**Independent University** **Bangladesh (IUB)**

**IUB Academic Repository**

Graduate Thesis 2026

2026-08

# Declared failure over silent failure: abstention, explain ability and an abstention-aware evaluation framework

Setu, Jahanggir Hossain

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1547

*Downloaded from IUB Academic Repository*




# Declared Failure over Silent Failure: Abstention, Explainability and an Abstention-Aware Evaluation Framework

By

**Jahanggir Hossain Setu**

Student ID: **2312719**

**Summer 2026**

Supervisor:

**Ashraful Islam, PhD**

**Assistant Professor**

Department of Computer Science & Engineering

Independent University, Bangladesh

Co-Supervisor:

**M. Ashraful Amin, PhD**

**Professor**

Department of Information Sciences & Technology

George Mason University, USA

A thesis submitted in partial fulfillment of the requirements for the degree of

**Master of Science in Computer Science**

**Department of Computer Science & Engineering**

**Independent University, Bangladesh**

# Attestation

I declare that this thesis, titled **"Declared Failure over Silent Failure: Abstention, Explainability and an Abstention-Aware Evaluation Framework"**, is the result of my own original work and research. I completed this work as part of the requirements for the degree of Master of Science in Computer Science and Engineering at Independent University, Bangladesh (IUB).

I also confirm that this work, or any part of it, has not been submitted before for any other degree or qualification at any other university or institution. All of the sources and references that I used in writing this thesis have been properly acknowledged.

I understand that academic dishonesty is a serious offense. I accept that if I have violated the university's rules, I will be subject to disciplinary action.

| | |
|---|---|
| Signature | Date |

**Jahanggir Hossain Setu**

Name

# Acknowledgement

I want to show my heartfelt gratitude to my supervisor. Without his guidance and feedback, I would not have been able to complete this study. He is the person who motivated me and kept motivating me to finish this work. On the days I felt I was falling behind and could not believe in myself, he would ask about my progress again and again.

Not just for this thesis, he was always there from the beginning of my master's journey. He even motivated me to do the master's in the first place. He is like a mentor to me; his presence in my life is invaluable. Thanking him is never enough. He brought me to the light I always hoped for.

Most importantly, I am grateful to my family. Without their support, I would not have been able to come this far. My study is about stress, but at the times when I felt stressed, they were always there for me.

I would also like to acknowledge the use of AI to refine the tone and grammar of this writing, and I carefully reviewed whether the meaning remains the same.

# Letter of Transmittal

Dr. Ashraful Islam
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Dear Dr. Ashraful Islam,

I am pleased to submit my thesis titled **"Declared Failure over Silent Failure: Abstention, Explainability and an Abstention-Aware Evaluation Framework"**, prepared in partial fulfillment of the requirements for the degree of *Master of Science in Computer Science and Engineering* at Independent University, Bangladesh.

This thesis is the result of my own research conducted under your supervision. The work examines a SelectiveNet architecture under adversarial perturbation and explainable AI applied to the selection head so that the abstention decision is itself explained. The principal contribution is the Abstention-Aware Evaluation Framework which treats abstention as an explicit prediction outcome whose theoretical properties are established and verified.

I would like to express my sincere gratitude for your invaluable guidance, feedback, and encouragement throughout the duration of this project.

I respectfully submit this thesis for your review and hope it fulfills the academic standards and expectations of the department.

Sincerely,

Jahanggir Hossain Setu
ID: 2312719
Department of Computer Science and Engineering
Independent University, Bangladesh

# Evaluation Committee

Supervision Panel

| | |
|---|---|
| ..... | ..... |
| Dr. Ashraful Islam (Supervisor) | Prof. M. Ashraful Amin (Co-supervisor) |

Examiners

| | |
|---|---|
| ..... | ..... |
| Dr. Rashedur Rahman (Internal) | Dr. AKM Mahbubur Rahman (Internal) |

.....

Prof. Moinul Zaber, Institute of Development Studies, UK (External Examiner)

Office Use

| | |
|---|---|
| ..... | ..... |
| Dr. Ashraful Islam | Dr. AKM Mahbubur Rahman |
| Program Coordinator | Head of the Department |

.....

Prof. M. Arshad Momen

Director, Office of Graduate Studies, Research and Industry Relations (GSRIR)

.....

Md. Shahajada Masud Anowarul Haque, Librarian

# Abstract

Deep image classifiers are constructed in a way so that every input produces a label, whatever the model has learned about that input. The resulting failures are silent, e.g., a classifier confronted with an Out-Of-Distribution (OOD) or adversarially perturbed image returns a class assignment with no signal that the assignment should not be trusted. Selective classification yields a partial remedy by permitting the model to decline but its standard evaluation practice, e.g., computing performance over the accepted subset and reporting coverage separately, charges nothing for that decline. Therefore, a reported score improves monotonically as the model abstains more. This study makes two contributions. The first is empirical where a SelectiveNet architecture is compared against a matched baseline sharing the same convolutional backbone across four datasets (MNIST, CIFAR-10, chest X-ray for pneumonia detection and Cats vs. Dogs), evaluated, under Fast Gradient Sign Method (FGSM) perturbation and with Gradient-weighted Class Activation Mapping (Grad-CAM) applied to the selection head, as well as, the prediction head so that the abstention decision is itself explained. The second is the Abstention-Aware Evaluation Framework (AAEF) which extends the confusion matrix with an abstention column preserving every sample and defines three metrics on it. Boundedness, exact reduction to the classical metrics at zero abstention and correct extreme-case behaviour are established and verified. Absolute error counts fell on all four datasets with missed pneumonia cases dropping from 17 to 7, though 4 to 14 deferrals were required per error avoided. Adversarial results were consistent, as well. As perturbation strength rises the baseline CNN degrades steadily, while SelectiveNet's selective accuracy recovers because it lowers coverage and rejects the samples it cannot resolve. The optimism introduced by discarding abstentions ranged from 0.069 to 0.26 in balanced accuracy. Charging abstention as an outcome rather than an exclusion changes the reported performance of a selective classifier which is large enough to reverse a comparison that conventional evaluation would decide the other way.

***Keywords***— SelectiveNet, evaluation metrics, responsible AI, OOD, adversarial attack

# Contents

**Attestation** i

**Acknowledgement** ii

**Letter of Transmittal** iv

**Evaluation Committee** v

**Abstract** v

**List of Abbreviations** x

**List of Figures** x

**List of Tables** x

**1 Introduction** 1

1.1 Background and Context . . . 1

1.1.1 The forced-decision constraint . . . 1

1.1.2 The reject option and its history . . . 2

1.1.3 Responsible AI as a design constraint . . . 3

1.2 Problem Statement . . . 4

1.2.1 Silent failure . . . 4

1.2.2 Partial transparency . . . 5

1.2.3 Detection machinery as an added component . . . 5

1.2.4 Evaluation that does not charge for abstention . . . 6

1.3 Motivation . . . 6

1.3.1 Changing the failure mode . . . 7

1.3.2 Intrinsic detection . . . 7

1.3.3 Explaining the refusal . . . 7

1.3.4 Penalizing for abstention . . . 7

1.4 Research Objectives . . . 7

1.4.1 Implementation and baselines . . . 8

1.4.2 Adversarial and detection behaviour . . . 8

1.4.3 Explainability . . . 8

1.4.4 The evaluation frameworks . . . 8
1.4.5 Application and synthesis . . . 9
1.5 Research Questions . . . 9
1.6 Contribution of the Study . . . 11
1.6.1 A conceptual and empirical contribution . . . 11
1.6.2 A methodological contribution . . . 11
1.6.3 A theoretical and practical contribution . . . 11
1.7 Structure of the Thesis . . . 11

**2 Literature Review** **13**
2.1 Deep Convolutional Classifiers and the Confidence Problem . . . 13
2.2 Classification with a Reject Option . . . 15
2.3 Adversarial Machine Learning . . . 17
2.4 Explainable AI . . . 17
2.5 Responsible AI as a Normative Frame . . . 18
2.6 Evaluation of Selective Classifiers . . . 19
2.7 Identified Research Gaps . . . 20

**3 The Abstention-Aware Evaluation Framework (AAEF)** **22**
3.1 Notation and Setup . . . 22
3.2 The Abstention-Aware Confusion Matrix . . . 23
3.2.1 Abstention-Aware Recall (AAR) . . . 24
3.2.2 Abstention-Aware Specificity (AAS) . . . 25
3.2.3 Abstention-Aware Balanced Accuracy (AABA) . . . 25
3.3 Multiclass Generalisation . . . 25
3.4 Theoretical Properties . . . 26
3.4.1 Boundedness . . . 26
3.4.2 Reduction . . . 26
3.4.3 Extreme cases . . . 27
3.5 Limitations of the Framework . . . 27

**4 Research Methodology** **28**
4.1 Overview and Design Principles . . . 28
4.2 Computational Environment and Software . . . 28
4.3 Datasets and Preprocessing . . . 29
4.3.1 MNIST . . . 29
4.3.2 CIFAR-10 . . . 29
4.3.3 Chest X-ray (Normal vs. Pneumonia) . . . 30
4.3.4 Cats vs. Dogs . . . 30
4.4 Baseline Architectures . . . 30
4.4.1 MNIST . . . 30
4.4.2 CIFAR-10 . . . 31

4.4.3 Chest X-ray . . . 31
4.4.4 Cats vs. Dogs . . . 31
4.5 SelectiveNet Implementation . . . 33
4.5.1 Architecture . . . 33
4.5.2 Loss function . . . 36
4.5.3 Hyperparameters . . . 36
4.5.4 Training Protocol . . . 37
4.6 Adversarial Evaluation Protocol . . . 37
4.7 Explainability Protocol . . . 38
4.8 Evaluation Protocol . . . 39
4.9 Threats to Methodological Validity . . . 39

**5 Results and Analysis** **40**
5.1 MNIST . . . 40
5.1.1 Baseline training dynamics . . . 40
5.1.2 Baseline performance . . . 40
5.1.3 Selective model training dynamics . . . 43
5.1.4 Selective model performance . . . 44
5.1.5 Behaviour under FGSM . . . 44
5.1.6 Dual-head explanations . . . 47
5.1.7 Abstention-aware evaluation . . . 47
5.2 CIFAR-10 . . . 53
5.2.1 Baseline training dynamics . . . 53
5.2.2 Baseline performance . . . 53
5.2.3 Selective model training dynamics . . . 55
5.2.4 Selective model performance . . . 56
5.2.5 Behaviour under FGSM . . . 56
5.2.6 Dual-head explanations . . . 58
5.2.7 Abstention-aware evaluation . . . 58
5.3 Chest X-ray . . . 63
5.3.1 Baseline training dynamics . . . 63
5.3.2 Baseline performance . . . 63
5.3.3 Selective model training dynamics . . . 64
5.3.4 Selective model performance . . . 65
5.3.5 Behaviour under FGSM . . . 65
5.3.6 Dual-head explanations . . . 66
5.3.7 Abstention-aware evaluation . . . 66
5.4 Cats vs Dogs . . . 70
5.4.1 Baseline training dynamics . . . 70
5.4.2 Baseline performance . . . 71
5.4.3 Selective model training . . . 72
5.4.4 Selective model performance . . . 72

5.4.5 Behaviour under FGSM . . . 73
5.4.6 Dual-head explanations . . . 74
5.4.7 Abstention-aware evaluation . . . 74
5.5 Behaviuor under OOD . . . 79

**6 Discussion** **82**
6.1 RQ1 — behaviour under adversarial perturbation . . . 82
6.2 RQ2 — detection and error reduction . . . 82
6.3 RQ3 — explaining the selection decision . . . 83
6.4 RQ4 — what the framework reveals . . . 83

**7 Conclusion and Future Work** **86**
7.1 Summary of the Work . . . 86
7.2 Limitations . . . 87
7.3 Future Work . . . 87
.1 Code Availability . . . 89
.2 Data Availability . . . 89

**Bibliography** **96**

# List of Abbreviations

**AABA** Abstention-Aware Balanced Accuracy
**AAEF** Abstention-Aware Evaluation Framework
**AAR** Abstention-Aware Recall
**AAS** Abstention-Aware Specificity
**AN** Abstained Negative
**AP** Abstained Positive
**API** Application Programming Interface
**BA** Balanced Accuracy
**CNN** Convolutional Neural Network
**ECE** Expected Calibration Error
**EU** European Union
**FGSM** Fast Gradient Sign Method
**TP** True Positive
**TN** True Negative
**FN** False Negative
**FP** False Positive
**GPU** Graphics Processing Unit
**Grad-CAM** Gradient-weighted Class Activation Mapping
**JFIF** JPEG File Interchange Format
**LIME** Local Interpretable Model-agnostic Explanations
**ODIN** Out-of-DIstribution detector for Neural networks
**OOD** Out-Of-Distribution
**RAI** Responsible Artificial Intelligence
**ReLU** Rectified Linear Unit
**ROC** Receiver Operating Characteristic
**RQ** Research Question
**SHAP** SHapley Additive exPlanations
**TCAV** Testing with Concept Activation Vectors
**VGG** Visual Geometry Group

# List of Figures


1.1 Comparison of the conventional forced-decision pipeline (top) with the selective pipeline adopted in this work (bottom). . . . . 4

2.1 Approaches to classification with rejection. . . . . 16
2.2 The risk–coverage tradeoff, sketched. . . . . 19

3.1 The abstention-aware confusion matrix. The four classical cells are computed over the accepted set and the two additional cells hold the abstained samples which conventional selective evaluation discards. . . . . 24
3.2 Partition of the evaluation set under the two conventions. . . . . 25
3.3 The augmented $n \times (n + 1)$ confusion matrix. . . . . 26

4.1 The architecture of CNN used to classify MNIST dataset. . . . . 31
4.2 The architecture of CNN used to classify CIFAR-10 dataset. . . . . 32
4.3 The architecture of CNN used to classify Chest X-ray dataset. . . . . 32
4.4 The architecture of CNN used to classify Cat vs. Dog dataset. . . . . 33
4.5 SelectiveNet as implemented. . . . . 34
4.6 The architecture of SelectiveNet used to classify MNIST dataset. . . . . 34
4.7 The architecture of SelectiveNet used to classify CIFAR-10 dataset. . . . . 35
4.8 The architecture of SelectiveNet used to classify Chest X-ray dataset. . . . . 35
4.9 The architecture of SelectiveNet used to classify Cat vs. Dog dataset. . . . . 36
4.10 The dual-head Grad-CAM procedure. . . . . 39

5.1 Training and validation accuracy and loss for the MNIST baseline CNN. . . . . 41
5.2 Risk–coverage curve for the MNIST SelectiveNet. . . . . 41
5.3 Confusion matrix for the baseline CNN on the MNIST test set. . . . . 42
5.4 Per-digit recall for the baseline CNN on MNIST dataset. . . . . 43
5.5 Training dynamics for the SelectiveNet on MNIST. . . . . 43
5.6 Confusion matrix for the MNIST SelectiveNet computed over accepted samples only. . . . . 45
5.7 Adversarial behaviour of the baseline CNN and the SelectiveNet under FGSM on MNIST (n = 2,000). . . . . 47
5.8 Grad-CAM attributions on clean MNIST digits comparing accepted and rejected samples. . . . . 48

5.9 The same digit before and after FGSM perturbation with the selection-head attribution shown for both. . . . . 49
5.10 Conventional (accepted-only) scores against abstention-aware scores for both MNIST models. . . . . 50
5.11 Abstention-aware confusion matrix for CNN on MNIST dataset. . . . . 50
5.12 Abstention-aware confusion matrix for SelectiveNet on MNIST dataset. . . . . 52
5.13 Training and validation accuracy and loss for the CIFAR-10 baseline CNN. . . . 53
5.14 Confusion matrix for the baseline CNN on CIFAR-10. . . . . 54
5.15 Per-class recall for the baseline CNN on CIFAR-10. . . . . 55
5.16 Risk–coverage curve for the CIFAR-10 SelectiveNet. . . . . 55
5.17 Training dynamics for the CIFAR-10 SelectiveNet. . . . . 56
5.18 Confusion matrix for the CIFAR-10 SelectiveNet over accepted samples only. . . 57
5.19 Degradation under FGSM on CIFAR-10 (n = 50). . . . . 58
5.20 Grad-CAM attributions on clean CIFAR-10 images, accepted against rejected. . 59
5.21 Conventional against abstention-aware scores for both CIFAR-10 models. . . . . 61
5.22 Abstention-aware confusion matrix for CNN on CIFAR-10 dataset. . . . . 61
5.23 Abstention-aware confusion matrix for SelectiveNet on CIFAR-10 dataset. . . . . 62
5.24 Training and validation accuracy and loss for the chest x-ray baseline. . . . . 63
5.25 Confusion matrix for the baseline CNN on the chest x-ray test set. . . . . 64
5.26 Training dynamics for the chest x-ray SelectiveNet. . . . . 64
5.27 Risk–coverage curve for the chest x-ray SelectiveNet. . . . . 65
5.28 Confusion matrix for the chest x-ray SelectiveNet over accepted scans only. . . . 66
5.29 Degradation under FGSM on the chest x-rays (n = 50). . . . . 67
5.30 Grad-CAM attributions on chest x-rays, accepted against rejected. . . . . 68
5.31 Conventional against abstention-aware scores for both chest x-ray models. . . . . 69
5.32 Abstention-aware confusion matrix for CNN on Chest X-ray dataset. . . . . 69
5.33 Abstention-aware confusion matrix for SelectiveNet on Chest X-ray dataset. . . . 69
5.34 Training and validation accuracy and loss for the Cats vs Dogs baseline CNN. . . 70
5.35 Confusion matrix for the baseline CNN on the Cats vs Dogs test set. . . . . 71
5.36 Training dynamics for the Cats vs Dogs SelectiveNet over twenty epochs. . . . . 72
5.37 Risk–coverage curve for the Cats vs Dogs SelectiveNet. . . . . 72
5.38 Confusion matrix for the SelectiveNet on the Cats vs Dogs test set. . . . . 73
5.39 Degradation under FGSM on Cats vs Dogs (n = 1,000). . . . . 74
5.40 Grad-CAM attributions on Cats vs Dogs, accepted against rejected. . . . . 75
5.41 Conventional against abstention-aware scores for both Cats vs Dogs models. . . . 76
5.42 Abstention-aware confusion matrix for CNN on Cats vs Dogs dataset. . . . . 76
5.43 Abstention-aware confusion matrix for SelectiveNet on Cats vs Dogs dataset. . . 79
5.44 OOD handwritten signature evaluated using the MNIST-trained SelectiveNet and rejected by the selection head. . . . . 79
5.45 OOD car image evaluated using the MNIST-trained SelectiveNet and rejected by the selection head. . . . . 80

5.46 In distribtion digit '7' using the MNIST-trained SelectiveNet and accepted by the selection head. . . . . 80
5.47 OOD panther image evaluated using the MNIST-trained SelectiveNet and rejected by the selection head. . . . . 81

# List of Tables


4.1 Comparison of datasets used for selective classification experiments . . . . . 29
4.2 Selective classification configuration across datasets. . . . . 37
4.3 Training protocol used across datasets . . . . . 37
4.4 Adversarial perturbation settings across datasets . . . . . 38

5.1 Classification performance of baseline CNN on MNIST. . . . . 42
5.2 Performance evaluation of SelectiveNet on MNIST (accepted samples). . . . . 44
5.3 MNIST under FGSM perturbation over 2,000 randomly drawn test digits. . . . . 46
5.4 Consolidated MNIST evaluation, macro-averaged over the ten classes. The baseline's abstention-aware values equal its conventional values, as the reduction property requires. . . . . 51
5.5 Classification performance on baseline CNN model on CIFAR-10 dataset. . . . . 54
5.6 Performance evaluation of SelectiveNet on accepted samples from CIFAR-10 dataset. . . . . 56
5.7 Comparison of plain CNN and SelectiveNet performance on CIFAR-10 dataset. . 60
5.8 Class-wise performance of the classification model on Chest X-ray dataset. . . . . 63
5.9 Class-wise performance of SelectiveNet on accepted samples on Chest X-ray dataset. 65
5.10 Comparison of classification performance for Chest X-ray dataset. . . . . 67
5.11 Per-class classification report for the Cats vs Dogs baseline. . . . . 71
5.12 Per-class classification report for the Cats vs Dogs SelectiveNet over the 1,291 accepted images only. . . . . 73
5.13 Consolidated Cats vs Dogs evaluation over the 1,760-image test set. . . . . 77
5.14 Comparison of classification performance Cats vs Dogs metrics under both conventions. . . . . 78

6.1 Absolute error counts for both models on each dataset, with the number of deferrals required to obtain the reduction. . . . . 83
6.2 The optimism gap across the four datasets. For the binary datasets the middle column is AAR and the accepted-only comparator is accepted-set accuracy; for the multiclass datasets it is micro AAR. . . . . 84

# Chapter 1

# Introduction

## 1.1 Background and Context

### 1.1.1 The forced-decision constraint

Deep convolutional neural networks have become the default tool for image classification and the accuracy figures they achieve on standard benchmarks are no longer in dispute [1, 2, 3, 4]. What is in dispute is what those figures mean once a model leaves the benchmark and enters a setting where its outputs influence decisions [5, 6, 7]. A classifier trained with a softmax output layer and a cross-entropy objective is constructed so that it must assign a label to every input it receives [8]. The architecture provides no mechanism for the model to indicate that a particular input falls outside the region where it has learned anything useful. Whatever arrives at the input layer produces a probability vector at the output layer and the argmax of that vector becomes a decision [9].

The constraint is worth stating because it is a property of the training objective and not merely of the inference code. Cross-entropy rewards the model for placing probability mass on the correct class and penalises it for placing mass elsewhere [10]. Nothing in the objective rewards the model for recognising that an input is unfamiliar. The softmax layer normalises whatever logits it is given, so the output is a probability distribution regardless of how far the input lies from the training manifold [11]. The resulting quantity, e.g., the maximum softmax probability, is read as a confidence estimate, but it was never trained to be one. It is a normalised score over classes, conditioned on the assumption that one of those classes is correct [9, 10].

This is easy to overlook when the test set is drawn from the same distribution as the training set. It becomes harder to overlook in the situations that motivate deployment. Guo et al. [12] showed that modern deep networks are poorly calibrated and that the confidence values they emit tend to overstate the probability that the corresponding prediction is correct. It is a tendency that grew worse, not better, as architectures deepened and as batch normalisation and reduced weight decay became standard [12]. Nguyen, Yosinski and Clune [4] demonstrated that networks assign high confidence to synthetically generated images that a human observer would not recognise as belonging to any class at all. Szegedy et al. [13] and later Goodfellow,

Shlens and Szegedy [14] established that small perturbations, constructed from the model's own gradients and imperceptible to a human viewer, are enough to flip a prediction while the reported confidence remains high. Hendrycks and Dietterich [15] showed that accuracy degrades under mundane corruptions, e.g., blur, noise, compression artefacts, weather effects, that carry no adversarial intent. Researchers examined how uncertainty estimates themselves degrade under distributional shift and found that the methods which look well-calibrated in-distribution are not the ones that remain so as the shift grows [16, 17, 18, 19].

These results do not describe rare or exotic failures. They describe a structural property of the forced-decision design. The model produces an answer and it produces a confidence score but the two are only loosely coupled and neither carries a dependable signal about whether the input was one the model was equipped to handle. Thus, the system fails and it fails without saying so.

### 1.1.2 The reject option and its history

The alternative has a long history. Chow [20] analysed the error–reject tradeoff for Bayes-optimal classifiers and derived the relationship between the rejection rate and the error rate on the remainder. The technique shows that permitting a classifier to withhold a decision on a bounded fraction of inputs lowers the error rate on the inputs it does act upon. Chow's rule reject when the maximum posterior falls below a threshold determined by the relative cost of an error and a rejection, is optimal when the posteriors are known, which is the condition that does not hold for a trained neural network [20].

The idea has been revisited in several forms since. Cordella et al. [21] applied a reject option to multilayer perceptrons using a reliability measure derived from the output activations. Bartlett and Wegkamp [22] developed the statistical theory of classification with a reject option, the latter introducing a modified hinge loss that yields a consistent estimator. Cortes, DeSalvo and Mohri [23] formulated learning with rejection as the joint learning of a predictor and a rejector with generalisation guarantees for the pair. El-Yaniv and Wiener [24] placed selective classification on a formal footing for the noise-free case characterising the risk–coverage tradeoff and introducing new vocabularies, e.g., selective risk, coverage, the selection function, that the subsequent deep learning literature adopted.

What changed more recently is that the reject option was moved inside the network rather than bolted onto it after training. Geifman and El-Yaniv [25] constructed selective classifiers from already-trained deep networks using softmax response and Monte-Carlo dropout as confidence functions and calibrating a threshold to guarantee a target risk with high probability. This works but the confidence function is derived from a network that was never trained to produce one. Their subsequent architecture, SelectiveNet [26], addressed that gap. A dedicated selection head $g(x)$ is trained jointly with the prediction head $f(x)$, under a loss combining empirical selective risk with a quadratic penalty on deviation from a target coverage and an auxiliary head $h(x)$ trained with standard cross-entropy keeps the shared representation from degrading on the samples the selection head suppresses [26]. The reject decision becomes a learned function of the input rather than a threshold applied to a quantity.

A parallel line of work frames the same mechanism differently. Rather than treating rejection as abstention from a decision, it treats it as 'deferral' to another agent [27, 28]. Madras, Pitassi and Zemel [29] formulated learning to defer and showed that the resulting policy interacts with fairness in ways that a simple rejection rule does not capture. Collins et al. [30] derived consistent surrogate losses for the deferral problem. Gonzalez et al. [31] examined the allocation of effort between model and human as a triage problem. This reframing matters for the present work because in a clinical setting an abstention is not a null output. It is a referral and it consumes a human's time.

### 1.1.3 Responsible AI as a design constraint

The question of what makes an automated system trustworthy has moved over the past several years from an academic concern to a regulatory one. More recently these commitments have acquired legal weight. Responsible Artificial Intelligence (RAI), widely known as Responsible AI, is a practical framework of ethics, rules and technical steps used to build and deploy AI safely. Running alongside this technical literature is a policy literature concerned with what makes an automated system fit for deployment. Surveys of the field, e.g., Jobin, Ienca and Vayena [32] reviewed eighty-four sets of guidelines. Floridi and Cowls [33] consolidated the landscape into five principles, converge on a small set of recurring commitments, e.g., fairness, transparency, accountability, reliability and safety or security. The European Commission's High-Level Expert Group [34] framed a similar list under the heading of Trustworthy AI and the requirement for human oversight was subsequently written into Article 14 of the European Union (EU) AI Act (Regulation (EU) 2024/1689) for systems classified as high-risk.

These documents are, for the most part, statements of what a system ought to satisfy rather than descriptions of how an architecture might satisfy it. A guideline that a system must support human oversight does not specify the mechanism by which the system identifies the cases a human should see. A guideline that a system must be transparent does not specify what is to be made transparent when the system's output is a refusal. The gap between the principle and the mechanism that would realise it is where this study situates itself. The claim is not that selective classification satisfies these principles, no architecture does that by itself, but that it supplies concrete mechanisms that several of them require and that those mechanisms can be examined empirically rather than asserted.

Figure 1.1 contrasts the two pipelines at the level of what leaves the system.

In the conventional case every input, whether in-distribution, out-of-distribution or adversarially perturbed, produces a class label (classification task) or a value within a range (regression task). In the selective case the learned selection head partitions inputs into an accepted set on which the prediction head answers and an abstained set, which is routed to a human decision-maker. The two pipelines share the same convolutional backbone in the experiments reported in this study so that any difference in behaviour is attributable to the selection mechanism rather than to representational capacity.

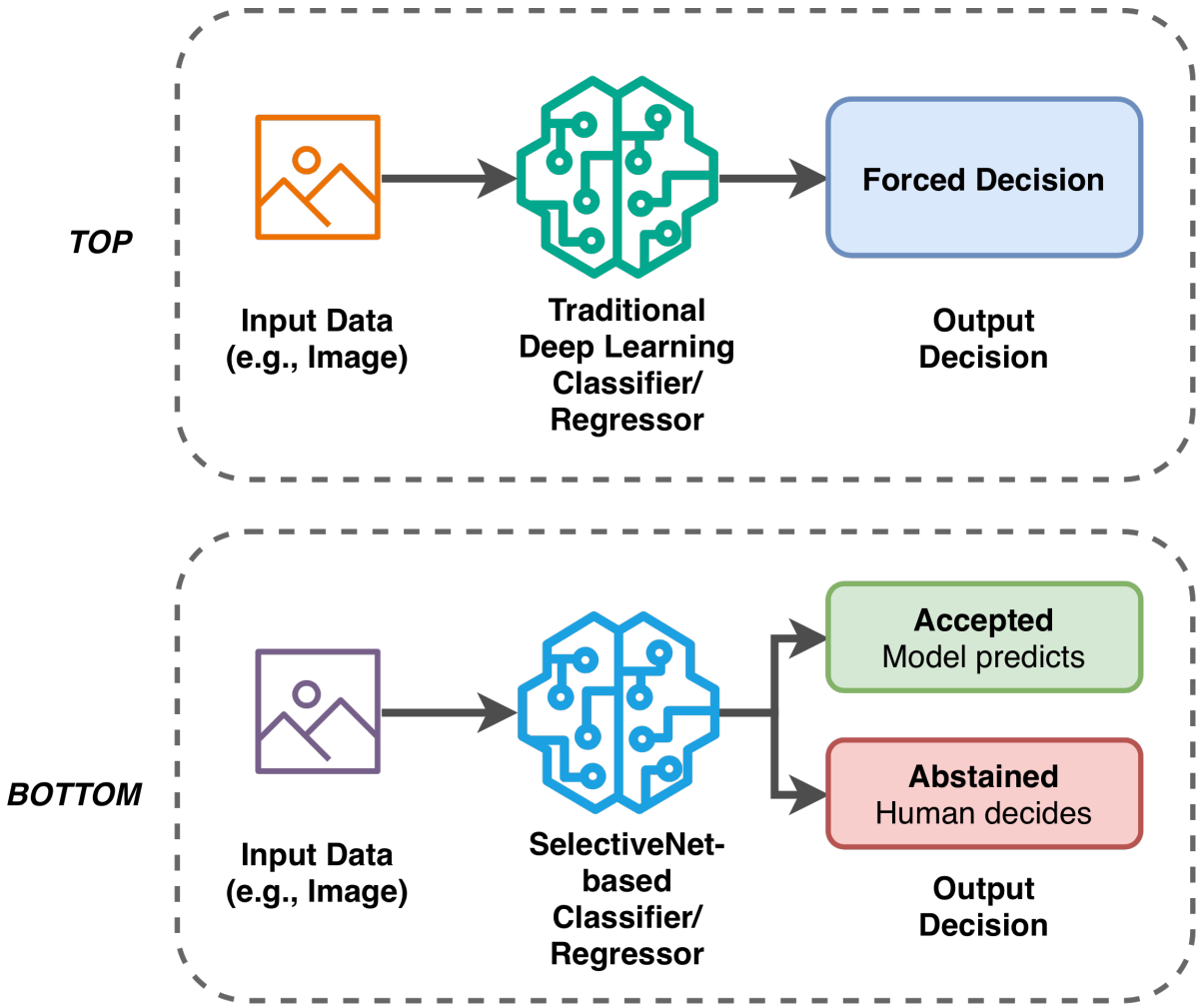



Figure 1.1: Comparison of the conventional forced-decision pipeline (top) with the selective pipeline adopted in this work (bottom).

## 1.2 Problem Statement

### 1.2.1 Silent failure

When a conventional deep learning model, e.g., Convolutional Neural Network (CNN) encounters an input drawn from outside its training distribution or one that has been perturbed to exploit its gradients, it does not signal difficulty. It produces a label and often it produces a high softmax value alongside it [25]. The problem is not that the model is wrong. Every model is wrong sometimes and an error rate is something a deployment can be designed around. The problem is that the wrongness is indistinguishable from correctness at the point of use. There is no channel through which the system can communicate that this particular case was one it should not have been asked to handle.

Muoka et al. [35] discussed the specific concern this raises in medical imaging, where a system that fails without announcing failure may be trusted precisely in the cases where trust is least warranted. Dong et al. [36] examined the characteristics of adversarial attacks on medical image analysis systems in more detail and found that medical images are in some respects more

susceptible than natural images owing to their standardised acquisition and limited variation. Kompa, Snoek and Beam [37] argued from the clinical side that the absence of a mechanism for the model to say 'I do not know' is one of the barriers to clinical adoption, since a clinician cannot integrate an output whose reliability is unstated into a decision process that requires reliability to be stated.

### 1.2.2 Partial transparency

Explainability methods in the CAM family [38] produce spatial attributions for a predicted class and they have become common in medical imaging work as a way of showing that a model attends to regions a clinician would consider relevant, the approach taken in CheXNet [39] and much of the work that followed [40]. But if a model is given a reject option, the explanation of the class prediction is no longer the whole story. A user confronted with an abstention needs to know why the system declined and a saliency map over the prediction head does not answer that question. The abstention is generated by a different head, with different learned weights, responding to different features of the input [26].

This matters for a reason beyond completeness. An abstention imposes a cost on a human and a system that abstains without explanation gives the human no basis for judging whether the abstention was well-founded. If the selection head is declining because the image is ambiguous, that is useful information. If it is declining because of an artefact at the image border that is a defect the operator should be able to see. Without an explanation of selection head, the two are indistinguishable.

The critical literature adds a caution here. Cheong [41] argued that post-hoc explanation of black-box models in high-stakes settings yield a weaker guarantee than it appears, since the explanation is itself a model of the model and may be faithful to neither. Raji et al. [42] showed that certain saliency methods produce visually convincing maps that survive randomisation of the network weights, which is to say that they were not explaining the network at all. Hassija [43] noted the absence of agreed criteria for evaluating interpretability, which makes claims about explanation quality difficult to adjudicate. Whatever one concludes from these critiques, they suggest that an explanation covering only part of the decision process is an incomplete account of it and that any claim made on the basis of such an explanation should be stated with corresponding caution.

### 1.2.3 Detection machinery as an added component

Out-Of-Distribution (OOD) detection is well studied, and several effective approaches exist. Huang [44] established the maximum-softmax-probability baseline. ODIN improved on it with temperature scaling and input preprocessing [16]. Mahalanobis distance was used in the feature space of a trained network [45]. Bayesian and ensemble approaches, e.g., Monte-Carlo dropout, deep ensembles, and the decomposition of aleatoric and epistemic uncertainty, estimate predictive uncertainty from multiple forward passes or multiple models.

Each of these researches adds something to the pipeline. A hyperparameter to tune on

OOD data that may not be available at development time, a distance computation over cached class statistics or a multiplication of inference cost by the ensemble size. Each also introduces a component that must itself be validated and whose failure modes must be understood. In a deployment with latency or memory constraints and in a development process where every additional component is an additional thing to get wrong, there is a case for a detection signal that emerges from the same forward pass as the prediction. Whether such a signal performs as well as a dedicated detector is a separate question, one this study does not attempt to settle, but the absence of additional machinery is itself a property worth examining.

### 1.2.4 Evaluation that does not charge for abstention

This is the problem the framework proposed in this study addresses and it is the one the literature has left least resolved.

Conventional selective evaluation reports two quantities. The first is a performance measure computed on the accepted subset, e.g., selective accuracy, selective risk or accepted-sample precision, recall and F1. The second is coverage, reported alongside as a separate figure [46]. The abstained samples leave the performance measure entirely.

The consequence follows from the arithmetic. Let the model abstain on a set of samples chosen because they are the ones it finds hardest. Removing those samples from the denominator raises the reported score. In the general case the score, therefore, rises monotonically as coverage falls and the limiting behaviour is degenerate. A model that accepts a single sample and answers it correctly reports perfect selective accuracy. Coverage is supposed to guard against this, but it does so only by requiring the reader to hold two numbers in mind and to combine them by judgement rather than by arithmetic. Two models operating at different coverage levels cannot be compared on their headline numbers at all and no single quantity expresses what the system delivers over the full population of inputs it will encounter.

The risk–coverage curve is the usual response, and it is a good one for characterising a single model's tradeoff across thresholds. It is less useful as a reporting instrument because it does not reduce to a number that can be placed in a table beside a non-abstaining baseline and because the area under it weights all coverage levels equally when a deployment will operate at one.

The problem has been recognised. Condessa, Bioucas-Dias and Kovačević [47] proposed performance measures for classification with rejection built on the relationship between the rejected region and the misclassified region, e.g., non-rejected accuracy, classification quality, rejection quality. These are useful and they move in the right direction. What this review did not find is a metric that treats abstention as a prediction outcome on the same footing as a class prediction reduces exactly to the classical metric when the model never abstains (full coverage). That is the gap the Abstention-Aware Evaluation Framework (AAEF) is intended to fill.

## 1.3 Motivation

The argument this study advances is that these four problems have a shared partial remedy and that the remedy is worth examining as a coherent design rather than as four separate

techniques.

### 1.3.1 Changing the failure mode

If the model can decline, a misclassification that would have been silent becomes a declared abstention. This does not make the model more robust in the sense the adversarial robustness literature uses the term. The underlying representation is no less susceptible to gradient-based perturbation. Adversarial training is the technique that addresses that and it is not used here. What changes is what the system does when the representation fails. Detecting an attack and refusing to act on it is a different guarantee from resisting the attack and for many deployment settings since a refusal can be escalated whereas a confident error cannot be caught.

### 1.3.2 Intrinsic detection

A trained selection head produces its accept or reject score in the same forward pass that produces the class posterior. The score is a learned function of the input, optimised against a coverage-constrained risk objective, rather than a transformation of logits that were optimised for something else [26]. This does not guarantee that it will be a good OOD detector and the study reports what is observed rather than assuming the result. But the absence of additional machinery is a property that matters for deployment cost and for the number of components that require separate validation.

### 1.3.3 Explaining the refusal

Applying Gradient-weighted Class Activation Mapping (Grad-CAM) to the selection head is a small extension of an existing method. It yields two attributions per input, e.g., one answering why this class, the other answering why this was accepted or declined. Comparing the second yield a way to look at what shifts in the selector's spatial focus when the model withdraws. This is an exploratory element of the work and the claims made from it are correspondingly modest.

### 1.3.4 Penalizing for abstention

If abstention carries a cost inside the metric, the reported number can no longer be inflated by declining. The AAEF treats abstention as a third prediction category.

## 1.4 Research Objectives

The aim of this study is to examine selective classification, in the SelectiveNet formulation, as a design for responsible deployment of image classifiers and to propose and verify an evaluation framework that treats abstention as a prediction outcome rather than as grounds for excluding a sample from the evaluation.

The objectives below are grouped by the aspect of the work they serve and each is stated with the criterion.

### 1.4.1 Implementation and baselines

**O1. Construct matched model pairs for each dataset:** Implement a baseline CNN and a SelectiveNet sharing the identical convolutional backbone for MNIST, CIFAR-10, Chest X-ray and Cats versus Dogs datasets. The SelectiveNet adds the prediction, selection and auxiliary heads on top of the shared feature extractor.

**O2. Train both models to a comparable standard** Train each pair with matched optimisers, batch size, learning rate and comparable epoch budgets applying early stopping where the validation curve warrants it.

### 1.4.2 Adversarial and detection behaviour

**O3. Implement Fast Gradient Sign Method (FGSM) and evaluate both models across an epsilon sweep** Generate adversarial examples from each model's own gradients and record, for each epsilon, the baseline accuracy, the selective model's overall accuracy, its coverage and its selective accuracy.

**O4. Quantify the change in error composition** Compare the absolute counts of FP and FN between the baseline and the selective model alongside the counts of abstained positives and abstained negatives.

### 1.4.3 Explainability

**O5. Implement Grad-CAM for both heads** Extend the standard Grad-CAM formulation so that the target scalar can be either the predicted-class probability from prediction head or the selection score from selection head taking gradients with respect to the final convolutional feature maps of the shared backbone in both cases.

### 1.4.4 The evaluation frameworks

**O6. Formalise the framework** Define the extended confusion matrix, establish the conservation identity, define Abstention-Aware Recall (AAR), Abstention-Aware Specificity (AAS) and Abstention-Aware Balanced Accuracy (AABA).

**O7. Verify the theoretical properties** Establish boundedness, reduction, monotonicity and extreme-case correctness where the argument is short and by systematic numerical testing.

### 1.4.5 Application and synthesis

**O8. Apply the framework across all four datasets** Compute the conventional and abstention-aware metrics for both models on each dataset, at a pinned selection threshold that does not depend on execution order and report the optimism gap.

**O9. Map findings to the RAI principles** Set out, for each of transparency, accountability, reliability and safety and security, which experimental result bears on it and how strongly.

## 1.5 Research Questions

Four Research Questions (RQs) structure the study. Each is stated with the reasoning that led to it.

**RQ1: How does a SelectiveNet-based classifier behave under adversarial perturbation compared with a traditional classifier sharing the same backbone and does its coverage fall as the perturbation strength increases? Rationale:** The claim that selective classification improves safety rests on an assumption that is easy to state and easy to leave untested. The selection head withdraws from inputs it cannot handle. If the selection head were to remain confident on adversarial inputs, e.g., accepting them at the same rate as clean inputs while the prediction head misclassifies them, then the abstention mechanism would provide no protection at all and the architecture's safety claim would collapse. The question is, therefore, a falsification test rather than a demonstration.

Two design decisions follow from this reasoning. First, the baseline must share the backbone with the selective model so that any difference observed is attributable to the selection mechanism and not to a difference in representational capacity or training. Second, coverage must be tracked alongside accuracy across the epsilon sweep because selective accuracy alone would be uninterpretable. A selective accuracy that holds steady under attack means one thing if coverage holds steady and something quite different if coverage has collapsed to a handful of samples. The pairing of the two curves is the actual observation.

**RQ2: Does the learned selection head provide a usable OOD and low-confidence signal without an auxiliary detection module and does the abstention mechanism reduce the absolute counts of false positives and false negatives relative to the non-abstaining baseline? Rationale:** This question has two parts because a plausible objection to selective classification must be addressed alongside the claim itself.

The first part concerns cost. If the selection head detects difficult and OOD inputs at a usable rate then the detection capability comes free with the architecture which is the practical argument for preferring it over a pipeline with a bolted-on detector. The word 'usable' is deliberate. The question is not whether selection head outperforms Out-of-DIstribution detector for Neural networks (ODIN) or a Mahalanobis detector which would require a benchmark this study does not construct.

The second part addresses the objection. A sceptic could argue that abstention does not reduce errors but relabels them that the samples which would have been false negatives are now abstentions and that nothing has been gained except a change in vocabulary. This objection has force and it is the reason the question asks about absolute counts rather than rates. If the False Positive (FP) and False Negative (FN) counts fall in absolute terms then the samples removed from those cells were removed by a mechanism that also improved the quality of the decisions retain. The extended confusion matrix under (AAEF) is what makes this distinction visible, since it displays both the error cells and the abstention cells for the same population.

**RQ3: What does Grad-CAM applied to the selection head reveal about the basis of the accept or reject decision and does the selector's spatial focus shift in an observable way?** **Rationale:** If abstention is the mechanism by which the architecture supports accountability then the abstention must be inspectable. A human asked to adjudicate a deferred case is entitled to know on what basis it was deferred. The standard application of Grad-CAM to the prediction head does not supply this because it explains a decision the model declined to stand behind.

The second clause of the question is exploratory and is framed as such. A visible shift would be consistent with the selection head responding to the perturbation rather than to a coincidental property of the image.

**RQ4: Can an evaluation framework that treats abstention as an explicit prediction outcome be defined so that it is bounded, reduces to the classical metrics at zero abstention, decreases monotonically in the abstention count, and behaves correctly at the extremes and what does it reveal about selective classifier performance that conventional accepted-only evaluation conceals?** **Rationale:** This question carries the theoretical weight of the study and each of the four required properties is there for a reason.

Boundedness is a usability requirement. A metric that can exceed one or fall below zero cannot be read against the conventions practitioners already hold and would be discarded on that ground alone.

Reduction is the property that makes comparison possible. Without it, an abstention-aware score and a classical score would live on different scales and placing a non-abstaining baseline beside a selective model in one table would be meaningless.

Extreme-case correctness guards against a metric that satisfies the other three properties while behaving strangely at the boundaries. A measure that assigns a non-zero score to a model rejecting every input, or a score below one to a perfect non-abstaining classifier, is not measuring what it claims to. The second clause asks what the framework reveals, and this is where the work becomes empirical rather than definitional.

## 1.6 Contribution of the Study

The contributions are of three kinds and they differ in weight.

### 1.6.1 A conceptual and empirical contribution

The study assembles selective classification, adversarial evaluation and dual-head explainability into a single account and evaluates it across four datasets spanning two task structures and two class counts. The individual components are drawn from existing literature and none is claimed as novel. What is provided is the framing that connects them to the RAI principles at the level of mechanism rather than aspiration.

### 1.6.2 A methodological contribution

Grad-CAM is applied to the selection head as well as the prediction head, yielding paired attributions that explain both the classification and the decision to classify. The standard explanation explains a decision the model declined to make. It is not something the existing selective classification literature reports.

### 1.6.3 A theoretical and practical contribution

The AAEF is the principal contribution. It comprises the followings: an extended confusion matrix with abstention as a third prediction category, three seperate metrics defined on that matrix, e.g., AAR, AAS, AABA and the optimism gap as a reportable diagnostic quantity giving the amount by which conventional accepted-only evaluation overstates a selective model's performance at a given operating point.

## 1.7 Structure of the Thesis

The chapter sequence departs from the conventional order in one respect.

Chapter 2 reviews the literature across deep convolutional classifiers and their calibration behaviour, selective classification and learning with a reject option, uncertainty estimation and OOD detection, adversarial machine learning and the distinction between robustness and detection and explainable AI. The chapter closes with a review of evaluation practice for selective classifiers.

Chapter 3 develops the AAEF. It sets out the arithmetic of the problem with accepted-only evaluation, defines the extended confusion matrix and the three metrics, generalises them to the multiclass case through one-vs-rest decomposition, states and verifies the four theoretical properties, introduces the optimism gap as a diagnostic and identifies the AAS degeneracy as a known limitation.

Chapter 4 describes the experimental methodology, e.g., the datasets with their preprocessing and splitting procedures, the shared backbone and the baseline and SelectiveNet architec-

tures, the selective loss and its hyperparameters, the FGSM protocol, the dual-head Grad-CAM implementation, and the evaluation procedure including threshold pinning.

Chapter 5 reports results dataset by dataset, e.g., baseline performance, selective behaviour, adversarial degradation, explainability observations and AAEF results for each and closes with a cross-dataset synthesis addressing the four research questions in turn.

Chapter 6 discusses the findings, maps them onto the RAI principles and sets out threats to validity.

Chapter 7 concludes, restates the contributions against the research questions, and identifies directions for further work.

# Chapter 2

# Literature Review

This chapter surveys the work on which the dissertation builds. The review is organised around six strands. The strands are not independent and the review tries to make the connections visible rather than treating each as a separate compartment. A note on the boundaries of the review. The literature on adversarial machine learning and on OOD detection is large enough that a complete survey would fill a dissertation on its own. What follows is selective and the selection is guided by relevance to the architecture under study rather than by an attempt at coverage.

## 2.1 Deep Convolutional Classifiers and the Confidence Problem

The architectures used in this study are conventional and require little introduction. The lineage runs from LeNet-5, through AlexNet, which established deep convolutional networks on large-scale image classification, to VGG, which showed that depth built from small kernels was productive, and ResNet, whose residual connections made much deeper networks trainable [48]. DenseNet extended the idea of shortcut connectivity. Alongside these came the regularisation and normalisation techniques that made training stable, e.g., dropout and batch normalisation [49].

The models trained in this study are small by contemporary standards, e.g., few convolutional blocks with pooling and dropout, rather than a deep residual network. Because the object of study is the selection mechanism rather than the backbone. A larger backbone would raise accuracy on all measures for both models in a comparison, without altering the question being asked.

Deep networks can fit random labels on standard datasets to zero training error, which implies that their capacity is sufficient to memorise the training set outright and that the classical bounds relating capacity to generalisation do not explain why they generalise when the labels are real [50]. A model whose generalisation is not well understood is a model whose behaviour off the training distribution cannot be predicted from first principles. The empirical question, e.g., what does the model do when it encounters something unfamiliar, has to be asked

empirically.

Hermann et al. [51] framed a related concern as shortcut learning. Networks frequently latch onto features that are predictive within the dataset but not causally connected to the label. In medical imaging the consequences have been documented in some detail. A pneumonia detection model had partly learned to identify the hospital system from image markers, and that its performance dropped when transferred across institutions [52]. DeGrave, Janizek and Lee [53] reached a similar conclusion for COVID-19 detection from chest radiographs, where models were found to rely on acquisition artefacts rather than pathology. Bani [54] described hidden stratification, in which aggregate performance conceals systematic failure on a clinically important subgroup. These results matter for the present work because they describe a class of failure that a reject option might plausibly catch and a class it will not since a model that has learned a shortcut and encounters that shortcut will be confident and wrong at the same time.

The confidence problem has been studied under the heading of calibration. A classifier is calibrated when its stated confidence matches its empirical accuracy. Among the predictions made with confidence 0.8, roughly 80% should be correct. Platt [55] introduced the scaling method that bears his name for producing probabilistic outputs from margin classifiers. A study [56] compared calibration across model families and found that some produce well-calibrated probabilities and others do not.

Guo et al. [12] brought the question to deep networks and reported that modern architectures are miscalibrated in a particular direction. They are overconfident and the miscalibration grew with depth, width and the removal of regularisation. They measured this with Expected Calibration Error (ECE). Subsequent work complicated the picture. ECE is sensitive to binning choices and that conclusions can shift with the estimator [57]. Also, label smoothing improves calibration while impairing knowledge distillation [58]. A study [59] revisited the question for more recent architectures and found that the trend Guo et al. [12] identified does not hold uniformly which suggests the overconfidence finding is architecture-dependent rather than an inevitable property of scale.

Two observations from this literature bear on the study. First, calibration is a property of the aggregate, e.g, a temperature-scaled model produces confidence values that match accuracy on average across the test distribution which is not the same as producing a trustworthy signal for an individual input and it says nothing about behaviour under shift. Ovadia et al. [16] made this explicit showing that calibration measured in-distribution degrades as the distribution shifts and that the ranking of methods changes with it. Second, calibration corrects the scale of a confidence score without changing what that score is derived from. The maximum softmax probability remains a normalised quantity over classes conditioned on the assumption that one of the classes is correct. Temperature scaling makes it a better-behaved number. This is the point at which the reject-option literature departs from the calibration literature. Rather than improving the confidence estimate, it introduces a separate decision.

## 2.2 Classification with a Reject Option

Chow [20] provides the reference point. Working with Bayes-optimal classifiers and known posteriors, [20] derived the relationship between the rejection rate and the error rate among accepted samples and showed that the optimal rule rejects when the maximum posterior probability falls below a threshold determined by the relative cost of an error against the cost of a rejection. The result is clean and its assumption is the difficulty. The posteriors must be known. For a trained neural network they are estimated by a quantity as an unreliable proxy.

The statistical learning treatment developed later. It has been studied that classification with a reject option in the plug-in framework and established rates of convergence [60]. Bartlett and Wegkamp [22] introduced a modified hinge loss yielding a consistent estimator for the reject-option problem, so that the rejection rule could be learned by empirical risk minimisation rather than imposed afterwards. A study [61] embedded reject options in support vector machines. Cortes, DeSalvo and Mohri [23] formulated learning with rejection as the joint learning of a predictor $h$ and a rejector $r$ giving calibrated surrogate losses and generalisation bounds for the pair. Franc et al. [62] revisited the optimal strategies for reject-option classifiers and clarified the conditions under which different formulations coincide.

El-Yaniv and Wiener [24] introduced the vocabulary this study uses. A selective classifier is a pair $(f, g)$, where $f$ is a prediction function and $g$ a binary selection function. Coverage is the expected value of $g$ and selective risk is the risk of $f$ conditioned on selection. Their analysis of the noise-free case established the risk–coverage tradeoff as the object of study.

Geifman and El-Yaniv [25] brought the framework to deep networks. Their approach takes an already-trained network, defines a confidence-rate function, e.g., softmax response or the negative variance of Monte-Carlo dropout samples and calibrates a threshold on a held-out set so that the selective risk is bounded with a stated probability. The method is practical and requires no retraining and their empirical results showed large risk reductions at moderate coverage on CIFAR and ImageNet.

The limitation is the one already noted. The confidence function is read off a network that was trained for a different objective. Geifman and El-Yaniv [26] addressed this with SelectiveNet which trains the selection function as part of the network. The architecture places three heads on a shared body, e.g., a prediction head $f$, a selection head $g$ producing a scalar in [0,1] and an auxiliary head $h$. The loss combines an empirical selective risk term, e.g., the prediction loss weighted by $g$ and normalised by empirical coverage, with a quadratic penalty on the deviation of empirical coverage from a target $c$ and adds the auxiliary head's standard cross-entropy loss with weight $\alpha$. The auxiliary head exists for a specific reason. Because the selective risk term down-weights the loss on rejected samples, the shared representation receives attenuated gradient from those samples and without $h$ the body can degrade on precisely the region the selector has abandoned. Their reported results showed improvements over the post-hoc approach at low coverage [26].

Several alternatives followed. Deep Gamblers is proposed, which derives an abstention output from portfolio theory [63]. the model is given an extra class whose payoff is a fixed value $o$ and the resulting loss induces abstention without a separate head. Self-adaptive training has

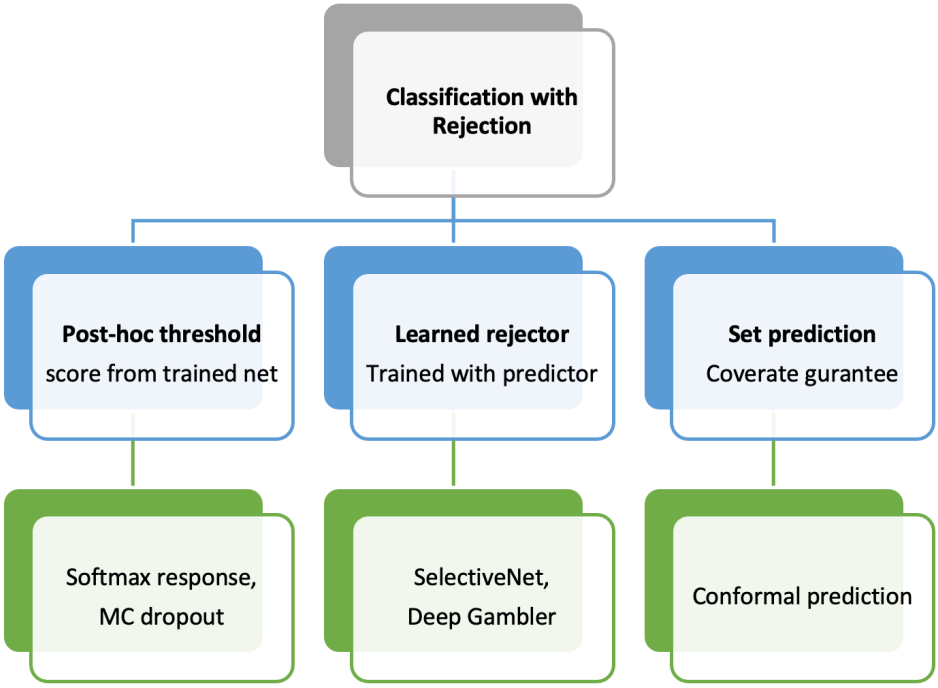



Figure 2.1: Approaches to classification with rejection.

been applied to selective classification and reported gains over both SelectiveNet and Deep Gamblers [63]. Galil et al. [64] evaluated selective prediction across a large collection of pretrained ImageNet classifiers and found substantial variation in selective performance that was not predictable from accuracy alone, which suggests the ranking of models under selective evaluation is not the ranking under standard evaluation.

A separate line treats rejection as deferral. Madras, Pitassi and Zemel [29] framed the problem as learning to defer to a downstream decision-maker and showed that the fairness properties of the combined system differ from those of the model alone. Mozannar and Sontag [65] derived consistent surrogate losses for learning to defer to an expert. The distinction from pure abstention matters for the medical setting considered in this study since a deferral is a referral and consumes a reviewer's attention.

A third framework deserves mention for contrast. Conformal prediction provides distribution-free coverage guarantees by outputting prediction sets rather than single labels with the set size expressing uncertainty [66]. It yields a finite-sample guarantee that selective classification does not. It is not used in this study, but the contrast is instructive. Conformal prediction and selective classification address overlapping problems with different guarantees and different failure conditions.

Figure 2.1 organises these approaches by where the rejection decision is produced. Post-hoc methods threshold a score derived from an already-trained network. Integrated methods learn the rejector jointly with the predictor. SelectiveNet, used in this study, belongs to this family. Distribution-free methods replace the reject decision with a prediction set whose size expresses uncertainty. The families differ in what they guarantee and in what they assume.

Two matters relevant to this study are left open by the work reviewed above. The first

is behaviour under adversarial perturbation. The selective classification papers evaluate on clean test data and characterise the risk–coverage tradeoff [26]. Whether the selection function withdraws under attack is a separate question and it is not one the standard evaluation answers.

The second is what the selection function has learned. SelectiveNet's $g$ is a small network on a shared representation and the literature reports what it achieves without examining what it responds to. The papers do not, as far as this review found, attribute the selection score to input regions. This is the gap that motivates the dual-head explainability component of the present work.

## 2.3 Adversarial Machine Learning

Imperceptible perturbations could reliably change a network's prediction and that the perturbations transferred across models trained on different subsets [13]. Another study [14] provided an explanation grounded in the local linearity of these models and derived the FGSM, a single-step attack that perturbs each pixel by $\varepsilon$ in the direction of the sign of the loss gradient. FGSM is remains widely used as a first-line evaluation.

Adversarial vulnerability arises from non-robust features that are genuinely predictive on the data distribution but imperceptible to humans [43]. On this account adversarial examples are a consequence of the model learning what the data supports rather than a defect in the training procedure which suggests that detection may be more tractable than elimination.

Adversarial training, e.g., augmenting the training set with adversarial examples, remains the most reliable defence [14] and it is costly, both computationally and in clean accuracy. This tradeoff is not an artefact of current methods but reflects a tension between robust and standard accuracy [45].

The detection literature is closer to the mechanism examined here. A detector is attached to subnetwork to intermediate representations [67]. Kernel density is used to estimate and Bayesian uncertainty to separate adversarial inputs from clean ones [68]. A study [69] applied Mahalanobis detector.

This literature bears on the present work in a way that should be stated. If the selection head of a SelectiveNet withdraws under FGSM, that is a detection result and the history above indicates that detection results which have not been tested against an adaptive adversary are provisional. An attacker with knowledge of the architecture could optimise a perturbation that flips $f$ while keeping $g$ above threshold. Whether such an attack succeeds is not tested in this study. The claim available from this study is that the withdrawal behaviour exists under a standard non-adaptive attack across four datasets.

## 2.4 Explainable AI

The gradient of a class score is computed with respect to the input to produce saliency maps [70]. Class Activation Mapping is introduced, using global average pooling to produce class-discriminative localisation maps, at the cost of requiring a specific architecture [71]. This was

generalised as Grad-CAM, weighting the feature maps of a convolutional layer by the gradient of the target score, which removed the architectural constraint and made the method applicable to any CNN [72]. Extensions followed by Grad-CAM++ for multiple instances of a class, and Score-CAM, which replaces gradients with forward-pass perturbation scores [73].

Other families take different routes. Local Interpretable Model-agnostic Explanations (LIME) fits a local surrogate model around an instance. SHapley Additive exPlanations (SHAP) grounds attribution in Shapley values from cooperative game theory [74]. Integrated Gradients accumulates gradients along a path from a baseline, satisfying axioms that plain gradients do not. Testing with Concept Activation Vectors (TCAV) works at the level of human-specified concepts rather than pixels. SmoothGrad averages over noisy inputs to reduce visual noise [75].

Grad-CAM is used in this study for reasons of fit rather than superiority. It requires one backward pass, applies to any convolutional architecture and the property that matters in this study, takes an arbitrary scalar as its target. That last point is what permits the extension to the selection head. Two conclusions follow for this study. First, the visualisations reported are observations rather than evidence of what the network computes since no faithfulness test is applied. Second, the argument for applying Grad-CAM to the selection head does not depend on Grad-CAM being a good explanation method. It depends on the observation that whatever explanation one applies to a selective classifier ought to cover the selection decision as well as the prediction. If a better method than Grad-CAM is available, the same argument applies to it.

## 2.5 Responsible AI as a Normative Frame

The guideline literature is extensive and has been surveyed. Jobin, Ienca and Vayena [32] analysed eighty-four documents and found convergence on transparency, justice and fairness, non-maleficence, responsibility and privacy, alongside divergence in how these were interpreted and in what was recommended to achieve them. Floridi and Cowls [33] consolidated the landscape into five principles, adapting four from bioethics and adding explicability. The European Commission's High-Level Expert Group [34] set out requirements for Trustworthy AI including human agency and oversight, technical robustness and safety, transparency, and accountability and elements of this framing were carried into Regulation (EU) 2024/1689, whose Article 14 requires human oversight for high-risk systems. A study [76] approached the same territory from an engineering direction listing concrete problems in AI safety. Two of their categories are relevant here, e.g., robustness to distributional shift and scalable oversight, which concerns how a system operating faster than human review can nonetheless be supervised. A mechanism that identifies which cases require human attention is a partial answer to the second.

Fairness has the most developed technical literature of the principles listed and the least connection to the work reported here. Madras, Pitassi and Zemel [29] connect this literature to the present subject, e.g., a deferral policy has fairness consequences of its own because the burden of deferred cases may fall on a particular subgroup even where the accepted-case performance appears balanced.

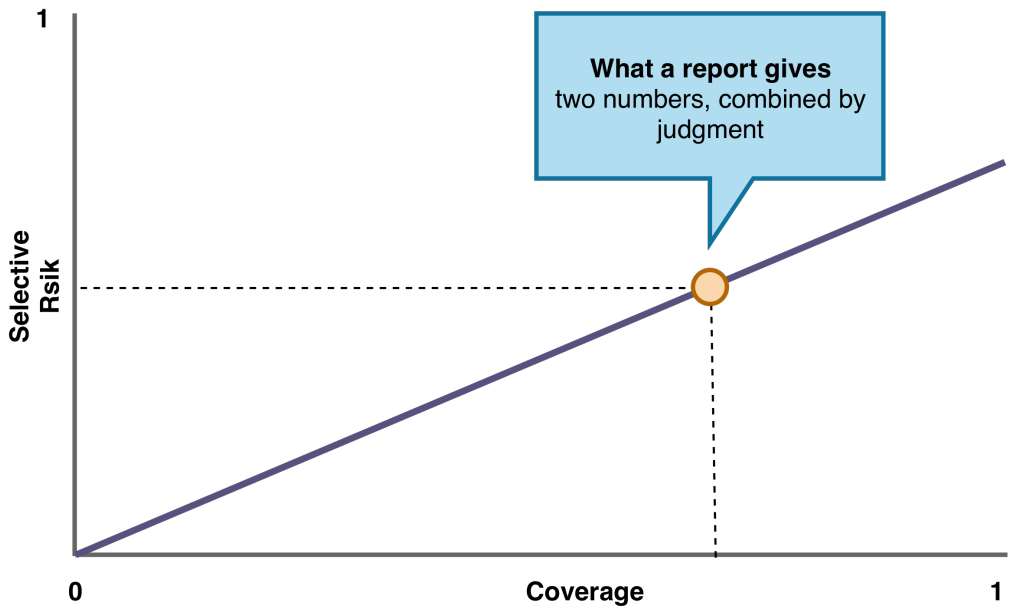



Figure 2.2: The risk–coverage tradeoff, sketched.

The observation that motivates this study's framing is that the guideline literature specifies outcomes and rarely specifies mechanisms. A requirement for human oversight does not say how the system should identify the cases a human ought to see. A requirement for transparency does not say what should be made transparent when the output is a refusal. Bridging that gap requires connecting each principle to an architectural component and then testing whether the component behaves as the connection assumes which is what the empirical chapters of this study attempt for four of the five principles.

## 2.6 Evaluation of Selective Classifiers

Classification evaluation rests on the confusion matrix and the measures derived from it, e.g., accuracy, precision, recall, specificity, F1. Naidu et al. [77] reviews these and their relationships. Balanced accuracy, the mean of recall and specificity, is the appropriate aggregate under class imbalance, and precision–recall curves are more informative than Receiver Operating Characteristic (ROC) curves on imbalanced data [78].

For selective classifiers the standard reporting adds coverage and computes the performance measures on the accepted subset [46]. Selective risk is the risk of f conditioned on $g$ accepting, and the risk–coverage curve traces this across thresholds [24, 25]. Area under the risk–coverage curve is sometimes reported as a scalar summary. Accuracy–rejection curves present the same tradeoff with accuracy on the vertical axis.

Figure 2.2 sketches the curve and marks what accepted-only reporting takes from it.

Selective risk falls as the model accepts fewer inputs since the samples it declines are the ones it was least likely to classify correctly. A conventional report selects one operating point on this curve and gives the risk and the coverage as two separate figures leaving the reader to

combine them. The curve characterises a single model well but it does not yield a scalar that can be placed beside a non-abstaining baseline and it does not prevent a low-coverage model from reporting a favourable risk.

The limitation has been noticed and addressed in part. Condessa, Bioucas-Dias and Kovačević [47] proposed three measures for classification with rejection, e.g., non-rejected accuracy (the accuracy on accepted samples), classification quality (the proportion of samples either correctly classified and accepted or misclassified and rejected) and rejection quality, comparing the concentration of misclassified samples in the rejected region against their concentration overall. Classification quality is the closest existing measure to the framework proposed in this study, since it evaluates over the whole sample set and credits a rejection when it falls on a sample that would have been misclassified.

The difference is in what is being asked. Classification quality treats a rejection of a would-be error as a success, which measures how well the rejector identifies the classifier's mistakes. It is not the question this study asks. From the standpoint of a system that must deliver a decision, a rejected sample is a decision not delivered, whether or not the classifier would have got it wrong. The Abstention-Aware framework charges the rejection in both cases. The two measures answer different questions and Chapter 3 states the difference rather than claiming that one supersedes the other. Franc, Průša and Voráček [62] work within an expected-cost formulation where rejection carries an explicit cost parameter which is a principled treatment when that cost is known. The framework proposed here fixes an implicit equal weighting instead.

Gathering these observations, the position this study takes is as follows. Accepted-only metrics discard samples and, therefore, charge nothing for abstention, with the arithmetic consequence that the reported score improves as coverage falls. Coverage is reported alongside but does not enter the calculation. Risk–coverage curves characterise a model across thresholds but do not yield a scalar for comparison with a non-abstaining baseline at a chosen operating point. The measures of Condessa et al. [47] do evaluate over the full sample set, but they credit rejections that fall on would-be errors, which answers a question about gate quality rather than about delivered performance. Cost-sensitive formulations require a cost parameter that is often unavailable.

What appears to be missing is a measure that treats abstention as a prediction outcome rather than as grounds for exclusion, charges it uniformly and reduces exactly to the classical metric when the model does not abstain, with those properties stated and verified rather than assumed. The reduction property is the one that makes a single-table comparison of a baseline CNN and a selective model possible and it is not, as far as this review established, a property that existing rejection-aware measures were designed to have.

## 2.7 Identified Research Gaps

The review supports four observations, each corresponding to a research question set out in Chapter 1.

Adversarial behaviour of learned selection functions is under-examined. The selective classi-

fication literature evaluates on clean data. The adversarial detection literature evaluates detectors built for that purpose. Whether SelectiveNet's selection head withdraws under attack and whether that behaviour is consistent across datasets, is not something either literature reports.

The cost of intrinsic detection has not been weighed against the relabelling objection. OOD detection methods add machinery, a learned selection head does not. But the argument that abstention improves a system requires ruling out the possibility that it only moves errors into a different column and that comparison requires an evaluation that displays error cells and abstention cells for the same population.

Explanation covers the prediction and not the refusal. Grad-CAM and its relatives are applied to class scores. When the model can decline, the decline is generated by a different head and is left unexplained, which weakens the accountability argument that abstention is supposed to support.

Evaluation practice does not charge for abstention. This is the central gap. Existing measures either discard the abstained samples or, where they retain them, credit rejections that fall on would-be errors. A measure that keeps every sample, charges abstention uniformly and reduces to the classical metric at zero abstention appears not to exist in the form required and its properties have not been established.

Chapter 3 addresses the fourth gap directly by developing the AAEF and Chapters 4 and 5 address the first three through the experimental programme.

# Chapter 3

# The Abstention-Aware Evaluation Framework (AAEF)

This chapter develops the principal theoretical contribution of the study. Chapter 2 concluded that the evaluation practice for selective classifiers rests on a convention, computing performance over the accepted subset and reporting coverage as a separate figure.

The Abstention-Aware Evaluation Framework, referred to throughout as AAEF, is an attempt to correct that convention while remaining compatible with the measures practitioners already use.

The chapter is placed before the methodology for two reasons. The framework can be read on its own since it is a statement about evaluation rather than about any particular architecture and it applies to any classifier with a reject option regardless of how that option is implemented. It also determines what the experimental chapters record, e.g., the quantities in Chapter 3's evaluation protocol are the quantities the framework requires and reading the methodology before the framework would leave several of those choices unmotivated.

## 3.1 Notation and Setup

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote an evaluation set of $N$ samples drawn from an input space $\mathcal{X}$ and a label space $\mathcal{Y}$. For the binary case, $\mathcal{Y} = \{0, 1\}$, with class 1 designated positive and class 0 negative. The multiclass case, $\mathcal{Y} = \{1, \ldots, n\}$, is treated in sub-sequent section.

Following [24], a *selective classifier* is a pair $(f, g)$ where $f : \mathcal{X} \to \mathcal{Y}$ is a prediction function and $g : \mathcal{X} \to \{0, 1\}$ a selection function. The selective classifier outputs

$$(f, g)(x) = \begin{cases} f(x), & g(x) = 1, \\ \varnothing, & g(x) = 0, \end{cases} \tag{3.1}$$

where $\varnothing$ denotes abstention. In practice $g$ is realised as a soft score $\tilde{g} : \mathcal{X} \to [0, 1]$ thresholded at some $\tau$. In SelectiveNet [26], $\tilde{g}$ is the output of a learned head. In post-hoc methods it is a confidence score read from a trained network. The framework developed here does not depend

on which, and the notation is deliberately agnostic about the source of $\tilde{g}$.

*Empirical coverage* is the fraction of samples accepted,

$$\hat{c} \;=\; \frac{1}{N}\sum_{i=1}^{N} g(x_i), \tag{3.2}$$

and *empirical selective risk* is the risk of $f$ conditioned on acceptance,

$$\hat{r}(f,g) \;=\; \frac{\sum_{i=1}^{N} \ell\big(f(x_i), y_i\big)\, g(x_i)}{\sum_{i=1}^{N} g(x_i)}, \tag{3.3}$$

for a loss $\ell$, taken here as the 0–1 loss. The accepted set is written

$\mathcal{A} = \{i : g(x_i) = 1\}$ and the abstained set $\mathcal{R} = \{i : g(x_i) = 0\}$, so that $\mathcal{A} \cup \mathcal{R} = \{1, \ldots, N\}$ and $\mathcal{A} \cap \mathcal{R} = \varnothing$.

Throughout, $P$ denotes the number of positive samples in $\mathcal{D}$ and $N_{neg}$ the number of negative samples, with $P + N_{neg}$. Classical measures are written with their usual definitions, e.g., recall (sensitivity) $= TP/(TP + FN)$, specificity $= TN/(TN + FP)$ and balanced accuracy as their mean. Where a measure is computed over $\mathcal{A}$ alone it carries the subscript *acc*, thus, recall $_{acc}$, spec $_{acc}$, BA $_{acc}$, to distinguish it from the abstention-aware version.

The standard response is the risk-coverage curve, which traces $\hat{r}$ against $\hat{c}$ across thresholds, sometimes summarised by its area [25]. The curve is the right instrument for characterising a single model's tradeoff and this study reports it.

It does not resolve the reporting problem. It yields no scalar comparable with a non-abstaining baseline, which has no curve at all, only a point at $\hat{c} = 1$. Its area weights all coverage levels equally, whereas a deployed system operates at one. [46] argued that common evaluation practice for selective classification systems contains flaws that cause published comparisons to be misleading, among them the confounding of classifier accuracy with the quality of the confidence score, which suggests that the difficulties are not confined to how the curve is summarised.

In a survey of machine learning with a reject option [61], note the absence of a settled evaluation protocol across the subfield as one of its open problems. The framework proposed below is an attempt at one part of that problem, not the whole of it.

## 3.2 The Abstention-Aware Confusion Matrix

The framework's starting move is to stop treating abstention as an exclusion criterion and treat it instead as a third prediction outcome, on the same footing as a positive or negative prediction. The binary confusion matrix acquires a third column, as shown in Figure 3.1.

The four familiar cells retain their meanings and are computed over $\mathcal{A}$ only. The two new cells are

- *AP — Abstained Positive*: samples with $y_i = 1$ and $g(x_i) = 0$;
- *AN — Abstained Negative*: samples with $y_i = 0$ and $g(x_i) = 0$.

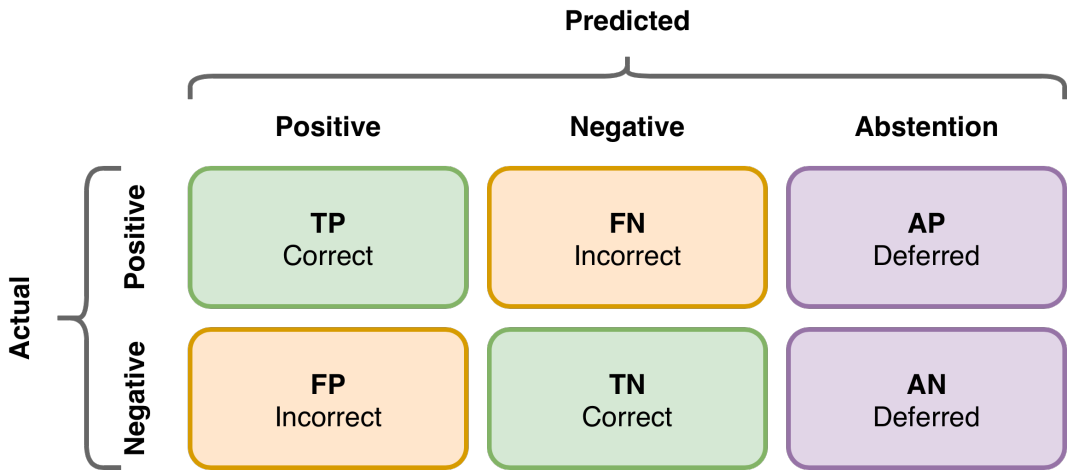



Figure 3.1: The abstention-aware confusion matrix. The four classical cells are computed over the accepted set and the two additional cells hold the abstained samples which conventional selective evaluation discards.

Since every sample has exactly one label and is either accepted or rejected, the six cells partition $\mathcal{D}$:

$$TP + FN + AP + FP + TN + AN = N. \tag{3.4}$$

Equation (3.4) is the *conservation identity* and it is the property that separates the framework from accepted-only evaluation. It also serves as an implementation check, e.g., any discrepancy in the sum indicates a bug in the cell assignment which is not a trivial matter in practice given that abstention is often implemented by masking rather than by explicit assignment.

Two marginal relations follow and are used later:

$$TP + FN + AP = P; \qquad FP + TN + AN = N_{neg}. \tag{3.5}$$

The row sums are the class supports and they do not depend on the model. This is what makes the metrics comparable across models operating at different coverage levels.

Figure 3.2 shows how the six cells are populated and where conventional evaluation draws its boundary. The selection function splits the $N$ samples into an accepted set which populates the four classical cells and an abstained set which populates $AP$ and $AN$. Accepted-only evaluation computes its scores from the upper path and reports the size of the lower path as coverage. AAEF computes its scores from both keeping the lower path in the denominator so that abstention lowers the reported value rather than leaving it untouched.

Three following measures are defined on the matrix.

### 3.2.1 Abstention-Aware Recall (AAR)

$$AAR = \frac{TP}{TP + FN + AP} = \frac{TP}{P}. \tag{3.6}$$

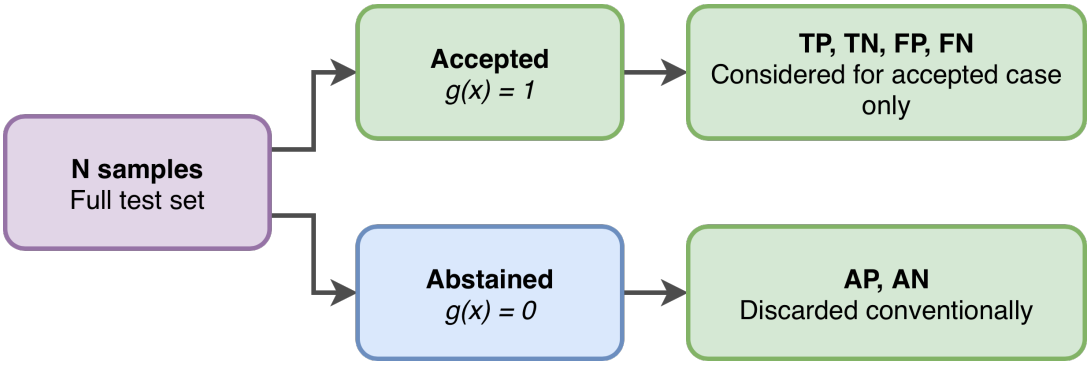



Figure 3.2: Partition of the evaluation set under the two conventions.

$AAR$ is the proportion of actually positive cases that the system classified as positive without abstaining. The denominator is the class support which does not vary with the operating point.

### 3.2.2 Abstention-Aware Specificity (AAS)

$$AAS = \frac{TN}{TN + FP + AN} = \frac{TN}{N_{neg}}. \tag{3.7}$$

$AAS$ is the corresponding quantity for the negative class.

### 3.2.3 Abstention-Aware Balanced Accuracy (AABA)

$$AABA = \frac{AAR + AAS}{2}. \tag{3.8}$$

$AABA$ weights the two classes equally irrespective of prevalence, following the rationale for balanced accuracy under class imbalance. This matters for the imbalanced dataset and an unweighted aggregate would be dominated by the majority.

## 3.3 Multiclass Generalisation

For $\mathcal{Y} = \{1, \ldots, n\}$ the extension is direct. The $n \times n$ confusion matrix acquires an $(n+1)$-th column holding abstentions giving an augmented matrix $M \in \mathbb{Z}^{n \times (n+1)}$. Row sums give the class supports, and the total sums to $N$ (shown in Figure 3.3).

The per-class metrics $AAR_k$ and $AAS_k$ follow from (3.6) and (3.7), and the macro averaged versions are their means over $k$:

$$AAR_{\text{macro}} = \frac{1}{n} \sum_{k=1}^{n} AAR_k, \qquad AAS_{\text{macro}} = \frac{1}{n} \sum_{k=1}^{n} AAS_k. \tag{3.9}$$

Macro averaging is used rather than weighted averaging (it treats classes as equally important) which is the intent when reporting balanced measures. Classes with zero support give 0/0

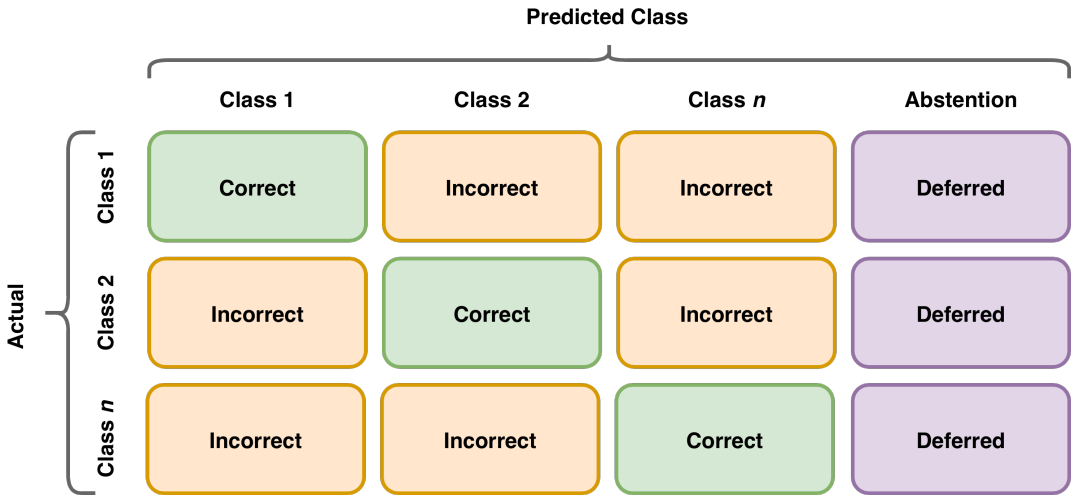



Figure 3.3: The augmented $n \times (n+1)$ confusion matrix.

and are excluded from the average with the exclusion noted. This does not arise on the datasets used here but is stated for completeness.

Micro-averaged has a direct reading:

$$AAR_{\text{micro}} = \frac{\sum_{k=1}^{n} TP_k}{\sum_{k=1}^{n} P_k} = \frac{\text{number correct and accepted}}{N}. \tag{3.10}$$

This is accuracy over the full test set with abstentions counted as errors. It is the most interpretable single number the framework produces for a multiclass problem and it is the quantity to compare against a non-abstaining baseline's accuracy.

## 3.4 Theoretical Properties

Four properties are established. Each was chosen because a measure lacking it would be unusable for the purpose at hand.

### 3.4.1 Boundedness

$0 \leq AAR \leq 1$, $0 \leq AAS \leq 1$, and $0 \leq AABA \leq 1$.

*Proof.* All six cells are non-negative counts. By (3.5), $AAR = TP/P$ with $0 \leq TP \leq P$, giving $0 \leq AAR \leq 1$; the argument for $AAS$ is identical. $AABA$ is the mean of two quantities in $[0, 1]$ and lies in $[0, 1]$. □

### 3.4.2 Reduction

If $AP = AN = 0$, then $AAR =$ recall, $AAS =$ specificity, and $AABA =$ balanced accuracy. In the multiclass case, $AAR_{\text{macro}}$ reduces to macro recall (and, thus, to balanced accuracy), and $AAR_{\text{micro}}$ reduces to accuracy.

*Proof.* Setting the abstention terms to zero in (3.6) and (3.7) recovers the classical definitions; (3.8) then recovers balanced accuracy. In the multiclass case, $AP_k = 0$ for all $k$ gives $AAR_k = TP_k/(\ TP_k + \ FN_k) = \text{recall}_k$, and (3.10) becomes the proportion correct. □

This carries most of the practical weight. Without it, the abstention-aware scores and the classical scores would occupy different scales, and placing a non-abstaining baseline beside a selective model in one table would have no defensible meaning.

### 3.4.3 Extreme cases

*(i)* A classifier that is correct on every sample and abstains on none attains $AAR = \ AAS = AABA = 1$. *(ii)* A classifier that abstains on every sample attains $AAR = \ AAS = \ AABA = 0$. *(iii)* A classifier that predicts incorrectly on every sample and abstains on none attains $AAR = \ AAS = \ AABA = 0$.

*Proof.* (i) $TP = P$, $TN = \ N_{neg}$, abstention cells zero. (ii) $AP = P$, $AN = \ N_{neg}$, so $TP = \ TN = 0$. (iii) $TP = \ TN = 0$ with zero abstention. □

Case (ii) is the one that distinguishes AAEF from accepted-only evaluation, where a model rejecting everything yields an undefined score rather than a zero and where the limiting behaviour as coverage approaches zero is towards one rather than towards zero.

## 3.5 Limitations of the Framework

Several limitations are recorded rather than deferred since they bear on how the results of they study should be read.

**Equal weighting of abstention and error:** AAEF charges an abstention the same as a misclassification of that class. Where the true costs differ and in most deployments they do since a referral consumes review time while a misclassification produces a wrong outcome, the framework understates the value of abstention.

**Scope:** The framework assumes a classification task with a single abstention action and does not extend as stated to regression or segmentation.

**No account of downstream human performance:** AAEF scores an abstention as a non-delivery which implicitly assumes the deferred case receives no benefit from the system. In a deployment with a human reviewer the case is decided and often decided well. Measuring the combined human–model system is a different problem.

# Chapter 4

# Research Methodology

## 4.1 Overview and Design Principles

This chapter describes the experimental programme through which the research questions of Chapter 1 are addressed. The design is organised around four datasets, each treated as a self-contained experiment with an identical structure, e.g., a baseline convolutional classifier is trained, a SelectiveNet sharing the same convolutional body is trained, both are evaluated on clean test data, both are subjected to a fast-gradient adversarial sweep, the selective model's two heads are explained with Grad-CAM and both models are scored under the conventional and abstention-aware conventions.

Three principles guided the design. Those are:

The first is matched comparison. Any difference in behaviour between the two models should be attributable to the selection mechanism rather than to representational capacity. So, the SelectiveNet body reproduces the baseline body layer for layer, with the heads added on top. The models are trained separately rather than sharing weights. Therefore, the comparison is between two independently trained networks of matched architecture rather than between two readouts of the same network.

The second is variation across task structure. The four datasets differ in class count, image resolution, colour depth, class balance and the operational meaning of a deferral. A behaviour observed on one dataset might be an artefact of that task. A behaviour observed on all four is more likely to be a property of the mechanism.

The third is reproducibility of the reported numbers. Seeds are fixed, splits are derived from a stated procedure rather than an ad hoc partition.

## 4.2 Computational Environment and Software

All experiments were implemented in Python and executed in Google Colab notebooks with Graphics Processing Unit (GPU) acceleration [79]. Models were built and trained with TensorFlow through the Keras Application Programming Interface (API) [80]. Numerical work used NumPy, evaluation metrics were computed with scikit-learn where classical counterparts

Table 4.1: Comparison of datasets used for selective classification experiments

| **Property** | **MNIST** | **CIFAR-10** | **Chest X-ray** | **Cats vs. Dogs** |
|---|---|---|---|---|
| **Task** | 10-class | 10-class | Binary | Binary |
| **Input Size** | $28 \times 28 \times 1$ | $32 \times 32 \times 3$ | $150 \times 150 \times 3$ | $128 \times 128 \times 3$ |
| **Class Balance** | Near-uniform | Uniform | Imbalanced | Near-balanced |
| **Source** | Kaggle [82] | Kaggle [83] | Kaggle [84] | Kaggle [85] |
| **Deferral Meaning** | Abstain | Abstain | Referral | Abstain |

were required and figures were produced with Matplotlib [81].

One notebook was written per dataset, each following the same section ordering so that results are comparable across them. A shared plotting and palette module is defined at the top of each notebook to keep figure styling consistent which matters for the visual comparisons in sub-sequent chapter.

Random seeds were fixed where the API permits, e.g., dataset splitting, shuffling and the random generators but GPU kernel non-determinism means that a rerun will not reproduce weights bit for bit.

## 4.3 Datasets and Preprocessing

The four datasets and their handling are summarised in Table 4.1.

### 4.3.1 MNIST

MNIST [82] has standard 60,000 training and 10,000 test images. Pixel values were scaled to the unit interval and the images reshaped to include an explicit channel dimension. Labels were one-hot encoded which the selective loss requires since it computes cross-entropy directly rather than through a Keras loss object.

MNIST is included as a low-difficulty reference point. A small CNN reaches high accuracy on it. So, the accepted-set metrics have little room to vary and the interesting behaviour lies in the pattern of abstentions and in how the selection head responds to perturbation.

### 4.3.2 CIFAR-10

CIFAR-10 [83] has 50,000 training and 10,000 test colour images across ten classes, scaled to the unit interval, labels one-hot encoded, with a tenth of the training set reserved for validation. No data augmentation was applied. This limits the accuracy the baseline reaches and is a deliberate simplification. Augmentation would improve both models and complicate the attribution of any difference between them but it does mean the CIFAR-10 results sit below what the literature reports for well-tuned models of comparable depth.

### 4.3.3 Chest X-ray (Normal vs. Pneumonia)

The chest radiograph dataset was obtained through Kaggle [84]. The response was to pool all images from the distributed train, validation and test folders and re-split them. Splitting was performed per class, so that the proportions are preserved in each partition, at ratios of 80% training, 10% validation and 10% test with a fixed random seed. Splitting was at the file level before any images were loaded.

Images were loaded at $150 \times 150$ pixels through the directory-based dataset API, with class names given so that NORMAL maps to 0 and PNEUMONIA to 1 across all splits. Pixel values were rescaled to the unit interval through a preprocessing layer and datasets were cached and prefetched.

The class imbalance, e.g., pneumonia cases outnumber normal cases by roughly three to one in this dataset, was addressed with class weights computed as the total training count divided by twice the count for that class, the standard inverse-frequency weighting.

### 4.3.4 Cats vs. Dogs

The Cats vs. Dogs dataset, originating from the Asirra CAPTCHA work, was downloaded from the Microsoft distribution point as a compressed archive [85]. The pipeline checks the archive size and tests its integrity before extraction since a truncated download produces confusing downstream errors.

The dataset contains a number of corrupt or non-image files. Each file was opened in binary mode and checked for the JPEG File Interchange Format (JFIF) marker in its header. Files failing the check were deleted, and the count of removed files is reported in the notebook output. Without this step the training pipeline raises decoding errors partway through an epoch.

Images were loaded at $128 \times 128$ pixels. Pixel values were scaled to the unit interval through a mapping function, and labels were one-hot encoded to match the two-unit softmax output.

## 4.4 Baseline Architectures

A baseline CNN was defined per dataset, scaled to the task rather than held constant across tasks. Constant architecture across four datasets of differing resolution and difficulty would have meant either an over-parameterised model on MNIST or an under-parameterised one on the higher-resolution sets.

### 4.4.1 MNIST

Three convolutional layers with 32, 64 and 64 filters of size $3 \times 3$ with Rectified Linear Unit (ReLU) activation, the first two followed by $2 \times 2$ max pooling, then flattening, a 64-unit dense layer, dropout at 0.5 and a ten-unit softmax output. The architecture is illustrated in Figure 4.1.

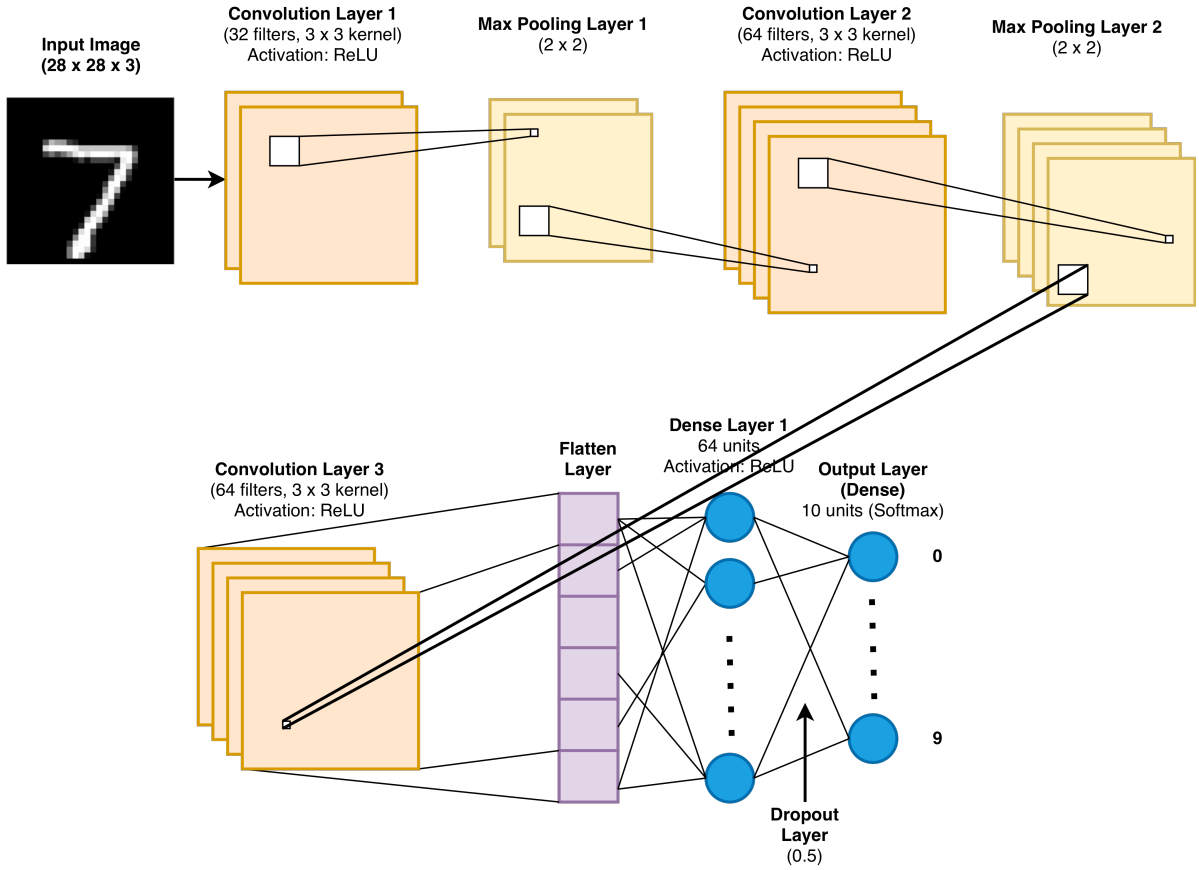



Figure 4.1: The architecture of CNN used to classify MNIST dataset.

### 4.4.2 CIFAR-10

A deeper Visual Geometry Group (VGG)-style stack: two 32-filter convolutions with batch normalisation and 2 × 2 pooling followed by dropout at 0.25; two 64-filter convolutions with the same structure and dropout at 0.3; one 128-filter convolution with pooling and dropout at 0.4; then flattening, a 128-unit dense layer, dropout at 0.5 and a ten-unit softmax output. All convolutions use same padding. The architecture is illustrated in Figure 4.2.

### 4.4.3 Chest X-ray

Four convolutional layers with 32, 64, 128 and 128 filters, each followed by 2 × 2 max pooling, then flattening, a 128-unit dense layer, dropout at 0.5 and a single sigmoid output for binary classification. The architecture is illustrated in Figure 4.3.

### 4.4.4 Cats vs. Dogs

Four convolutional blocks with 32, 64, 128 and 128 filters, each with batch normalisation and 2 × 2 pooling, followed by global average pooling rather than flattening, a 128-unit dense layer, dropout at 0.5, and a two-unit softmax output. Global average pooling was used here because flattening a 128-channel feature map at this input resolution produces a dense layer large enough to dominate the parameter count. The architecture is illustrated in Figure 4.4.

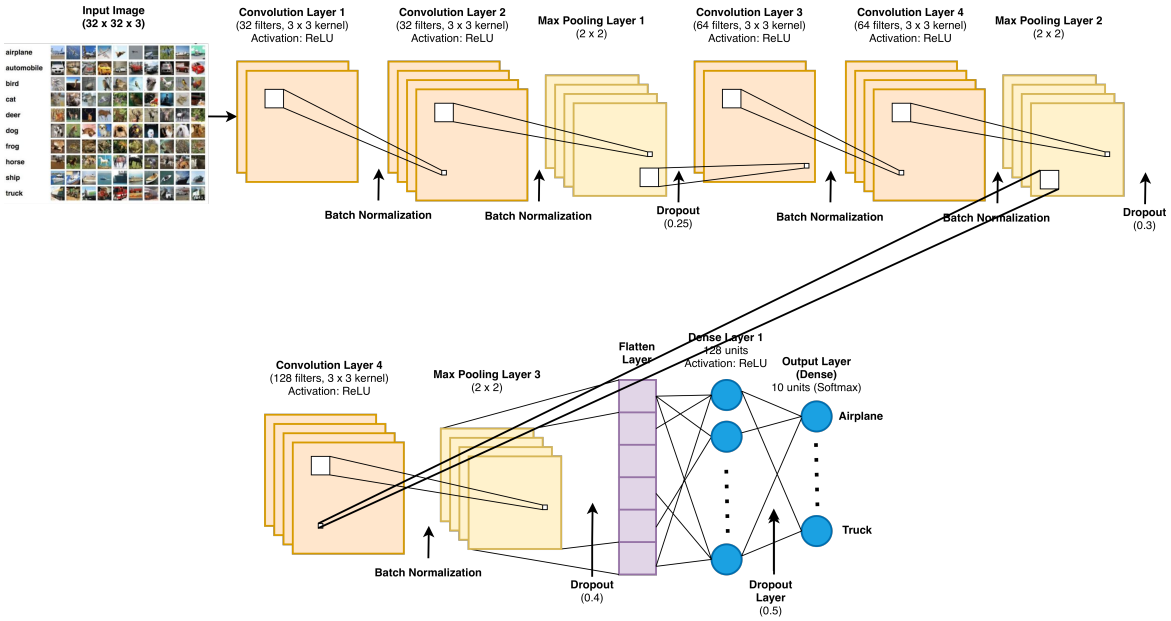



Figure 4.2: The architecture of CNN used to classify CIFAR-10 dataset.

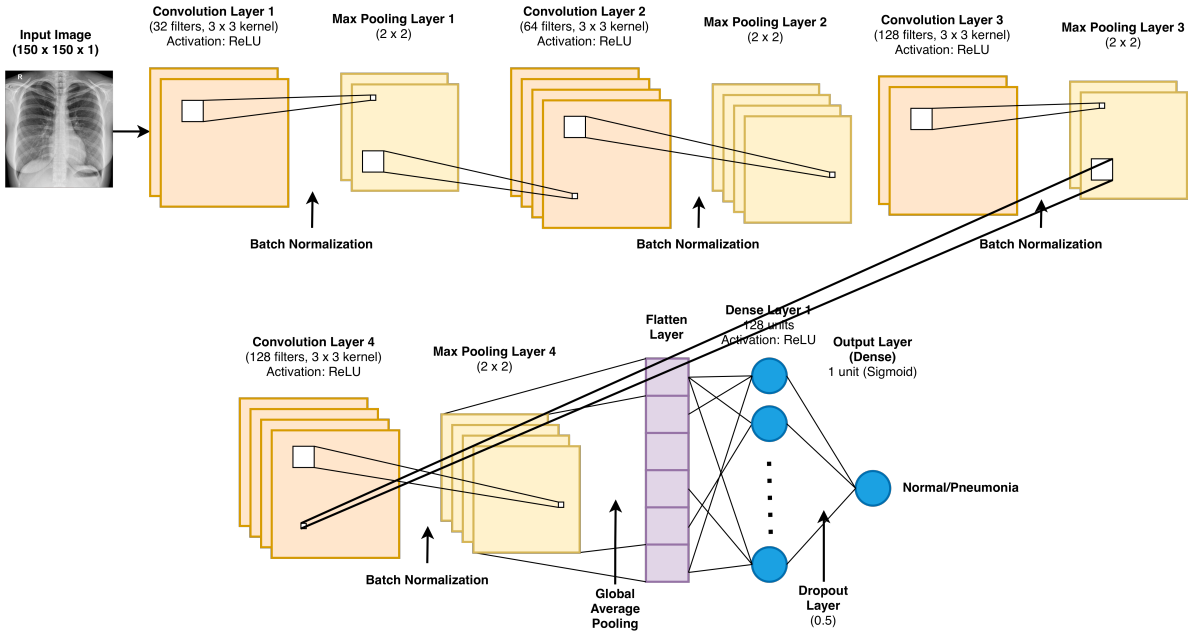



Figure 4.3: The architecture of CNN used to classify Chest X-ray dataset.

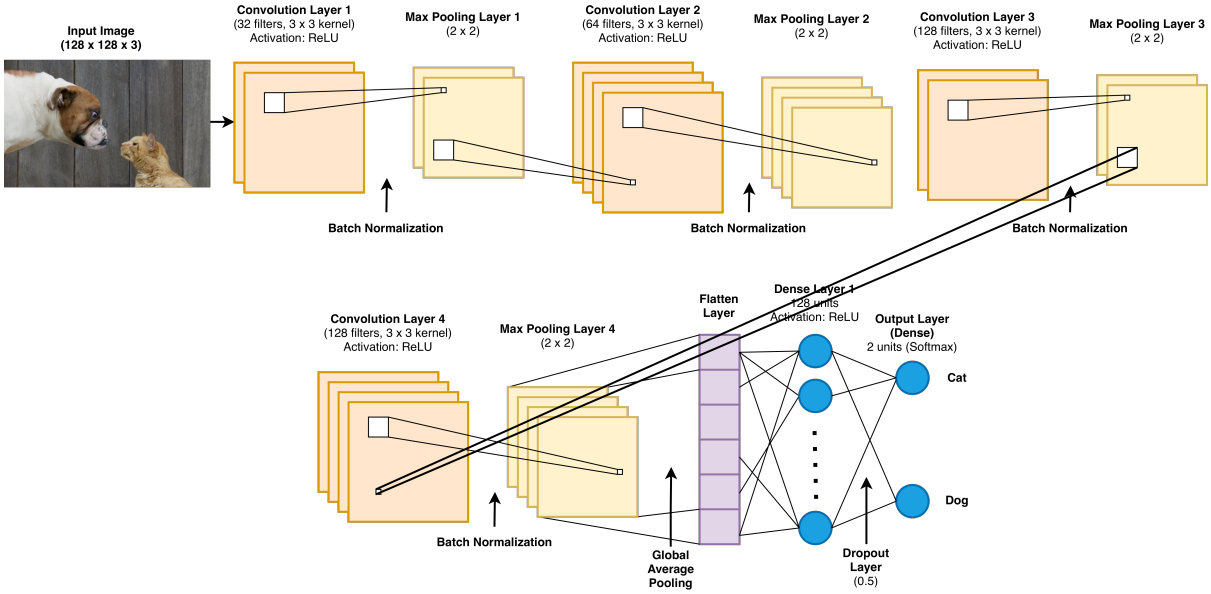



Figure 4.4: The architecture of CNN used to classify Cat vs. Dog dataset.

All four were compiled with the Adam optimiser at its default settings. The multiclass models use categorical cross-entropy. The chest model uses binary cross-entropy with class weighting. The Cats vs. Dogs model uses categorical cross-entropy over a two-class softmax rather than a sigmoid, which keeps its output shape consistent with the multiclass notebooks and simplifies the shared AAEF code.

## 4.5 SelectiveNet Implementation

### 4.5.1 Architecture

Each selective model reproduces its baseline's convolutional body and dense feature layer then attaches three heads to the resulting feature vector, following [26] (also depicted in Figures 4.6,4.7, 4.8 and 4.9):

- **Prediction head** $f(x)$ : a 64-unit ReLU dense layer followed by a softmax (or sigmoid, for the chest model) output over the classes.
- **Selection head** $g(x)$ : a 64-unit ReLU dense layer, batch normalisation, then a single sigmoid unit producing a score in the unit interval.
- **Auxiliary head** $h(x)$ : a 64-unit ReLU dense layer followed by a class-sized softmax output.

The batch normalisation layer in the selection head follows the original paper [26], where it stabilises the selection score during the early epochs in which the coverage penalty dominates the gradient.

The three outputs are concatenated into a single tensor, e.g., the prediction distribution, the selection score and the auxiliary distribution in that order, as shown in Figure 4.5. The

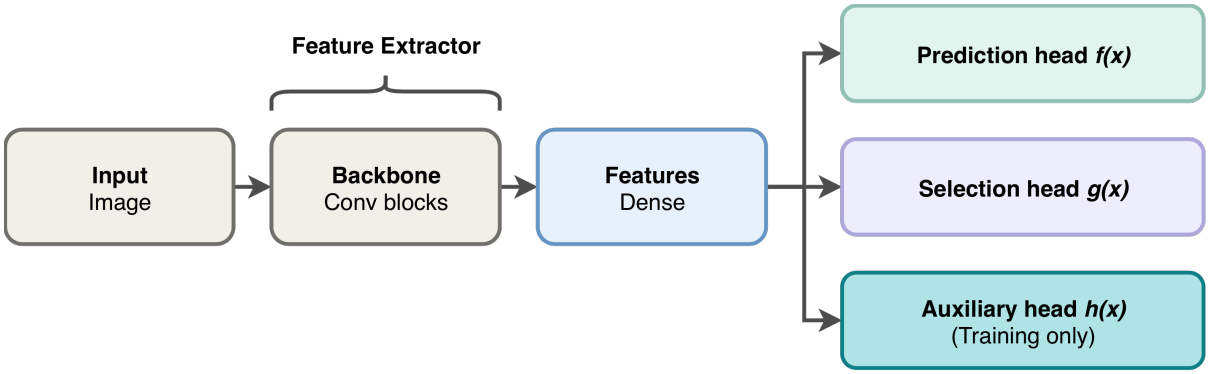



Figure 4.5: SelectiveNet as implemented.

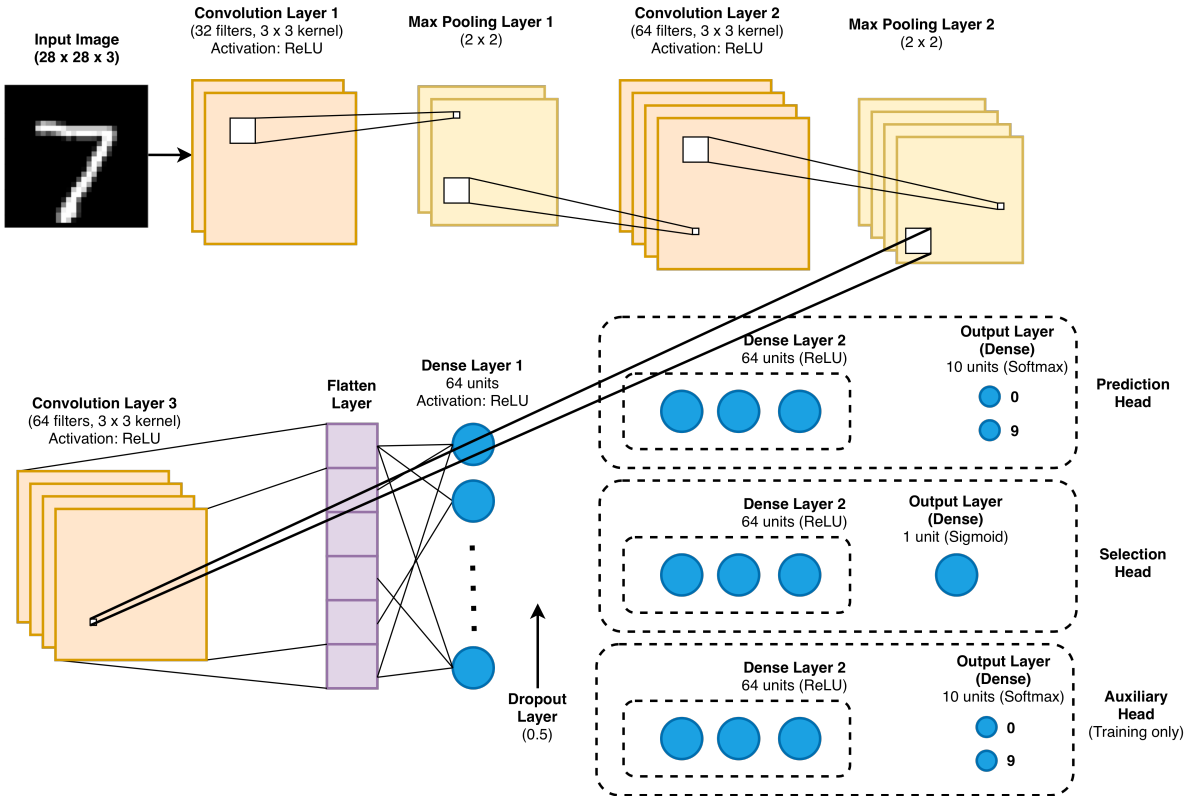



Figure 4.6: The architecture of SelectiveNet used to classify MNIST dataset.

convolutional backbone and the dense feature layer reproduce the corresponding baseline model exactly. The three heads are added on top. The prediction head supplies the class distribution, the selection head a scalar accept score and the auxiliary head a second class distribution that participates in training but is not used at inference. This is an implementation convenience rather than a design choice. Keras loss functions receive a single prediction tensor and concatenation avoids the multi-output loss plumbing but it means every loss and metric begins by slicing the tensor into its three parts and the slicing indices are one of the places where a silent error would be easy to introduce.

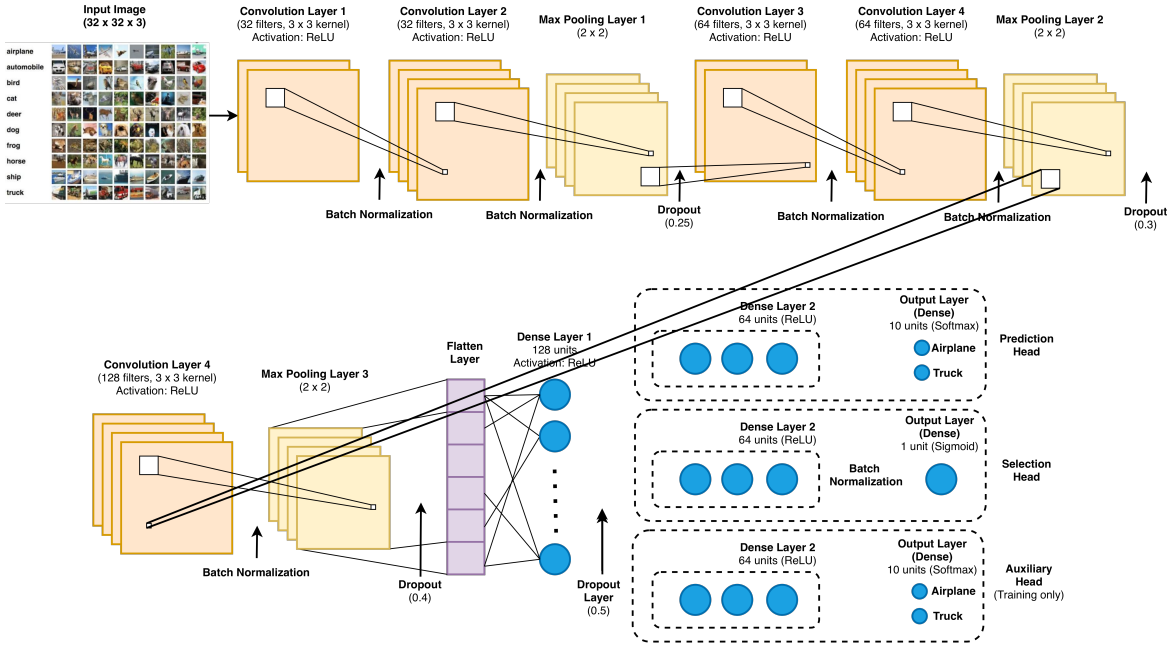



Figure 4.7: The architecture of SelectiveNet used to classify CIFAR-10 dataset.

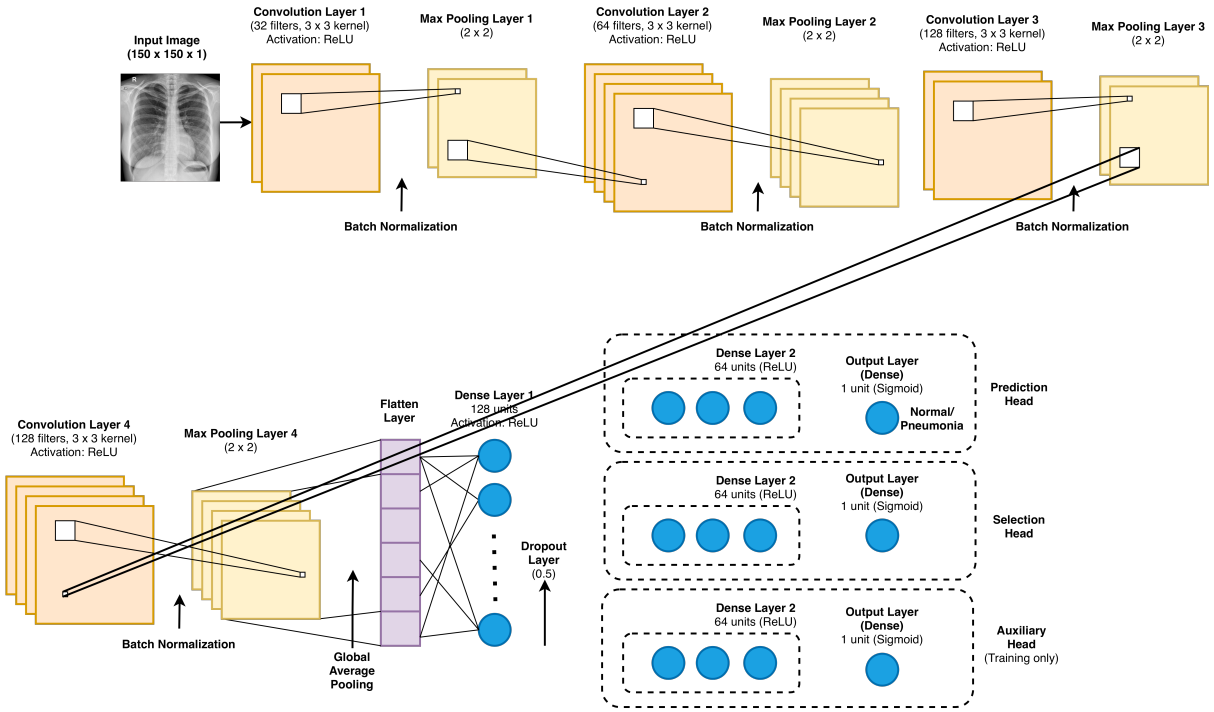



Figure 4.8: The architecture of SelectiveNet used to classify Chest X-ray dataset.

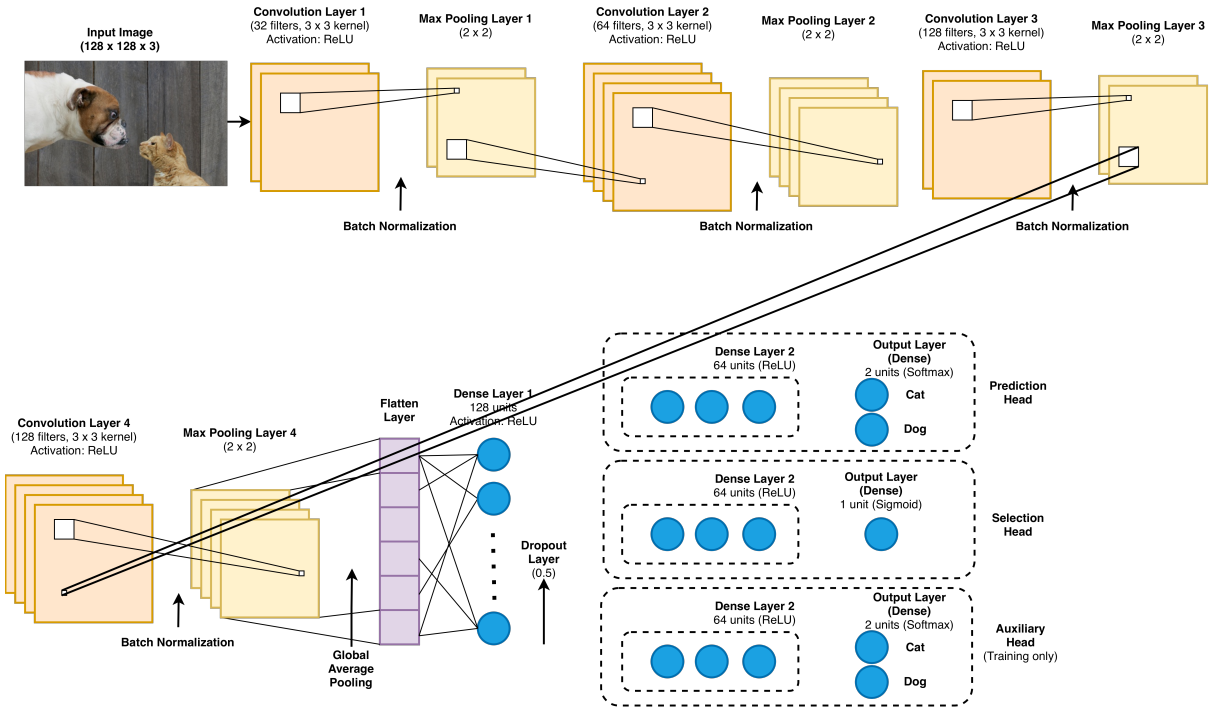



Figure 4.9: The architecture of SelectiveNet used to classify Cat vs. Dog dataset.

### 4.5.2 Loss function

The combined objective follows the original formulation [26]. Writing $\mathcal{L}_{\text{sel}}$ for the selective loss, $\mathcal{L}_{\text{aux}}$ for the auxiliary loss, $c$ for the target coverage and $\lambda$ for the penalty weight. Empirical coverage is the mean of the selection score over the batch. Selective risk is the sum of per-sample cross-entropy weighted by the selection score, divided by the sum of the selection scores with a small constant added to the denominator for numerical stability. $\mathcal{L}_{\text{sel}}$ is selective risk plus coverage penalty. $\mathcal{L}_{\text{aux}}$ is ordinary categorical cross-entropy computed on the auxiliary head. Combined loss is as follows [26]:

$$\mathcal{L} \;=\; \alpha\,\mathcal{L}_{\text{sel}} \;+\; (1-\alpha)\,\mathcal{L}_{\text{aux}} \tag{4.1}$$

The coverage penalty activates when empirical coverage falls below the target and vanishes when coverage exceeds it. The original formulation uses a two-sided quadratic that penalises deviation in either direction. The one-sided version places a floor under coverage without discouraging the model from accepting more than the target which in practice pushes coverage above the target rather than towards it. The consequence, that empirical coverage exceeds the nominal target in the trained models, is visible. Also, the coverage and selective risk are computed per batch rather than over the epoch. Both are ratios, and the average of per-batch ratios is not the ratio computed over the epoch.

### 4.5.3 Hyperparameters

Table 4.2 shows that the penalty weight of 32 and the equal weighting between the selective and auxiliary terms follow the values used in the original SelectiveNet work. The target coverage

Table 4.2: Selective classification configuration across datasets.

| Dataset | Target Coverage | $\lambda$ | $\alpha$ | Selection Threshold |
|---|---|---|---|---|
| MNIST | 0.90 | 32 | 0.5 | 0.7 |
| CIFAR-10 | 0.90 | 32 | 0.5 | 0.7 |
| Chest X-ray | 0.90 | 32 | 0.5 | 0.7 |
| Cats vs. Dogs | 0.70 | 32 | 0.5 | 0.5 |

of 0.9 and 0.7 were chosen as a moderate operating point at which the model still declines a meaningful fraction of inputs.

No systematic hyperparameter search was performed. Values were fixed at the outset and adjusted only where training failed to produce a usable selection head.

### 4.5.4 Training Protocol

All models used Adam [86] at its default learning rate. The selective models were given a modest additional epoch budget in each case on the reasoning that they must learn the selection function alongside the classifier and that the coverage penalty dominates the early epochs. Whether this constitutes an unfair advantage is arguable. The counter-argument is that a matched budget would disadvantage a model solving a harder optimisation problem. Early stopping on validation loss with weight restoration was applied across the datasets. Training and validation curves for accuracy, loss, coverage and selective accuracy were recorded for every run. The coverage curve is the diagnostic of most interest for the selective models since a coverage trace that flattens near 1.0 within the first few epochs indicates a selection head that has collapsed to accepting everything.

## 4.6 Adversarial Evaluation Protocol

The FGSM was implemented for adversarial attack. For a batch of inputs and their labels, the loss is computed under a gradient tape, the gradient with respect to the input is taken, its element-wise sign is scaled by epsilon and added to the input and the result is clipped to the valid pixel range [87].

For the baseline models the loss is the model's training loss. For the selective models the attack targets the prediction head only. The concatenated output is sliced and cross-entropy is computed against the prediction distribution. Therefore, the perturbation is constructed to

Table 4.3: Training protocol used across datasets

| Dataset | Baseline Epochs | Selective Epochs | Batch Size |
|---|---|---|---|
| MNIST | 10 | 15 | 128 |
| CIFAR-10 | 25 | 30 | 128 |
| Chest X-ray | 20 | 20 | 32 |
| Cats vs. Dogs | 15 | 20 | 32 |

Table 4.4: Adversarial perturbation settings across datasets

| **Dataset** | $\epsilon$ **Values** | **Sample Size** |
|---|---|---|
| MNIST | 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30 | 2,000 |
| CIFAR-10 | 0, 0.1, 0.2, 0.3, 0.5, 0.8, 1.0 | 50 |
| Chest X-ray | 0, 0.1, 0.2, 0.3, 0.5, 0.8, 1.0 | 50 |
| Cats vs. Dogs | 0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0 | 1,000 |

induce a misclassification without reference to the selection head. This is the correct construction for the question being asked. The interest is in whether the selection head withdraws from inputs perturbed to fool the classifier which requires that the attack not target the selector. It is also, as Chapter 1 acknowledged, a non-adaptive attack.

Samples were drawn at random from the test set. At each epsilon three quantities are recorded for the selective model, e.g., overall accuracy over the full sample, coverage and selective accuracy over accepted samples alongside the baseline's accuracy. The three selective curves are reported together since selective accuracy alone is uninterpretable. A flat selective accuracy means one thing at stable coverage and something quite different if coverage has collapsed.

The sample sizes for CIFAR-10 and the chest dataset are small at fifty images each. This was a compute-time decision made during development and it is a weakness. The epsilon ranges differ across datasets, as shown in Table 4.4.

## 4.7 Explainability Protocol

Grad-CAM was implemented so that the target scalar can be drawn from either head [40]. A gradient model is constructed with two outputs, e.g., the activations of the final convolutional layer, located by scanning the model for the last layer of that type and the concatenated model output, as shown in Figure 4.10. Both attributions are computed with respect to the activations of the same final convolutional layer in the shared backbone, differing only in the scalar the gradient is taken from. This means the pair of heatmaps produced for a given input can be compared against each other, since the extraction path is identical and only the target differs. Within a gradient tape, the concatenated output is sliced into its prediction and selection components and the target is set to either the predicted class probability from $f(x)$ or the selection score from $g(x)$. The gradient of the target with respect to the convolutional activations is taken, pooled by averaging over the spatial dimensions and used to weight the feature maps. The weighted sum is passed through a ReLU and normalised to its maximum, then resized to the input resolution.

Because the two attributions are taken with respect to the same layer of the same shared backbone in the same forward pass, any difference between them reflects the head rather than the extraction procedure. This is the property that makes the comparison meaningful and it is the reason the last shared convolutional layer is used rather than a layer within either head.

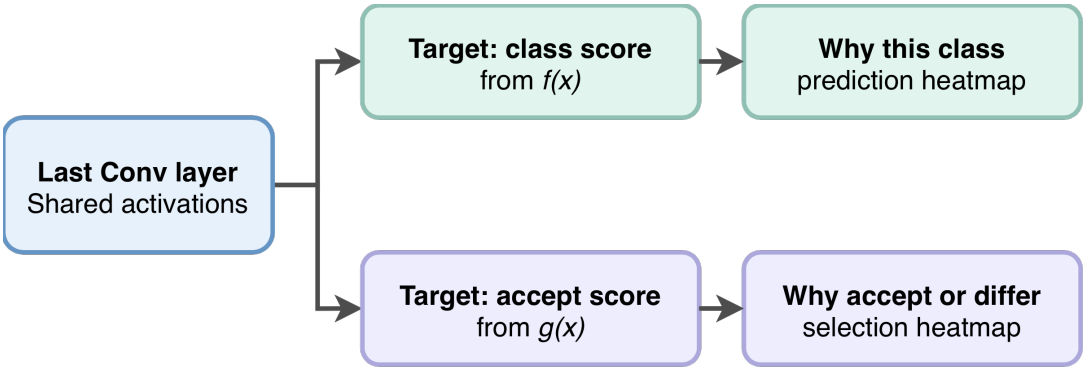



Figure 4.10: The dual-head Grad-CAM procedure.

## 4.8 Evaluation Protocol

For each dataset, the following are computed on the held-out test set:

For the baseline model, accuracy, per-class precision, recall and F1, the confusion matrix and balanced accuracy [77]. For the selective model, the same conventional quantities computed over the full test set, empirical coverage at the operating threshold, selective accuracy and selective risk over the accepted subset, the risk–coverage curve obtained by sweeping the threshold across points spanning the unit interval with the operating point marked and the distribution of selection scores [46].

For both models under AAEF, the six-cell confusion matrix (binary) or the augmented matrix with an abstention column (multiclass), AAR, AAS and AABA, the per-class breakdown for the multiclass datasets, micro-averaged AAR, and the optimism gaps.

## 4.9 Threats to Methodological Validity

Several aspects of the design limit what the results can support and they are gathered here rather than distributed through the chapter.

The baseline and selective models share an architecture specification but not weights. Some of the observed difference may reflect initialisation and optimisation variance rather than the selection mechanism. Weight sharing, e.g., training the baseline, freezing the backbone and attaching heads, would isolate the mechanism better and would introduce its own confound since the frozen representation was optimised without the selective objective.

Target coverage, lambda, alpha and the threshold were set at the outset. The reported operating points are, therefore, not optimal under any stated criterion and a different configuration might change the ordering of results.

FGSM with full gradient access, no adversarial training, no iterative or adaptive attack, the findings concern a behaviour under one attack.

# Chapter 5

# Results and Analysis

This chapter reports the experimental results. Each dataset in turn described in the following the same internal order, e.g., training dynamics and the baseline classifier, the selective model, behaviour under FGSM perturbation, the dual-head explanations and the abstention-aware evaluation.

## 5.1 MNIST

### 5.1.1 Baseline training dynamics

The baseline was trained for ten epochs at a batch size of 128 with a 10% validation split. Training accuracy, as shown in Figure 5.1, rose from 0.8569 in the first epoch to above 0.988 by the eighth and validation accuracy reached 0.9908 by the sixth epoch and then moved little. Validation loss fell from 0.0773 to a minimum near 0.0295 around the seventh epoch. After which it began to drift upward while training loss continued to fall. The curves indicate the model was trained properly. Early stopping did not invoke on this dataset. Therefore, the final weights are those of epoch ten.

The selection score distribution spans the full unit interval on the test set, from 0.0000 to 1.0000. A selection head that had collapsed towards a constant output would produce a narrow range and a flat coverage curve during training. Neither occurred, therefore, the gate is discriminative rather than degenerate.

Risk–coverage curve, in Figure 5.2, for the MNIST SelectiveNet, obtained by sweeping the selection threshold. The marked operating point corresponds to the pinned threshold of 0.7. Selective risk falls as coverage is reduced, which is the behaviour the framework identifies as the source of score inflation under accepted-only evaluation.

### 5.1.2 Baseline performance

On the 10,000-image test set the baseline reached 99.34 per cent accuracy, misclassifying 66 digits, as shown in Table 5.1. Macro-averaged recall was 0.9933 and macro-averaged specificity 0.9993.

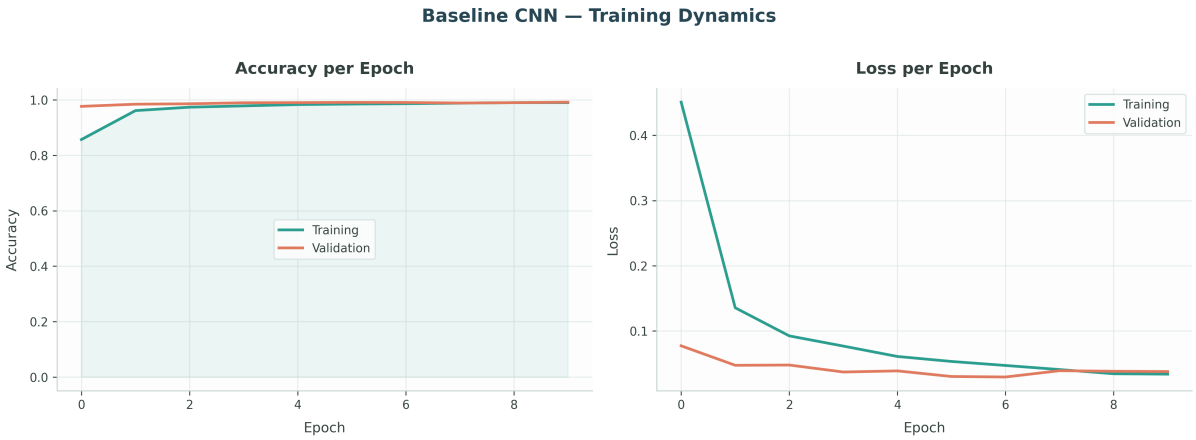



Figure 5.1: Training and validation accuracy and loss for the MNIST baseline CNN.

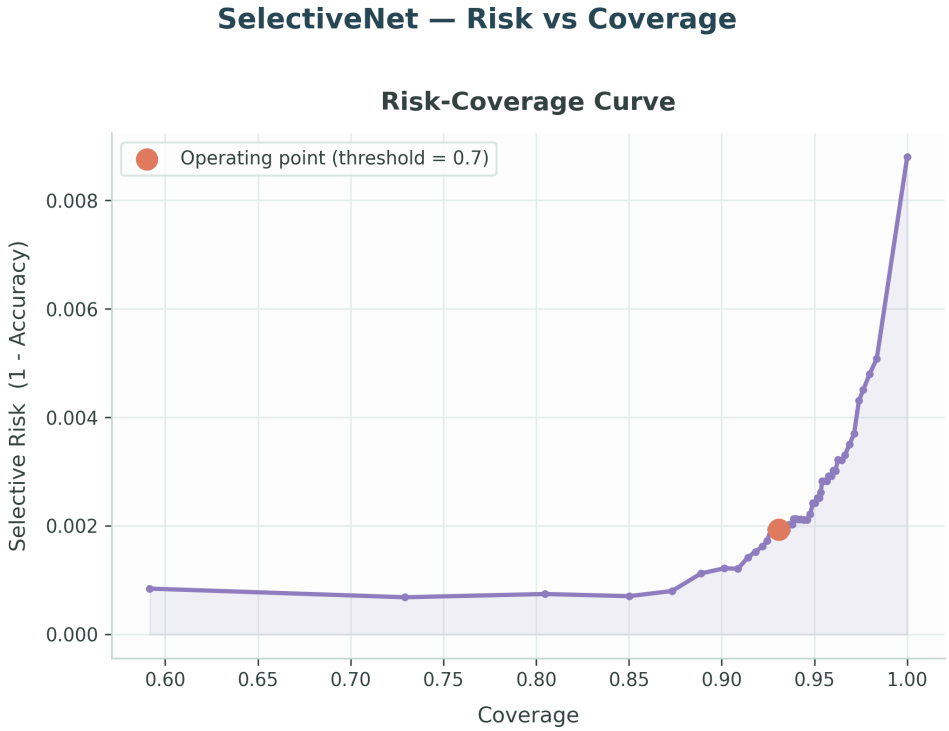



Figure 5.2: Risk–coverage curve for the MNIST SelectiveNet.

Table 5.1: Classification performance of baseline CNN on MNIST.

| **Quantity** | **Plain CNN** |
|---|---|
| Test samples evaluated | 10,000 |
| Correct | 9,934 |
| Incorrect | 66 |
| Accuracy | 0.9934 |
| Macro recall | 0.9933 |
| Macro specificity | 0.9993 |
| Balanced accuracy (macro recall) | 0.9933 |
| Composite column reported in notebook | 0.9963 |
| Selective risk | Not applicable |
| Empirical coverage | 1.0000 |

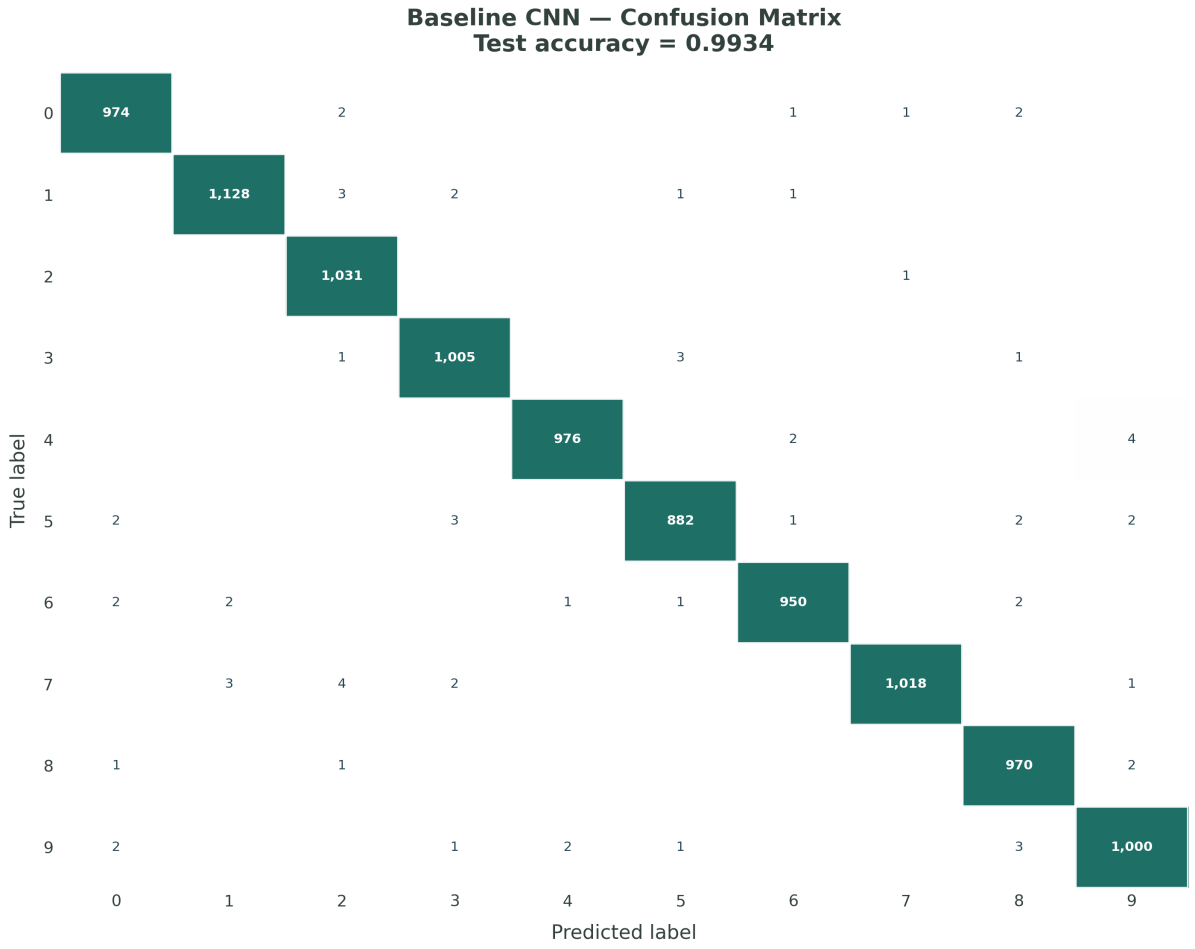



Figure 5.3: Confusion matrix for the baseline CNN on the MNIST test set.

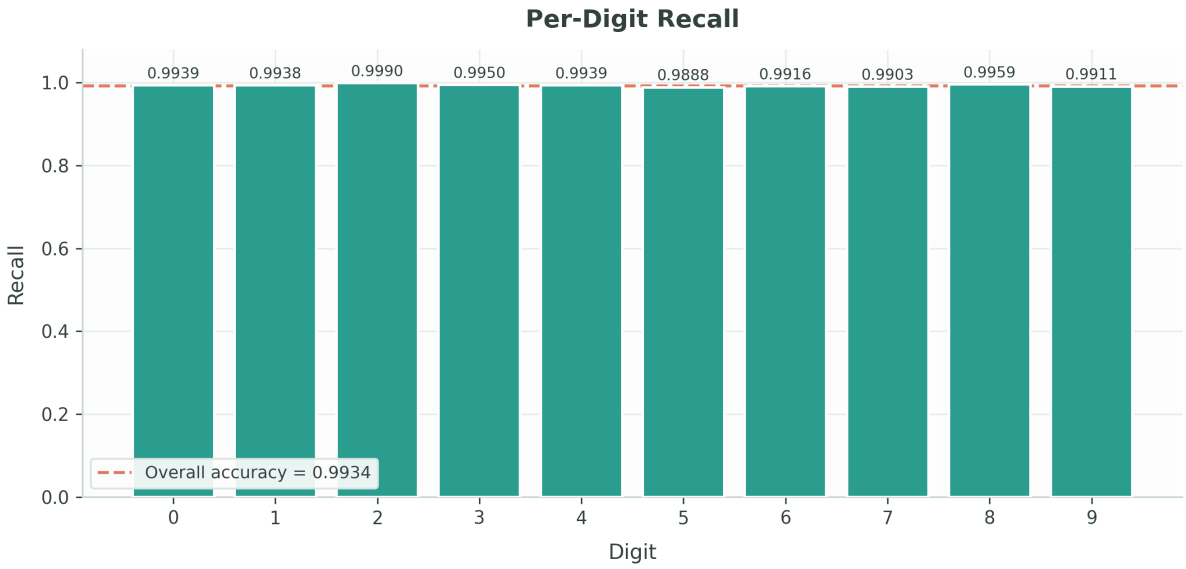



Figure 5.4: Per-digit recall for the baseline CNN on MNIST dataset.

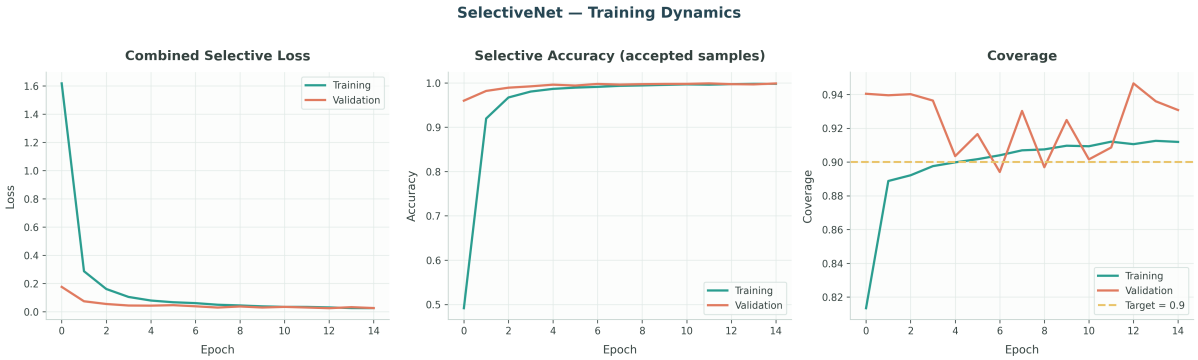



Figure 5.5: Training dynamics for the SelectiveNet on MNIST.

The confusion matrix, illustrated in Figure 5.3, shows no dominant failure mode. With 66 errors spread over off-diagonal cells there is little structure to interpret, and this is part of why MNIST is a useful starting point. Because the classifier is almost never wrong.

Figure 5.4 shows per-digit recall for the baseline CNN. All ten classes sit above 0.98, with digit 5 the weakest at 0.9888.

### 5.1.3 Selective model training dynamics

The selective model was trained for fifteen epochs on the same backbone. Its training curves differ in kind from the baseline's, because the quantities being tracked are different, e.g., coverage and selective accuracy rather than accuracy alone, as depicted in Figure 5.5. The coverage trace is the diagnostic of interest. It settles close to the target within the first few epochs and stays there rather than climbing towards one which is considered as the failure mode the coverage penalty makes possible and which would indicate a selection head that had collapsed to accepting everything. Selective accuracy rises alongside it. The two curves moving together is what one wants to see.

Table 5.2: Performance evaluation of SelectiveNet on MNIST (accepted samples).

| **Quantity** | **SelectiveNet (Accepted Only)** |
|---|---|
| Test samples evaluated | 9,309 |
| Correct | 9,291 |
| Incorrect | 18 |
| Accuracy | 0.9981 |
| Macro recall | 0.9980 |
| Macro specificity | 0.9998 |
| Balanced accuracy (macro recall) | 0.9980 |
| Composite column reported in notebook | 0.9989 |
| Selective risk | 0.0019 |
| Empirical coverage | 0.9309 |

From Table 5.2, it is evident that the SelectiveNet accepted 9,309 of the 10,000 test digits at the pinned threshold of 0.7, an empirical coverage of 0.9309 against a target of 0.90. Coverage exceeding its target here is consistent with the one-sided penalty which places a floor under coverage without discouraging the model from accepting more. On the accepted subset the model was correct on 9,291 samples and wrong on 18, an accepted-set accuracy of 0.9981 and a selective risk of 0.0019. Set against the baseline's 66 errors (shown in Figure 5.6), the selective model commits 18, a reduction of roughly three quarters in the absolute count of misclassifications, obtained at the cost of 691 deferrals.

### 5.1.4 Selective model performance

### 5.1.5 Behaviour under FGSM

The adversarial sweep was run over 2,000 randomly drawn test digits at seven perturbation strengths. The baseline degrades as the literature would suggest, falling from 99.45 per cent accuracy on clean inputs to 34.85 per cent at an epsilon of 0.30, as shown in Table 5.3 and Figure 5.7.

Coverage falls from 0.9355 on clean inputs to 0.4615 at the strongest perturbation and the fall is monotone across the sweep. Selective accuracy, by contrast, barely moves. It begins at 0.9995 and ends at 0.9783, dipping to 0.9710 in the middle of the range before recovering. The two curves say that the model is withdrawing from the perturbed inputs rather than misclassifying them. By an epsilon of 0.15 it has declined roughly half the sample and the half it retains it still classifies at above 97 per cent accuracy.

This is the clearest instance in the study of the behaviour the safety argument depends on. The prediction head is no less susceptible to the perturbation than the baseline. Its ungated accuracy falls to 0.5015, and the delivered figure in the final column of Table 5.3 falls from 0.9350 to 0.4515. At epsilon 0.30 the baseline asserts a wrong answer on roughly 65 per cent of inputs, whereas the selective model asserts a wrong answer on about 1 per cent. Whether that trade is worth making depends on what a deferral costs in the deployment concerned.

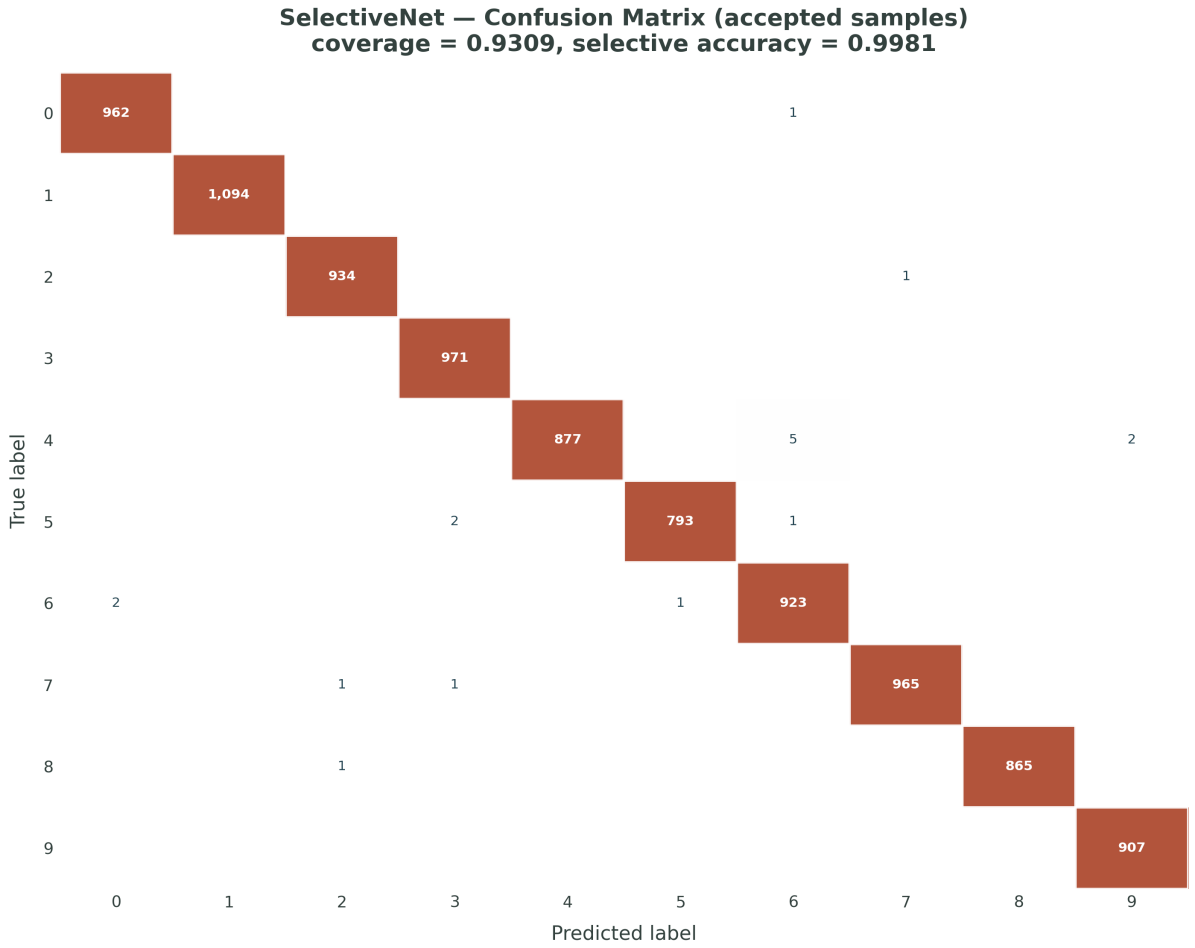



Figure 5.6: Confusion matrix for the MNIST SelectiveNet computed over accepted samples only.

Table 5.3: MNIST under FGSM perturbation over 2,000 randomly drawn test digits.

| $\epsilon$ | **CNN Accuracy** | **SelNet Overall Acc. (Ungated)** | **SelNet Coverage** | **SelNet Selective Accuracy** | **Delivered (Cov. $\times$ Sel. Acc.)** |
|---|---|---|---|---|---|
| 0.00 | 0.9945 | 0.9930 | 0.9355 | 0.9995 | 0.9350 |
| 0.05 | 0.9740 | 0.9695 | 0.7980 | 0.9893 | 0.7895 |
| 0.10 | 0.9140 | 0.9320 | 0.5790 | 0.9750 | 0.5645 |
| 0.15 | 0.8040 | 0.8350 | 0.4900 | 0.9724 | 0.4765 |
| 0.20 | 0.6315 | 0.6900 | 0.4655 | 0.9710 | 0.4520 |
| 0.25 | 0.4705 | 0.5800 | 0.4630 | 0.9752 | 0.4515 |
| 0.30 | 0.3485 | 0.5015 | 0.4615 | 0.9783 | 0.4515 |

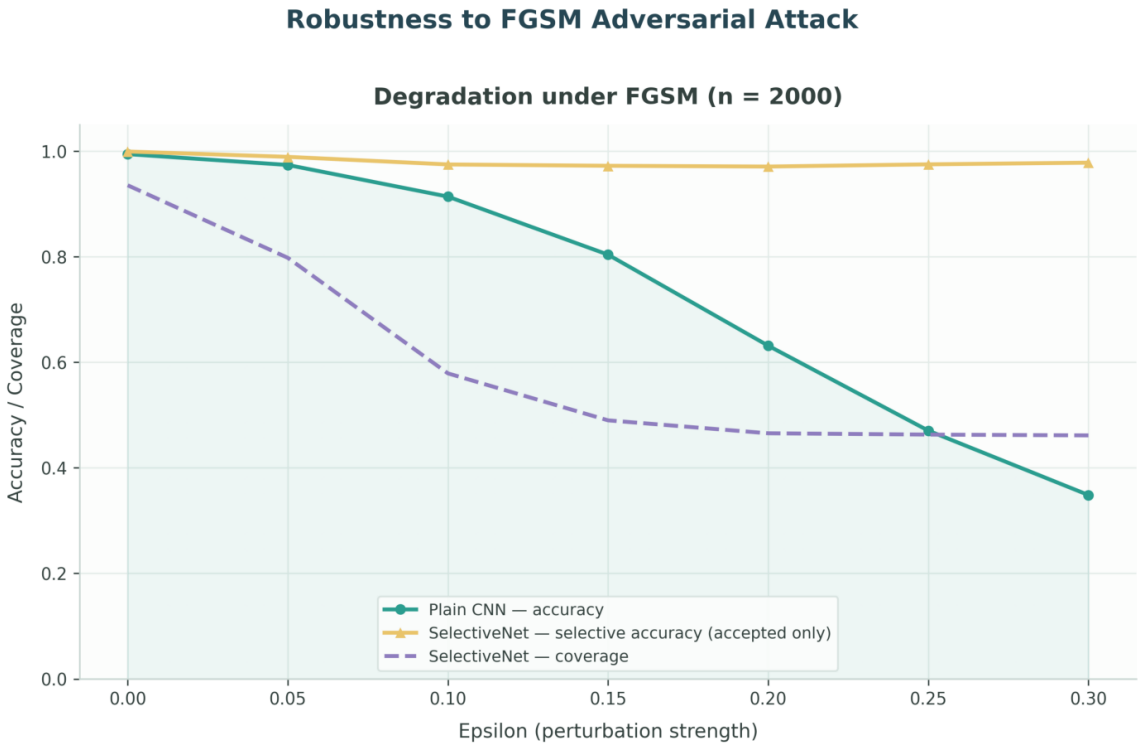



Figure 5.7: Adversarial behaviour of the baseline CNN and the SelectiveNet under FGSM on MNIST (n = 2,000).

### 5.1.6 Dual-head explanations

Grad-CAM was applied to the last shared convolutional layer with the target drawn from either head. The explanation section enumerated 9,291 clean accepted and 691 clean rejected samples for illustration.

As shown in Figure 5.8, on accepted digits the two maps tend to overlap on the stroke structure of the digit. On rejected digits the selection-head attribution is more diffuse and often shifts away from the region the prediction head attends to. It would be convenient to conclude that the selector is responding to stroke ambiguity.

The clean-versus-adversarial comparison in Figure 5.9 shows a visible shift in the selection head's focus consistent with the coverage collapse reported above. It remains an observation rather than a mechanism. A saliency map that moves does not establish what moved it.

### 5.1.7 Abstention-aware evaluation

The contrast between the two conventions is the point of the table, displayed in Table 5.4 and Figure 5.10. Under accepted-only evaluation the SelectiveNet outperforms the baseline on every measure, e.g., macro recall of 0.9980 against 0.9933, accuracy of 0.9981 against 0.9934. Under the abstention-aware convention the ordering reverses, e.g., AAR of 0.9283 against 0.9933, micro AAR of 0.9291 against 0.9934. Neither reading is wrong. They answer different questions, e.g., how well the model performs on the cases it chooses to decide and how many cases it decides

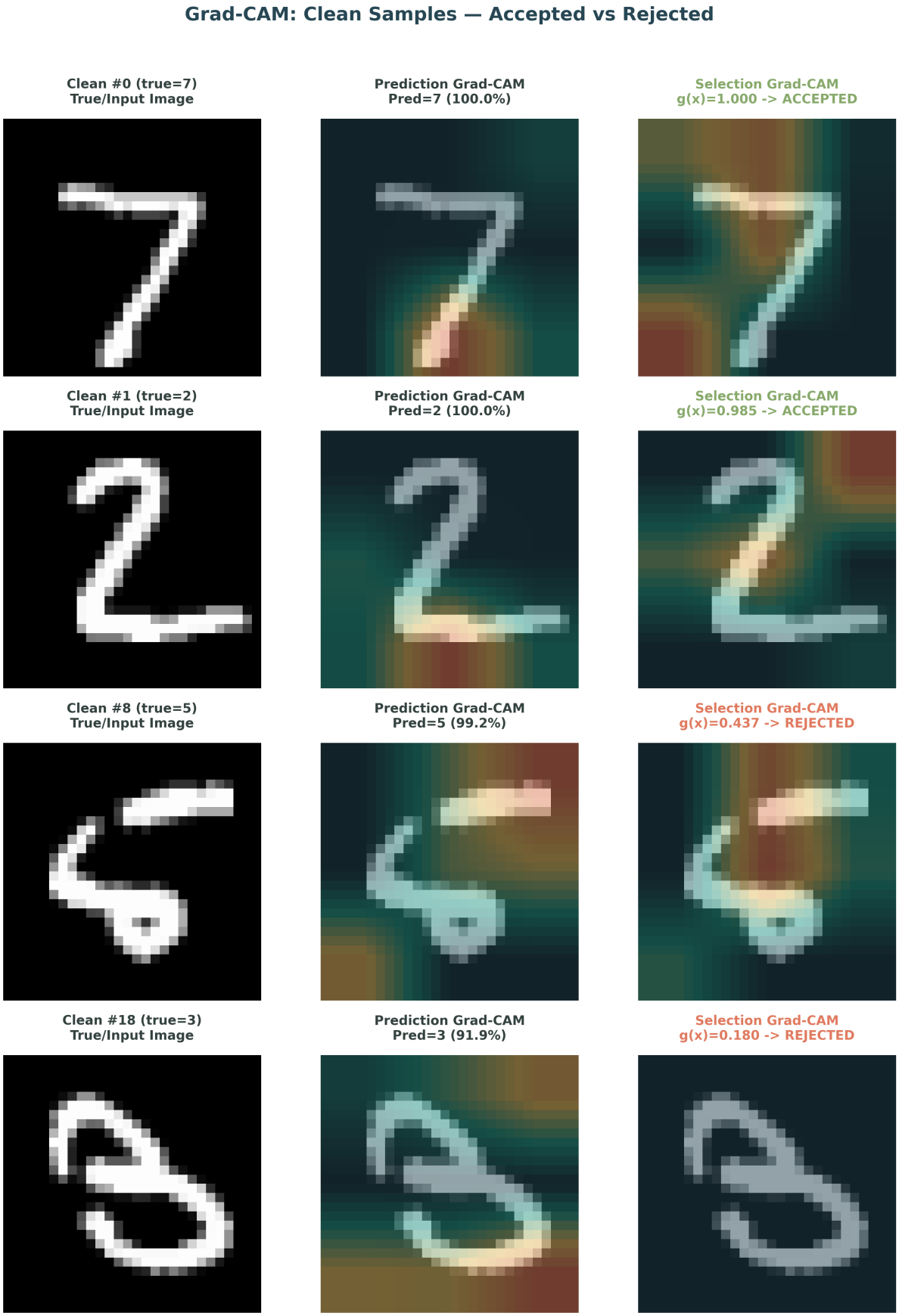



Figure 5.8: Grad-CAM attributions on clean MNIST digits comparing accepted and rejected samples.




Figure 5.9: The same digit before and after FGSM perturbation with the selection-head attribution shown for both.

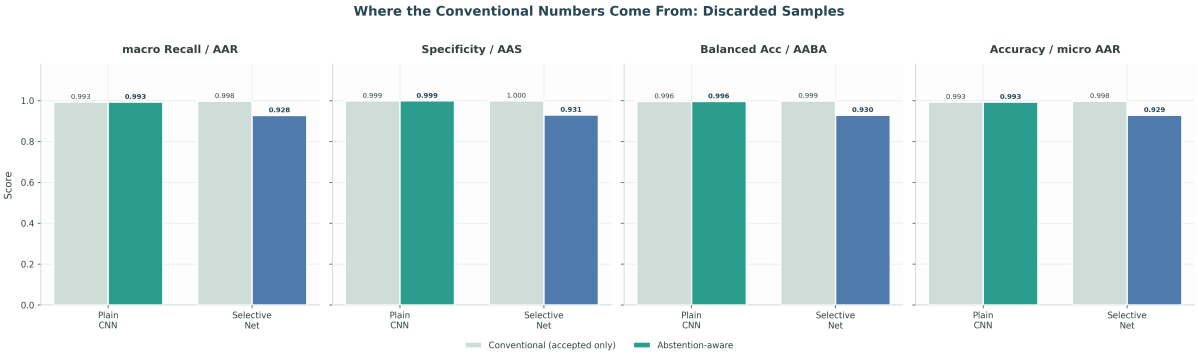



Figure 5.10: Conventional (accepted-only) scores against abstention-aware scores for both MNIST models.

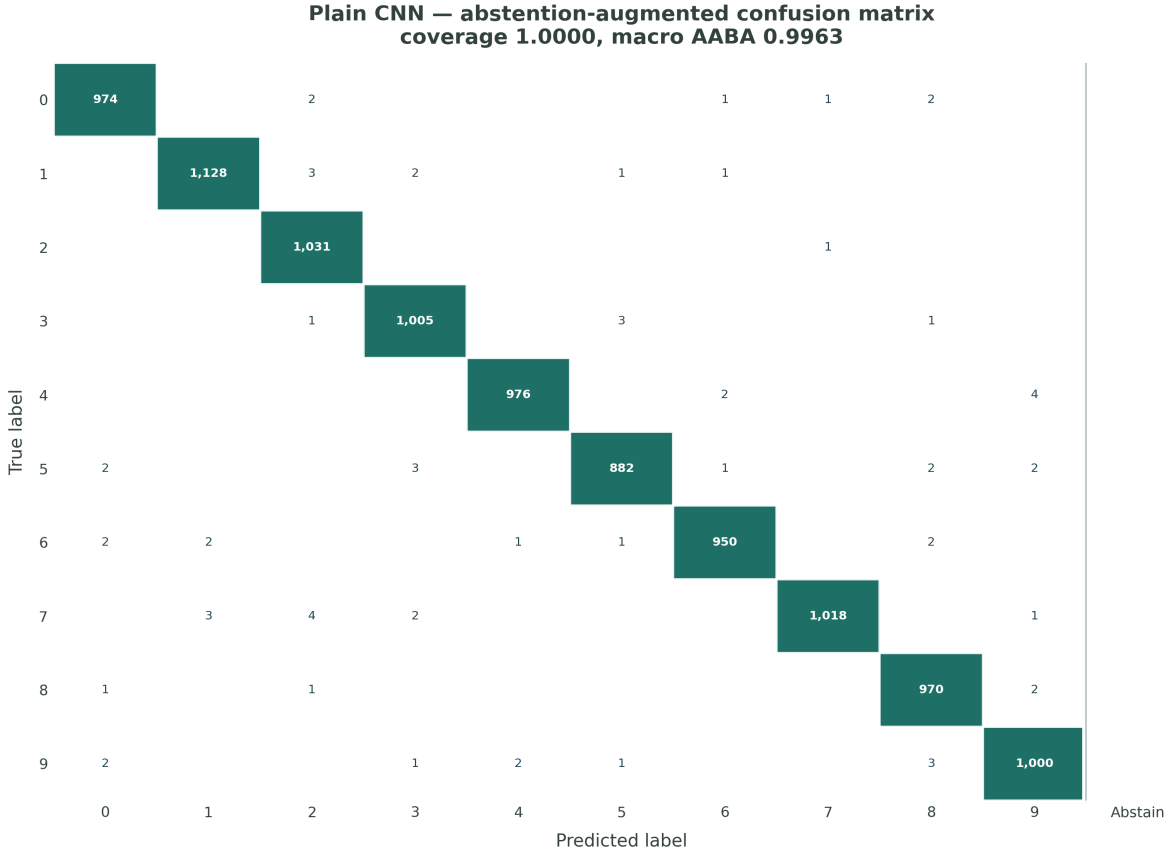



Figure 5.11: Abstention-aware confusion matrix for CNN on MNIST dataset.

correctly out of all the cases it was given. The optimism gap is 0.0690 in accuracy for the selective model and zero for the baseline by construction.

Digit '8' is the hardest class for the selective model. Figure 5.12 shows that the model misclassified a single '8' and deferred 108 of them. The same pattern holds for digits 2, 4, 5 and 9, each combining a high abstention count with a low single-digit error count as comapared to baseline CNN, as shown in Figure 5.11.

This is the diagnostic value the framework was designed to provide. Accepted-only evaluation would report the selective model's recall on digit 8 as high because the 108 deferred samples

Table 5.4: Consolidated MNIST evaluation, macro-averaged over the ten classes. The baseline's abstention-aware values equal its conventional values, as the reduction property requires.

| **Model** | **Correct** | **Incorrect** | **Abstained** | **Coverage** | **Acc. (Accepted)** | **Macro Recall (Accepted)** | **AAR** | **AAS** | **AABA** | **Micro AAR** |
|---|---|---|---|---|---|---|---|---|---|---|
| Plain CNN | 9,934 | 66 | 0 | 1.0000 | 0.9934 | 0.9933 | 0.9933 | 0.9993 | 0.9963 | 0.9934 |
| SelectiveNet | 9,291 | 18 | 691 | 0.9309 | 0.9981 | 0.9980 | 0.9283 | 0.9307 | 0.9295 | 0.9291 |

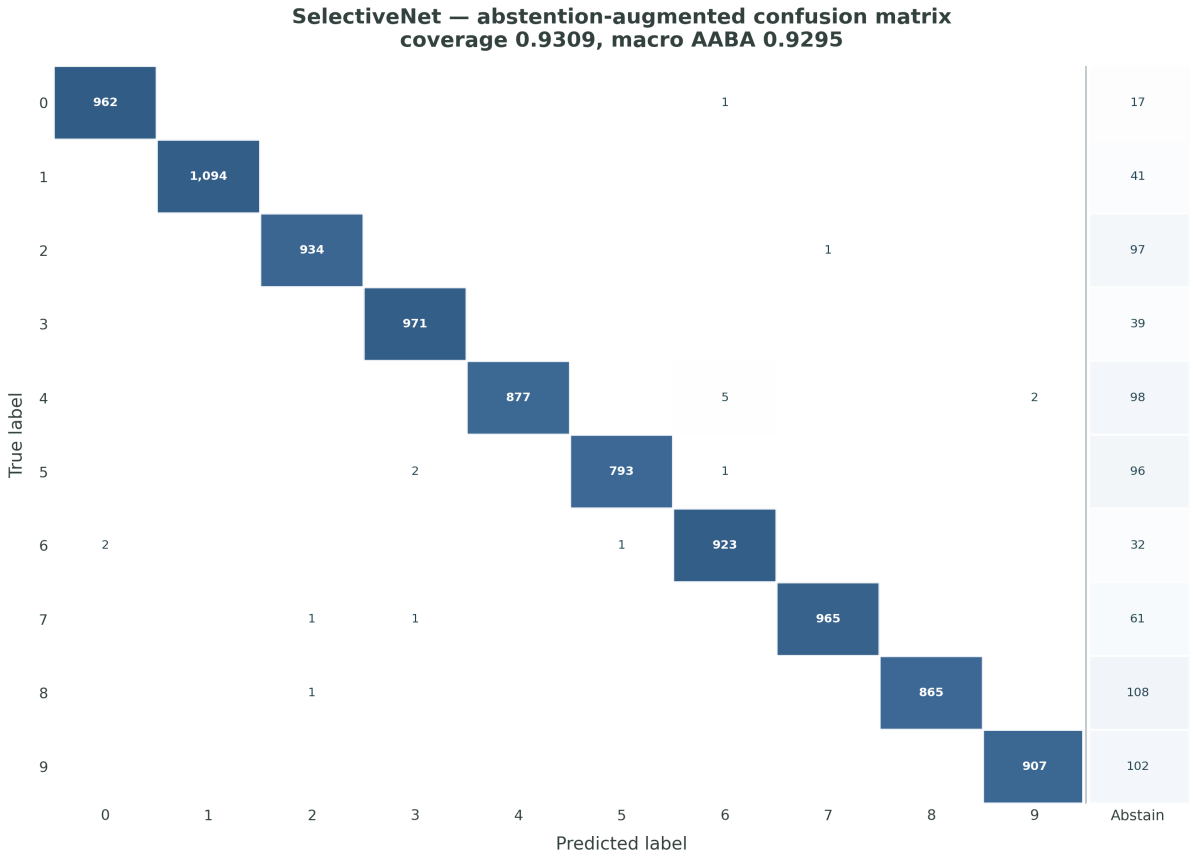



Figure 5.12: Abstention-aware confusion matrix for SelectiveNet on MNIST dataset.

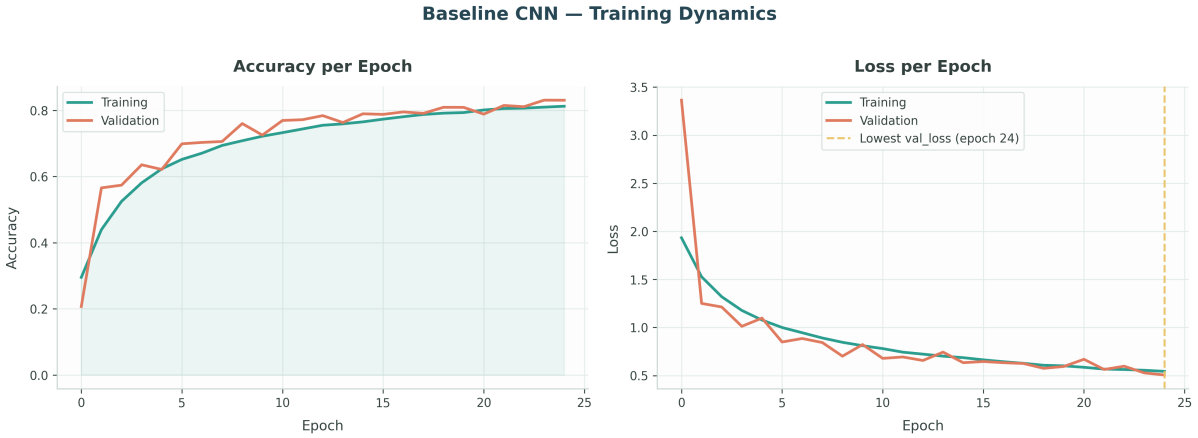



Figure 5.13: Training and validation accuracy and loss for the CIFAR-10 baseline CNN.

would have left the denominator. Charging abstention as an error would report it as low and give no way to tell whether the model was wrong or absent. The framework reports it as low and shows that the cause is deferral. In a deployment where deferral is inexpensive, that is a benign finding. In one where it is costly, it identifies exactly where the cost is being incurred.

## 5.2 CIFAR-10

### 5.2.1 Baseline training dynamics

The baseline was trained for 25 epochs at batch size 128 with a 10 percent validation split on the deeper batch-normalised stack described in Chapter 4 and without data augmentation.

The curves, in Figure 5.13, show the behaviour one expects of an unaugmented model of this capacity on CIFAR-10. Training accuracy continues to climb while validation accuracy plateaus and the loss curves separate. The final test accuracy of 82.74 percent sits below what an augmented model of comparable depth would reach and this is a deliberate simplification rather than an attempt at a competitive result.

### 5.2.2 Baseline performance

Classification performance on CIFAR-10 dataset is displayed in Table 5.5. On the 10,000-image test set the baseline reached 82.74 percent accuracy, misclassifying 1,726 images. Macro-averaged recall was 0.8274 and macro-averaged specificity 0.9808.

Confusion matrix for the baseline CNN on CIFAR-10 is illustrated in Figure 5.14 annotated with the test accuracy of 0.8274. Errors concentrate among the animal classes.

Per-class recall for the baseline CNN on CIFAR-10 is shown in Figure 5.15. Cat is the weakest class at 0.627 and frog the strongest at 0.914. The error structure is the familiar one for this dataset. Confusions concentrate among the animal classes with cat and dog exchanging mass and bird drawing errors from several directions. The vehicle classes separate more cleanly.

Table 5.5: Classification performance on baseline CNN model on CIFAR-10 dataset.

| **Quantity** | **Plain CNN** |
|---|---|
| Test samples evaluated | 10,000 |
| Correct | 8,274 |
| Incorrect | 1,726 |
| Accuracy | 0.8274 |
| Macro recall | 0.8274 |
| Macro specificity | 0.9808 |
| Balanced accuracy (macro recall) | 0.8274 |
| Composite column reported in notebook | 0.9041 |
| Selective risk | Not applicable |
| Empirical coverage | 1.0000 |

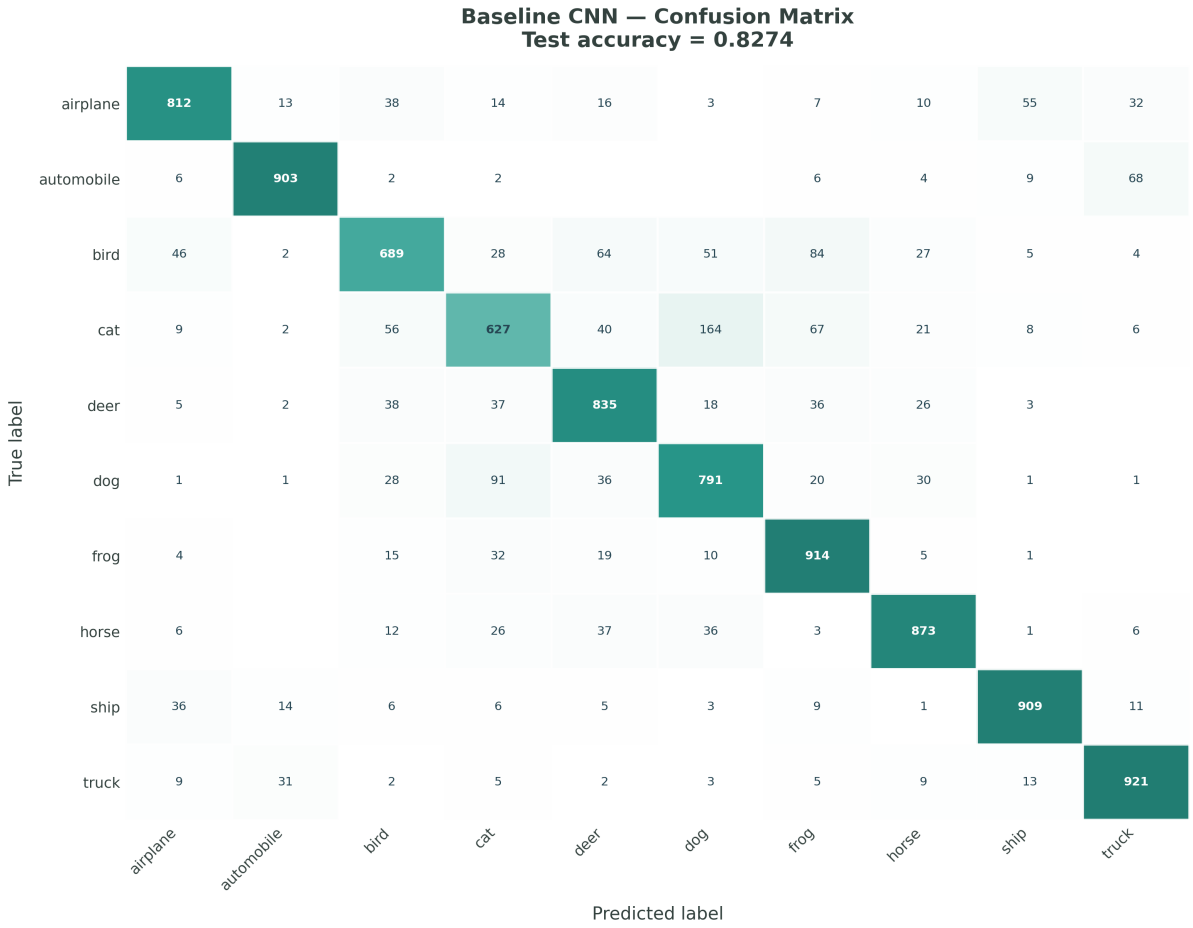



Figure 5.14: Confusion matrix for the baseline CNN on CIFAR-10.

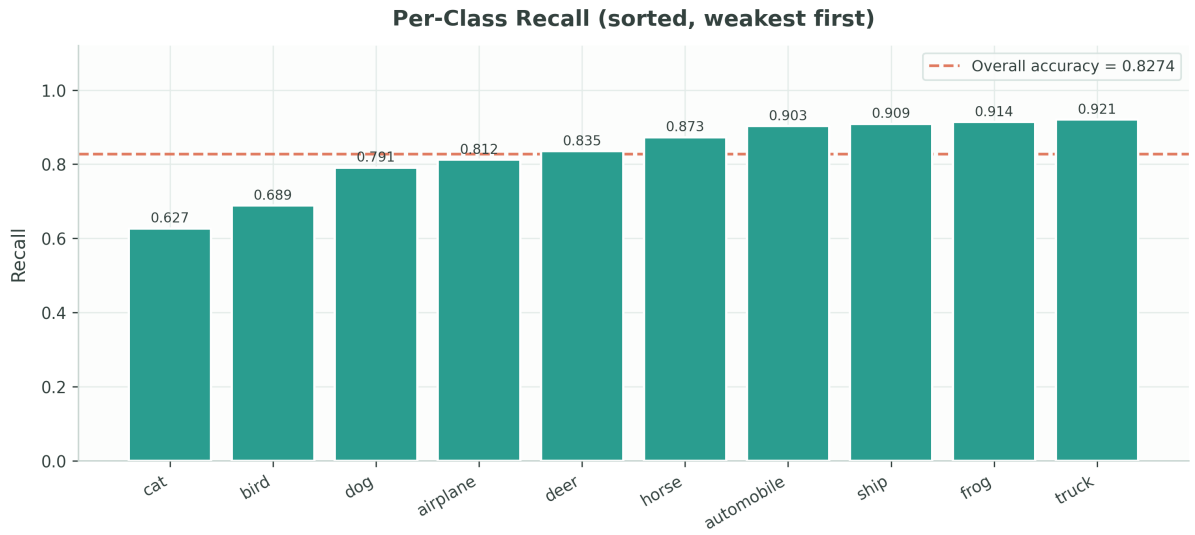



Figure 5.15: Per-class recall for the baseline CNN on CIFAR-10.

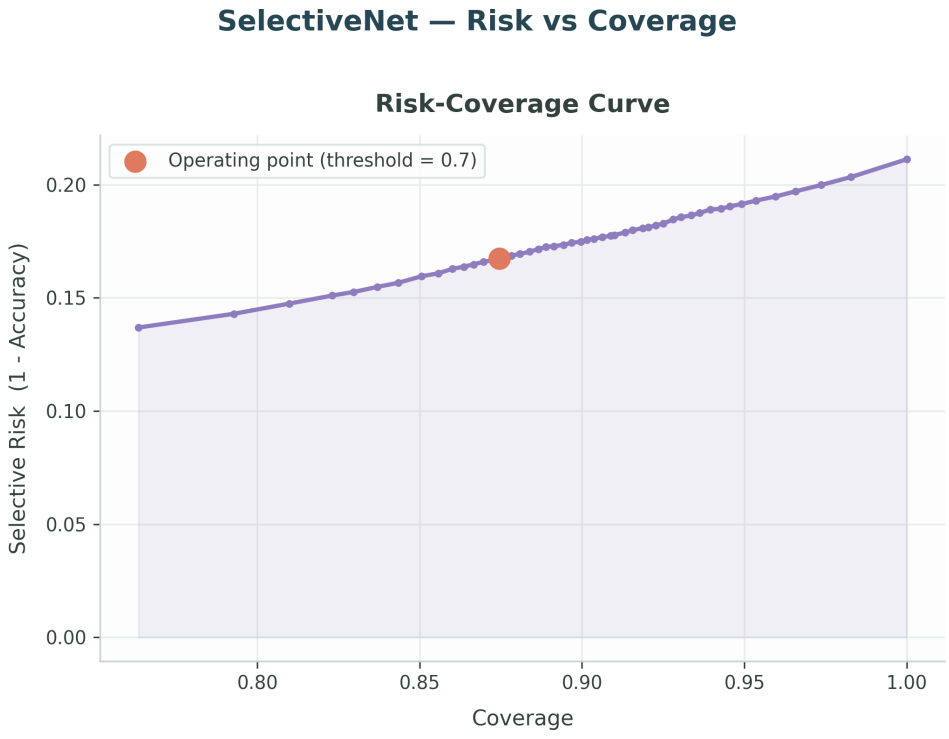



Figure 5.16: Risk–coverage curve for the CIFAR-10 SelectiveNet.

### 5.2.3 Selective model training dynamics

Coverage approaches the target from below and settles slightly under it ending at 0.8746 on the test set, as shown in Figure 5.16. Selective accuracy climbs steadily through training, as presented in Figure 5.17. Unlike MNIST, coverage here does not overshoot the target. On a harder task the selective risk term evidently pulls coverage down enough to offset the missing upper penalty.

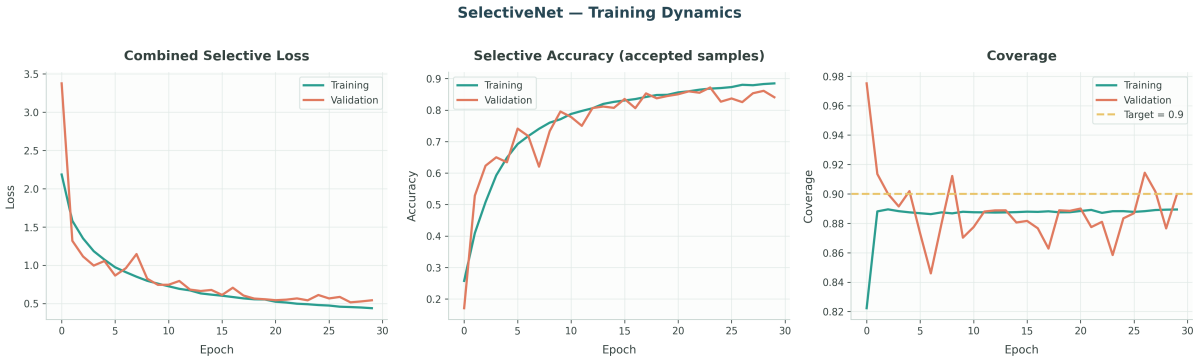



Figure 5.17: Training dynamics for the CIFAR-10 SelectiveNet.

Table 5.6: Performance evaluation of SelectiveNet on accepted samples from CIFAR-10 dataset.

| **Quantity** | **SelectiveNet (Accepted Only)** |
|---|---|
| Test samples evaluated | 8,746 |
| Correct | 7,283 |
| Incorrect | 1,463 |
| Accuracy | 0.8327 |
| Macro recall | 0.8251 |
| Macro specificity | 0.9814 |
| Balanced accuracy (macro recall) | 0.8251 |
| Composite column reported in notebook | 0.9032 |
| Selective risk | 0.1673 |
| Empirical coverage | 0.8746 |

### 5.2.4 Selective model performance

At the pinned threshold the model accepted 8,746 of 10,000 images, as presented in Table 5.6. On the accepted subset it was correct on 7,283 and wrong on 1,463, an accepted-set accuracy of 0.8327 against the baseline's 0.8274. The improvement is roughly half a percentage point.

The absolute error count is the more informative comparison, as shown in Figure 5.18. 1,463 against the baseline's 1,726, a reduction of 263 misclassifications obtained at the cost of 1,254 deferrals. Roughly one error avoided for every five deferrals. The selection head is declining a large number of samples the classifier would have handled correctly.

### 5.2.5 Behaviour under FGSM

The CIFAR-10 adversarial sweep was run over 50 randomly drawn test images. At a coverage of around 0.5, a proportion estimated from 50 samples carries a standard error of roughly seven percentage points. The pattern resembles the MNIST result, e.g., selective accuracy rises and selective coverage falls with it. Therefore, it is abstaining more adversarial images to retain the selective accuracy high, as shown in Figure 5.19.

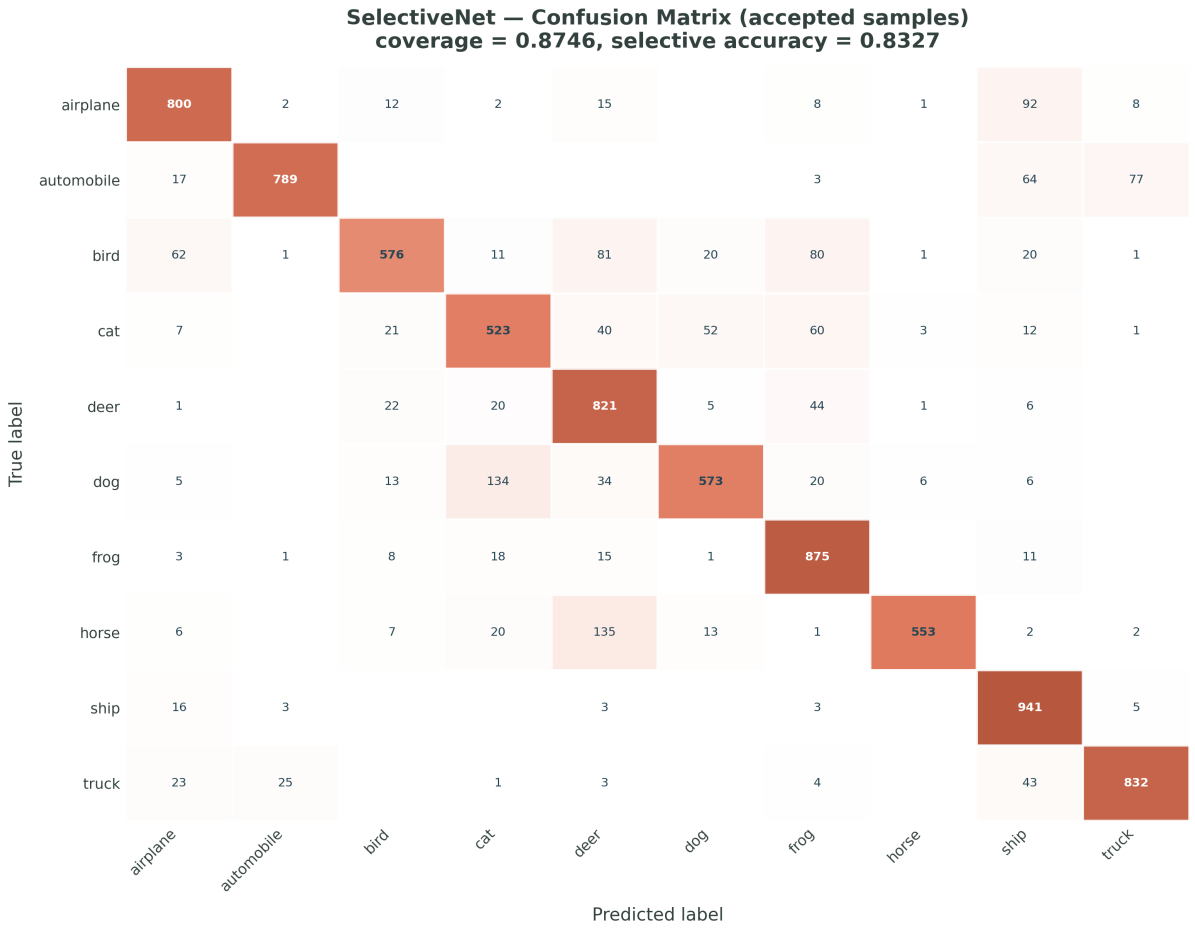



Figure 5.18: Confusion matrix for the CIFAR-10 SelectiveNet over accepted samples only.

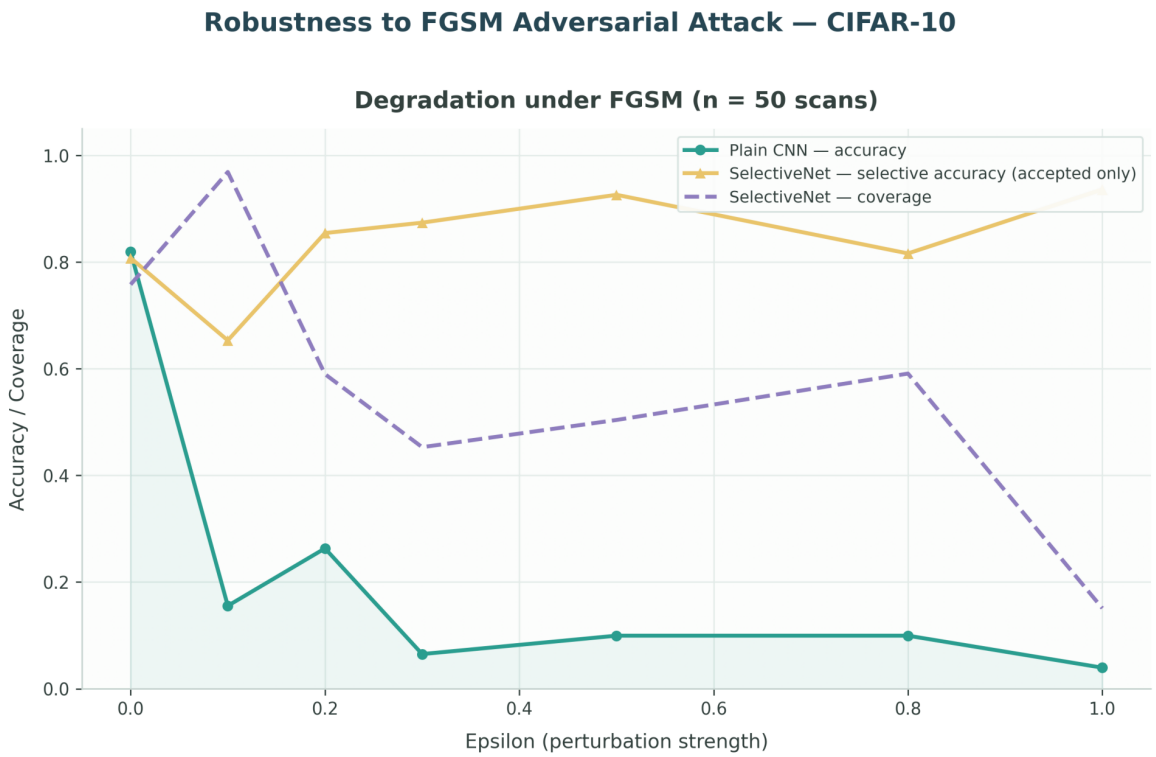



Figure 5.19: Degradation under FGSM on CIFAR-10 (n = 50).

### 5.2.6 Dual-head explanations

The explanation section reports 7,200 accepted and 1,408 rejected clean samples. Figure 5.20 is, therefore, at a stricter threshold than the tables. The attributions here are less informative than on the other datasets and the reason is resolution. The final convolutional feature map is 8 x 8, and upsampling that to 32 x 32 produces a coarse map in which a highlighted region covers a substantial fraction of the object. Distinguishing 'the selector attends to the animal's head' from 'the selector attends to the upper half of the image' is not possible at this granularity. The CIFAR-10 explanations carry the least weight of the four sets.

### 5.2.7 Abstention-aware evaluation

The optimism gap in accuracy is 0.1044 (shown in Table 5.7 and Figure 5.21), larger than MNIST's. Under accepted-only evaluation the two models look near-identical with accuracies of 0.8274 and 0.8327 separated by half a point. Under the abstention-aware convention they are separated by ten points. The entire difference is the 1,254 deferred images which the first convention removes and the second retains.

Cat is the hardest class under AABA (shown in Figure 5.22) which is unsurprising given the baseline confusion matrix. What the abstention column adds is that the selective model deferred 281 cats and misclassified 196 (presented in Figure 5.23), deferral and error are of comparable magnitude, unlike the MNIST case where deferral dominated. Horse is the more interesting entry. The baseline achieves 0.873 on it while the selective model reaches 0.553 and the gap

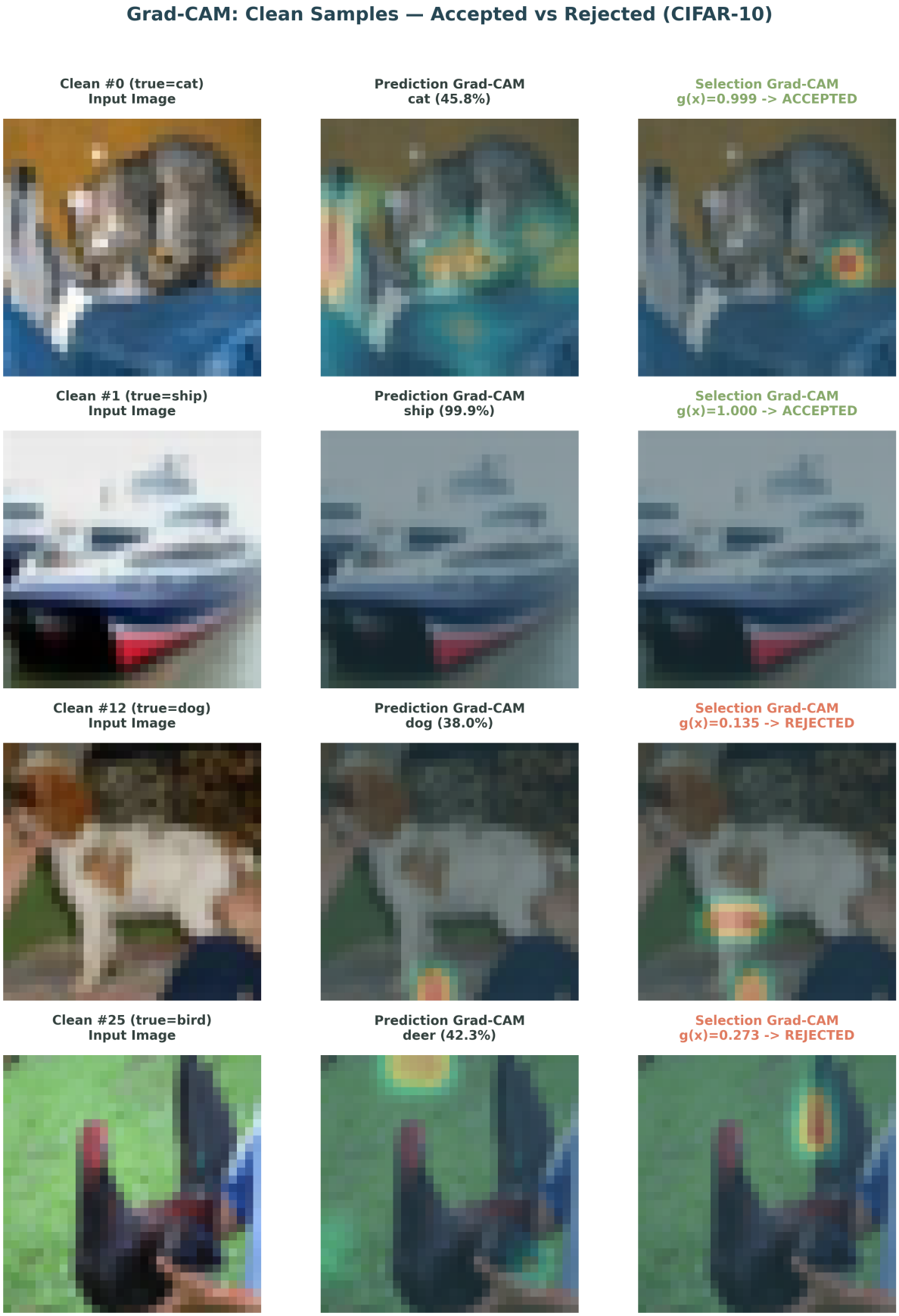



Figure 5.20: Grad-CAM attributions on clean CIFAR-10 images, accepted against rejected.

Table 5.7: Comparison of plain CNN and SelectiveNet performance on CIFAR-10 dataset.

| **Model** | **Correct** | **Incorrect** | **Abstained** | **Coverage** | **Acc. (Accepted)** | **Macro Recall (Accepted)** | **AAR** | **AAS** | **AABA** | **Micro AAR** |
|---|---|---|---|---|---|---|---|---|---|---|
| Plain CNN | 8,274 | 1,726 | 0 | 1.0000 | 0.8274 | 0.8274 | 0.8274 | 0.9808 | 0.9041 | 0.8274 |
| SelectiveNet | 7,283 | 1,463 | 1,254 | 0.8746 | 0.8327 | 0.8251 | 0.7283 | 0.8583 | 0.7933 | 0.7283 |

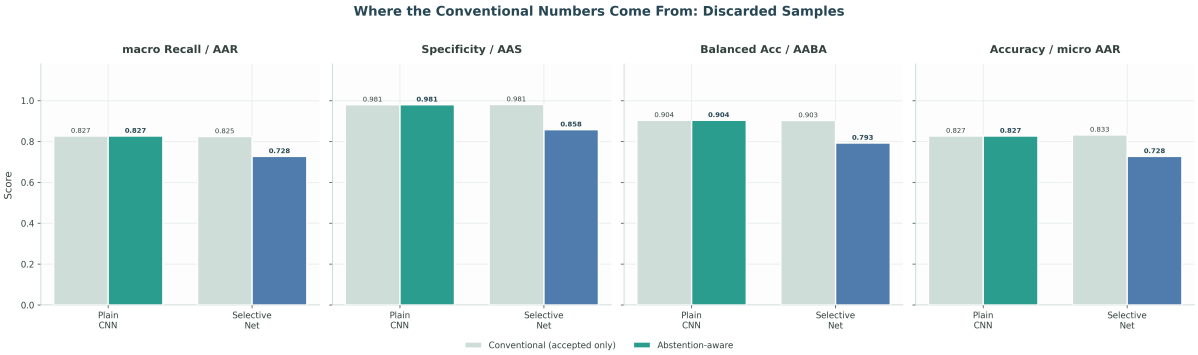



Figure 5.21: Conventional against abstention-aware scores for both CIFAR-10 models.

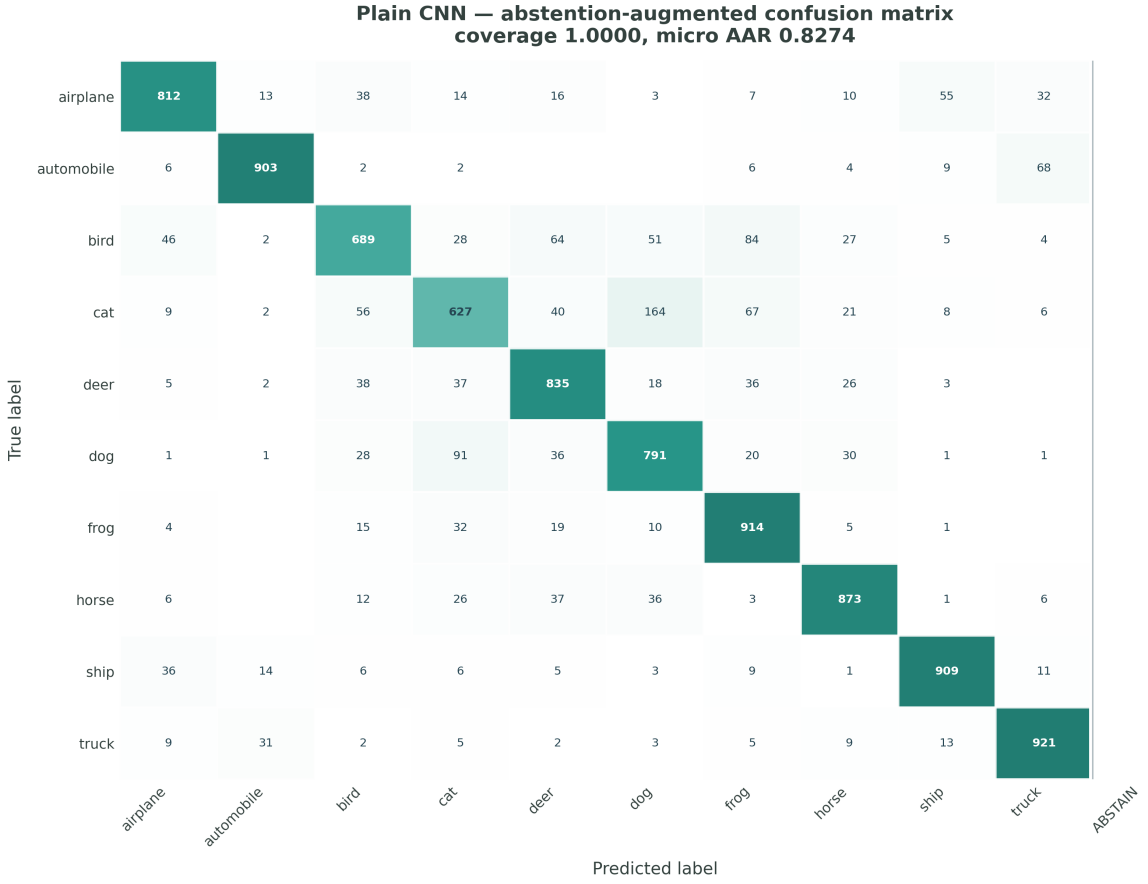



Figure 5.22: Abstention-aware confusion matrix for CNN on CIFAR-10 dataset.

is made up of 261 deferrals against 186 errors. A class the baseline handles well is one the selector treats as difficult. Ship is the only class where the selective model's abstention-aware recall exceeds the baseline's at 0.941 against 0.909 with 29 deferrals and 30 errors. On classes the model finds easy, the selector accepts nearly everything and the two readings converge.

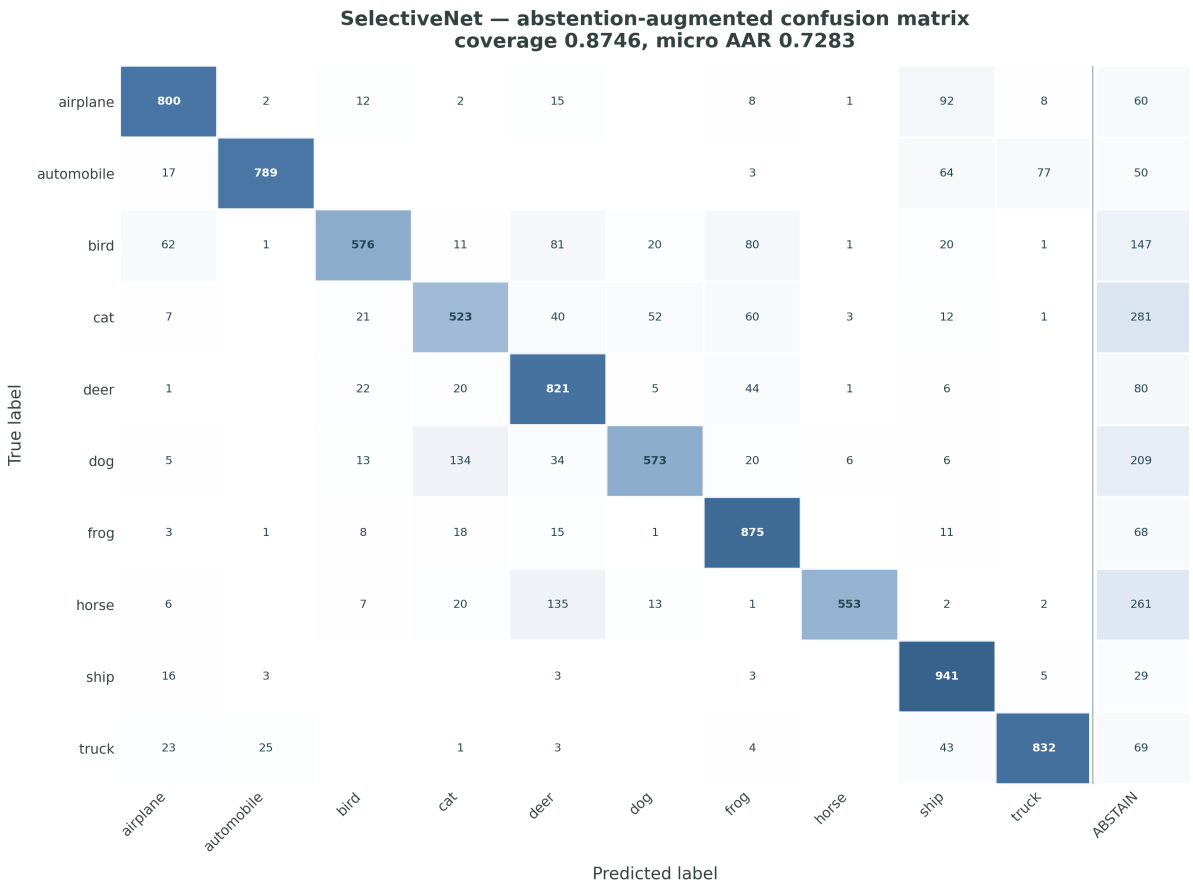



Figure 5.23: Abstention-aware confusion matrix for SelectiveNet on CIFAR-10 dataset.

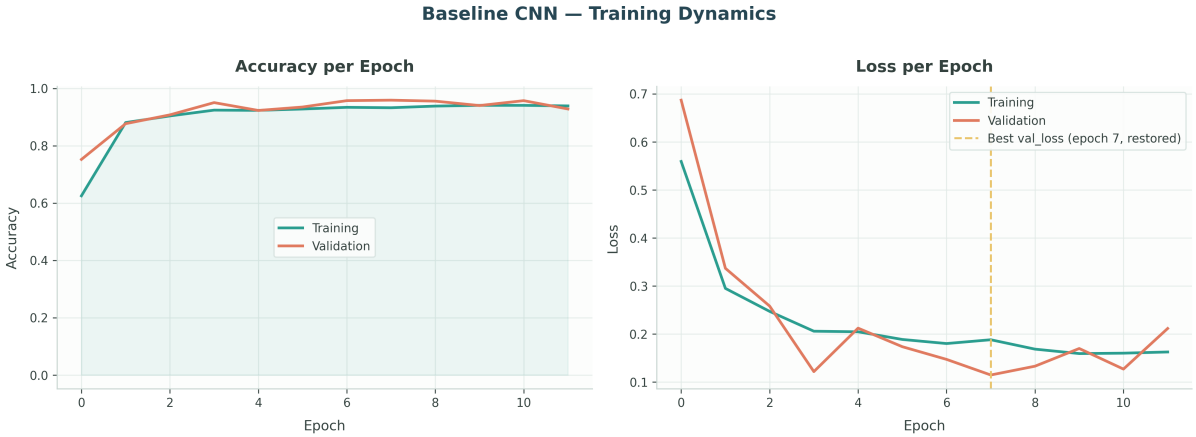



Figure 5.24: Training and validation accuracy and loss for the chest x-ray baseline.

Table 5.8: Class-wise performance of the classification model on Chest X-ray dataset.

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| NORMAL | 0.8970 | 0.9308 | 0.9136 | 159 |
| PNEUMONIA | 0.9739 | 0.9603 | 0.9671 | 428 |
| Macro average | 0.9355 | 0.9455 | 0.9403 | 587 |
| Weighted average | 0.9531 | 0.9523 | 0.9526 | 587 |
| Accuracy | | 0.9523 | | 587 |

## 5.3 Chest X-ray

### 5.3.1 Baseline training dynamics

Pooling the distributed folders gave 1,583 normal and 4,273 pneumonia images. The stratified re-split produced 4,684 training, 585 validation and 587 test images, the test set containing 159 normal and 428 pneumonia scans. Class weights of 1.85 for normal and 0.685 for pneumonia were applied during training.

The curves illustrated in Figure 5.24 are noisier than the MNIST and CIFAR-10 equivalents which follows from the smaller dataset and the batch size of 32. Validation loss falls to 0.1147 at epoch 8, rises, recovers and rises again and early stopping terminated training at epoch 12 with the epoch-8 weights restored. The oscillation in validation loss across epochs 9 to 12, e.g., 0.1330, 0.1698, 0.1270, 0.2116, is large relative to the differences being measured.

### 5.3.2 Baseline performance

The baseline reached 95.23 percent test accuracy, as presented in Table 5.8. In terms of the confusion cells in Figure 5.25, it produced 411 True Positives (TPs), 17 FNs, 11 FPs and 148 True Negatives (TNs). The 17 FNs are the figure that matters clinically. Each is a pneumonia case the system declared normal and it did so without any signal that the case was difficult.

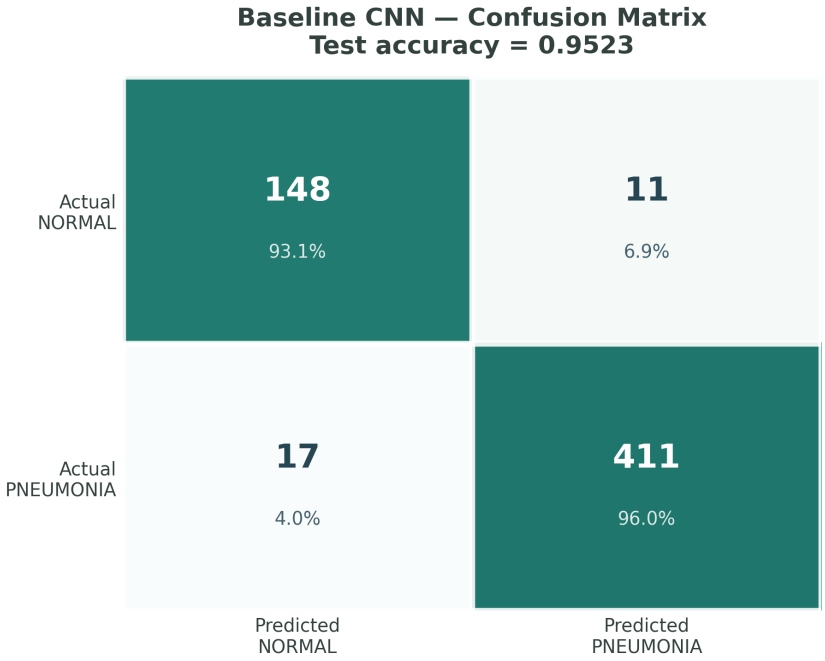



Figure 5.25: Confusion matrix for the baseline CNN on the chest x-ray test set.

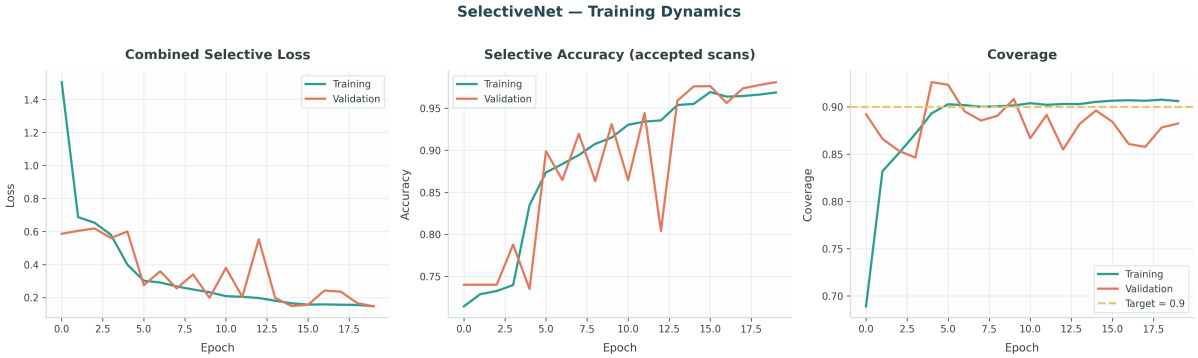



Figure 5.26: Training dynamics for the chest x-ray SelectiveNet.

### 5.3.3 Selective model training dynamics

The coverage trace, presented in Figure 5.26, is the most instructive of the four. It begins at 0.689 in the first epoch, climbs through 0.83 and 0.85 and settles around 0.90 from roughly epoch 6 onward tracking the target closely. Selective accuracy lags behind it, e.g., 0.714 in the first epoch, still 0.733 by the third then rising through 0.874 at epoch 6 to 0.969 by epoch 16. The ordering is informative. The coverage penalty dominates the early epochs pushing the gate open before the classifier has learned much and the selective risk term only begins to shape behaviour once coverage is near target. This is the training dynamic the SelectiveNet formulation is designed to produce.

The validation selective accuracy oscillates through the middle epochs, e.g., 0.899, 0.865, 0.919, 0.863, 0.931, 0.864, before stabilising above 0.95.

Table 5.9: Class-wise performance of SelectiveNet on accepted samples on Chest X-ray dataset.

| Class | Precision | Recall | F1-score | Support (Accepted) |
|---|---|---|---|---|
| NORMAL | 0.9510 | 0.9444 | 0.9477 | 144 |
| PNEUMONIA | 0.9789 | 0.9815 | 0.9802 | 378 |
| Macro average | 0.9650 | 0.9630 | 0.9640 | 522 |
| Weighted average | 0.9712 | 0.9713 | 0.9712 | 522 |
| Accuracy | | 0.9713 | | 522 |

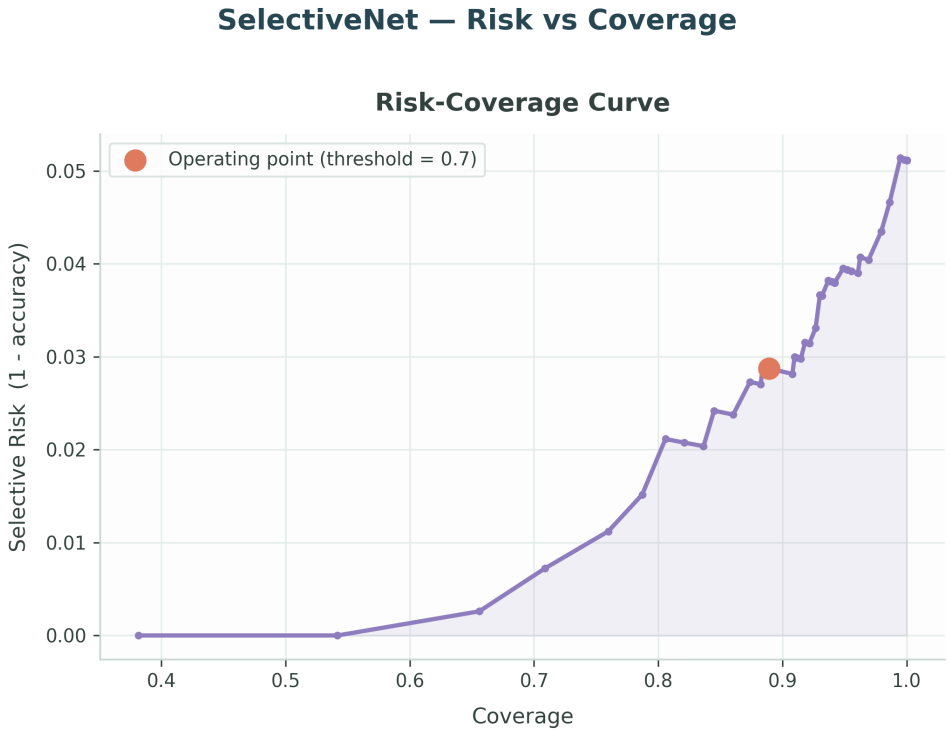



Figure 5.27: Risk–coverage curve for the chest x-ray SelectiveNet.

### 5.3.4 Selective model performance

The trained model accepted 522 of 587 scans (presented in Table 5.9), a coverage of 0.8893 against a target of 0.90 (shown in Figure 5.27), slightly below target, as with CIFAR-10. Selective accuracy on the accepted subset was 0.9713 with a selective risk of 0.0287. Ungated accuracy over all 587 scans, ignoring the reject decision, was 0.9489.

In terms of the confusion matrix in Figure 5.28, it produced 371 TPs, 7 FNs, 8 FPs and 136 TNs. The FNs reduced from 17 to 7, as well as, the FP from 11 to 8, compared to the baseline CNN.

### 5.3.5 Behaviour under FGSM

The chest adversarial sweep was run over 50 randomly drawn scans. A similar pattern can be observed in the Chest X-ray results where selective accuracy increases as selective coverage decreases. This inverse relationship indicates that the SelectiveNet abstains from uncertain or

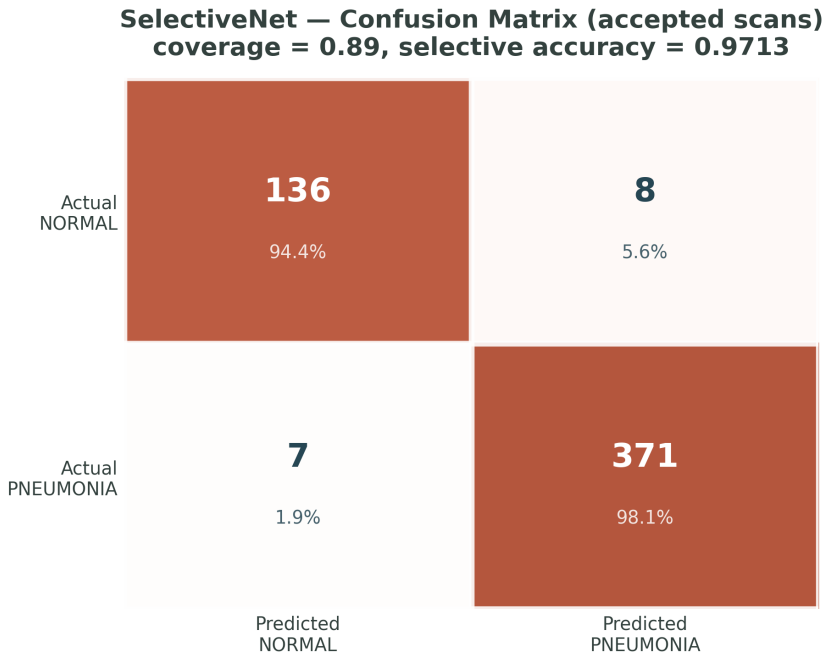



Figure 5.28: Confusion matrix for the chest x-ray SelectiveNet over accepted scans only.

adversarial samples as the perturbation strength increases. By excluding potentially misclassified adversarial instances from the accepted set, the model maintains a high selective accuracy despite the degradation in classification performance. This behavior demonstrates the model's ability to trade coverage for reliability under adversarial conditions, as illustrated in Figure 5.29.

### 5.3.6 Dual-head explanations

The chest attributions are legible which follows from the input resolution. On accepted scans the prediction-head map tends to fall on the lung fields (shown in Figure 5.30) which is the region a clinician would attend to and is the observation usually reported as evidence that a model has learned something reasonable.

### 5.3.7 Abstention-aware evaluation

The optimism gap in accuracy is 0.1019 (shown in Table 5.10 and Figure 5.31). Under accepted-only evaluation the two models look near-identical with accuracies of 0.9630 and 0.9455. Under the abstention-aware convention they are separated by around ten points.

As per the confusion matrices presented in Figure 5.32 and 5.33, FNs fall from 17 to 7 and FPs from 11 to 8. Both fall in absolute terms over the same test population. Therefore, the reduction is not an artefact of a shrinking denominator. 17 pneumonia cases that the baseline declared normal are, under the selective model, either classified correctly or referred and the abstention column shows that 50 pneumonia cases were referred. The honest statement is that whether that exchange is favourable depends on the cost of a referral against the cost of a

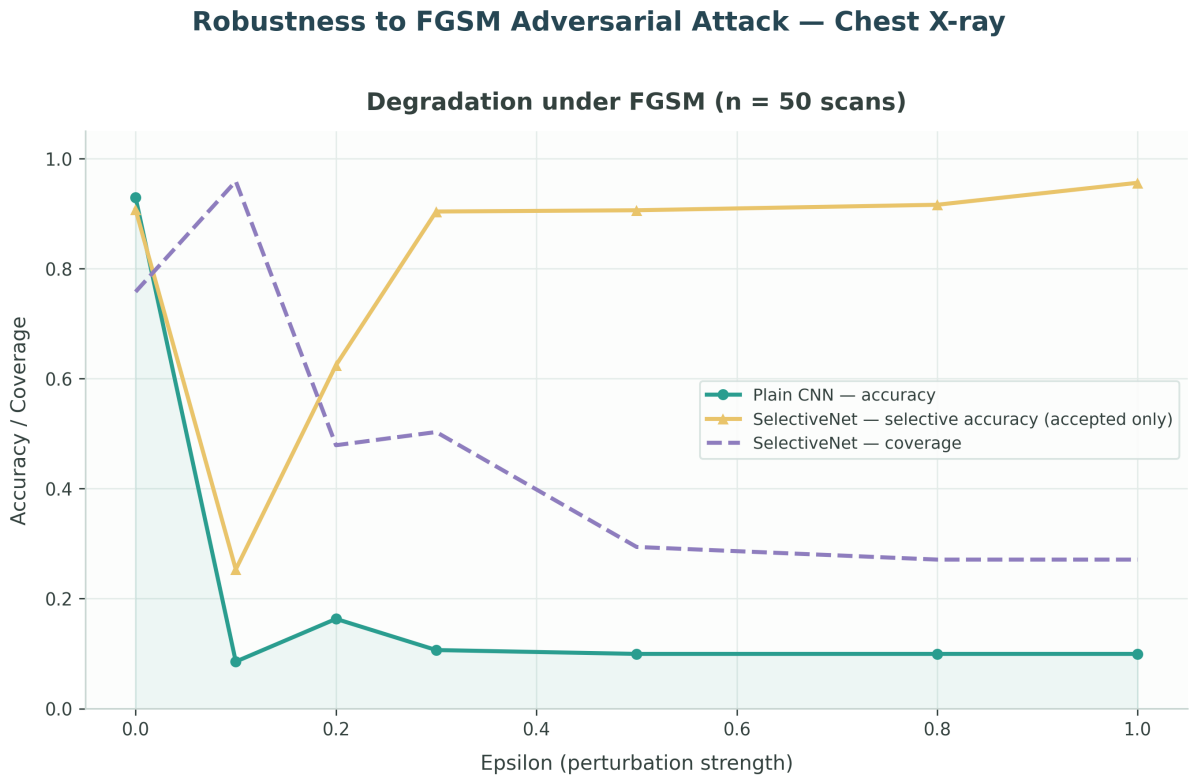



Figure 5.29: Degradation under FGSM on the chest x-rays (n = 50).

Table 5.10: Comparison of classification performance for Chest X-ray dataset.

| **Quantity (Pneumonia as Positive)** | **Plain CNN** | **SelNet Accepted-Only** | **SelNet Abstention-Aware** |
|---|---|---|---|
| Precision | 0.9739 | 0.9789 | 0.9789 |
| Recall / AAR | 0.9603 | 0.9815 | 0.8668 |
| Specificity / AAS | 0.9308 | 0.9444 | 0.8553 |
| F1-score | 0.9671 | 0.9802 | 0.9195 |
| Balanced Accuracy / AABA | 0.9455 | 0.9630 | 0.8611 |

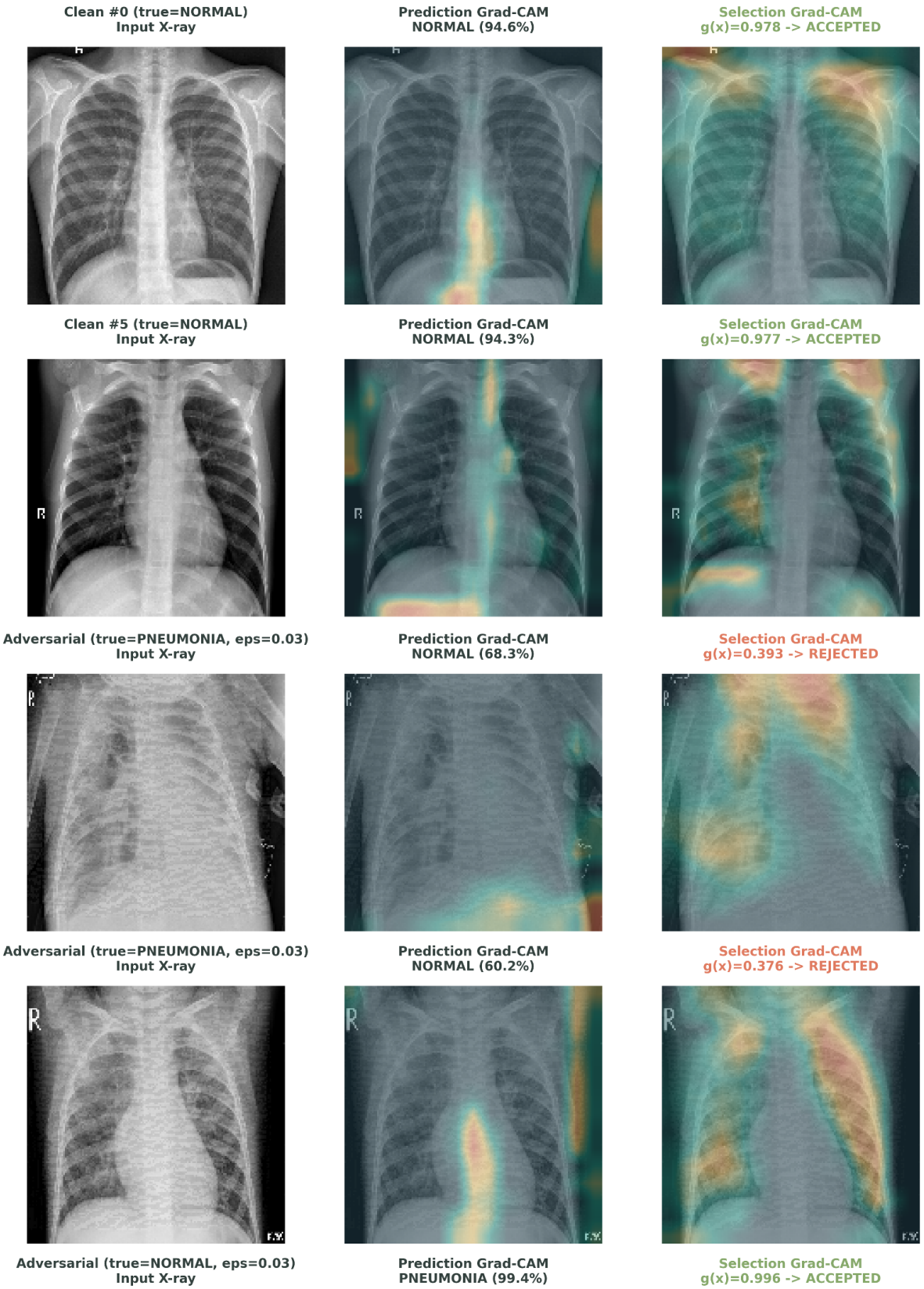



Figure 5.30: Grad-CAM attributions on chest x-rays, accepted against rejected.

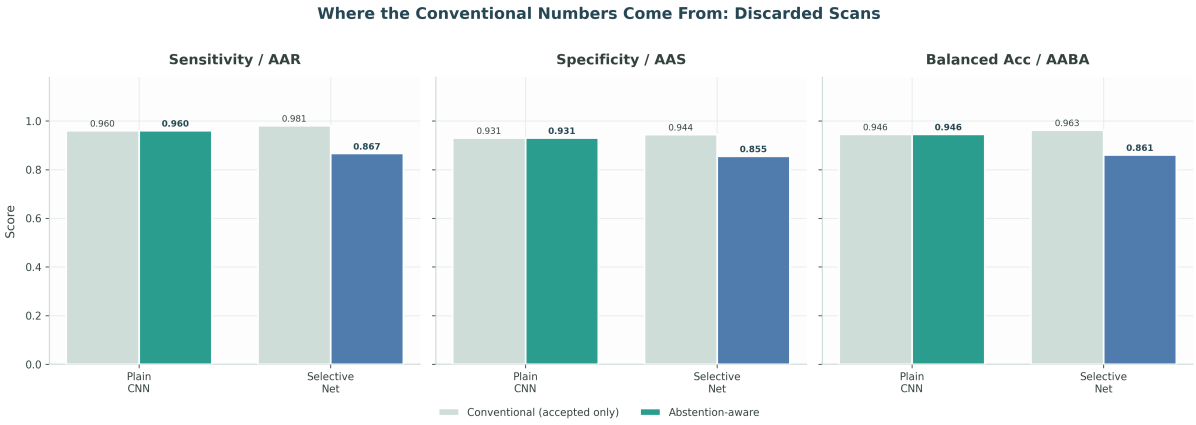



Figure 5.31: Conventional against abstention-aware scores for both chest x-ray models.

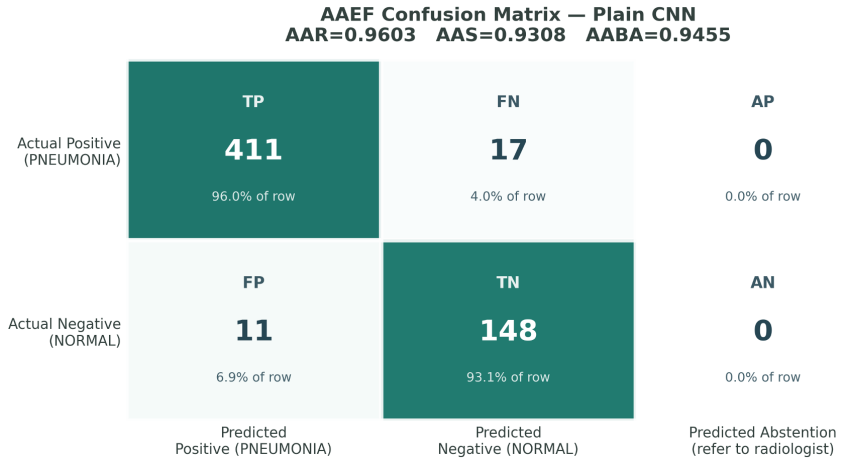



Figure 5.32: Abstention-aware confusion matrix for CNN on Chest X-ray dataset.

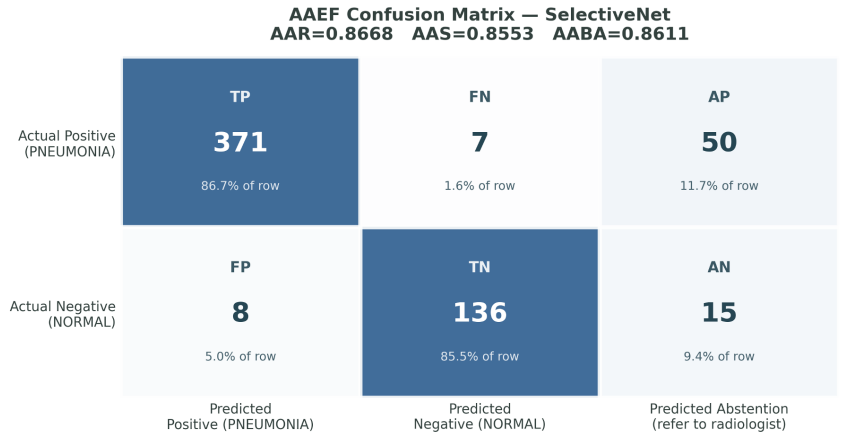



Figure 5.33: Abstention-aware confusion matrix for SelectiveNet on Chest X-ray dataset.

missed diagnosis.

The second concerns, presented in Table 5.10 the two conventions. Accepted-only balanced accuracy is 0.9630 for the selective model against 0.9455 for the baseline. Therefore, on the cases it decides it is the better classifier. Abstention-aware balanced accuracy is 0.8611 against

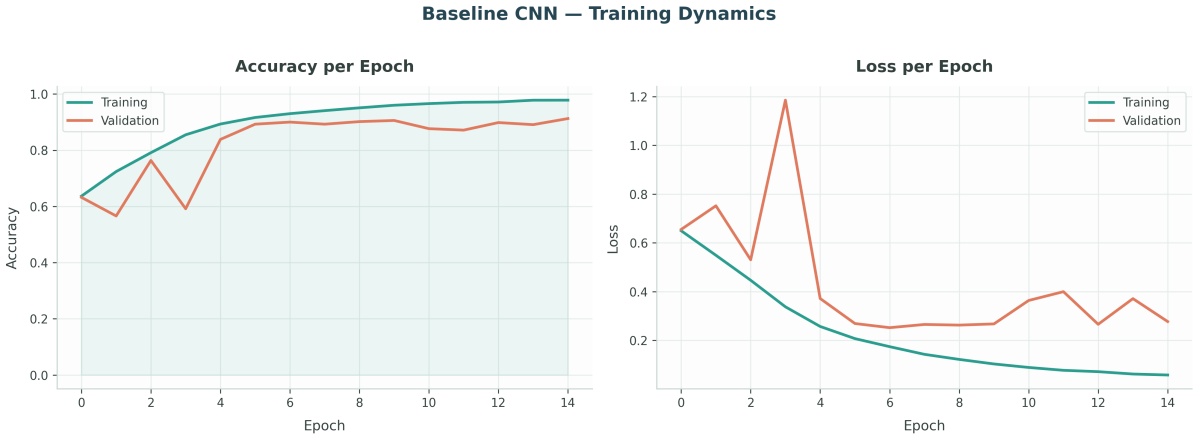



Figure 5.34: Training and validation accuracy and loss for the Cats vs Dogs baseline CNN.

0.9455, so over the full population of scans presented to it, it resolves fewer correctly. The optimism gap of 0.1019 is the size of the discrepancy and a report giving only the first figure would overstate delivered performance by that amount. The third concerns which class is being deferred. Coverage within the pneumonia class is (428 - 50) / 428 = 0.883 and within the normal class (159 - 15) / 159 = 0.906. The selector defers pneumonia cases at the higher rate. This asymmetry is invisible in the aggregate coverage figure of 0.8893. Whether it is desirable is arguable in both directions. Deferring the class where a missed decision carries more consequence routes those cases to a human, which is defensible. It also means the system is least autonomous precisely where it would be most useful. The framework does not adjudicate this. It makes the asymmetry visible.

## 5.4 Cats vs Dogs

### 5.4.1 Baseline training dynamics

The downloaded archive contained 12,501 images per class, of which 1,590 were removed as corrupt or non-decodable, leaving 23,410 usable files. These were split into 19,899 training and 3,511 held-out images. The latter divided into validation and test portions. The materialised test set contained 1,760 images.

The validation curves, presented in Figure 5.34, track the training curves more closely in this case indicating a smaller gap between training and validation performance. This behavior is consistent with the characteristics of the binary classification task which uses higher-resolution input images and a larger training set. The closer alignment suggests that the model generalizes more consistently to unseen validation samples and exhibits less pronounced overfitting during training.

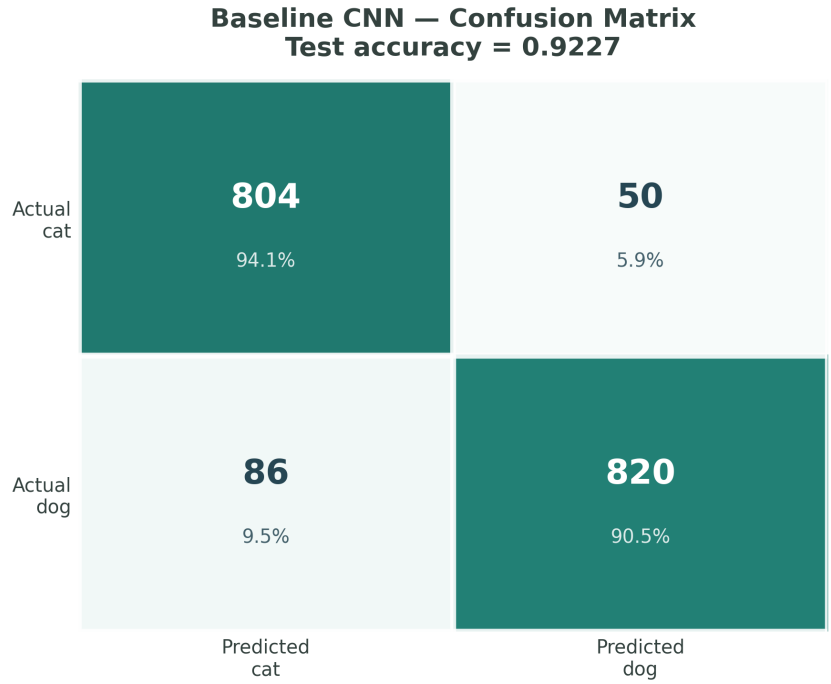



Figure 5.35: Confusion matrix for the baseline CNN on the Cats vs Dogs test set.

Table 5.11: Per-class classification report for the Cats vs Dogs baseline.

| Class | Precision | Recall | F1-score | Specificity | Support |
|---|---|---|---|---|---|
| Dog (Positive) | 0.9425 | 0.9051 | 0.9234 | 0.9415 | 906 |
| Cat (Negative) | 0.9034 | 0.9415 | 0.9220 | 0.9051 | 854 |
| Macro Average | 0.9230 | 0.9233 | 0.9227 | — | 1,760 |
| Accuracy | | 0.9227 | | | 1,760 |

### 5.4.2 Baseline performance

As per the confusion matrix, presented in Figure 5.35, the baseline model produced 820 TPs, 86 FNs, 50 FPs and 804 TNs on the test set.

The baseline CNN achieved an overall accuracy of 92.27% indicating strong classification performance on the Cats vs. Dogs test set, as shown in Table 5.11. Performance was relatively balanced between the two classes with macro precision, recall and F1-score of 0.9230, 0.9233, and 0.9227, respectively. The model performed slightly better in identifying dogs as positive samples achieving a recall of 0.9051 and specificity of 0.9415. For cats, the recall was higher at 0.9415, although specificity decreased to 0.9051. The small differences between the class-wise metrics suggest that the model does not exhibit substantial class-specific bias. The results provide a strong baseline against which the SelectiveNet model can be evaluated, in terms of, whether abstention can improve reliability on uncertain samples.

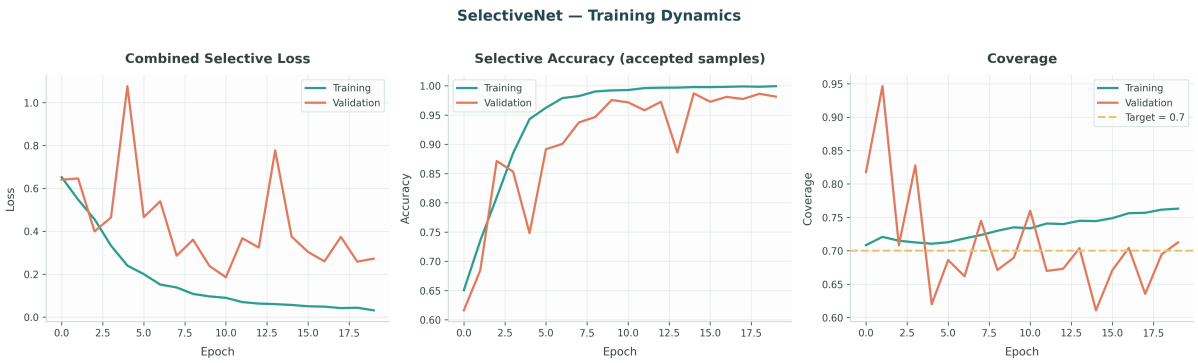



Figure 5.36: Training dynamics for the Cats vs Dogs SelectiveNet over twenty epochs.

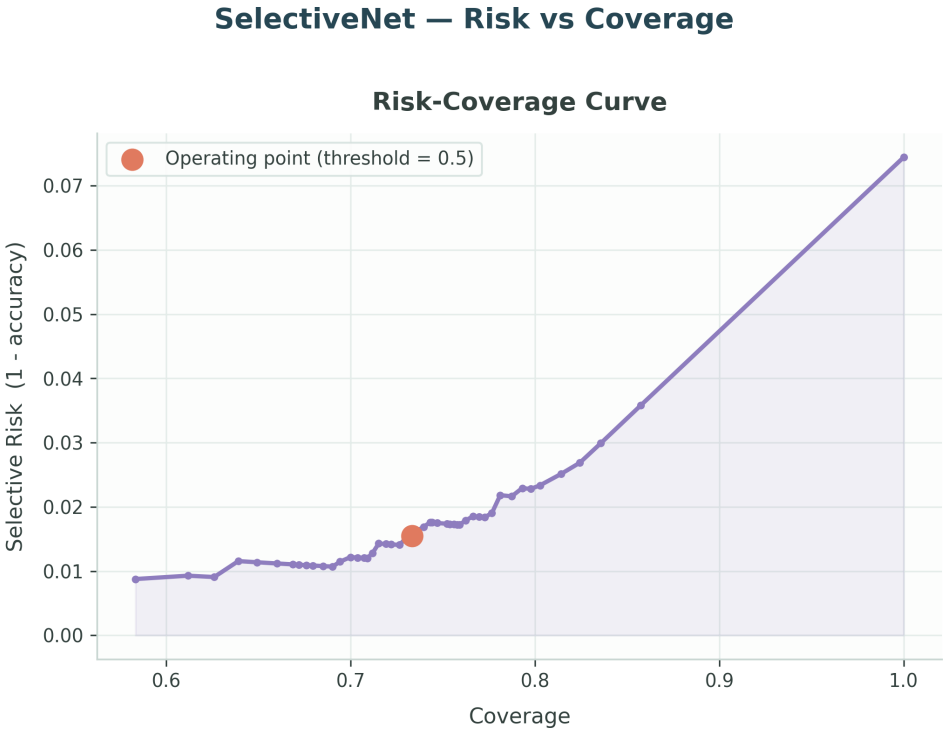



Figure 5.37: Risk–coverage curve for the Cats vs Dogs SelectiveNet.

### 5.4.3 Selective model training

The training curves, presented in Figure 5.36, showed consistent convergence in both loss and accuracy. Validation accuracy remained stable and closely aligned with training accuracy throughout the training process indicating good generalization. Although the validation loss fluctuated during the early stages, it gradually stabilized and converged toward the end of training. Empirical coverage on the test set was 0.7335, above the target of 0.70, as shown in Figure 5.37.

### 5.4.4 Selective model performance

The absolute error reduction is the largest observed anywhere in this work, as shown in Figure 5.38. Total errors fall from 136 to 20, with FNs dropping from 86 to 16 and FPs from

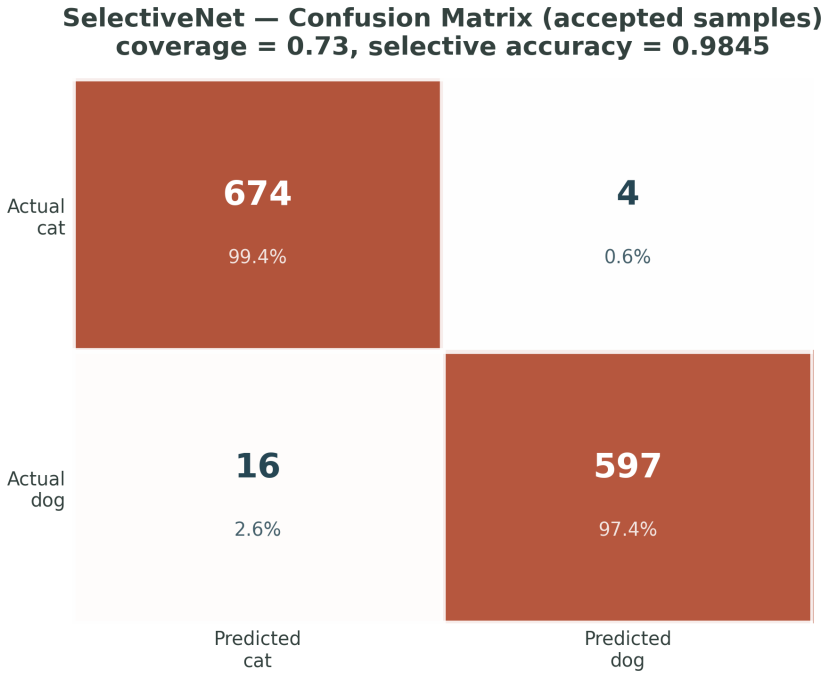



Figure 5.38: Confusion matrix for the SelectiveNet on the Cats vs Dogs test set.

50 to 4. The cost is 469 deferrals, more than a quarter of the test set.

The selector is trading a great deal of coverage for a small residual error rate which the accepted-set accuracy of 0.9845, as presented in Table 5.12. The macro precision, recall, and F1-score were 0.9851, 0.9840, and 0.9845, respectively, indicating highly balanced classification performance. The model achieved slightly higher recall for cats (0.9941) than for dogs (0.9739), while dog precision was higher (0.9933 vs. 0.9768). The high specificity for both classes (0.9941 for dogs and 0.9739 for cats) further indicates effective class discrimination. The results suggest that SelectiveNet improves prediction reliability among the samples it accepts, although this improvement should be interpreted together with its coverage, since the reported metrics exclude abstained samples.

Table 5.12: Per-class classification report for the Cats vs Dogs SelectiveNet over the 1,291 accepted images only.

| Class | Precision | Recall | F1-score | Specificity | Support (Accepted) |
|---|---|---|---|---|---|
| Dog (Positive) | 0.9933 | 0.9739 | 0.9835 | 0.9941 | 613 |
| Cat (Negative) | 0.9768 | 0.9941 | 0.9854 | 0.9739 | 678 |
| Macro Average | 0.9851 | 0.9840 | 0.9845 | — | 1,291 |
| Accuracy | | 0.9845 | | | 1,291 |

### 5.4.5 Behaviour under FGSM

Figure 5.39 illustrated the sweep, run over 1,000 samples. Adversarial perturbations negatively affect the performance of the baseline CNN. As the perturbation strength ($\epsilon$) increases,

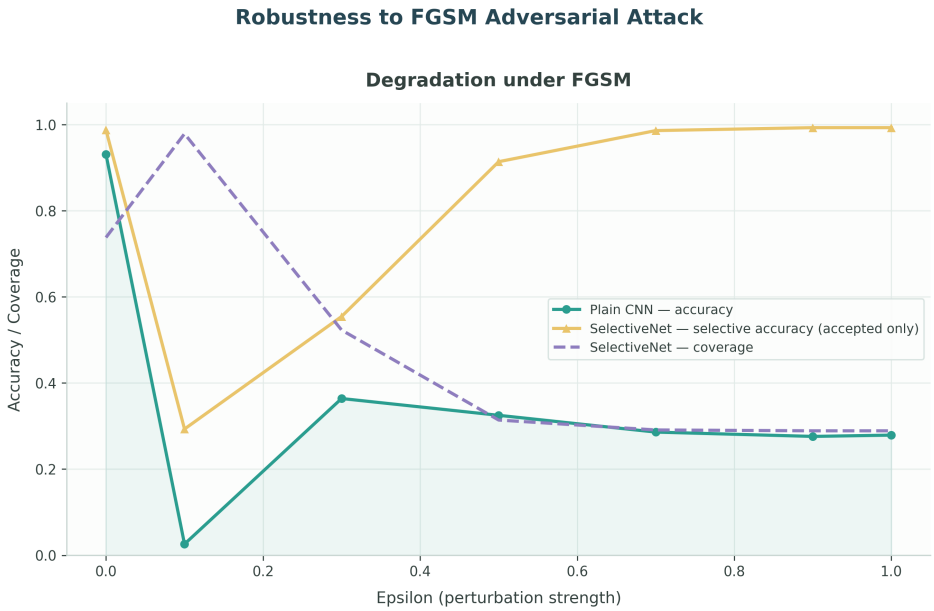



Figure 5.39: Degradation under FGSM on Cats vs Dogs (n = 1,000).

the model's classification performance gradually deteriorates. In contrast, SelectiveNet exhibits the opposite trend, with selective accuracy improving as $\epsilon$ increases. This behavior can be attributed to the dataset containing a mixture of clean and adversarial images. As the perturbation strength increases, SelectiveNet reduces its coverage by abstaining from more uncertain or adversarial samples. Consequently, the accepted set contains a higher proportion of clean and confidently classified images leading to improved selective accuracy. Therefore, the increase in selective accuracy does not indicate that the model becomes more accurate on adversarial images. Rather, it reflects its ability to reject difficult samples and maintain higher reliability on the samples it accepts.

### 5.4.6 Dual-head explanations

At 128 x 128 input resolution (shown in Figure 5.40), these are among the most detailed attributions in the study. The comparison shows a shift in the selection head's focus in the last image. Otherwise, it was consistent with the prediction head.

### 5.4.7 Abstention-aware evaluation

SelectiveNet outperformed the plain CNN on the accepted samples achieving an F1-score of 0.9835 and balanced accuracy of 0.9840, compared with 0.9234 and 0.9233 for the baseline, respectively, as shown in Table 5.13. Its precision and specificity also increased to 0.9933 and 0.9941, indicating reliable predictions among accepted images.

However, when abstentions are incorporated into the evaluation, as presented in Table 5.13 and Figure 5.41, performance decreases with balanced accuracy falling to 0.7241. This reduction reflects the cost of abstention rather than poor classification of accepted samples.




Figure 5.40: Grad-CAM attributions on Cats vs Dogs, accepted against rejected.

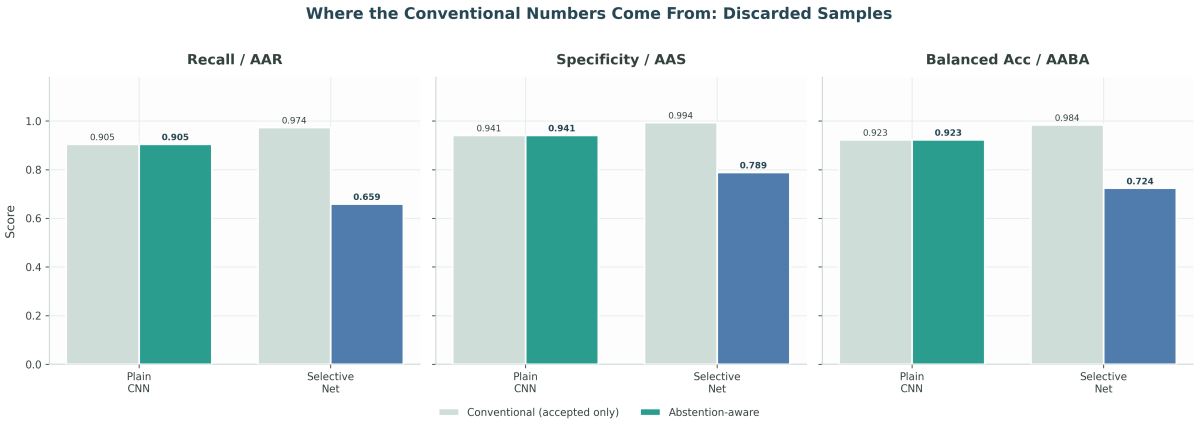



Figure 5.41: Conventional against abstention-aware scores for both Cats vs Dogs models.

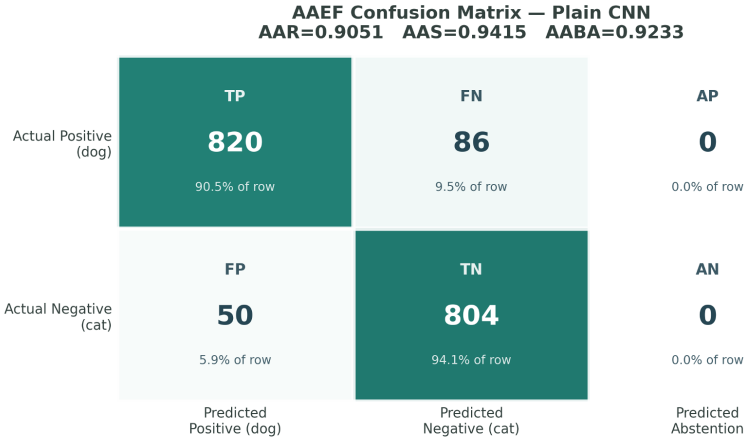



Figure 5.42: Abstention-aware confusion matrix for CNN on Cats vs Dogs dataset.

The results demonstrate the key trade-off of selective classification. SelectiveNet improves reliability by rejecting uncertain samples. The results demonstrate a clear accuracy–coverage trade-off introduced by SelectiveNet. The plain CNN classifies all samples achieving 100% coverage with a balanced accuracy of 0.9233. In contrast, SelectiveNet accepts only 73.35% of the samples and abstains from 26.65% but its accepted predictions achieve a higher balanced accuracy of 0.9840. Among the accepted samples, SelectiveNet improves recall from 0.9051 to 0.9739 and specificity from 0.9415 to 0.9941. Figure 5.42 and 5.43 illustrate that it also reduces FPs from 50 to 4 and FNs from 86 to 16. These results indicate that the abstention mechanism effectively filters out difficult or uncertain samples. However, the abstention-aware metrics (AAR = 0.6589, AAS = 0.7892, AABA = 0.7241) are lower than the corresponding accepted-only metrics. This is expected because abstained samples are treated as unresolved cases rather than correct predictions. This dataset produces the largest optimism gap in the study, at 0.2599.

Table 5.13: Consolidated Cats vs Dogs evaluation over the 1,760-image test set.

| **Model** | **TP** | **FN** | **AP** | **FP** | **TN** | **AN** | **Coverage** | **Recall (Acc.)** | **Spec. (Acc.)** | **BA (Acc.)** | **AAR** | **AAS** | **AABA** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Plain CNN | 820 | 86 | 0 | 50 | 804 | 0 | 1.0000 | 0.9051 | 0.9415 | 0.9233 | 0.9051 | 0.9415 | 0.9233 |
| SelectiveNet | 597 | 16 | 293 | 4 | 674 | 176 | 0.7335 | 0.9739 | 0.9941 | 0.9840 | 0.6589 | 0.7892 | 0.7241 |

Table 5.14: Comparison of classification performance Cats vs Dogs metrics under both conventions.

| **Quantity (Dog as Positive)** | **Plain CNN** | **SelNet Accepted-Only** | **SelNet Abstention-Aware** |
|---|---|---|---|
| Precision | 0.9425 | 0.9933 | 0.9933 |
| Recall / AAR | 0.9051 | 0.9739 | 0.6589 |
| Specificity / AAS | 0.9415 | 0.9941 | 0.7892 |
| F1-score | 0.9234 | 0.9835 | 0.7924 |
| Balanced Accuracy / AABA | 0.9233 | 0.9840 | 0.7241 |

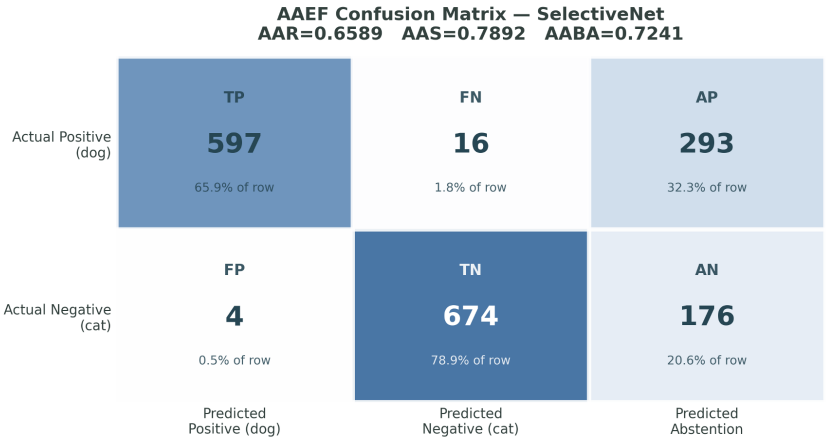



Figure 5.43: Abstention-aware confusion matrix for SelectiveNet on Cats vs Dogs dataset.

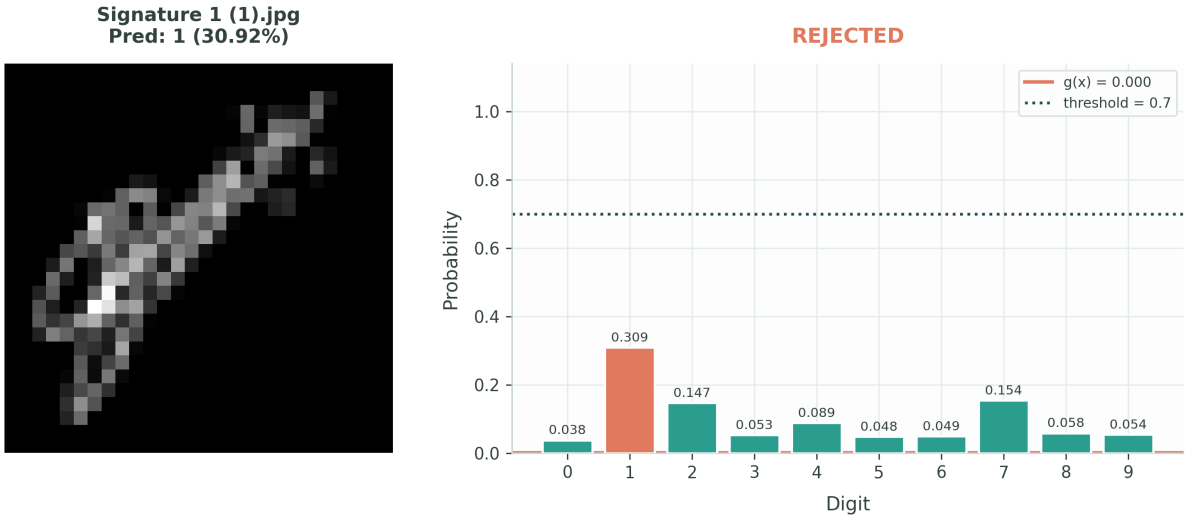



Figure 5.44: OOD handwritten signature evaluated using the MNIST-trained SelectiveNet and rejected by the selection head.

## 5.5 Behaviuor under OOD

Four images were randomly selected to evaluate the model's ability to handle OOD inputs. The selected images consisted of a handwritten signature, a car, the digit '7' and a panther. The test was performed using the SelectiveNet model trained on MNIST. Among the four images, only the digit '7' was accepted by the selection head and correctly classified. The remaining three OOD images were rejected demonstrating the model's ability to abstain from inputs that differ from the training distribution.

The signature image was incorrectly classified as digit 1 with a prediction confidence of 30.92%, as depicted in Figure 5.44. However, its selection score was only 0.309 causing the selection head to reject the sample.

Similarly, the car image was classified as digit 5 with a prediction confidence of 35.9%, but the selection head assigned it a score of 0.359 and, therefore, rejected it, as illustrated in Figure 5.45. In contrast, the digit '7' image was correctly classified with both the prediction confidence

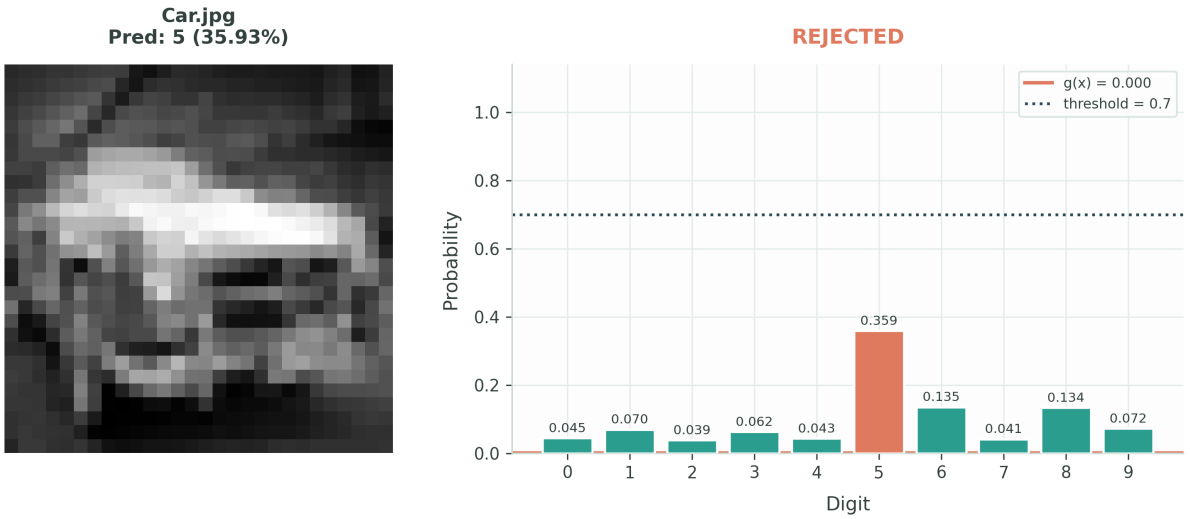



Figure 5.45: OOD car image evaluated using the MNIST-trained SelectiveNet and rejected by the selection head.

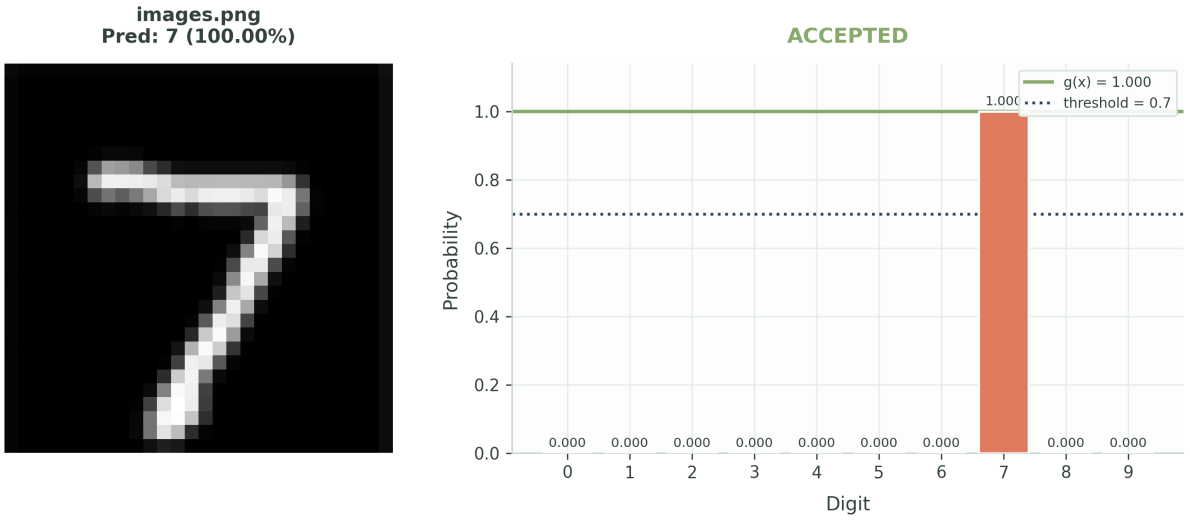



Figure 5.46: In distribtion digit '7' using the MNIST-trained SelectiveNet and accepted by the selection head.

and selection score reaching 100%, resulting in its acceptance, as shown in Figure 5.46.

The panther image was incorrectly classified as digit '0' with a low prediction confidence of 17.04%. Its selection score was also low (0.17) leading the selection head to reject the sample, as presented in Figure 5.47.

These observations suggest that the selection head can identify several OOD inputs and abstain from making unreliable predictions while accepting an in-distribution sample with high confidence.

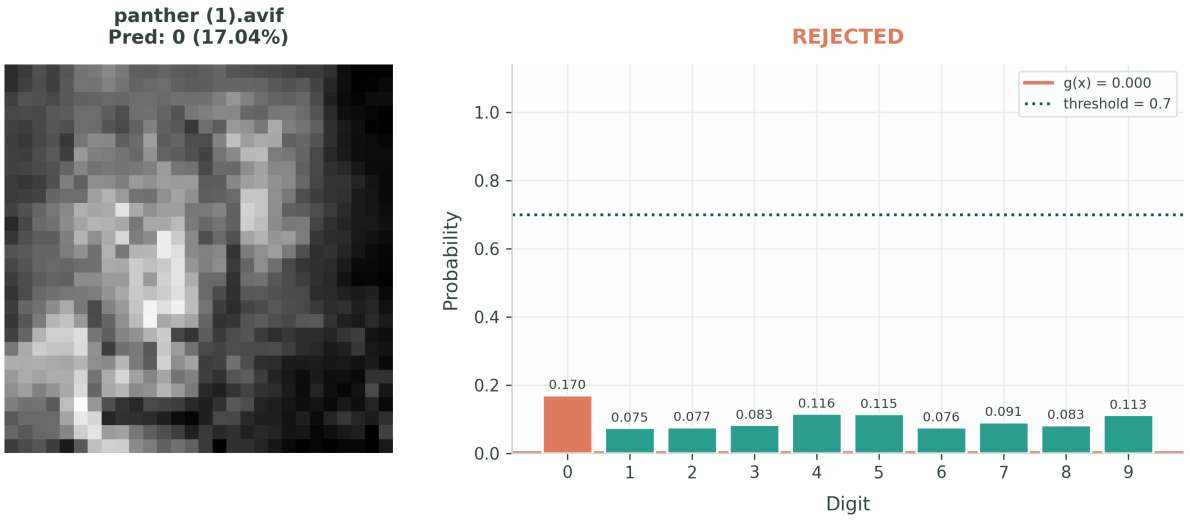



Figure 5.47: OOD panther image evaluated using the MNIST-trained SelectiveNet and rejected by the selection head.

# Chapter 6

# Discussion

## 6.1 RQ1 — behaviour under adversarial perturbation

On MNIST the selection head behaves as the safety argument requires. Coverage falls monotonically from 0.9355 to 0.4615 as epsilon rises from 0 to 0.30, while selective accuracy holds between 0.97 and 1.00 throughout. The model withdraws from the perturbed inputs rather than misclassifying them and the failure mode changes from silent error to declared abstention. On Cats vs. Dogs the pattern holds at moderate and high perturbation. The selector is most permissive at the perturbation strength where the prediction head is least reliable. The answer to RQ1 is, therefore, conditional. Withdrawal under attack is observable and, on one dataset, pronounced and monotone. However, it is not a property that can be assumed to hold across datasets or across the whole range of perturbation strengths.

## 6.2 RQ2 — detection and error reduction

The error-reduction of this question has a consistent answer across all four datasets, as shown in Table 6.1. Absolute error counts fall in every case and on the two binary datasets both error types fall together.

The relabeling objection is not sustained. In every case the reduction in errors is real, in the sense that the errors are fewer over an unchanged test population and it is not merely the result of moving misclassified samples into an abstention column. But the final column shows the price. Between four and roughly fourteen deferrals are required per error avoided. On MNIST the ratio is worst which is a consequence of the baseline being nearly perfect and leaving few errors to remove. On the chest data the breakdown by error type is the most consequential. False negatives fall from 17 to 7 and false positives from 11 to 8 over the same 587 scans. On Cats vs. Dogs, false negatives fall from 86 to 16 and false positives from 50 to 4. The MNIST probes showed the selection head assigning scores between 0.0000 and 0.0012 to non-digit images while assigning 0.9997 to a genuine digit which is the behaviour a detector should exhibit. This

Table 6.1: Absolute error counts for both models on each dataset, with the number of deferrals required to obtain the reduction.

| Dataset | Baseline Errors | Selective Errors | Reduction | Abstentions | Deferrals per Error Avoided |
|---|---|---|---|---|---|
| MNIST | 66 | 18 | 48 | 691 | 14.4 |
| CIFAR-10 | 1,726 | 1,463 | 263 | 1,254 | 4.8 |
| Chest X-ray | 28 | 15 | 13 | 65 | 5.0 |
| Cats vs. Dogs | 136 | 20 | 116 | 469 | 4.0 |

is suggestive and it is not a benchmark. No standard out-of-distribution dataset was used and no detection metric was computed. The correct conclusion is that the intrinsic signal exists and appears usable with its quality relative to dedicated detectors left open.

## 6.3 RQ3 — explaining the selection decision

The dual-head extension works as a procedure. Attributions were produced for both heads on all four datasets, from the same forward pass and the same layer, and the two maps differ from one another in every case examined. On the higher-resolution datasets, e.g., chest x-ray at 150 x 150 and Cats vs. Dogs at 128 x 128, the maps localise well enough to be examined. on CIFAR-10, with an 8 x 8 final feature map upsampled to 32 x 32, they do not. The clean-versus-adversarial comparisons show a shift in the selection head's spatial focus on MNIST and Cats vs. Dogs. The Cats vs. Dogs case is instructive in the negative direction. The attribution shifted visibly under perturbation at a point where the selection decision was going wrong. Applying Grad-CAM to the selection head makes the abstention decision inspectable, where before it was a bare number. Inspectability is a precondition for the accountability argument of Chapter 1.

## 6.4 RQ4 — what the framework reveals

The framework's properties were verified and the framework was applied to eight models across four datasets. The quantity of interest is the optimism gap and it is substantial everywhere, as presented in Table 6.2.

The balanced-accuracy gap ranges from 0.069 to 0.260 and tracks the abstention rate closely. These are large numbers by the standards of the differences reported as improvements in the selective classification literature. On three of the four datasets the gap exceeds the entire difference between the baseline's balanced accuracy and a perfect score. The ordering of the two models reverses under the two conventions on all four datasets. Under accepted-only evaluation the selective model is the better classifier everywhere. Under the abstention-aware convention, the baseline is ahead everywhere. Both statements are true of the same experiments and the framework's contribution is to make the second visible in a single number rather than leaving it to be inferred from a coverage figure reported elsewhere. The per-class breakdowns add a second

Table 6.2: The optimism gap across the four datasets. For the binary datasets the middle column is AAR and the accepted-only comparator is accepted-set accuracy; for the multiclass datasets it is micro AAR.

| **Dataset** | **Coverage** | **Accuracy (Accepted)** | **AAR AAR / AAR** | **Accuracy Gap** | **BA (Accepted)** | **AABA** | **BA Gap** |
|---|---|---|---|---|---|---|---|
| MNIST | 0.9309 | 0.9981 | 0.9291 | 0.0690 | 0.9989 | 0.9295 | 0.0694 |
| CIFAR-10 | 0.8746 | 0.8327 | 0.7283 | 0.1044 | 0.9032 | 0.7933 | 0.1099 |
| Chest X-ray | 0.8893 | 0.9713 | 0.8668 | 0.1045 | 0.9630 | 0.8611 | 0.1019 |
| Cats vs. Dogs | 0.7335 | 0.9845 | 0.6589 | 0.3256 | 0.9840 | 0.7241 | 0.2599 |

kind of information. On MNIST, digit 8 has the lowest abstention-aware recall and the six-cell decomposition shows the cause to be 108 deferrals against a single misclassification (a class the model declines rather than one it gets wrong). Accepted-only evaluation would not have shown the weakness at all. Charging abstention as an error without keeping the third column would have shown it and misattributed it. On CIFAR-10 the horse class shows the same structure in more extreme form. The AAS degeneracy anticipated in Chapter 3 was confirmed on both multiclass datasets.

# Chapter 7

# Conclusion and Future Work

## 7.1 Summary of the Work

This study set out to examine whether selective classification constitutes a more responsible design for image classifiers than the conventional forced-decision architecture and to address a defect in how selective classifiers are evaluated. The two aims are connected, e.g., the case for abstention cannot be assessed properly under an evaluation convention that charges nothing for abstaining.

Chapter 2 established that three literatures, e.g., selective classification, out-of-distribution detection and explainable AI, have converged on overlapping mechanisms while measuring their success against incompatible criteria and that the evaluation practice for classifiers with a reject option computes performance over the accepted subset and reports coverage as a separate figure alongside. Chapter 3 developed the AAEF in response, e.g., an extended confusion matrix that treats abstention as a third prediction outcome and preserves every evaluation sample. Together with three metrics defined on it and theoretical properties established. Chapter 4 described the experimental design, e.g., a baseline CNN and a matched SelectiveNet per dataset, sharing the same backbone so that any difference is attributable to the selection mechanism, evaluated on test data, under FGSM perturbation, with Grad-CAM applied to both the prediction head and the selection head. Chapter 5 reported the results across MNIST, CIFAR-10, chest X-ray and Cats vs. Dogs. Chapter 6 read those results against the RAI principles and set out the threats to validity.

On test data the results are consistent. Absolute error counts fell on all four datasets, e.g., from 66 to 18 on MNIST, 1,726 to 1,463 on CIFAR-10, 28 to 15 on the chest X-ray, and 136 to 20 on Cats vs. Dog, and on the two binary datasets both error types fell together, with missed pneumonia cases dropping from 17 to 7 and false alarms from 11 to 8 over the same 587 scans. The objection that abstention relabels errors rather than reducing them is not sustained. But the price is also visible. Between 4 and 14 deferrals were required per error avoided which means the selection head declines many samples the classifier would have handled correctly.

On adversarial behaviour the results was consistent. As the perturbation strength ($\epsilon$) increases, the model's classification performance deteriorates. In contrast, SelectiveNet exhibits

the opposite trend, with selective accuracy improving as $\epsilon$ increases. This behavior can be attributed to the dataset containing a mixture of clean and adversarial images. As the perturbation strength increases, SelectiveNet reduces its coverage by abstaining from more uncertain or adversarial samples. Consequently, the accepted set contains a higher proportion of clean and confidently classified images leading to improved selective accuracy. Therefore, the increase in selective accuracy does not indicate that the model becomes more accurate on adversarial images. Rather, it reflects its ability to reject difficult samples and maintain higher reliability on the samples it accepts.

On explainability the outcome is procedural. Attributions were produced for both heads on all four datasets from a shared extraction path and the two maps differ in every case examined. Whether the attributions are faithful was not tested.

On evaluation the results are the firmest in the study. The framework's properties held under numerical verification and the optimism gap, e.g., the amount by which accepted-only evaluation overstates delivered performance, ranged from 0.069 on MNIST to 0.260 on Cats vs. Dogs. On all four datasets the ordering of the two models reverses between the two conventions, e.g., under accepted-only evaluation the selective model is the better classifier everywhere and under the abstention-aware convention the baseline is ahead everywhere.

## 7.2 Limitations

The limitations were stated where they arose and are gathered here so that the strength of each claim can be judged as a whole.

There was only one attack, e.g., FGSM, non-adaptive, white-box, with no adversarial training. Two of four sweeps used fifty samples, which is too few for the coverage estimates to be interpretable. One attribution method was used, e.g., GradCAM, no faithfulness testing of the kind, no comparison against alternatives, no human evaluation. On CIFAR-10 the $8 \times 8$ final feature map upsampled to $32 \times 32$ produces attributions too coarse to interpret. The baseline and selective models share an architecture specification but not weights. Therefore, some observed difference may reflect initialisation and optimisation variance. No hyperparameter search was conducted. AAEF charges abstention at the same rate as a misclassification of the same class which is a conservative default rather than a cost model. It assumes a classification task with a single deferral action and it scores an abstention as a non-delivery which implicitly assumes the deferred case receives no benefit. Fairness was not addressed. None of the four datasets carries demographic metadata, so the question could not have been investigated within this design even in principle.

## 7.3 Future Work

The abstention cells AP and AN can be computed per subgroup as readily as in aggregate and a selector that defers one group at a systematically higher rate would show as a divergence in class-conditional coverage across groups while the accepted-set metrics remained balanced.

The RQ1 result should be tested against an adversary who knows the architecture. The obvious construction optimises a perturbation that flips the prediction head while holding the selection score above threshold. The informal probing in Chapter 5 should be replaced with a controlled evaluation using standard protocols and metrics, e.g., the near-OOD and far-OOD distinction, against the established baselines, e.g., maximum softmax probability, ODIN, Mahalanobis distance and energy-based scoring. AAEF's equal weighting of abstention and error is its most limiting simplification. A weighted variant in which a deferral is charged at a factor $c$ relative to a misclassification would give $AAR(c) = TP/(TP + FN + cAP)$ with $c = 1$ recovering the framework as it stands and $c = 0$ recovering accepted-only evaluation. In the clinical case the parameters are measurable in principle, e.g., radiologist review time against the cost of a missed diagnosis which makes the chest X-ray setting a good candidate for a worked example. Also, AAEF as defined covers classification with one deferral action. Extension to regression, where abstention would apply to an interval rather than a label. To segmentation, where a model might abstain per pixel or per region and to settings with several distinct deferral targets, refer to a specialist, refer to a generalist, request another scan, would each require rethinking the confusion matrix rather than relabeling it.

# Appendix: Code and Data Availability

## .1 Code Availability

The complete implementation of the proposed experiments, e.g., model training, SelectiveNet implementation, evaluation, abstention-aware metrics and adversarial robustness analysis, is available through the following Google Colab notebook:

- **MNIST:**
  https://colab.research.google.com/drive/1NIr40b0Rs_tSWYWOe7ULwb03dbXVQ3sC?
- **Cat vs Dogs:**
  https://colab.research.google.com/drive/1uWkxR-OvSblpEM_KgURNyKmphcL_hAaH?
- **Chest X-ray:**
  https://colab.research.google.com/drive/1J8v8I41g2-GV0AsMX51gqvuX_Kuuysuj?
- **CIFAR-10:**
  https://colab.research.google.com/drive/1bsl6XBZFG3srBK8RPAFBULwzU_JHauIN?

## .2 Data Availability

The datasets used in this study and their respective sources are described in Chapter 4.

# Bibliography


[1] C. Chen, N. A. M. Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Computers in biology and medicine*, vol. 185, p. 109507, 2025.

[2] X. Deng, M. Shi, B. Khan, Y. H. Choo, F. Ghaffar, and C. P. Lim, "A lightweight cnn model for uav-based image classification," *Soft Computing*, vol. 29, no. 4, pp. 2363–2378, 2025.

[3] W. Ilham and A. Ahmad, "A comprehensive review of convnext architecture in image classification: Performance, applications, and prospects," *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 2, pp. 108–114, 2026.

[4] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

[5] J. Strong, Q. Men, and J. A. Noble, "Trustworthy and practical ai for healthcare: A guided deferral system with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 28413–28421, 2025.

[6] D. Jenavani, R. Deepa, *et al.*, "Learning to decide when not to decide: Coverage-constrained selective classification with abstention," in *2026 IEEE International Conference on AI Engineering and Innovations (AIEI)*, pp. 1–5, IEEE, 2026.

[7] R. S. Ferreira, J. Guérin, K. Delmas, J. Guiochet, and H. Waeselynck, "Safety monitoring of machine learning perception functions: a survey," *Computational Intelligence*, vol. 41, no. 2, p. e70032, 2025.

[8] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. Al Mamun, M. A. Rahman, S. Shatabda, and M. S. H. Mukta, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Decision analytics journal*, vol. 11, p. 100470, 2024.

[9] P. Pawara, E. Okafor, M. Groefsema, S. He, L. R. Schomaker, and M. A. Wiering, "One-vs-one classification for deep neural networks," *Pattern Recognition*, vol. 108, p. 107528, 2020.

[10] G. Taricco, "Entropy regularization in deep reinforcement learning: A structured review across classical control, generative policies, and reasoning language models," *Entropy*, vol. 28, no. 7, p. 811, 2026.

[11] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International conference on machine learning*, pp. 23631–23644, PMLR, 2022.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[15] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[16] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[17] K. Wang, C. Shen, X. Li, and J. Lu, "Uncertainty quantification for safe and reliable autonomous vehicles: A review of methods and applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 3, pp. 2880–2896, 2025.

[18] Y. Xiao, H. Shao, and B. Liu, "Evaluating calibration of deep fault diagnostic models under distribution shift," *Computers in Industry*, vol. 171, p. 104334, 2025.

[19] L. Tamang, M. R. Bouadjenek, R. Dazeley, and S. Aryal, "Handling out-of-distribution data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[20] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on information theory*, vol. 16, no. 1, pp. 41–46, 1970.

[21] L. P. Cordella, C. De Stefano, F. Tortorella, and M. Vento, "A method for improving classification reliability of multilayer perceptrons," *IEEE transactions on neural networks*, vol. 6, no. 5, pp. 1140–1147, 1995.

[22] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss.," *Journal of Machine Learning Research*, vol. 9, no. 8, 2008.

[23] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," in *International conference on algorithmic learning theory*, pp. 67–82, Springer, 2016.

[24] R. El-Yaniv *et al.*, "On the foundations of noise-free selective classification.," *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[25] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," in *International conference on machine learning*, pp. 2151–2159, PMLR, 2019.

[27] T. Albalawi and A. Alzahrani, "Learning selective deferral policies for reliable medical text classification," *Technologies*, vol. 14, no. 6, p. 359, 2026.

[28] G. DeSalvo, C. Mohri, M. Mohri, and Y. Zhong, "Budgeted multiple-expert deferral," *arXiv preprint arXiv:2510.26706*, 2025.

[29] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: improving fairness and accuracy by learning to defer," *Advances in neural information processing systems*, vol. 31, 2018.

[30] K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, *et al.*, "Building machines that learn and think with people," *Nature human behaviour*, vol. 8, no. 10, pp. 1851–1863, 2024.

[31] C. Gonzalez and H. Heidari, "A cognitive approach to human–ai complementarity in dynamic decision-making," *Nature Reviews Psychology*, vol. 4, no. 12, pp. 808–822, 2025.

[32] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[33] L. Floridi and J. Cowls, "A unified framework of five principles for ai in society," *Machine learning and the city: Applications in architecture and urban design*, pp. 535–545, 2022.

[34] M. Lognoul, "Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act–ai act)," *Revue du Droit des Technologies de l'information*, no. 3-4, pp. 145–189, 2025.

[35] G. W. Muoka, Y. Ding, C. C. Ukwuoma, A. Mutale, C. J. Ejiyi, A. K. Mzee, E. S. Gyarteng, A. Alqahtani, and M. A. Al-antari, "A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense," *Mathematics*, vol. 11, no. 20, p. 4272, 2023.

[36] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Survey on adversarial attack and defense for medical image analysis: Methods and challenges," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–38, 2024.

[37] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *NPJ Digital Medicine*, vol. 4, no. 1, p. 4, 2021.

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[39] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[40] A. M. Fayyaz, S. J. Abdulkadir, N. Talpur, S. M. Al-Selwi, S. U. Hassan, and E. H. Sumiea, "Grad-cam (gradient-weighted class activation mapping): A systematic literature review," *Computers in Biology and Medicine*, vol. 198, p. 111200, 2025.

[41] B. C. Cheong, "Transparency and accountability in ai systems: Safeguarding wellbeing in the age of algorithmic decision-making," *Frontiers in Human Dynamics*, vol. 6, p. 1421273, 2024.

[42] I. D. Raji and R. Dobbe, "Concrete problems in ai safety, revisited," *arXiv preprint arXiv:2401.10899*, 2023.

[43] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024.

[44] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, *et al.*, "Position: Trustllm: Trustworthiness in large language models," in *International Conference on Machine Learning*, pp. 20166–20270, PMLR, 2024.

[45] M. Mohamed and F. AlJuaid, "Armor: A multi-layered adaptive defense framework for robust deep learning systems against evolving adversarial threats," *Computer Standards & Interfaces*, p. 104117, 2025.

[46] J. Traub, T. J. Bungert, C. T. Lüth, M. Baumgartner, K. H. Maier-Hein, L. Maier-Hein, and P. F. Jäger, "Overcoming common flaws in the evaluation of selective classification systems," in *Advances in Neural Information Processing Systems*, vol. 37, pp. 2323–2347, 2024.

[47] F. Condessa, J. Bioucas-Dias, and J. Kovačević, "Performance measures for classification systems with rejection," *Pattern Recognition*, vol. 63, pp. 437–450, 2017.

[48] S. Hasan, A. Shekh, M. M. Ali, and S. A. Sakib, "A deep learning approach to identifications of vehicle type using sign hybrid (cnn) and fine tuned vgg-16 model," in *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, pp. 1–6, IEEE, 2026.

[49] N. Aziz, N. Minallah, J. Frnda, M. Sher, M. Zeeshan, and A. H. Durrani, "Precision meets generalization: Enhancing brain tumor classification via pretrained densenet with global average pooling and hyperparameter tuning," *PLOS ONE*, vol. 19, no. 9, p. e0307825, 2024.

[50] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[51] K. Hermann, H. Mobahi, T. Fel, and M. Mozer, "On the foundations of shortcut learning," in *International Conference on Learning Representations*, vol. 2024, pp. 43832–43868, 2024.

[52] G.-R. Han, A. Goncharov, M. Eryilmaz, S. Ye, B. Palanisamy, R. Ghosh, F. Lisi, *et al.*, "Machine learning in point-of-care testing: Innovations, challenges, and opportunities," *Nature Communications*, vol. 16, no. 1, p. 3165, 2025.

[53] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[54] M.-A. Bani, "Smart lies and sharp eyes: Pragmatic artificial intelligence for cancer pathology: Promise, pitfalls, and access pathways," *Cancers*, vol. 18, no. 3, p. 421, 2026.

[55] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: A step-by-step guide," *BMJ*, vol. 386, 2024.

[56] C. Wang, "Calibration in deep learning: A survey of the state-of-the-art," *arXiv preprint arXiv:2308.01222*, 2023.

[57] R. O. Lane, "A comprehensive review of classifier probability calibration metrics," *Data Science*, vol. 9, no. 1, p. 24518492261436231, 2026.

[58] Z. Zhou, S. Wei, X. Zhang, W. Dou, M. Qu, and Y. Cai, "Training deep neural networks with virtual smoothing classes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 23036–23044, 2025.

[59] M. Chidambaram and R. Ge, "Reassessing how to compare and improve the calibration of machine learning models," in *International Conference on Learning Representations*, vol. 2025, pp. 61542–61570, 2025.

[60] S. Ruggieri and A. Pugnana, "Things machine learning models know that they don't know," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 28684–28693, 2025.

[61] K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, and J. Davis, "Machine learning with a reject option: A survey," *Machine Learning*, vol. 113, no. 5, pp. 3073–3110, 2024.

[62] V. Franc, D. Prusa, and V. Voracek, "Optimal strategies for reject option classifiers," *Journal of Machine Learning Research*, vol. 24, no. 11, pp. 1–49, 2023.

[63] M. M. Hasan, M. Abdar, A. Khosravi, U. Aickelin, P. Lio, I. Hossain, A. Rahman, and S. Nahavandi, "Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1–35, 2025.

[64] I. Galil, M. Dabbah, and R. El-Yaniv, "What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers," *arXiv preprint arXiv:2302.11874*, 2023.

[65] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *International conference on machine learning*, pp. 7076–7087, PMLR, 2020.

[66] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[67] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 538–566, 2023.

[68] P. G. Arce, R. Naveiro, and D. Ríos Insua, "A unifying bayesian framework for adversarial robustness," *arXiv preprint arXiv:2510.09288*, 2025.

[69] H. Anthony and K. Kamnitsas, "On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging," in *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, (Cham), pp. 136–146, Springer Nature Switzerland, 2023.

[70] E. Şahin, N. N. Arslan, and D. Özdemir, "Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Computing and Applications*, vol. 37, no. 2, pp. 859–965, 2025.

[71] A. P. T. Minh, "Overview of class activation maps for visualization explainability," *arXiv preprint arXiv:2309.14304*, 2023.

[72] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–39, 2025.

[73] A. Akrout, A. Echtioui, and M. Ghorbel, "Explainable ai techniques with vgg-16 for detecting covid-19," in *2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 1015–1020, IEEE, 2025.

[74] A. M. Salih, Z. Raisi-Estabragh, I. Boscolo Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: Shap and lime," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025.

[75] D. Bhusal, *Towards Reliable and Trustworthy Deep Learning Through Explainability and Interpretability*. PhD thesis, Rochester Institute of Technology, 2026.

[76] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, *et al.*, "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023.

[77] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Computer Science On-line Conference*, (Cham), pp. 15–25, Springer International Publishing, 2023.

[78] H. Shang, J.-M. Langlois, K. Tsioutsiouliklis, and C. Kang, "Precision/recall on imbalanced test data," in *International Conference on Artificial Intelligence and Statistics*, pp. 9879–9891, PMLR, 2023.

[79] Google, "Google colaboratory." `https://colab.research.google.com/`. Accessed: June 17, 2026.

[80] TensorFlow, "Tensorflow python api documentation." `https://www.tensorflow.org/api_docs/python/tf`. Accessed: June 24, 2026.

[81] scikit-learn, "scikit-learn api reference." `https://scikit-learn.org/stable/api/index.html`. Accessed: Accessed: June 29, 2026.

[82] Kaggle, "Mnist dataset." `https://www.kaggle.com/datasets/hojjatk/mnist-dataset`. Kaggle dataset.

[83] Kaggle, "Cifar-10." `https://www.kaggle.com/c/cifar-10`. Kaggle competition dataset.

[84] Kaggle, "Chest x-ray images: Normal and pneumonia." `https://www.kaggle.com/datasets/ghost5612/chest-x-ray-images-normal-and-pneumonia`. Kaggle dataset.

[85] Kaggle, "Dogs vs. cats." `https://www.kaggle.com/c/dogs-vs-cats`. Kaggle competition dataset.

[86] A. R. Ismail, M. Z. R. Azhary, and N. A. Hitam, "Evaluating adan vs. adam: An analysis of optimizer performance in deep learning," in *International Symposium on Intelligent Computing Systems*, (Cham), pp. 251–263, Springer Nature Switzerland, 2024.

[87] J. Sen, "The fgsm attack on image classification models and distillation as its defense," in *International Conference on Advances in Distributed Computing and Machine Learning*, (Singapore), pp. 347–360, Springer Nature Singapore, 2024.