs documented in the repository README.

Documentation

- `README.md` with installation instructions, dependencies, and usage examples
- `requirements.txt` specifying exact Python package versions
- Code comments and docstrings throughout implementation
- Reproducibility instructions for all experimental results

Software Dependencies

The experimental framework requires the following software environment:

Required software dependencies.

Package	Version	Purpose
PyTorch	1.12.1	Neural network framework
Cirq	1.0.0	Quantum circuit simulation
Torchattacks	3.3.0	Adversarial attack generation
NumPy	1.21.0+	Numerical computing
pandas	1.3.0+	Data analysis and results processing
matplotlib	3.4.0+	Visualization and chart generation
openpyxl	3.0.0+	Excel file manipulation
CUDA	11.6	GPU acceleration (optional)
Ubuntu	20.04 LTS	Operating system

Reproducibility

To reproduce the experimental results presented in this dissertation:

1. Clone the repository: `git clone https://github.com/ericyoc/dissertation_code_and_data`
2. Install dependencies: `pip install -r requirements.txt`
3. Follow instructions in `README.md` for dataset preparation and model training
4. Execute evaluation scripts as documented in the repository

Complete step-by-step instructions are provided in the repository README file.

Data Availability Statement

All research artifacts described in this appendix are publicly available under the MIT License at `https://github.com/ericyoc/dissertation_code_and_data`. The repository is maintained by the author and will remain publicly accessible. For questions regarding artifact access or reproducibility issues, please open an issue on the GitHub repository or contact the author at ericyoc@gmail.com.

Long-term Preservation

To ensure long-term availability:

- The GitHub repository is publicly accessible with permanent URL
- Source code is archived in Dakota State University institutional repository with persistent identifier
- All datasets are sourced from established public repositories (torchvision, Stanford CS231n, GTSRB)
- Repository is mirrored to ensure redundancy and persistence beyond the author's active maintenance

Ethical Considerations

All datasets used in this research are publicly available and do not contain personally identifiable information or sensitive data requiring IRB approval. The adversarial attack methodologies are intended solely for defensive research purposes and are documented to enable the research community to develop more robust machine learning systems. The open-source release supports transparency and reproducibility in adversarial machine learning research.