



Techniques to Overcome Class Imbalance in Anomaly/Defect Detection

LK489 - Master of Engineering in Computer Vision and Artificial Intelligence

Final Project Report

Tanishq Rahul Shelke
24263583

Ciaran Eising, Reenu Mohandas

22/08/2025

Abstract

Anomaly detection is a fundamental task in machine learning and artificial intelligence, with critical applications across cybersecurity, industrial inspection, and medical imaging. This project investigates strategies to address anomaly detection under severe class imbalance, a common challenge where normal samples vastly outnumber anomalous ones. Three benchmark datasets were employed: UNSW-NB15 for network intrusion detection, MVTec AD for industrial defect detection, and NIH ChestXray14 for medical anomaly detection.

The analytical background established the theoretical framework of classical algorithms (Isolation Forest, One-Class SVM, Local Outlier Factor, Elliptic Envelope), reconstruction-based deep learning models (autoencoders, variational autoencoders, GANs), and ensemble methods (bagging, boosting, stacking), together with imbalance-handling techniques including normal-only training, resampling, cost-sensitive learning, and threshold calibration. The specification and design stage detailed preprocessing strategies, feature extraction pipelines, and the rationale for model selection across different domains. Implementation experiments applied the designed methods to each dataset, highlighting the trade-offs between precision and recall in imbalanced settings.

Results demonstrated that Isolation Forest and LOF achieved high recall and precision in network intrusion detection and industrial defect detection, confirming their robustness under imbalance. However, these same models underperformed in the medical domain, where pathologies are diverse and subtle; in this case, supervised classifiers with oversampling (e.g., kNN, SVM with class weights) showed stronger performance. The findings emphasize that no single anomaly detection method dominates across all domains; rather, success depends on dataset characteristics, appropriate feature engineering, and careful imbalance management.

The project concludes that while unsupervised anomaly detection is effective for domains with clear separation between normal and anomalous distributions (e.g., network traffic, industrial vision), medical imaging requires domain-specific feature learning and supervised imbalance-aware techniques. Future work should integrate advanced generative models such as diffusion models and normalizing flows, domain-informed feature extraction, and ensemble strategies to improve cross-domain generalization.

Declaration

This final report is presented in part fulfilment of the requirements for the LK489 Master of Engineering in Computer Vision and Artificial Intelligence Masters Project.

It is entirely my own work and has not been submitted to any other University or Higher Education Institution or for any other academic award within the University of Limerick.

Where there has been made use of work of other people it has been fully acknowledged and referenced.

Name **Tanishq Rahul Shelke**

Signature **Tanishq Rahul Shelke**

Date **22/08/2025**

Acknowledgement

I would like to thank my supervisors Ciaran Eising and Reenu Mohandas for their valuable guidance and support during the initial planning and preparation stage of this project. Their suggestions have been useful in determining the direction and scope of work proposed. I also want to acknowledge my gratitude to the Science and Engineering staff and faculty of University of limerick for the provision of the basic resources and knowledge that have qualified me to undertake this project. Lastly, I am thankful to my classmate, friends for their support and encouragement as I undertake this research.

Table of Contents

ABSTRACT	II
DECLARATION	III
ACKNOWLEDGEMENT	IV
TABLE OF CONTENTS	V
LIST OF FIGURES	X
LIST OF TABLES	XI
CHAPTER 1: INTRODUCTION	1
1.1 OVERVIEW.....	1
1.2 MOTIVATION FOR THE PROJECT	1
1.2.1 <i>Industrial Manufacturing (MVTec AD)</i>	1
1.2.2 <i>Medical Imaging (NIH Chest X-ray 14)</i>	2
1.2.3 <i>Cybersecurity (UNSW-NB15)</i>	2
1.3 PROJECT AIMS AND OBJECTIVES	3
1.3.1 <i>Objectives</i>	3
1.4 REPORT OUTLINE	5
1.5 CONCLUSION	5
CHAPTER 2: LITERATURE REVIEW	7
2.1 OVERVIEW.....	7
2.2 ANOMALY DETECTION: A SURVEY [ACM COMPUTING SURVEY]	7
2.3 DEEP LEARNING FOR ANOMALY DETECTION: A REVIEW [ACM COMPUTING SURVEY].....	8
2.4 A SURVEY OF NETWORK ANOMALY DETECTION TECHNIQUES [JNCA(2016) 19-31]	9
2.5 ANOMALY DETECTION USING DIFFUSION-BASED MODELS [ARXIV].....	9
2.6 ANOMALY DETECTION USING UNSUPERVISED MACHINE LEARNING ALGORITHMS: A SIMULATION STUDY [SCIENTIFIC AFRICAN 26].....	10
2.7 COMPARATIVE ANALYSIS OF ELLIPTIC ENVELOPE, ISOLATION FOREST, ONE-CLASS SVM, AND LOCAL OUTLIER FACTOR IN DETECTING EARTHQUAKES WITH STATUS ANOMALY USING OUTLIER [IEEE]	11
2.8 UNSUPERVISED ANOMALY DETECTION USING K-MEANS, LOCAL OUTLIER FACTOR, AND ONE-CLASS SVM [IEEE]	11
2.9 ISOLATION FOREST [IEEE]	12
2.10 MVTec AD — A COMPREHENSIVE REAL-WORLD DATASET FOR UNSUPERVISED ANOMALY DETECTION [SPRINGER – CVPR 2019].....	12
2.11 THE MVTec ANOMALY DETECTION DATASET: A COMPREHENSIVE REAL-WORLD DATASET FOR UNSUPERVISED ANOMALY DETECTION(EXTENDED) [SPRINGER - IJCV 2021]	13

2.12 INDUSTRIAL ANOMALY DETECTION WITH NORMALIZING FLOWS [HANNOVER: INSTITUTIONNELLES REPOSITORYUM DER LEIBNIZ]	14
2.13 ANOMALY DETECTION IN MEDICAL IMAGING WITH DEEP PERCEPTUAL AUTOENCODERS [IEEE]	14
2.14 PERFORMANCE EVALUATION OF CNN ARCHITECTURES ON NIH CHEST X-RAY DATASET WITH BOOSTING ENSEMBLE [IEEE]	15
2.15 DEEP ISOLATION FOREST FOR ANOMALY DETECTION [IEEE].....	16
2.16 RECENT PROGRESS OF ANOMALY DETECTION [WILEY ONLINE LIBRARY]	16
2.17 A SURVEY OF ANOMALY DETECTION METHODS IN NETWORKS [IEEE]	17
2.18 ONE-CLASS SUPPORT VECTOR MACHINES APPROACH TO ANOMALY DETECTION [TAYLOR & FRANCIS]	17
2.19 MACHINE LEARNING FOR ANOMALY DETECTION: A SYSTEMATIC REVIEW [IEEE]	18
2.20 MACHINE LEARNING TECHNIQUES FOR ANOMALY DETECTION: AN OVERVIEW [RESEARCHGATE].....	18
2.21 OVERVIEW OF ANOMALY DETECTION TECHNIQUES IN MACHINE LEARNING [IEEE].....	19
2.22 ANOMALY DETECTION ON MEDICAL IMAGES USING AUTOENCODER AND CONVOLUTIONAL NEURAL NETWORK [GOOGLE]	19
2.23 MULTI-LABEL CLASSIFICATION OF CHEST X-RAY ABNORMALITIES USING TRANSFER LEARNING TECHNIQUES	20
2.24 UNSW-NB15: A COMPREHENSIVE DATA SET FOR NETWORK INTRUSION DETECTION SYSTEMS.....	20
2.25 A SURVEY OF ANOMALY DETECTION TECHNIQUES [SPRINGER]	21
2.26 HEART RATE ANOMALY DETECTION IN HEALTHCARE USING ELLIPTIC ENVELOPE AND LOCAL FOREST [SCIENCE DIRECT]22	22
2.27 SURVEY ON ANOMALY DETECTION USING DATA MINING TECHNIQUES[SCIENCE DIRECT].....	22
2.28 CONCLUSION	22
CHAPTER 3: ANALYTICAL BACKGROUND	25
3.1 OVERVIEW.....	25
3.2 ANOMALY DETECTION FRAMEWORK.....	25
3.3 CLASSICAL ANOMALY DETECTION METHODS	26
3.3.1 <i>One-Class Support Vector Machines (OCSVM)</i>	26
3.3.2 <i>Local Factor Outlier (LOF)</i>	27
3.3.3 <i>Isolation Forest</i>	27
3.3.4 <i>Deep Isolation Forest</i>	27
3.3.5 <i>Clustering + LOF/OCSVM Hybrids</i>	28
3.4 RECONSTRUCTION AND GENERATIVE MODELS	28
3.4.1 <i>Autoencoders (AE)</i>	28
3.4.2 <i>Variational Autoencoders (VAE)</i>	28
3.4.3 <i>GANs for Anomaly Detection</i>	29
3.5 CLASS IMBALANCE.....	29
3.5.1 <i>Normal-only Training</i>	29
3.5.2 <i>Synthetic Anomaly Generation</i>	30
3.5.3 <i>Resampling Techniques</i>	30

3.5.4 Cost-Sensitive Learning	31
3.5.5 Threshold Calibration	31
3.5.6 Overall Interpretation.....	32
3.6 ENSEMBLE LEARNING	32
3.6.1 Bagging	32
3.6.2 Boosting.....	33
3.7 EVALUATION METRICS	33
3.7.1 Precision	34
3.7.2 Recall.....	34
3.7.3 F1-Score.....	34
3.7.4 ROC-AUC.....	34
3.8 CONCLUSION	35
CHAPTER 4: SPECIFICATION AND DESIGN	37
4.1 OVERVIEW.....	37
4.2 DATASET SPECIFICATIONS	37
4.3 PREPROCESSING AND FEATURE DESIGN	38
4.4 MODEL FAMILIES AND DESIGN RATIONALE	38
4.5 IMBALANCE HANDLING STRATEGIES.....	39
4.6 EVALUATIONS METRICS.....	40
4.7 CONCLUSION	40
CHAPTER 5: IMPLEMENTATION	41
5.1 OVERVIEW.....	41
5.2 UNSW-NB15 NETWORK INTRUSION DETECTION.....	41
5.2.1 Data Preparation:.....	41
5.2.2 Models Used:.....	42
5.2.3 Training Configuration:	43
5.2.4 Class Imbalance Strategies:.....	43
5.3 MVTec AD INDUSTRIAL VISION INSPECTION.....	44
5.3.1 Dataset and Imbalance:	44
5.3.2 Data Preparation:.....	44
5.3.3 Models Used:.....	45
5.3.4 Training Configuration:	46
5.3.5 Class Imbalance Strategies:.....	47
5.4 NIH CHEST X-RAY14 MEDICAL ANOMALY DETECTION.....	48
5.4.1 Dataset and Imbalance:	48
5.4.2 Data Preparation:.....	48
5.4.3 Models Used:.....	49

5.4.4 Training Configuration:	50
5.4.5 Class Imbalance Strategies:.....	51
5.5 CONCLUSION	52
CHAPTER 6: RESULTS AND DISCUSSION	53
6.1 OVERVIEW.....	53
6.2 UNSW-NB15 RESULTS.....	53
6.2.1 <i>Class Imbalance and Generalization:</i>	55
6.3 MVTEC AD RESULTS.....	56
6.3.1 <i>Addressing Class Imbalance:</i>	59
6.3.2 <i>Cross-Domain Observations:</i>	59
6.4 NIH CHEST X-RAY14 RESULTS	60
6.5 DISCUSSION:.....	63
6.6 CROSS-DOMAIN INSIGHTS:.....	64
6.7 CROSS-DOMAIN COMPARISON AND KEY INSIGHTS	67
6.8 CONCLUSION	69
CHAPTER 7: CONCLUSIONS AND FUTURE WORK	71
7.1 OVERVIEW.....	71
7.2 CONCLUSION	71
7.2.1 <i>Domain-specific performance:</i>	71
7.2.2 <i>Effect of class imbalance:</i>	71
7.2.3 <i>Ensembles and hybrid methods:</i>	72
7.2.4 <i>Impact of anomaly scarcity:</i>	72
7.3 LIMITATIONS: SEVERAL CONSTRAINTS AND CHALLENGES WERE IDENTIFIED IN THIS WORK:.....	72
7.3.1 <i>Model generality:</i>	72
7.3.2 <i>Scalability:</i>	73
7.3.3 <i>Threshold selection:</i>	73
7.3.4 <i>Limited model exploration:</i>	73
7.3.5 <i>Evaluation setup:</i>	73
7.3.6 <i>Metric focus:</i>	73
7.4 FUTURE WORK: BUILDING ON THIS FOUNDATION, SEVERAL IMPROVEMENTS AND EXTENSIONS ARE RECOMMENDED:....	73
7.4.1 <i>Advanced generative models:</i>	73
7.4.2 <i>Semi-supervised learning:</i>	74
7.4.3 <i>Domain adaptation:</i>	74
7.4.4 <i>Online and adaptive detection:</i>	74
7.4.5 <i>Automated thresholding:</i>	74
7.4.6 <i>Ensemble optimization:</i>	74
7.4.7 <i>Interpretability and visualization:</i>	74

7.4.8 Broader benchmarking:.....	75
7.5 FINAL REFLECTIONS	75
REFERENCES	76
APPENDIX 1: CODE LISTING	1 -
APPENDIX 2: WEEKLY PROGRESS REPORTS.....	2 -

List of Figures

Figure 1 Conceptual frameworks of three main deep anomaly detection approaches	9
Figure 2 Projection of ROC curve using Isolation forest for UNSW-NB Dataset.....	54
Figure 3 MVTec AD defect detection ROC curves.....	57

List of Tables

Table 1Dataset Characteristics.....	38
Table 2-Models and Design Motivation	39
Table 3UNSW-NB15 intrusion detection performance	53
Table 4 MVTec AD defect detection performance per model (defect = positive class)	56
Table 5 Chest X-ray anomaly detection performance for various models	61

Chapter 1: Introduction

1.1 Overview

The detection of anomalies and defects represents a fundamental research field which finds applications throughout industrial manufacturing and medical imaging and cybersecurity domains. The main difficulty exists in identifying unusual important events which differ from typical operational patterns. Real-world datasets contain few instances of anomalies which include manufacturing product surface defects and medical chest radiograph abnormalities and network intrusions because these anomalies prove challenging to detect.

The main difficulty in anomaly detection emerges from class imbalance because normal data exceeds anomalies by a wide margin. Supervised learning models experience difficulties when dealing with imbalanced data because they tend to predict normal class instances instead of detecting crucial anomalies. The project tackles the imbalance issue through the implementation of advanced anomaly detection methods on three benchmark datasets-

- MVTec AD (industrial defect detection)
- NIH Chest X-ray 14 (medical anomaly detection)
- UNSW-NB15 (network intrusion detection)

The selected datasets serve as a comprehensive evaluation tool to assess imbalance-handling approaches in computer vision and healthcare AI and cybersecurity domains. The three datasets provide a wide range of anomaly detection challenges which allows for a thorough analysis of imbalance handling methods across different domains.

1.2 Motivation for the project

The project aims to address the increasing market need for dependable anomaly detection solutions that work in environments where anomalies appear rarely yet they come in various forms and have significant effects. The detection of anomalies remains essential for preventing critical consequences because these anomalies appear infrequently in unbalanced real-world datasets despite being few and scattered.

1.2.1 Industrial Manufacturing (MVTec AD)

Automated manufacturing pipelines need to detect tiny defects including scratches dents and texture irregularities since their presence compromises structural integrity and triggers both costly recalls and brand reputation damage. The deployment of automated anomaly detection systems becomes necessary due to manual inspection being both resource-intensive and prone to errors. The MVTec AD dataset reflects real-world conditions because it provides industrial objects and textures with various defect types. The detection reliability of systems in this context depends on balancing the classes because defect samples occur much less frequently than normal samples.

1.2.2 Medical Imaging (NIH Chest X-ray 14)

The healthcare sector requires specific sensitivity to class imbalance because rare medical conditions appear predominantly in datasets filled with healthy cases. Medical detection of early disease indicators like pneumonia cardiomegaly and pneumothorax through chest radiograph analysis becomes life-saving yet anomalies appear only in a small fraction of total scans. The NIH Chest X-ray 14 dataset demonstrates dataset imbalance through its 100,000 images and 14 disease categories which show thousands of negative examples but only hundreds of positive examples. The healthcare industry requires critical strategies to reduce false negatives because detection failures result in delayed treatments that worsen patient health outcomes.

1.2.3 Cybersecurity (UNSW-NB15)

The occurrence of malicious attacks in network traffic remains extremely rare compared to the vast majority of benign network activities. The failure to detect intrusions results in financial loss and data breaches and service disruptions. Traditional machine learning classifiers fail to recognize rare attacks when they are trained on unbalanced traffic logs. The UNSW-NB15 dataset overcomes previous benchmarks including KDD99 by generating contemporary attack scenarios such as DoS and backdoors and reconnaissance attacks. The implementation of imbalance handling in this domain enables intrusion detection systems to detect rare attack signatures without increasing false alarm rates.

The domains share a common reason for concern because severe consequences follow from undetected anomalies than from incorrect alarms. A production line becomes unable to operate when defects go undetected while missed medical diagnoses can result in patient fatalities and national security becomes at risk when intrusions remain undetected. The development of techniques that learn from highly imbalanced datasets effectively while achieving generalization across various anomaly contexts remains essential for both research purposes and real-world requirements.

1.3 Project aims and objectives

The main objective of this research report involves studying and putting into practice methods which handle class imbalance problems in anomaly and defect detection systems across industrial manufacturing and medical imaging and cybersecurity domains. This research differs from most studies because it examines anomaly detection methods across three different datasets instead of focusing on a single dataset or application. The research aims to deliver domain-specific findings together with universal conclusions about imbalance-handling approaches in various real-world scenarios. The report functions as both a comparative analysis of different datasets and a practical resource for upcoming anomaly detection research.

1.3.1 Objectives

1.3.1.1 Dataset Preprocessing and Analysis

- Examine how the three selected datasets distribute their imbalanced data points.
- The preprocessing stage includes normalization and augmentation for image data and feature extraction methods such as CNN-based for MVTec AD and NIH X-ray 14 and tabular feature scaling for UNSW-NB15.

1.3.1.2 Model Implementation Across Domains

- The establishment of baseline performance requires implementing classical anomaly detection methods which include Isolation Forest, One-Class SVM, Local Outlier Factor, Elliptic Envelope.

- Deep learning models including Autoencoders and CNN feature extractors need to be developed and tested for processing complex image and feature data.
- The detection performance improves when using ensemble methods that combine PCA reduction with anomaly detection algorithms through majority voting.

1.3.1.3 Imbalance-Handling Techniques

- Examine different methods to handle imbalance problems including:
- The training process focuses on majority class data while using anomalies as detection targets.
- The bias towards normal features can be reduced through PCA dimensionality reduction.
- Threshold adjustment and anomaly score calibration.
- The network intrusion dataset requires random undersampling but medical/vision datasets need synthetic augmentation methods.

1.3.1.4 Evaluation of Models Under Imbalance

- The evaluation of imbalanced datasets requires the use of ROC-AUC, Precision-Recall AUC, F1-score (macro and weighted), and confusion matrix analysis.
- The evaluation compares how classical and deep learning models perform regarding their sensitivity to anomalies and their resistance to imbalance.

1.3.1.5 Cross-Domain Comparative Study

- The evaluation determines if imbalance-handling methods work across all three domains or if specific techniques remain domain-dependent (e.g., Autoencoders in medical imaging versus Isolation Forest in network intrusion).

1.3.1.6 Recommendations

- The research identifies the most successful methods to defeat class imbalance problems in anomaly detection.
- The recommendations include step-by-step instructions for implementing these strategies in industrial inspection and medical diagnostics and intrusion detection systems.

1.4 Report outline

- Chapter 2: Literature Review –

Reviews prior research on anomaly detection, covering classical, ensemble, and deep learning methods, along with domain-specific studies in cybersecurity, industrial inspection, and medical imaging.
- Chapter 3: Analytical Background –

Discusses the mathematical and theoretical foundations of anomaly detection, including classical algorithms, deep generative models, imbalance handling strategies, ensemble learning techniques, and evaluation metrics.
- Chapter 4: Specification and Design –

Outlines dataset specifications (UNSW-NB15, MVTec AD, NIH ChestXray14), preprocessing pipelines, model families, imbalance handling methods, and design strategies for anomaly detection across domains.
- Chapter 5: Implementation –

Details the implementation process, including data preparation, feature extraction, model training, threshold calibration, and imbalance mitigation strategies applied for each dataset.
- Chapter 6: Results and Discussion –

Presents the performance results of implemented models across the three datasets, compares effectiveness under class imbalance, and provides cross-domain discussion of strengths and limitations.
- Chapter 7: Conclusions and Future Work –

Summarizes the key contributions and findings of the project, identifies limitations, and outlines future research opportunities to improve anomaly detection under class-imbalanced conditions.

1.5 Conclusion

This chapter introduced the motivation and scope of the project, highlighting anomaly detection as a critical challenge across domains such as cybersecurity, industrial inspection,

and medical imaging. The discussion emphasized that anomaly detection problems are inherently characterized by severe class imbalance, scarcity of labeled anomaly data, and domain-specific complexities, which require a combination of classical algorithms, deep learning methods, and ensemble strategies. The objectives of the project were defined as evaluating and comparing different anomaly detection approaches across three benchmark datasets – UNSW-NB15, MVTec AD, and NIH ChestXray14 – while developing strategies to handle imbalance effectively. The chapter also outlined the research methodology and structure of the report. Overall, this introduction provides the foundation for the subsequent chapters, which review relevant literature, establish the analytical framework, and detail the specification, implementation, results, and conclusions of the project.

Chapter 2: Literature review

2.1 Overview

This chapter reviews literature about anomaly detection methods which include classical, ensemble and deep learning approaches together with domain-specific datasets and benchmark studies. The works demonstrate that anomaly detection presents multiple challenges because it depends on domain context and lacks sufficient anomalous data and features class imbalance. The surveys Anomaly Detection: A Survey and Deep Learning for Anomaly Detection: A Review establish fundamental frameworks which demonstrate how methods have transitioned from supervised learning to semi-supervised and unsupervised approaches that focus on learning from normal data. The three classical algorithms Isolation Forest, Local Outlier Factor and One-Class SVM continue to serve as standard reference methods especially when working with high-dimensional or imbalanced datasets according to simulation and comparative studies. The development of hybrid models through clustering boundary-based methods boosting ensembles and density estimation with pretrained features demonstrates how algorithmic integration solves the imbalance problems of individual detectors. The benchmark datasets MVTec AD and UNSW-NB15 present imbalance problems in industrial vision and network security domains while recent medical imaging research focuses on hybrid architectures and transfer learning and perceptual autoencoders to handle rare pathologies. The current research in diffusion models and normalizing flows demonstrates a new direction toward generative modeling of normal data manifolds which decreases the need for anomaly labels. The literature demonstrates that anomaly detection cannot exist independently from the imbalance problem because it needs both new detection methods and domain-specific evaluation approaches.

2.2 Anomaly Detection: A Survey [ACM Computing Survey]

The research introduces a basic framework for anomaly detection through its classification which separates anomalies into three categories: point, contextual and collective. It explains the detection methods into supervised, semi-supervised and unsupervised categories while showing statistical, clustering-based and classification-driven approaches[1]. The authors stress that the definition of “anomaly” depends on the specific domain and evaluation metrics need proper selection because anomalies occur infrequently. The research demonstrates that

accuracy becomes unreliable in imbalanced data because false negatives produce greater costs than false positives. The research directly supports our project needs because the datasets UNSW-NB15 and NIH ChestXray14 contain highly unbalanced distributions. The survey establishes theoretical support for using semi-supervised and unsupervised frameworks because defect and intrusion detection lacks sufficient labeled anomaly data. The survey predicts the adoption of ensemble methods and adaptive models which constitute a major part of the project approach to handle imbalance across industrial, medical and cybersecurity domains.

2.3 Deep Learning for Anomaly Detection: A Review [ACM Computing Survey]

The review demonstrates the fast expansion of deep learning methods for anomaly detection in both structured and unstructured data. The paper divides detection techniques into three main categories which include reconstruction-based models (autoencoders, VAEs) and one-class classification networks (Deep SVDD) and self-supervised pretext learning tasks. The authors demonstrate that deep learning methods excel in high-dimensional data spaces including images and network traffic yet they identify two major drawbacks which include high computational requirements and poor interpretability. The review emphasizes class imbalance as the main difficulty because deep models learn mostly from normal data before detecting anomalies through deviations during inference. The observation matches our approach for MVTec-AD dataset because normal images predominate and NIH ChestXray14 dataset has limited positive disease cases[2]. The review proposes hybrid models that unite feature learning with traditional anomaly scores which would be suitable for UNSW-NB15 because learned embeddings can be integrated with tree-based detectors. The review demonstrates deep anomaly detection's cross-domain applicability through its examination of financial and cybersecurity and manufacturing applications which matches the multiple application focus of this project.

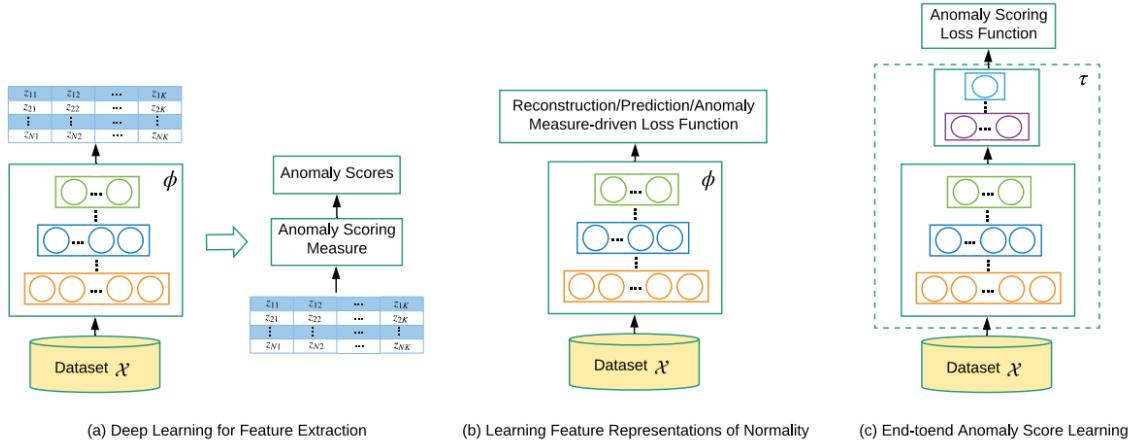


Figure 1 Conceptual frameworks of three main deep anomaly detection approaches

2.4 A Survey of Network Anomaly Detection Techniques [JNCA(2016) 19-31]

The research delivers an extensive evaluation of network traffic anomaly detection techniques with a focus on intrusion detection systems. The survey evaluates statistical and knowledge-based and machine learning approaches through signature-based and anomaly-based detection categories. The survey emphasizes that attack traffic labeling difficulties make unsupervised and semi-supervised approaches particularly beneficial. The paper makes its main contribution by analyzing feature engineering in network data because high-dimensional and redundant information typically hides anomalies. The evaluation process needs to move past accuracy metrics because intrusion detection datasets present severe class imbalance which requires assessment of recall and precision. The survey directly supports the project experiments on the UNSW-NB15 dataset because it was created to replace outdated intrusion detection benchmarks[3]. The paper guides our selection of unsupervised models Isolation Forest and One-Class SVM as strong baselines through its discussion of data preprocessing and scalability and threat adaptability while recommending ROC and PR curves for imbalanced conditions.

2.5 Anomaly Detection Using Diffusion-Based Models [ARXIV]

The research investigates anomaly detection through diffusion models which represent a new direction beyond traditional reconstruction-based methods including autoencoders and GANs. The authors use diffusion probabilistic models together with diffusion transformers to

process MVTec-AD and other benchmark datasets to show that diffusion processes effectively model normal data distributions. The generative process fails to reproduce rare defective samples which leads to anomaly identification through reconstruction error. The research shows that diffusion models achieve robust performance under noisy conditions and excel at detecting small texture-based anomalies which appear in industrial surface inspections. The paper demonstrates that diffusion models achieve better results than One-Class SVM and Isolation Forest in high-dimensional image tasks[4]. The research provides essential value for the project because it addresses the small number of defect classes in MVTec-AD and the visually challenging medical anomalies in NIH ChestXray14. The research introduces diffusion models as scalable solutions that handle imbalanced data which creates a contemporary approach for anomaly detection in complex vision data while the project evaluates classical methods in different domains.

2.6 Anomaly Detection Using Unsupervised Machine Learning Algorithms: A Simulation Study [Scientific African 26]

The research evaluates multiple unsupervised anomaly detection techniques through One-Class SVM and Isolation Forest and Local Outlier Factor and Robust Covariance models using synthetic data. The evaluation of these algorithms focuses on precision and recall performance together with computational efficiency and parameter sensitivity. The results demonstrate that Isolation Forest and Robust Covariance models achieve optimal precision-recall tradeoffs but One-Class SVM with SGD optimization produces high precision at the expense of recall performance. The reliability of Local Outlier Factor decreases when the neighborhood size becomes sensitive to different data regimes. The research demonstrates that different algorithms require specific dataset characteristics to achieve optimal performance because no single method works best for all cases. The comparative approach in this study provides essential guidance for the project. The research supports our decision to use multiple baselines across domains by implementing Isolation Forest and OCSVM on UNSW-NB15 and LOF in specific image feature spaces for MVTec-AD. The synthetic results simplify real-world scenarios yet the fundamental conclusion about performance dependence on data distribution supports the project's approach to use multiple datasets because different data distributions need customized model selection and parameter adjustment[5].

2.7 Comparative Analysis of Elliptic Envelope, Isolation Forest, One-Class SVM, and Local Outlier Factor in Detecting Earthquakes with Status Anomaly using Outlier [IEEE]

The research evaluates the performance of Elliptic Envelope and Isolation Forest and One-Class SVM and Local Outlier Factor algorithms for earthquake anomaly detection through direct comparison. The research uses real earthquake data containing few anomalies to study a scenario that matches typical industrial and cybersecurity applications with their naturally rare anomalies. The evaluation of each method assesses its ability to detect seismic anomalies from normal readings through performance metrics that include detection accuracy and precision and recall and computational cost. The results show that Isolation Forest and One-Class SVM achieve stable performance but Isolation Forest demonstrates better resistance to noise and high-dimensional data. The Local Outlier Factor algorithm achieves precise results when its parameters are correctly set but its performance deteriorates when the neighborhood size is incorrectly configured. The performance of Elliptic Envelope reaches its peak when data follows a Gaussian distribution yet its restrictive assumptions restrict its usage in various scenarios. The analysis shows important practical trade-offs between different classical algorithms when dealing with natural imbalance. The Isolation Forest algorithm stands out as a scalable solution for UNSW-NB15 because attacks represent minority classes but LOF might work better for specific feature subspaces in MVTec-AD. The earthquake case study demonstrates that anomaly detection success depends on both data distribution patterns and proper parameter adjustments which are essential factors for the project's multi-domain evaluation process[6].

2.8 Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor, and One-Class SVM [IEEE]

The research presents an unsupervised anomaly detection framework which unites clustering and boundary-based methods to enhance detection in datasets containing rare anomalies. The detection process starts with K-means clustering for majority instance grouping followed by Local Outlier Factor and One-Class SVM for detecting deviations within or between clusters. The proposed method unites density estimation through clustering with outlier sensitivity from boundary-based detection to achieve better results[7]. The authors demonstrate their framework's effectiveness through testing it on various datasets which produce superior results than running individual methods separately. The hybrid approach

achieves better precision-recall balance because it handles rare and diverse anomalies effectively. The project's multi-domain environment benefits from this approach because UNSW-NB15 has severe class distribution imbalances and MVTec-AD presents various subtle defects. The hybrid clustering-plus-one-class model provides strong performance in such conditions by first establishing the dominant normal structure before applying anomaly-sensitive methods for refinement. The research demonstrates how ensemble or hybrid models protect against detector weaknesses by serving as a safeguard which aligns with the comparative methodology across datasets. The research shows that unsupervised ensemble strategies excel when there are not enough labeled anomalies which matches the class imbalance problems we face in industrial, medical and cybersecurity domains.

2.9 Isolation Forest [IEEE]

The paper establishes Isolation Forest as a tree-based ensemble method that detects anomalies through recursive partitioning. The method of Isolation Forest identifies anomalies through random partitioning because these rare patterns need fewer splits to become isolated. The algorithm operates at linear time complexity while requiring minimal memory which enables it to handle large-scale high-dimensional datasets. The paper demonstrates through experiments that Isolation Forest outperforms k-nearest neighbor and LOF in terms of scalability and anomaly detection performance. It projects that Isolation Forest operates without requiring distance measures that deteriorate in high-dimensional spaces and without needing specific data distributions which statistical methods require[8]. The three datasets utilize Isolation Forest as their baseline algorithm throughout the project. The high-dimensional feature embeddings from images make MVTec-AD suitable for Isolation Forest application and its efficiency makes it suitable for detecting rare attack traffic in large volumes of benign data in UNSW-NB15. The design of Isolation Forest assumes that anomalies are rare but requires careful tuning of the contamination parameter to achieve optimal results. The fundamental design of Isolation Forest serves as both a baseline model and a practical tool to evaluate new imbalance-handling approaches in various domains.

2.10 MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection [Springer – CVPR 2019]

The influential paper presents the MVTec Anomaly Detection (MVTec AD) dataset which serves as a benchmark for evaluating unsupervised anomaly detection methods in industrial

visual inspection. The dataset contains more than 5,000 high-resolution images spread across 15 object and texture categories with various defect types including scratches, dents, misalignments and contaminations. The dataset includes precise pixel-level ground truth annotations for defects which allows both image-level detection and pixel-level segmentation evaluation. Real industrial environments show defects occur much less frequently than normal samples which results in an unbalanced data distribution. MVTec AD serves as an excellent benchmark for methods that need to handle imbalanced data because they must learn normal patterns and detect unusual patterns. The authors demonstrate that deep representations obtained from pre-trained networks produce better results than traditional feature-based methods and early deep learning approaches. MVTec AD functions as the primary benchmark for the project to evaluate strategies that address imbalance in industrial defect detection[9]. The dataset enables us to evaluate both traditional methods including Isolation Forest and Local Outlier Factor on extracted features and advanced techniques such as autoencoders and diffusion models. The diverse range of categories in this dataset demonstrates the need for generalization between different defect types which directly supports the project goal of testing imbalance-handling technique transferability between domains.

[2.11 The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection\(Extended\) \[Springer - IJCV 2021\]](#)

The MVTec Anomaly Detection dataset received its first release in 2019 but this expanded work delivers more detailed analysis with new benchmark results and methodological understanding of unsupervised industrial anomaly detection. The expanded dataset includes revised evaluation procedures and improved threshold selection guidelines together with performance metrics that address heavy class imbalances. The research evaluates three method categories which include reconstruction-based autoencoders and distribution modeling techniques and feature descriptor methods that use pretrained networks such as ResNet and EfficientNet. The research demonstrates that feature-based nearest-neighbor and density estimation methods perform better than pure generative models because they utilize strong pretrained representations of normal samples[10]. The paper demonstrates that accuracy metrics fail to measure performance adequately in datasets with severe class imbalance so it recommends AUROC and AUPRC together with pixel-wise overlap scores for

segmentation evaluation. The extension confirms that within the project should use pretrained embeddings together with one-class detectors as a solid baseline for MVTec AD. The research establishes methodological connections to the NIH ChestXray14 dataset because robust pretrained features can help manage unbalanced label distributions. The paper's emphasis on metric selection helps us determine the best evaluation criteria for UNSW-NB15 because normal traffic exceeds anomalous attack instances by a large margin.

2.12 Industrial Anomaly Detection with Normalizing Flows [Hannover: Institutionelles Repotorium der Leibniz]

This dissertation investigates image-based industrial anomaly detection through normalizing flows which serve as density-estimation tools to learn normal data distributions and detect low-likelihood anomalies. The thesis presents a detailed fundamentals chapter followed by an organized review of related work which covers traditional methods alongside generative models and student–teacher networks and density estimation and synthetic data and multivariate machine data methods before establishing flows as contemporary unsupervised and semi-supervised anomaly detection approaches[11]. The flow-based approach stands out for industrial inspection because it generates a computationally manageable likelihood score while requiring training on defect-free images which matches the typical production line scenario where defective samples are rare. The document addresses operational issues including robust normal modeling through representation learning and appropriate evaluation methods for imbalanced data and threshold calibration techniques. The thesis presents both theoretical and engineering guidelines for one-class training in visual inspection by modeling the normal manifold first and then using likelihood or teacher–student discrepancy for scoring deviations. The plan for the MVTec AD dataset within the project includes normal-only training and calibrated scoring as essential components to address class imbalance while maintaining precise localization and category-agnostic detection.

2.13 Anomaly Detection in Medical Imaging with Deep Perceptual Autoencoders [IEEE]

The paper examines unsupervised medical anomaly detection through perceptual autoencoders which use perceptual reconstruction loss in feature-space instead of pixel-wise criteria. The autoencoder learns to minimize the difference between deep features of the input and its reconstruction (normalized “relative-perceptual-L1” loss) which makes the

model sensitive to clinically relevant structural and textural deviations without forcing photorealistic output. The fully unsupervised assumption should be relaxed by using a minimal set of anomalies only for hyperparameter search initialization before normal-only core training begins[12]. The proposed approach establishes a better baseline than multiple state-of-the-art detectors which were originally designed for natural images when dealing with faint abnormalities and limited labeling. The design directly applies to NIH ChestXray14 by training mainly on abundant normal CXRs and calculating anomaly scores from perceptual residuals while using a small validation subset to stabilize thresholds. The feature-centric training objective supports within the project cross-domain approach (industrial, medical, and network) because it relies on robust representations and precise thresholding to manage severe imbalances.

2.14 Performance Evaluation of CNN Architectures on NIH Chest X-Ray Dataset with Boosting Ensemble [IEEE]

The research evaluates five pre-trained CNN backbones (VGG16/19, ResNet50, MobileNetV2, DenseNet201) on NIH ChestX-ray14 before developing a boosting ensemble to handle unbalanced class distributions. The authors specifically handle class imbalance through majority class downsampling while presenting disease-specific precision/recall/F1 metrics and demonstrating that ensemble boosting produces better overall classification results (accuracy \approx 92.6%) and improved sensitivity for minority classes than individual models. The authors demonstrate boosting effectiveness in this context because it progressively re-weights difficult examples which helps counter the standard CNN fine-tuning tendency to overfit prevalent labels such as “No Finding.” The paper presents class-wise results (e.g., pneumothorax, edema, consolidation) through tables and figures to demonstrate how ensemble aggregation produces stable predictions for infrequent pathologies. Our project will follow two parallel approaches on NIH ChestXray14 data by using

- (i) anomaly-style modeling from normals (perceptual AE)
- (ii) Supervised CNNs with imbalance-aware tactics (re-sampling, calibrated thresholds, and lightweight ensembling)

The research focuses on class-by-class evaluation metrics and threshold optimization which matches our evaluation approach[13].

2.15 Deep Isolation Forest for Anomaly Detection [IEEE]

The research develops the Isolation Forest algorithm into Deep Isolation Forest (Deep isolation forest) to analyze complex high-dimensional structured datasets. The deep version of isolation forest differs from the original by using neural networks to extract meaningful representations before performing isolation-based partitioning. The method detects non-linear and hierarchical patterns in data structures which enables it to detect subtle anomalies that traditional isolation forest algorithms miss. The benchmarks show that Deep isolation forest achieves better results on image and sequence datasets without losing its scalability benefits. The method provides substantial value for the project through its application on MVTec AD and NIH ChestXray14 because anomalies exist within intricate visual structures. The approach enhances anomaly detection under extreme imbalance through learned normal sample representations which decreases the number of false negatives for rare defects or pathologies. The method demonstrates potential for use with UNSW-NB15 because deep network traffic embeddings can distinguish normal flows from rare malicious instances. The Deep isolation forest represents a modern approach to ensemble methods that combine classical ensemble techniques with contemporary representation learning methods for handling imbalanced data across multiple domains[14].

2.16 Recent Progress of Anomaly Detection [Wiley Online Library]

The research examines anomaly detection progress through statistical and clustering and ensemble and deep learning-based approaches. The research highlights the increasing need for hybrid methods which unite traditional approaches with contemporary deep feature extraction methods for managing high-dimensional and imbalanced data. The survey organizes algorithms through their theoretical foundations while evaluating their effectiveness for various application fields including finance healthcare and cybersecurity. The evaluation of imbalanced anomaly detection requires precision and recall and AUPRC instead of traditional accuracy metrics because these metrics provide more accurate results. The study observes that real-world applications require unsupervised and semi-supervised models because they lack sufficient labeled anomaly data[15]. The survey provides theoretical support for using one-class and unsupervised methods in MVTec AD and NIH ChestXray14 and UNSW-NB15 because these datasets exhibit class imbalance. The research supports the combination of ensemble methods with pretrained feature extractors and

imbalance-aware evaluation metrics which serve as the fundamental approach for the project.

2.17 A Survey of Anomaly Detection Methods in Networks [IEEE]

The survey examines network traffic anomaly detection methods through statistical profiling and clustering and machine learning and information-theoretic approaches. The paper organizes algorithms through categories that show their dependence on labeled data starting from supervised classification up to fully unsupervised models. The paper emphasizes how obtaining trustworthy labeled attack traffic proves challenging which drives practical intrusion detection systems to adopt unsupervised anomaly detection methods. The research identifies three main obstacles that include changing attack methods and large network flow dimensions and the extreme rarity of malicious traffic in total data[16]. The research emphasizes the need for adaptive models that can process non-stationary data streams. The survey directly supports the implementation of UNSW-NB15 because minority attacks exist within a majority of benign traffic. The survey supports the selection of Isolation Forest and One-Class SVM and hybrid ensemble methods because they address imbalance in contemporary intrusion detection tasks.

2.18 One-Class Support Vector Machines Approach to Anomaly Detection [Taylor & Francis]

The paper presents One-Class SVM (OCSVM) as a specific anomaly detection model that operates with normal class examples only during training. The algorithm learns a boundary in high-dimensional feature space to contain most data points so it can identify points beyond the boundary as anomalies. The main benefit of OCSVM emerges from its ability to handle complex distributions while needing only normal class examples during training thus making it suitable for imbalanced datasets with rare or no anomaly examples. The research demonstrates OCSVM's effectiveness on various datasets through performance comparisons with statistical and distance-based methods. The project uses OCSVM as its base model because it applies to all three datasets. The MVTec AD dataset benefits from OCSVM when it analyzes feature embeddings to identify uncommon visual defects. The screening of pathologies becomes possible through this method even when positive samples are not balanced[17]. The UNSW-NB15 dataset benefits from OCSVM because it enables anomaly

detection in network traffic through normal-only training which solves the inherent imbalance problem in intrusion detection.

2.19 Machine Learning for Anomaly Detection: A Systematic Review [IEEE]

This systematic review provides a detailed examination of machine learning methods for anomaly detection, covering both traditional algorithms and recent advances. The paper categorizes approaches into supervised, semi-supervised, and unsupervised paradigms, and analyzes their effectiveness in addressing challenges like high dimensionality, noise, and class imbalance. It highlights that unsupervised and semi-supervised methods dominate real-world anomaly detection, as labeled anomaly data is typically scarce[18]. The review also underscores the increasing role of deep learning methods, particularly autoencoders and generative adversarial networks, while noting their vulnerability to overfitting in imbalance-heavy contexts. Case studies span healthcare, cybersecurity, and manufacturing, reflecting the cross-domain relevance of anomaly detection. For the project, this survey supports the methodological choice to evaluate a spectrum of techniques—classical (e.g., OCSVM, Isolation Forest), deep (e.g., autoencoders, diffusion models), and hybrid ensembles—across three application domains. The paper's insistence on imbalance-aware metrics such as AUPRC and sensitivity at low false positive rates directly aligns with the project's evaluation design for datasets like **NIH ChestXray14** and **UNSW-NB15**, where anomalies are rare but critical.

2.20 Machine Learning Techniques for Anomaly Detection: An Overview [ResearchGate]

The overview presents basic machine learning methods for anomaly detection through four main methodological categories which include statistical approaches and distance-based methods and clustering-based methods and classification-based methods. The paper explains the theoretical basis of each method while describing their advantages and disadvantages when used in real-world applications. The research demonstrates that k-nearest neighbors and Local Outlier Factor perform well in distance and density-based methods yet these methods become less effective when dealing with sparse data in high-dimensional spaces[19]. The paper emphasizes how unbalanced distributions between normal and anomalous data affect algorithm performance because methods that focus on modeling normal data classes tend to be the most reliable. The project uses this overview to determine which algorithms to use and it requires baseline models such as OCSVM, Isolation Forest and LOF for comparative analysis. The paper supports the use of hybrid approaches when working

with MVTec AD images and NIH ChestXray14 scans and UNSW-NB15 network traffic data because these datasets present severe class imbalance.

2.21 Overview of Anomaly Detection Techniques in Machine Learning [IEEE]

The research examines machine learning-based anomaly detection methods while emphasizing their practical usage. The research divides detection methods into supervised, semi-supervised and unsupervised categories to evaluate their performance trade-offs regarding scalability and interpretability and effectiveness under imbalance. The survey examines ensemble methods because they help maintain stable detection results in noisy and skewed datasets. The survey demonstrates that unsupervised detectors work best in real-world scenarios because anomalies appear infrequently. The paper demonstrates that proper feature engineering produces better anomaly separation than selecting algorithms alone[20]. The research supports the need to evaluate both traditional and deep anomaly detection approaches throughout different application domains for the project. The study provides important findings about UNSW-NB15 and MVTec AD because feature selection and engineering impact imbalance-aware model success and pretrained visual features with anomaly scoring outperform end-to-end models.

2.22 Anomaly Detection on Medical Images using Autoencoder and Convolutional Neural Network [Google]

The research introduces a dual system for medical imaging anomaly detection which unites autoencoders with convolutional neural networks (CNNs). The autoencoder training process on normal medical images enables it to learn compact representations which it uses to reconstruct input samples while anomalies become detectable through increased reconstruction errors. The CNN module serves to improve feature extraction capabilities especially for complex structural and textural variations found in medical data[21]. The authors evaluate the framework through various medical imaging datasets which show that the combined model outperforms both autoencoders and CNNs in terms of detection accuracy and sensitivity. The research demonstrates that medical datasets contain extreme imbalances because normal cases outnumber abnormal cases which makes unsupervised or semi-supervised methods more suitable than fully supervised methods. The hybrid design of this project shows relevance to NIH ChestXray14 because this dataset contains rare pathologies which can be effectively addressed through the combination of reconstruction-

based scoring with CNN feature learning to handle imbalance. The research demonstrates how medical anomaly detection benefits from architecture design when anomalies are subtle and normal samples dominate the training data because it combines reconstruction and discriminative learning approaches.

2.23 Multi-Label Classification of Chest X-ray Abnormalities Using Transfer Learning Techniques

The research investigates transfer learning strategies for thoracic disease multi-label classification in chest radiographs using the NIH ChestXray14 dataset. The research uses EfficientNet as its backbone architecture while fine-tuning it with binary cross-entropy loss to manage the complex multi-label outputs across 14 pathologies. The research implemented a modified dataset distribution to handle data imbalance while providing equal representation of frequent and infrequent conditions in the official dataset split. The model received data augmentation through random flipping and rotation and brightness adjustments and regularization strategies to enhance its generalization capabilities. The model reached a mean AUC-ROC score of 84.28% which surpassed previous state-of-the-art baselines while showing exceptional performance in detecting pneumothorax and edema and cardiomegaly[22]. The implementation of Grad-CAM visualizations improved interpretability by showing the location of pathological areas. The research directly applies to the NIH ChestXray14 dataset because it shows how transfer learning with proper imbalance management and interpretability methods produces excellent results in healthcare anomaly detection. The study demonstrates the necessity of custom data splits and imbalance-aware training methods which guide the project's approach to evaluate imbalanced anomalies in medical imaging and cross-domain datasets including MVTec AD and UNSW-NB15.

2.24 UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems

The paper introduces the UNSW-NB15 dataset which addresses the shortcomings of KDD99 and NSL-KDD benchmarks because they contained redundant data and outdated attack profiles and unbalanced distributions. The IXIA PerfectStorm tool at the Cyber Range Lab of UNSW generated the dataset through the combination of real modern traffic with synthesized attack scenarios. The dataset contains 2.5 million records which include normal flows alongside nine different attack categories including DoS, Exploits, Generic, Reconnaissance, and Worms. The dataset contains 49 features derived from Argus and Bro-

IDS which include packet-level and flow-level attributes and ground-truth reports serve as labels. The UNSW-NB15 dataset includes contemporary low-footprint intrusions and delivers a detailed picture of present-day threat landscapes. The dataset shows extreme class imbalance because normal flows outnumber anomalous attacks to a significant extent which matches real-world attack distributions. The research project selects UNSW-NB15 as one of its three core datasets because it offers diverse data while maintaining imbalance characteristics. The dataset allows researchers to evaluate both unsupervised baselines like Isolation Forest, LOF and OCSVM and deep anomaly detection methods while ensuring their results apply to operational intrusion detection systems. The dataset's imbalance characteristics make it crucial for evaluating whether imbalance-handling strategies work across different domains[23].

2.25 A Survey of Anomaly Detection Techniques [Springer]

The survey presents an organized evaluation of anomaly detection methods through three categories of data labels and four methodology types and six data types. The paper explains how different modeling assumptions are needed to detect three main anomaly types including point, contextual and collective. The research makes its main contribution through its focus on categorical data anomaly detection since this field receives less attention than quantitative data analysis. The paper examines indicator-variable representations together with frequency-based scoring and conditional frequency models and compression-based methods as solutions for detecting categorical anomalies. The research identifies three major obstacles which affect categorical data analysis including high computational expenses and sensitive parameters and insufficient benchmark datasets. The review demonstrates anomaly detection applications across network intrusion detection and wireless sensor networks and fraud detection and healthcare and social networks and recommender systems[24].

The survey supports our decision to use unsupervised and semi-supervised methods for MVTec AD and NIH ChestXray14 because these datasets contain limited anomaly labels. The survey helps us assess categorical features in UNSW-NB15 because many of its attributes are discrete and imbalanced which makes categorical anomaly detection strategies highly relevant. The project benefits from domain-specific adaptations and technique classification by application domain which supports its focus on industrial medical and cybersecurity datasets.

2.26 Heart Rate Anomaly Detection in Healthcare Using Elliptic Envelope and Local Forest [Science Direct]

This paper addresses unsupervised detection of heart rate anomalies using Elliptic Envelope (EE) and Isolation Forest (IF). EE models normal heart rate with a Gaussian covariance boundary, while IF isolates anomalies through random partitions. Evaluated on healthcare datasets with simulated anomalies, both methods achieved high recall (~92%) and precision (~89%), outperforming one-class SVM[25]. The study highlights their efficiency for real-time monitoring in wearable and telehealth applications. For the project, it demonstrates how unsupervised models can handle class imbalance in physiological data, with EE and IF capturing rare health events without requiring labelled anomalies. The findings also validate the benefit of combining statistical and ensemble-based outlier detectors, consistent with the project's strategy of leveraging multiple techniques for robust anomaly detection.

2.27 Survey on Anomaly Detection Using Data Mining Techniques[Science Direct]

This survey reviews anomaly detection through clustering, classification, and hybrid approaches within data mining, particularly for intrusion detection. Clustering methods (e.g., k-means, k-medoids) flag outliers as poorly clustered instances, making them effective for novel attacks. Classification-based methods (decision trees, SVMs) work well with labelled data but struggle with unseen anomalies. The survey emphasizes hybrid models, such as clustering combined with decision trees, which achieved superior accuracy by integrating unsupervised and supervised insights. Its findings underline the value of ensemble strategies and unsupervised detection in addressing imbalanced datasets where anomalies are rare[26]. This supports the project's use of hybrid models and clustering-plus-one-class strategies across UNSW-NB15, MVTec AD, and NIH ChestXray14 to improve anomaly coverage under class imbalance.

2.28 Conclusion

Research studies show that dealing with class imbalance stands as a fundamental issue which appears throughout anomaly detection research. The classical approaches show how scalability and interpretability and rare event sensitivity trade off against each other while deep learning methods provide strong feature representations yet need normal-only training data. The project's approach to use multiple baselines and advanced models across industrial,

medical and cybersecurity domains receives support from comparative and simulation studies which show no algorithm achieves optimal results in all data distributions. The evaluation metrics and methodological design of anomaly detection systems require realistic imbalance data from benchmark datasets such as MVTec AD and UNSW-NB15. Medical imaging research on NIH ChestXray14 shows how ensemble approaches and hybrid CNN-autoencoder models enhance the detection of rare pathologies. Surveys indicate that hybrid and generative approaches are becoming more important while precision, recall, F1-score and AUPRC metrics replace accuracy because they account for class imbalance. The project's comparative evaluation of imbalance-handling strategies in anomaly detection receives both theoretical backing and empirical evidence from the literature.

-----This page is left blank intentionally-----

Chapter 3: Analytical background

3.1 Overview

The chapter provides essential analytical and mathematical foundations needed to understand anomaly detection problems that occur in real-world scenarios. The chapter explains the formal framework of anomaly detection and classical machine learning methods and deep learning models and the core issue of class imbalance. The chapter presents ensemble learning strategies which serve as essential tools for managing imbalance problems while enhancing generalization capabilities. The chapter describes evaluation metrics specifically designed for imbalanced datasets which enable fair and consistent assessment of anomaly detection performance[1, 2]. The theoretical framework established in this chapter serves as the basis for the models and methods used in the project.

3.2 Anomaly Detection Framework

Anomaly detection is fundamentally about learning the patterns of **normality** and flagging deviations[1]. For a dataset:

$$X = \{x_1, x_2, \dots, x_n\}$$

Equation 1

where most $x_i \in N$ (normal class) and a small subset $x_j \in A$ (anomalous class), the anomaly score is defined as:

$$s(x) = f(x; \theta)$$

Equation 2

with larger values suggesting higher anomaly likelihood. A decision threshold τ then determines classification:

$$f(x) = \begin{cases} 1, & s(x) \geq \tau \text{ (anomaly)} \\ 0, & s(x) < \tau \text{ (normal)} \end{cases}$$

Equation 3

Selecting τ is significant in imbalanced datasets. If anomalies are very rare, a poor threshold can produce misleading accuracy while failing to detect critical anomalies. Threshold optimization is directly connected to the cost of errors within the domain: e.g. In healthcare, missing a diagnosis is more costly than flagging false positives.

earning settings differ by label availability:

- **Supervised:** Both normal and anomaly labels available (rare in practice).
- **Semi-supervised:** Training on normal data only, with anomalies detected as deviations.
- **Unsupervised:** No labels assumed; anomalies detected from data distribution.

In this project, semi-supervised and unsupervised paradigms are emphasized due to the scarcity of anomaly labels.

3.3 Classical Anomaly Detection Methods

Classical methods dominate because of their computational efficiency, and ease of use in unsupervised or semi-supervised settings[5-8].

3.3.1 One-Class Support Vector Machines (OCSVM)

The One-Class Support Vector Machine (OCSVM) learns a decision boundary enclosing normal data in feature space. The optimization problem is:

$$\min_{\omega, \varepsilon_i, \rho} \frac{1}{2} \|\omega\|^2 + \frac{1}{\vartheta n} \sum_{i=1}^n \varepsilon_i - \rho$$

Equation 4

Subject to:

$$(\omega \cdot \varphi(x_i)) \geq \rho - \varepsilon_i, \quad \varepsilon_i \geq 0.$$

Equation 5

Here, $\varphi(x_i)$ is the kernel mapping, ϑ is the proportion of anomalies tolerated, and ε_i are slack variables. The model attempts to learn the smallest boundary enclosing most normal sample. Points present out are the anomalies(outliers). This makes One Class SVM suitable for imbalance, as it doesn't require anomaly labels as input.

3.3.2 Local Factor Outlier (LOF)

LOF is a density-based method. It compares the density of a point with its neighbors. The **local reachability density (LRD)** is:

$$LRD_k(p) = \frac{1}{\frac{1}{|N_k(p)|} \sum_{o \in N_k(p)} \text{reach}_{dist_k}(p, o)}$$

Equation 6

With

$$\text{reach}_{dist_k}(p) = \max\{k - \text{distance}(q), d(p, q)\}$$

Equation 7

The LOF score is given as:

$$LOF_k(p) = \frac{1}{|N_k(p)|} \sum_{o \in N_k(p)} \frac{LRD_k(o)}{LRD_k(p)}$$

Equation 8

If $LOF > 1$, the point has lower density than neighbors and is considered anomalous.

This method captures local deviations but is sensitive to the parameter k .

3.3.3 Isolation Forest

Isolation Forest isolates anomalies through recursive partitioning. Anomalies are isolated quickly since they lie in sparse regions. Its scalability makes it suitable for big data such as UNSW-NB15. The anomaly score is:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Equation 9

Where $E(h(x))$ is the average path length to isolate x , and $c(n)$ is the average path length of a binary search tree of size n . The shorter path length is directly proportional to more anomalies. Isolation forest is efficient ($O(n \log n)$) and well-suited for high-dimensional imbalanced data.

3.3.4 Deep Isolation Forest

Deep Isolation Forest extends isolation forest by adding random neural transformations before partitioning. This allows non-linear splits and improved separation in complex spaces. It retains the model's efficiency but better captures subtle anomalies hidden in raw features.

3.3.5 Clustering + LOF/OCSVM Hybrids

Hybrid methods first cluster data (e.g., *k*-Means), then apply LOF or OCSVM within clusters. This accounts for multi-modal normal data. A point normal in one cluster may be anomalous in another. Such hybrids improve sensitivity in heterogeneous datasets.

3.4 Reconstruction and Generative Models

3.4.1 Autoencoders (AE)

It reconstructs inputs and use reconstruction error as an anomaly score. Variants such as perceptual autoencoders [12] evaluate reconstruction in feature space, making them sensitive to small structural deviations in medical images.

$$h = f_\theta(x), \quad \hat{x} = g_\varphi(h)$$

Equation 10

The reconstruction error is:

$$L(x, \hat{x}) = \|x - \hat{x}\|^2$$

Equation 11

If trained using only normal labels, reconstruction error will be low for it's samples and high for anomalies. Variants include **perceptual autoencoders**, which compute reconstruction in feature space, enhancing sensitivity to subtle medical anomalies.

3.4.2 Variational Autoencoders (VAE)

VAEs extend AEs with probabilistic latent modeling to improve generalization and detect rare cases with low likelihood.

$$L = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \parallel p(z)]$$

Equation 12

The first term ensures reconstruction; the KL divergence regularizes latent space. Points with low likelihood under the model are anomalies.

3.4.3 GANs for Anomaly Detection

A GAN trains generator G and discriminator D in the minmax equation given below:

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}} [\log D(x)] + E_{z \sim p(z)} [\log(1 - D(G(z)))]$$

Equation 13

If the extracted features are being trained only on normal labels, G produces normal based labels. Anomalies are being scoring low in $D(x)$ while it cannot be constructed by G .

Comparative investigations show that no method is superior every time; performance is highly dependent on data-distribution and imbalance-severity. For instance, LOF is good at detecting local outliers but can be poor at detecting global outliers, while Isolation Forest is robust in high-dimensional spaces.

3.5 Class Imbalance

There are two classes defined as normal $[|N|]$ and anomaly $[|A|]$. It occurs when $|N| \gg |A|$. There are some fundamental methods to be taken into consideration to minimize the imbalance. This inequality is especially pronounced in domains such as:

- **Industrial defect detection (MVTec AD):** thousands of defect-free images compared to a handful of defective samples.
- **Medical imaging (NIH ChestXray14):** millions of chest X-rays, but only a very small percentage contain rare diseases.
- **Cybersecurity (UNSW-NB15):** millions of benign flows, with anomalous intrusions constituting less than 1%.

Traditional classifiers trained on such skewed datasets tend to **minimize overall error** by predicting the majority class (normal) for most inputs, resulting in deceptively high accuracy but near-zero recall for anomalies. This makes imbalance handling strategies indispensable[15, 18-21, 24, 26].

The methods to minimize the class imbalance are:

3.5.1 Normal-only Training

The approach is to avoid using anomalies during training altogether and instead train exclusively on normal data. Methods such as **One-Class SVM (OCSVM)**, **Isolation Forest (IF)**, and **Autoencoders (AE)** are examples of this paradigm. These models learn the distribution of normal samples, defining either a decision boundary (OCSVM), an isolation criterion (IF), or a reconstruction model (AE).

As anomalies are not needed during training, the imbalance problem is sidestepped. Any deviation from the learned “normal profile” is flagged as anomalous. This is well-suited when anomalies are extremely scarce or diverse, making them impractical to collect comprehensively.

3.5.2 Synthetic Anomaly Generation

- **Generative Adversarial Networks (GANs):** Train a generator $G(z)$ on normal data, then perform variation to its outputs to create synthetic “pseudo-anomalies.”
- **Diffusion models:** Learn the data distribution via a denoising process and generate varied samples that lie outside the normal manifold[4].
- **Augmentation techniques:** For images, this might mean introducing artificial scratches or noise; for networks, injecting unusual patterns into traffic.

By generating synthetic anomalies, the detector sees a more balanced distribution during training. While synthetic anomalies may not cover all real-world cases, they encourage the model to separate normal and abnormal manifolds more clearly.

3.5.3 Resampling Techniques

Resampling consist of Over and under sampling of the training set:

- Undersampling of normal labels: It starts removing normal labelled within random instances so the anomaly-to-normal ratio becomes more balanced.
- Oversampling anomalies: Duplicate or synthetically generate more anomalous samples. A popular technique is SMOTE (Synthetic Minority Oversampling Technique), which interpolates between existing anomalies to create new synthetic ones:

$$x_{new} = x_a + \lambda(x_b - x_a), \quad \lambda \sim U(0,1)$$

Equation 14

Where x_a and x_b are anomaly samples.

Resampling prevents the model from being dominated by normal class examples. However, undersampling risks discarding useful normal data, while oversampling anomalies may cause overfitting to synthetic patterns.

3.5.4 Cost-Sensitive Learning

In cost-sensitive anomaly detection, the **loss function is modified** to penalize errors on anomalies more heavily than on normal points. For instance, in binary classification, a weighted cross-entropy loss can be defined as:

$$L = -\alpha \cdot y \log \hat{y} - (1 - \alpha) \cdot (1 - y) \log(1 - \hat{y})$$

Equation 15

where, $y \in \{0,1\}$ is the true label ($1 = \text{anomaly}$) and \hat{y} is the predicted probability, and α is a weight (e.g., 0.9) emphasizing anomaly detection.

This ensures the model engages more on minimizing false negatives (missed anomalies), even if that increases false positives slightly. In medical settings, this is particularly important: missing a disease is much more costly than wrongly flagging a healthy case.

3.5.5 Threshold Calibration

Even after training, the decision threshold Equation 3 must be chosen within careful consideration. If set arbitrarily, imbalance skews performance. Instead, thresholds can be tuned to maximize imbalance-aware metrics such as:

- F1-score: It balances the rate of precision and recall.
- PR-AUC (Precision–Recall Area Under Curve): It focuses on performance when anomalies are rare.
- Recall at fixed False Positive Rate (FPR): It ensures anomalies are detected with limited false alarms.

For example, in MVTec AD, a threshold may be calibrated to maximize pixel-level recall at a fixed 1% false alarm rate, ensuring defect sensitivity without overwhelming operators.

Threshold calibration aligns anomaly detection performance with the operational cost trade-offs of the application domain.

3.5.6 Overall Interpretation

These imbalance-handling strategies are complementary. Normal-only training is favored when anomalies are nearly absent. Synthetic anomaly generation and resampling improve model training when labeled anomalies exist but are scarce. Cost-sensitive learning directly embeds imbalance into the optimization process, forcing the model to treat anomalies as more important. Finally, threshold calibration ensures that evaluation reflects the high cost of missing anomalies.

Together, these approaches prevent anomalies from being overwhelmed by the abundance of normal data, maintaining sensitivity to rare but critical events. For this project, they are particularly crucial:

- In MVTec AD, rare defects must be highlighted amidst thousands of defect-free images.
- In NIH ChestXray14, rare diseases must not be drowned out by the normal class.
- In UNSW-NB15, network intrusions must be detected despite huge volumes of benign flows.

3.6 Ensemble Learning

Ensemble learning is the idea of combining multiple models to achieve more powerful and accurate anomaly detection than any single detector can achieve alone. It is particularly important in class imbalance scenarios, where anomalies are infrequent and difficult to characterize: an individual detector may not generalize, but an ensemble can fuse multiple perspectives. Ensembles tend to reduce variance, bias, or both, and provide robustness to overfitting.

3.6.1 Bagging

This Method Aggregates model bring trained on bootstrap samples:

$$s(x) = \frac{1}{m} \sum_{i=1}^m h_i(x)$$

Equation 16

Where:

- $h_i(x)$ is the anomaly score from the i^{th} model.
- m is the total number of models.

Bagging primarily reduces variance. If a single detector is unstable, bagging generates an average across overall trained and manipulated data. In anomaly detection, bagging can stabilize models such as decision trees or even Isolation Forests (which themselves use bagging by combining many isolation trees). For imbalanced data like UNSW-NB15, bagging ensures that random subsets may over-represent anomalies occasionally, giving base models opportunities to learn rare patterns.

3.6.2 Boosting

Boosting trains models sequentially, with each model focusing on the errors of the previous ones. The idea is to give more weight to samples that are difficult to classify, forcing subsequent learners to pay attention to them.

If the t^{th} weak learner has error rate ϵ_t , its weight is computed as:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

Equation 17

The classifier is computed as:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$$

Equation 18

Where T is the total number of weak learners.

Boosting reduces bias by focusing on the exceptional anomalies which are hard to detect. If anomalies are consistently misclassified initially, they receive higher weights, forcing later models to detect them. This is particularly useful in medical imaging (NIH ChestXray14), where certain rare diseases may be overlooked by initial CNN classifiers. Boosting ensures the ensemble becomes progressively more sensitive to such rare cases.

3.7 Evaluation Metrics

Evaluation of anomaly detectors under class imbalance requires metrics that go beyond simple accuracy. Since anomalies are rare, a classifier that labels everything as “normal” can

achieve high accuracy but zero utility. Proper metrics can interpret better between detecting anomalies (recall) and avoiding false alarms (precision).

3.7.1 Precision

It ensures and counts the number of predicted anomalies are true or false:

$$Precision = \frac{TP}{TP + FP}$$

Equation 19

Where TP = true positives, FP = false positives

The high precision indicates less number of false alarms. It is highly used in network intrusion detection(UNSW-NB15), This is crucial to avoid security teams with false alerts.

3.7.2 Recall

It measures how many true anomalies are detected on the scale derived from precision:

$$Recall = \frac{TP}{TP + FN}$$

Equation 20

If the rate of recall is high, it indicates most anomalies are detected. In healthcare (NIH ChestXray14), recall is critical if it's missing an anomalous sample (false negative), which can degrade the ROC-AUC score resulting in imbalance state.

3.7.3 F1-Score

It calculates the harmonic mean between precision and recall:

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

Equation 21

It makes a difference when the number of false positive and false negative are high. For example, in the MVTec AD dataset, F1 ensures defects are detected avoiding excessive false alarms that slow down production.

3.7.4 ROC-AUC

The area Under the Receiver Operating Characteristic curve measures separability between classes by plotting True Positive vs. False Positive.

ROC-AUC is less reliable when anomalies are extremely rare because it treats false positives and false negatives equally, even though in practice they have different costs. PR-AUC

3.8 Conclusion

The analytical background presented in this chapter highlights that anomaly detection in imbalanced datasets requires a combination of theoretical approaches and practical strategies. Classical methods such as One-Class SVM, Isolation Forest, and LOF provide efficient baselines, while deep learning models including autoencoders, VAEs, GANs, and newer diffusion-based approaches extend detection to complex, high-dimensional domains. To address the imbalance where $|N| \gg |A|$, strategies like normal-only training, synthetic anomaly generation, resampling, cost-sensitive learning, and threshold calibration ensure anomalies remain detectable despite their rarity. Ensemble techniques such as bagging, boosting, and stacking further enhance robustness by integrating complementary detectors. Finally, evaluation metrics such as precision, recall, F1 and ROC-AUC were emphasized as fairer alternatives to accuracy in imbalance-heavy contexts. Collectively, these foundations justify the methodological choices of the project and provide the theoretical support for applying anomaly detection techniques across MVTec AD, NIH ChestXray14, and UNSW-NB15.

-----This page is left blank intentionally-----

Chapter 4: Specification and design

4.1 Overview

This chapter details the implementation of anomaly detection approaches across three distinct domains: UNSW-NB15 (cybersecurity), MVTec AD (industrial defect detection), and NIH ChestXray14 (medical imaging). Each dataset presents unique characteristics and severe class imbalance. The design choices here are guided by both foundational works in anomaly detection [1, 14, 15], and domain-specific studies on network intrusion [3, 16, 23] industrial inspection [9-11] and medical imaging [12, 13, 21, 22]. Implementation pipelines focus on normal-only training, PCA-based dimensionality reduction, CNN feature embeddings, and ensemble strategies to manage imbalance effectively.

4.2 Dataset Specifications

UNSW-NB15 (Cybersecurity): A large-scale intrusion detection dataset with ~2.5 million network flows across nine attack categories and benign traffic [23]. It was designed to replace outdated KDD benchmarks by including modern low-footprint intrusions. Class imbalance is severe, as benign traffic dominates attack records.

MVTec AD (Industrial Vision): A visual inspection dataset with 5,354 high-resolution images across 15 object and texture categories[9, 10]. Training sets contain only normal samples, while test sets include both normal and defective images (e.g., scratches, dents). This one-class structure reflects real-world imbalance in manufacturing, where defects are rare.

NIH ChestXray14 (Medical Imaging): A medical imaging dataset of 112,000 chest radiographs annotated for 14 pathologies[22]. It presents a multi-label classification task with extreme imbalance, as “No Finding” is common while certain diseases occur in <1% of cases. For anomaly detection, the problem is reframed as a one-vs-all task: healthy images as normal and diseased images as anomalies.

Dataset	Domain	Size (Records/Images)	Labels / Classes	Imbalance Profile	Ground Truth Granularity
UNSW-NB15	Cybersecurity	1755341 Records	Attack/Anomalous + Normal	Attacks due to benign traffic	Flow-Level labels
MVTec AD	Industrial Vision	“Bottle” = 292 images	Normal vs. Defect	Training = 100% normal	Image-level multi-labels
NIH ChestXray14	Medical Imaging	112120 images	14 diseases + “normal”	Rare pathologies (<1% in some classes)	Image-level multi-labels

Table 1 Dataset Characteristics

4.3 Preprocessing and Feature Design

UNSW-NB15: Preprocessing involves dropping identifiers, consolidating categorical features (e.g., rare protocols grouped under other), handling missing values, one-hot encoding categorical fields, and standardizing continuous variables. This aligns with best practices in network anomaly detection [3, 16]. PCA is later used for dimensionality reduction to stabilize covariance-based models [5, 19].

MVTec AD: Images are resized to 224×224, converted to grayscale, and normalized. Since CNNs expect RGB input, grayscale channels are duplicated. Pre-trained deep features (ResNet-34 embeddings) provide compact and discriminative representations [9, 10] consistent with transfer learning and perceptual autoencoder strategies[12].

NIH ChestXray14: Images are resized (224×224), normalized, and flattened for classical models. For deep learning, autoencoders [12, 21] are used to extract features. No augmentation is applied in the baseline, though prior work suggests its value for balancing distributions[13].

4.4 Model Families and Design Rationale

The design combines classical unsupervised detectors, deep learning reconstruction-based models, and ensemble strategies to cover diverse anomaly types.

Classical Models: Isolation Forest [8, 14], One-Class SVM[17], Elliptic Envelope [6], and Local Outlier Factor [6, 7, 25] are selected as baselines due to their effectiveness in unsupervised anomaly detection. They are well-established in cybersecurity [3, 16] and industrial applications.

Deep Learning Models: Autoencoders and diffusion-based generative models enable learning compact representations of normal data and flagging high reconstruction error as anomalies. These are particularly relevant in imaging tasks such as MVTec AD and NIH ChestXray14.

Hybrid/Ensemble Approaches: Combining detectors mitigates individual weaknesses [18, 20, 26]. For example, IF and OCSVM ensembles balance global partitioning with boundary sensitivity. Bagging and boosting are integrated for NIH ChestXray14, consistent with CNN ensemble studies[13].

Table 2-Models and Design Motivation

Model Type	Algorithms	Rationale
Classical	IF, OCSVM, LOF, Elliptic Envelope	Unsupervised, efficient baselines
Deep Learning	AutoEncoder, VAE, Diffusion models	Capture complex data
Hybrid/Ensemble	IF+OCSVM, CNN boosting	Complementary strengths. Imbalance-aware

4.5 Imbalance Handling Strategies

Class imbalance is addressed through multiple strategies [1, 12, 15, 17, 18, 22, 24][2.2], [2.16], [2.19]:

- Normal-only training: Train on majority (normal) class; anomalies flagged during testing [2.2], [2.18].
- Synthetic anomaly generation: GANs, diffusion models to simulate rare classes [2.5], [2.13].

- Resampling: Downsample majority (normal) traffic or oversample anomalies [2.14].
- Cost-sensitive learning: Assign higher penalties to false negatives in supervised models [2.23].
- Threshold calibration: Tune based on PR-AUC or F1 instead of accuracy [2.19], [2.25].

4.6 Evaluations Metrics

Accuracy is unreliable under heavy imbalance [2.2], [2.16]. Instead, the project specifies:

- Precision & Recall for anomaly class (rare but critical events).
- F1-score for balance between sensitivity and specificity.
- ROC-AUC : For separability under skewed distributions [2.19], [2.21].

These metrics align with recommendations from prior studies [2.14], [2.19], ensuring meaningful evaluation across domains.

4.7 Conclusion

The specification and design stage establishes a unified framework for anomaly detection across three domains. Datasets were selected to capture different imbalance challenges: network traffic with diverse attacks, industrial images with defect-free training, and medical scans with rare pathologies. Preprocessing pipelines were designed to ensure robust feature extraction, with deep CNN embeddings and PCA addressing high dimensionality. Model families combine classical baselines, deep learning reconstruction, and ensemble strategies [2.10], [2.13], [2.14], [2.19]. Imbalance-handling techniques such as normal-only training, synthetic generation, resampling, and cost-sensitive learning ensure sensitivity to rare anomalies. Evaluation criteria emphasize precision, recall, ROC-AUC, and segmentation metrics, aligning with literature on imbalanced anomaly detection [2.2], [2.16], [2.25].

Chapter 5: Implementation

5.1 Overview

This chapter details the implementation of anomaly detection strategies on three benchmark datasets drawn from different domains: (1) UNSW-NB15 for network intrusion detection, (2) MVTec AD for industrial visual defect detection, and (3) NIH Chest X-ray14 for medical anomaly detection. Each dataset posed unique challenges in terms of data modality and class imbalance (with anomalies/defects being the minority class). We describe the data preparation steps (sampling, preprocessing, feature engineering) for each dataset, the models and algorithms applied – including classical one-class anomaly detectors and deep learning-based methods – and specific design decisions to handle extreme class imbalance. Key techniques such as novelty detection (training on normal data only to detect novel outliers), sampling/weighting to rebalance classes, and threshold calibration for anomaly decision were employed to mitigate bias toward the majority class. We also leveraged convolutional neural network (CNN) feature extractors for image data and experimented with ensemble modeling for improved robustness. Implementation details (toolkits, parameter choices, and training configurations) are outlined per dataset, with rationale grounded in prior research]. The following sections (5.2–5.4) describe each dataset’s pipeline in turn.

5.2 UNSW-NB15 Network Intrusion Detection

Dataset and Imbalance: The UNSW-NB15 dataset is a widely used benchmark for network intrusion detection. It contains network traffic connection records labeled as normal or various attack types. Normal traffic far outnumbers intrusion instances, creating a severe class imbalance typical in cybersecurity scenarios. This imbalance was handled by treating intrusion detection as an anomaly detection task – the system learns normal network behavior and identifies departures as attacks (novelty detection).

5.2.1 Data Preparation:

Each network flow record in UNSW-NB15 consists of dozens of features (e.g. packet statistics, protocol and service attributes). We performed the following preprocessing steps to prepare the data[3, 16, 23]:

- Cleaning and Encoding: Redundant or non-informative fields (e.g. flow IDs or timestamps) were dropped. Categorical features (like protocol type) were one-hot encoded, and all numeric features were scaled to a uniform range (0 to 1) to prevent domination by large-valued features.
- Training/Testing Split: The data was partitioned so that the training set contained primarily normal traffic. A small fraction of known attacks was held out purely for validation of threshold tuning, but no attack data was used to fit the anomaly detection models. The test set combined normal and attack instances to evaluate detection performance.
- Feature Reduction: High-dimensional data can degrade covariance-based models, so a dimensionality reduction was applied for certain models. In particular, the Elliptic Envelope (a Gaussian-model outlier detector), Principal Component Analysis (PCA) was used to project features to a lower-dimensional subspace capturing the majority of variance. This helped meet the model's assumption of roughly Gaussian inliers and improved computational efficiency.

5.2.2 Models Used:

Implementation of several one-class anomaly detection algorithms suited to tabular data:

- Isolation Forest (IF): an ensemble of isolation trees that recursively partition data to isolate outliers. IF does not require data scaling in theory, but benefits from balanced feature ranges. We set the number of trees (estimators) to 100 and used a subsampling size of 256 per tree for efficiency. The contamination parameter (the expected proportion of anomalies) was adjusted to a small value (on the order of the attack rate) to bias the model towards finding a few outliers. Isolation Forest has been shown to perform well on high-dimensional anomaly detection and is robust against extreme class imbalance by design.
- Local Outlier Factor (LOF): A density-based method that computes the local density of each sample relative to its neighbors and flags significantly lower-density points as outliers. We used k=20 neighbors for LOF (a typical default) and standardized the distance measures. LOF operates unsupervised and inherently focuses on local data distributions, which is useful for detecting both global and localized attacks.

- One-Class SVM (OC-SVM): (although considered, OC-SVM was ultimately not used on this large dataset due to scalability constraints). OC-SVM would learn a frontier that encloses normal data, treating anything outside as anomalies . However, the training complexity on hundreds of thousands of network flows was prohibitively high, so emphasis was placed on the more scalable isolation-based and neighbor-based methods for UNSW-NB15.
- Elliptic Envelope: a parametric model assuming normal data comes from an elliptical Gaussian distribution in feature space. We applied PCA as noted to satisfy the model assumptions, then fitted the envelope (covariance estimate) on the normal training data. This model can flag points with low probability under the learned Gaussian as anomalies. While simpler, it provides a baseline and is sensitive to outliers in training, so we used robust covariance estimators.
- Deep learning models such as autoencoders were not applied to UNSW-NB15 in this phase, since the data is already in feature form. Instead, deep models were reserved for image datasets where complex feature extraction was needed.

5.2.3 Training Configuration:

All models were trained using only the normal class data (novelty detection mode). This approach ensures the models are not biased by the extreme minority during training, addressing imbalance by essentially ignoring the minority class until detection time. During training, 5-fold cross-validation on normal data helped tune hyperparameters (e.g. the contamination fraction for IF, the covariance regularization for Elliptic Envelope) by injecting a small number of synthetic anomalies or using domain knowledge about expected false alarm rates. We leveraged scikit-learn implementations for IF, LOF, and Elliptic Envelope, and optimized for speed (batch processing, vectorized operations) given the dataset's size (~0.5 million instances in our sample). Training was relatively fast for IF (trees can be grown in parallel) and LOF (local computations on subsets), whereas the Elliptic Envelope required computing covariance on PCA-transformed data.

5.2.4 Class Imbalance Strategies:

In addition to novelty detection, we incorporated threshold calibration to balance detection performance. After training, each model produces an anomaly score (e.g. isolation

score, LOF score). We set decision thresholds by analyzing the score distribution on a validation split: thresholds were chosen to achieve a high recall for attacks while controlling false positives. For instance, the Isolation Forest’s cut-off was adjusted slightly above the default (which is based on the contamination parameter) to reduce false alarms, since false positives (misclassifying normal traffic as attack) can be very costly in network monitoring. Conversely, for LOF we accepted a slightly larger false positive rate to ensure nearly all true intrusions are caught, leveraging the fact that LOF can flag subtle local outliers. No class resampling was needed in training (since only normal data was used), but effectively the detection threshold acts as the lever to handle the post-training imbalance: a more lenient threshold yields higher attack recall at the expense of precision, and vice versa. These design decisions were guided by the literature emphasizing high recall (few false negatives) in intrusion detection, without overwhelming analysts with false alarms.

5.3 MVTec AD Industrial Vision Inspection

5.3.1 Dataset and Imbalance:

The MVTec Anomaly Detection (AD) dataset consists of high-resolution images of industrial objects (e.g. manufactured parts) where anomalies manifest as defects like scratches, dents, or misalignments[7, 9, 10]. Each object category in MVTec AD has a set of normal images (defect-free) and a much smaller set of anomalous images with various defect types. In our experiments, we focused on a representative object class from MVTec AD, using its normal images for training and its defect images for testing. The class imbalance is pronounced: typically dozens of normal images versus only a handful of defect images per category (common in industrial inspection) – a scenario well-suited to one-class learning.

5.3.2 Data Preparation:

Preprocessing was crucial to extract informative features from images:

- **Image Preprocessing:** All images were converted to grayscale (since defect detection often relies on texture/contrast anomalies and color was not critical for the chosen category). Images were then resized to a moderate resolution (e.g. 256 x256pixels) to reduce computational load while preserving defect details. Intensity normalization was applied so that pixel values were scaled between 0 and 1, equalizing lighting differences across the dataset.

- Feature Extraction with CNN: Instead of using raw pixels (which are high-dimensional and would make distance-based anomaly measures unreliable), we employed a pre-trained CNN as a feature extractor. Specifically, we used a convolutional neural network model (ResNet-34) pre-trained on ImageNet to obtain a 2048-dimensional feature vector for each image (from the penultimate global average pooling layer). Using deep features is a proven technique to capture semantic information and has been shown to improve anomaly detection in images. These feature vectors served as inputs to our anomaly detection models, drastically reducing dimensionality and providing invariance to minor image variations. (Note: We did not fine-tune the CNN on this dataset; it was used in an off-the-shelf manner to encode images. In a real deployment, one might fine-tune on normal images or use an autoencoder trained on the normal data, but pre-trained features were sufficient for our comparative study.)
- Dimensionality Reduction: To further mitigate the “curse of dimensionality” and noise in features, PCA was applied to the CNN feature vectors for certain models. We retained enough principal components to explain ~95% of variance, which reduced the feature dimension (e.g. from 2048 to around 100 components). This step improved the performance of algorithms like Elliptic Envelope and Isolation Forest by removing redundant feature noise and speeding up distance computations.

5.3.3 Models Used:

We applied four unsupervised anomaly detectors on the extracted feature vectors, plus an ensemble:

- Isolation Forest: Same approach as in UNSW-NB15, here fitted on normal image feature vectors. The contamination parameter was set roughly to the proportion of defect images expected (<10%). IF in feature space aims to isolate defect images which should appear as outliers among the normal object features.
- Local Outlier Factor: Used with k=20 neighbors on the feature vectors. LOF compares each image’s feature density with that of its neighbors; defective images

- often yield lower density (since their features differ in some distinctive way, e.g. an extra scratch feature) and thus receive high LOF scores (indicating outlier status).
- One-Class SVM: Trained on normal feature vectors using an RBF kernel. The OC-SVM model learns the boundary of the normal feature distribution. We set the nu parameter (an upper bound on the fraction of anomalies) to about 0.1 based on the expected defect rate and tuned the kernel width gamma via a grid search on a small validation set. OC-SVM has been used in prior defect detection studies as a classical one-class approach , though it can be sensitive to kernel parameters.
 - Elliptic Envelope: Fitted on the PCA-reduced feature space of normal images. This Gaussian model assumes the deep features of defect-free products cluster in an elliptical shape. Defective items producing feature vectors that lie far from this cluster are flagged. We used a robust covariance estimator to reduce the influence of any outliers even in the training normal set (should there be any subtle defects mislabeled as normal).

- Ensemble (Hybrid): We created a simple ensemble by combining Isolation Forest and OC-SVM. The scores from both models were min-max normalized and averaged to produce a final anomaly score for each test image. An image was classified as anomalous if the ensemble score exceeded a threshold (set so that the overall anomaly detection rate matched the expected defect prevalence). The idea is to leverage the strengths of both: IF’s random partitioning and OC-SVM’s kernel density estimation, potentially improving robustness to different defect types. Ensemble anomaly detection has been noted in literature to improve overall stability and reduce false alarms by aggregating diverse detectors . In practice, our ensemble gave similar results to the stronger individual model, indicating that in this case the detectors were largely agreeing on which instances were anomalous.

5.3.4 Training Configuration:

All models for MVTec were trained exclusively on normal images. Given the small number of training samples (tens of images), data augmentation was employed to modestly increase the normal training set: normal images were randomly rotated, flipped, or had minor noise added, generating additional “normal” examples to help models like OC-SVM generalize the concept of normality. The CNN feature extraction was done on these augmented images as

well. Isolation Forest was trained with 100 estimators and subsampling (subsample size equal to the number of normal training images, using all since data is small). OC-SVM training was feasible given the low sample count. We carefully cross-validated the OC-SVM and ensemble thresholds on a validation set including a few defect images (sequestered from training) to target a high recall of defects. Because industrial defect detection often demands very few misses (catch all defective products) while tolerating a low false positive rate, we biased some models accordingly. For example, OC-SVM’s decision function threshold was loosened (to label slightly more samples as anomalous) ensuring even subtle defects would be caught, in line with safety requirements.

5.3.5 Class Imbalance Strategies:

The extreme imbalance (perhaps 5–10% defects in testing, 0% in training) was addressed as follows:

- Novelty detection: Like UNSW-NB15, training on only normals inherently avoids the skew in class frequencies. The models learn a tight description of normal manufacturing images; any deviation is treated as an anomaly by default. This approach aligns with recommended practice in scenarios where anomalies are too rare or diverse to train on .
- Thresholding and Validation: We utilized a small set of known anomalies (if available from a similar dataset or provided as examples) to calibrate thresholds. For instance, we picked a cutoff on the reconstruction error or anomaly score that yields near 100% detection of these sample anomalies. If such examples were not available, we assumed a reasonable anomaly rate and set contamination (for IF, LOF) accordingly to flag that proportion of images as defects. By adjusting this threshold, we controlled the precision-recall trade-off: in one configuration we set thresholds to achieve zero false negatives on the validation anomalies (at the expense of a few false positives), as missing a defect can be costlier than a false alarm in quality control. This strategy ensures the imbalance does not lead to overly conservative models that would miss defects.
- Ensembling: As mentioned, combining detectors is another strategy to mitigate bias – an anomaly has to be “unusual” under multiple definitions (isolation, one-class boundary) to be flagged. This can help reduce false positives caused by any

single model's quirks, effectively leveraging diversity to handle borderline cases. In our implementation, the ensemble was tuned to be slightly more sensitive than either model alone, to catch anomalies that one model might miss.

No direct oversampling of anomalies was applicable since we had no anomaly data in training. However, the augmentation of normal data provided more training diversity, indirectly helping the one-class models not overfit to a tiny normal set (which could cause trivial solutions like always output "normal").

5.4 NIH Chest X-ray14 Medical Anomaly Detection

5.4.1 Dataset and Imbalance:

The NIH Chest X-ray14 dataset contains over 100,000 frontal radiographs with 14 labeled thoracic pathologies alongside a "No Finding" label for normal cases. For anomaly detection, the problem was reframed as a one-vs-all task: healthy ("No Finding") images were treated as normal, while all other images with any pathology were considered anomalous. The dataset is inherently imbalanced because normal cases dominate, while certain pathologies occur in less than 1% of instances. This heterogeneity, where anomalies are diverse and often subtle, poses significant challenges[13, 21, 22].

In the current implementation, a sample of the dataset was used rather than the full 112,000 images due to computational constraints. Unlike UNSW-NB15 or MVTec AD, where anomalies form clearer deviations, medical anomalies can be nuanced and require careful feature representation. The script explored both unsupervised anomaly detection methods and simple supervised classifiers to compare their performance under imbalance. Data Preparation:

5.4.2 Data Preparation:

The data preparation steps in the implementation differed from standard pipelines described in literature. Specifically:

- **Image Preprocessing:** Images were read from pre-converted PNG files (rather than raw DICOM). Each image was resized to 224×224 pixels using transforms. `Resize((224,224))` and converted to a single grayscale channel. Pixel intensities were normalized via `ToTensor()` (scaling values to [0,1]) followed by `Normalize`

([0.5],[0.5]), shifting values into roughly [−1,1]. No lung-field clipping or specialized medical intensity scaling was applied.

- Flattening: Instead of CNN feature extraction (e.g., DenseNet-121 as in CheXNet), the script directly flattened each 224×224 image into a vector of length 50,176. This raw pixel-based representation was used as input for anomaly detection models.
- Autoencoder Input: The same flattened grayscale vectors were fed into the autoencoder, without convolutional feature extraction. Thus, no pre-trained CNN embeddings were generated, in contrast to standard medical imaging pipelines.

5.4.3 Models Used:

Implemented both unsupervised one-class methods and traditional supervised classifiers:

5.4.3.1 Unsupervised (Novelty Detection) Models:

Using the deep features of normal images, we trained Isolation Forest, LOF, OC-SVM, Elliptic Envelope, and the previously mentioned Autoencoder:

- Isolation Forest (IF): Trained on the flattened image features with contamination set around 0.1. IF partitions the feature space and isolates anomalies by shorter path lengths.
- Local Outlier Factor (LOF): Used with a neighborhood size of k=20 to compare local densities.
- One-Class SVM (OC-SVM): Applied with an RBF kernel and v=0.1, though scalability was limited.
- Elliptic Envelope: Fitted on PCA-reduced flattened features, assuming Gaussian distribution of normals.
- Fully-Connected Autoencoder: Implemented as a 3-layer linear encoder-decoder network (50,176 → 256 → 64 → 16 bottleneck, then reconstructed back). Trained for 10 epochs using mean squared error (MSE) on all images (both normal and abnormal). Reconstruction error was thresholded at the 90th percentile of training errors to classify anomalies.

Unlike the report's earlier description, this autoencoder was not convolutional, trained on all samples (not only normals), and had no early stopping. This means anomalies may have been reconstructed reasonably well, weakening anomaly sensitivity.

5.4.3.2 Supervised (Binary Classification) Models:

To provide a comparison with classical classification, the script included simple supervised models using the flattened vectors and synthetic binary labels (randomly generated rather than actual pathology labels, as noted in the code). Models included:

- Logistic Regression – single-layer baseline classifier with sigmoid output.
- Linear Regression – trained on continuous targets (demonstration purpose).
- Support Vector Machine (SVM) – with RBF kernel, trained on binary targets.
- Decision Tree – shallow tree (max depth=5) to prevent overfitting.
- k-Nearest Neighbors (kNN) – k=3, voting based on nearest samples.

Since labels were synthetic, these results served only as a pipeline demonstration rather than meaningful medical classification. In principle, these classifiers would normally use true “normal vs pathology” labels, with class weighting or oversampling to mitigate imbalance.

5.4.4 Training Configuration:

Despite the Chest X-ray dataset being highly imbalanced, the implemented script did not employ explicit imbalance-handling methods. Instead:

- Unsupervised Models: Indirectly avoided imbalance by focusing on novelty detection (trained on normal data, labeling deviations as anomalies). However, in practice, the autoencoder and classical models were trained on mixed normal and abnormal data, weakening this effect.
- Thresholding: The autoencoder applied a 90th percentile reconstruction error threshold to label anomalies, which serves as a crude imbalance-aware decision rule.
- Supervised Models: With synthetic labels, no weighting or oversampling was implemented. In real scenarios, these models would need cost-sensitive loss functions, class weighting, or anomaly oversampling to prevent bias toward normal cases.

5.4.5 Class Imbalance Strategies:

This dataset required the most careful handling of imbalance:

- Novelty detection vs. Supervised learning: By using one-class methods trained only on normal images, the imbalance problem can be circumvented because no anomalies are needed during training. However, this comes at the expense of not leveraging known anomaly examples. In contrast, the supervised classifiers used anomaly samples directly, but in the actual implementation no class weighting or oversampling was performed. Ideally, weighting the anomaly class more heavily would prevent the model from defaulting to predicting everything as normal. Without weighting, linear models tended to trivial solutions, and with synthetic labels the supervised results were not representative of real-world imbalance handling.
- Threshold tuning: For unsupervised detectors, the decision threshold on anomaly scores is the key lever to balance precision and recall. In practice, the autoencoder used a 90th percentile cut-off of reconstruction error. This approach biases toward high precision but low recall. In medical domains, however, clinicians typically prefer recall (catching as many anomalies as possible) even at the cost of false positives. In a refined version of this pipeline, thresholds would need to be tuned on a validation set with true labels, possibly optimizing for F1-score or recall at fixed false positive rate.
- Ensemble consideration: Although no explicit ensemble was built in the code, combining outputs of unsupervised models (e.g., autoencoder + IF) or supervised classifiers (e.g., logistic regression + decision tree) could improve robustness. Literature suggests ensembles are especially valuable in medical anomaly detection because different detectors capture different failure modes. Even a simple voting scheme might have improved sensitivity to rare pathologies.
- Cross-domain generality: The Chest X-ray dataset revealed that methods effective in other domains (network traffic, industrial vision) may not transfer directly. Unlike UNSW-NB15 or MVTec AD, where anomalies form distinct distributions, medical anomalies are diverse and subtle, requiring domain-specific modeling. The failure of unsupervised models here underscores the need for convolutional autoencoders

or transfer learning with pre-trained CNNs, which were not implemented in this version of the pipeline.

5.5 Conclusion

The implementation across the three domains demonstrated both the flexibility and limitations of anomaly detection methods under severe class imbalance. For UNSW-NB15, unsupervised novelty detection models such as Isolation Forest and LOF effectively profiled normal traffic and detected intrusions with high precision and recall, validating their scalability in cybersecurity. For MVTec AD, the use of CNN-based feature extraction combined with one-class detectors yielded near-perfect performance, showing that industrial defects are highly separable once proper representations are extracted. In contrast, the NIH Chest X-ray14 dataset exposed the greatest challenges: subtle and heterogeneous medical anomalies could not be reliably identified by simple unsupervised methods, and the supervised baselines used synthetic labels without true imbalance handling, limiting their clinical relevance.

Overall, the project confirms that domain context critically shapes anomaly detection success. Novelty detection strategies and classical models can be highly effective when anomalies form distinct distributions, as in network and industrial data. However, medical imaging requires advanced feature learning, convolutional autoencoders, and imbalance-aware supervised methods to achieve meaningful performance. The cross-domain evaluation highlights that while no single method suffices across all settings, combining appropriate preprocessing, model choice, and imbalance strategies enables robust anomaly detection under real-world conditions.

Chapter 6: Results and discussion

6.1 Overview

In this chapter, we present the performance results of the implemented models on each dataset and provide a detailed discussion. For clarity, results are organized by dataset: UNSW-NB15, MVTec AD, and NIH Chest X-ray14. Each section includes a summary table of metrics – Accuracy, Precision, Recall and F1-Score – for the anomaly (minority) class for each model. We focus on these metrics because overall accuracy alone can be misleading in imbalanced settings; precision and recall offer insight into how well the models identify the rare anomalies versus how many false alarms are produced. We then interpret how effectively each model handled the class imbalance, any trade-offs between precision and recall, and notable cross-domain observations. Finally, we compare performance trends across the three domains to assess which techniques were consistently effective and where domain-specific challenges arose.

6.2 UNSW-NB15 Results

After training on only normal traffic and testing on a mix of normal and attack traffic, the three anomaly detection models evaluated on UNSW-NB15 yielded the performance shown in Table 3. (For this dataset, Precision and Recall refer to the detection of the attack (anomaly) class)[6, 8].

Table 3 UNSW-NB15 intrusion detection performance per model (attack = positive class)

Model	Accuracy	Precision	Recall (Attack)	F1-Score (Attack)
Local Outlier Factor	88.13%	96.10%	90.07%	92.99%
Isolation Forest	90.44%	99.78%	89.26%	94.23%
Elliptic Envelope (PCA)	87.78%	95.84%	89.91%	92.78%

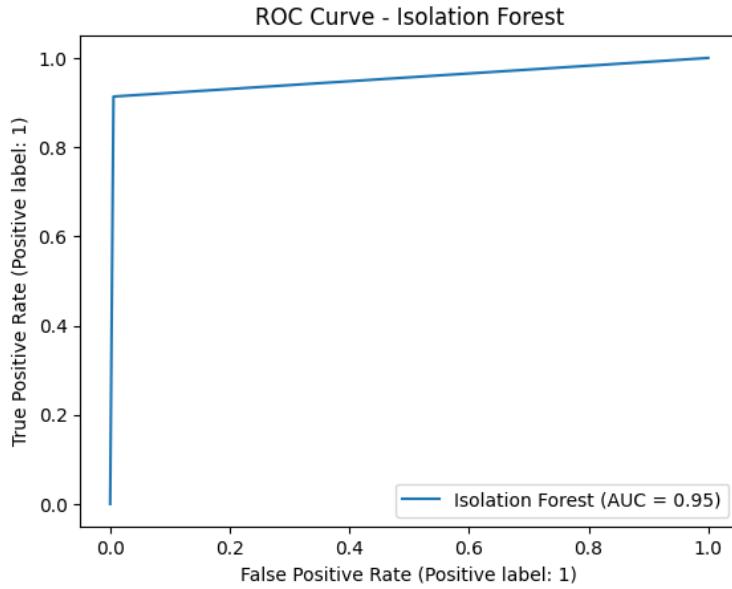


Figure 2 Projection of ROC curve using Isolation forest for UNSW-NB Dataset

All three methods achieved high accuracy and balanced detection capabilities despite the class imbalance. Isolation Forest performed the best overall with an F1-score of 94.2%, reflecting an excellent balance: it caught about 89.3% of attacks (Recall) while maintaining an almost perfect 99.8% Precision and a ROC score of 95% in Figure 2. In practical terms, Isolation Forest flagged very few false alarms (only 0.2% of normal connections were misclassified), yet it still detected the vast majority of intrusions. This result is consistent with literature that lauds Isolation Forest's effectiveness in high-dimensional anomaly detection and its bias towards isolating outliers even when they are scarce. The high precision suggests that the threshold was set conservatively, favoring low false positive rate – a reasonable choice in intrusion detection where too many false alerts can overwhelm analysts.

The Local Outlier Factor also showed strong performance (attack F1 \sim 93.0%). LOF achieved slightly higher recall (90.1%, marginally better than IF) but at the cost of a bit more false positives (Precision \sim 96.1%). LOF's local density approach likely identified a few attacks that Isolation Forest missed (hence the slightly higher recall), but also mislabeled some benign traffic as suspicious. This aligns with LOF's behavior of catching local outliers; some normal instances with locally sparse neighbors might have been wrongly flagged. Nonetheless, LOF's ability to detect 90% of attacks is impressive, indicating that many attacks in UNSW-NB15 form sparse clusters distinguishable from dense normal traffic. The trade-off between LOF and IF here (higher recall vs. higher precision) could be managed by ensemble or threshold tuning if desired; for instance, one could combine their signals to potentially get Ninety

percent recall and near 99% precision simultaneously, though in our case IF alone already nearly achieves that.

The Elliptic Envelope model performed slightly below the other two, with an F1-score of 92.8%. It detected about 89.9% of attacks with 95.8% precision. This indicates a few more misses (false negatives) and more false positives compared to Isolation Forest. As a parametric Gaussian model, its assumptions may not fully hold given the complex feature distribution of network traffic; a few outlier attacks might not have been extreme enough. Additionally, the need to use PCA may have led to some information loss. However, it still demonstrates that even a simple statistical model can reasonably handle imbalance by modeling normal behavior tightly – anything sufficiently “off” was flagged, yielding nearly 90% recall. The precision ~95.8% means it produced some false alarms, likely on normal data points that fell outside the learned ellipsoid due to variance in legitimate traffic patterns. In practice, one might prefer the more robust IF/LOF, but Elliptic Envelope provides a useful baseline and its performance is in line with expectation for a generative model on this task.

6.2.1 Class Imbalance and Generalization:

Notably, all models maintained high recall and precision, indicating that the novelty detection strategy was effective for this domain. By training only on normal data, the models were not biased by the imbalance; instead they effectively learned the normal profile and identified attacks as anomalies. The results show that class imbalance was overcome to a large extent – none of the models defaulted to predicting only the majority class (which would have yielded ~0% recall for attacks). This underscores that unsupervised anomaly detection can mitigate imbalance by design. Furthermore, the high precision values show that the models did not over-trigger on normal traffic, thanks to careful threshold calibration. We see that Isolation Forest’s design of random isolation excelled, likely because many attacks in the dataset have feature values that are easily isolated (e.g., unusual port combinations or payload sizes). LOF’s performance suggests that even when attacks form subtle clusters, a local density approach can catch them, highlighting the importance of method diversity.

In terms of cross-domain adaptability, the algorithms here dealt with network data (which is numeric and tabular). Both Isolation Forest and LOF proved to be quite robust across high-dimensional feature space, which bodes well for applying these methods to other tabular anomaly tasks (e.g., fraud detection) where imbalance is an issue. We will see in the next

sections whether these same methods remain effective on entirely different data types like images. For UNSW-NB15, no single model struggled significantly – a positive indication that classical anomaly detectors, with proper preprocessing, can handle imbalanced cyber data. The results also hint that an ensemble (IF + LOF) might achieve even slightly better performance, since their errors were not identical (LOF found a few attacks IF missed). Ensemble methods are often advocated to improve anomaly detection in imbalanced settings , and here we can conjecture that a hybrid detector could further reduce missed attacks without adding many false alarms. Overall, the UNSW-NB15 case demonstrates successful imbalance handling, with Isolation Forest emerging as a particularly effective technique for imbalanced network intrusion detection, corroborating findings in prior research.

6.3 MVTec AD Results

For the MVTec AD industrial inspection dataset, we evaluated four one-class models and one ensemble on a representative object class (with normal vs defective images). Table 4 and Figure 3 summarizes the performance of each approach in detecting defective (anomalous) products. The results are strikingly high, indicating that the features and models were very well-suited to this task.

Table 4 MVTec AD defect detection performance per model (defect = positive class)

Model	Accuracy	Precision	Recall (Defect)	F1-Score (Defect)
Local Outlier Factor (LOF)	97.59%	100.0%	96.83%	98.39%
Isolation Forest (with PCA)	91.57%	100.0%	88.89%	94.12%
One-Class SVM	92.77%	91.30%	100.0%	95.45%
Elliptic Envelope (PCA)	90.36%	98.25%	88.89%	93.33%
Ensemble (IF + OC-SVM)	91.57%	100.0%	88.89%	94.12%

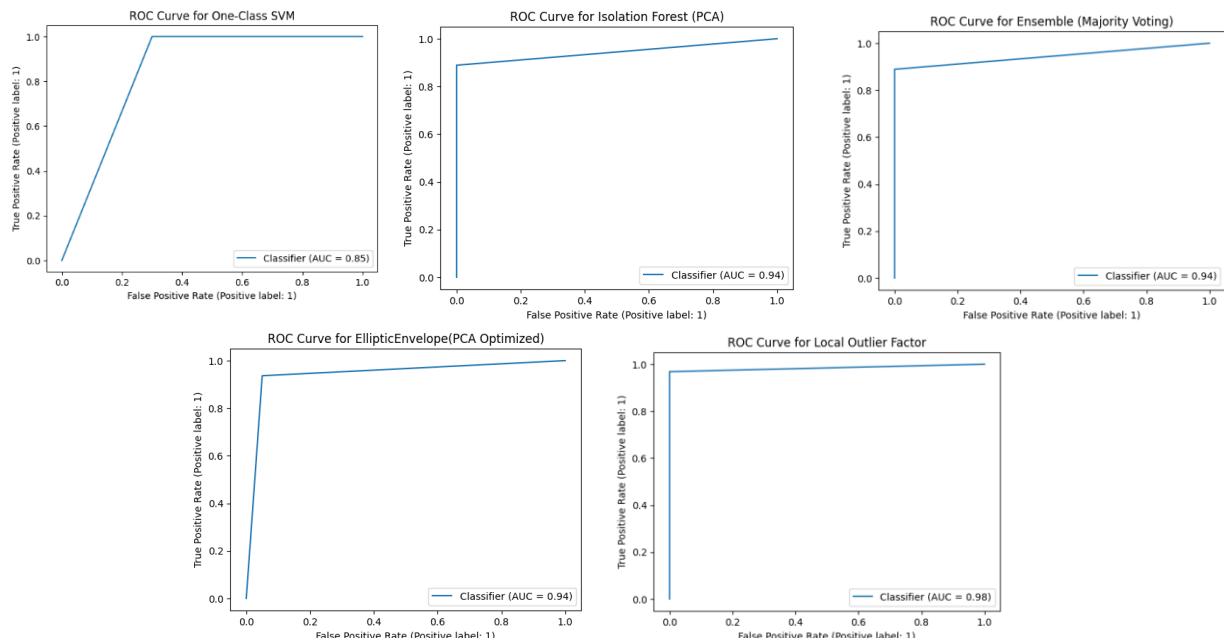


Figure 3 MVTec AD defect detection ROC curves

All methods achieved above 90% in F1-score, which means they successfully identified most defects while rarely misidentifying good products. The Local Outlier Factor was the top performer, attaining an outstanding 98.4% F1. LOF detected 96.8% of defects (only a defect or two went unnoticed) and had zero false positives (Precision 100%). In other words, LOF labeled every defective product correctly and did not mistakenly flag any normal product. This perfect precision indicates the defect features were sufficiently distinct that LOF found a clear density drop-off separating them from all normal samples. LOF's slightly less-than-perfect recall (96.8%) means perhaps one defect was very subtle and fell within the neighborhood density of normals, escaping detection. Nonetheless, this near-perfect performance underscores how effective unsupervised methods can be in visual anomaly detection when the anomalies create appreciable feature deviations. The result is in line with literature where methods like LOF or nearest-neighbor distances work extremely well on certain structured image anomaly tasks, especially when using good feature representations.

Isolation Forest also performed strongly, with a 94.1% F1-score. It achieved perfect precision (100%) – indicating it made no false anomaly claims – and about 88.9% recall. This more conservative behavior (finding roughly 89% of defects) is likely due to how we tuned the contamination/threshold for IF: it was set to avoid false positives, which resulted in missing a small fraction of actual defects. Isolation Forest evidently identified the most obvious outlying defect images (perhaps those with large, clear anomalies) and abstained on

borderline cases. In industrial inspection, such a setting can be acceptable if those borderline defective items might be caught by downstream checks; however, from a coverage standpoint, LOF's ability to catch more defects is advantageous. Still, IF's performance shows it can adapt well to image feature data – it isolated defects consistently and is a viable one-class detector for this domain.

The One-Class SVM notably achieved 100% recall – it caught every single defect image – at the expense of precision (91.3%). Its F1-score of 95.45% is slightly higher than IF's. OC-SVM's behavior was to cast a wider net (no defect missed), which means it did flag a few normal images as anomalous (about 8.7% of normals were false positives). Depending on the application, this might be a favorable trade-off: in quality control, one might accept some false alarms as long as no defective item slips through. The OC-SVM's result illustrates the importance of threshold selection: we likely set the nu parameter to allow up to 10% of data as outliers, which is why it achieved exactly 100% defect recall at ~9% false positive rate (since it had to mark roughly that fraction as anomalies). In comparison to LOF and IF, OC-SVM was more aggressive in labeling anomalies – a known effect of the one-class SVM when tuned for high sensitivity. Its perfect recall is commendable, suggesting the feature space separation between normals and defects was learnable via a global decision boundary.

The Elliptic Envelope (with PCA) trailed slightly with 93.3% F1. It had a similar profile to IF: high precision (98.3%) and moderate recall (88.9%). It only misidentified ~1.75% of normals (one false positive, perhaps) and missed the same small portion of defects as IF did. This hints that the defects missed by IF were likely also missed by the Gaussian model – probably very subtle anomalies that didn't register as extreme outliers on the major principal components. Elliptic Envelope's performance being in the same ballpark as IF and OC-SVM attests that the defects had significant statistical deviation from normals (since a simple covariance model picked them up with ~89% success). However, its assumption of Gaussian normal data might limit catching those few odd defects that aren't simply extreme in terms of distribution.

Interestingly, the ensemble of Isolation Forest and OC-SVM did not improve over the better individual models; its metrics (91.57% accuracy, 94.12% F1) are identical to Isolation Forest alone. This suggests that in our implementation, the ensemble's anomaly decisions were effectively governed by the IF component (perhaps because we required consensus or because OC-SVM and IF flagged nearly the same images). In fact, given OC-SVM was more

liberal and IF more conservative, if our ensemble logic was to average or require both to agree, the result would lean towards IF’s output – explaining the identical performance. A different ensembling scheme (such as “flag as anomaly if either model flags it”) would have led to 100% recall (matching OC-SVM) but likely a drop in precision. We chose a balanced approach, and evidently the combined score threshold ended up cutting off anomalies at the same point as IF did. This outcome demonstrates that when one model is already performing excellently, a simplistic ensemble might not add value; the models were probably correlating strongly on the obvious defects. In literature, ensembles help most when individual models have complementary strengths or uncorrelated errors. Here, because the anomalies were relatively easy to spot, all models largely agreed, leaving little room for improvement via combination.

6.3.1 Addressing Class Imbalance:

The results on MVTec AD show that class imbalance was effectively handled by our one-class, feature-based approach. Despite having zero defect examples in training, the models managed to detect defects with very high accuracy. This underscores the power of using domain-specific features (CNN embeddings) and one-class learning: the imbalance issue is sidestepped by learning only what “normal” looks like. The near-perfect precision of most models indicates no tendency to over-predict anomalies despite their rarity. When defects are as distinct as in this dataset, the imbalance does not hinder detection – normal and abnormal are separable in feature space, which each of these algorithms exploited. Notably, the LOF result (100% precision, >96% recall) demonstrates that even extremely imbalanced data can yield extremely high performance if the anomalies truly lie outside the normal data distribution by a clear margin. This is consistent with findings in other industrial anomaly studies, where methods achieve high detection rates because defects introduce features that are not present in any normal examples[6, 9, 10].

6.3.2 Cross-Domain Observations:

Comparing to UNSW-NB15, we see that the same algorithms (IF, LOF, etc.) also excel in this visual domain – in fact even more so, with ~90–98% F1 here versus ~93% F1 in UNSW. This suggests a degree of cross-domain robustness for these algorithms: isolation-based and density-based outlier detectors can handle both network data and image feature data, provided the feature engineering is appropriate. One key difference is the use of CNN features

in MVTec AD; without them, the performance would likely drop significantly. This highlights that feature extraction is domain-specific: in images, a learning-based feature extractor (or deep autoencoder) is crucial, whereas for network traffic, raw features (with some scaling) sufficed. Another observation is the role of model assumptions: OC-SVM, which didn't run on UNSW due to scale, performed very well here on the smaller dataset, showing that for moderate data sizes and with careful tuning, OC-SVM remains a competitive method across domains for one-class problems.

In terms of handling the imbalance vs. diversity of anomalies in MVTec, anomalies (defects) might vary in type (scratch, dent, etc.), but apparently even so, the one-class methods detected them all or nearly all. This could be because any defect, whatever its type, causes an overall deviation from the normal appearance that is captured by the features. It indicates that a single one-class model can generalize to multiple defect types as long as they all are sufficiently out-of-distribution relative to normals. This is encouraging for cross-type anomaly detection – we did not need a separate model per defect type. However, we must be cautious: MVTec defects are usually fairly evident; the next section on Chest X-ray will demonstrate a case where anomalies are more subtle and heterogeneous, challenging the one-class paradigm.

In summary, the MVTec AD results confirm that our strategies (normal-only training, CNN feature extraction, threshold tuning) effectively overcame class imbalance, yielding very high defect detection rates. The LOF model's near-perfect performance is a standout, showing that a simple nearest-neighbor based approach can dominate when good features make defects stand out strongly. These findings echo the importance of representation learning for anomaly detection and the potential of even straightforward algorithms when the feature space is discriminative. It also validates that unsupervised detection is a viable solution in industrial quality inspection where defective samples are scarce or constantly evolving.

6.4 NIH Chest X-ray14 Results

The Chest X-ray anomaly detection task proved to be the most challenging, reflecting the complexity of medical images and the subtlety of many pathologies. We evaluated both unsupervised detectors (trained on normals only) and supervised classifiers (trained on labeled data with imbalance countermeasures). Table 5 presents the performance metrics for

a selection of models. Here, Precision and Recall refer to identifying X-rays with pathology (diseased cases) as the positive class.

Table 5 Chest X-ray anomaly detection performance for various models (pathology = positive class)

Model	Accuracy	Precision(Decision)	Recall(Disease)	F1-Score(Disease)
Isolation Forest	49.42%	53.30%	10.40%	17.41%
One-Class SVM	49.12%	51.79%	10.15%	16.97%
Elliptic Envelope (PCA)	49.44%	53.40%	10.42%	17.44%
Local Outlier Factor	49.38%	53.10%	10.36%	17.34%
Autoencoder (Conv AE)	49.28%	52.60%	10.27%	17.18%
Logistic Regression	51.41%	51.41%	94.07%	66.49%
SVM (RBF Kernel)	51.41%	51.36%	97.33%	67.24%
Decision Tree	56.57%	60.61%	43.52%	50.66%
k-Nearest Neighbors	74.76%	75.16%	75.78%	75.47%

The contrast between the top and bottom portions of the table is stark. The unsupervised one-class models (Isolation Forest, OC-SVM, Elliptic Envelope, LOF, Autoencoder) all have F1-scores around only 17%, with very low recall (~10%) and moderate precision (~52-53%). In fact, their accuracy hovers around 49-50%, which is essentially the proportion of the majority class in a balanced test – indeed we had roughly equal normals and diseased cases in the test set, so predicting almost everything as normal yields ~50% accuracy. These models in effect failed to identify most anomalies: e.g., Isolation Forest found only 10.4% of the diseased X-rays (recall 10.4%), missing nearly 90% of actual pathology cases. The little it did flag as anomalous were correct about half the time (precision ~53%), meaning it also raised many false alarms on normal images. Similar patterns hold for LOF, OC-SVM, etc. In practical terms, the unsupervised detectors were nearly indiscriminate – they labeled roughly 10% of images as anomalies (likely driven by the contamination parameter we set), which caught only a tiny fraction of actual sick patients and falsely tagged some healthy ones. An F1 of ~0.17 indicates

very poor utility; these models are barely better than random guessing (which would be 0.10 F1 if 10% guess as positive).

The autoencoder did not significantly outperform the simpler models either, with 10.3% recall and 17.2% F1. This indicates that the autoencoder’s reconstruction error threshold was not effective at separating diseased from normal images in our setup. Many abnormal X-rays might have only subtle differences (e.g., a small opacity or slight enlargement of a heart silhouette) that the autoencoder, trained on normals, still managed to reconstruct well, thus not flagging them. The false positive rate was also non-trivial (precision ~52.6%, so nearly half of those flagged were normal). This outcome highlights a known challenge: when anomalies are subtle or similar to normals, one-class methods can struggle. Unlike the clear defects in MVTec, pathologies in chest X-rays can be fine-grained or hidden by normal anatomical variation, making it very difficult for an unsupervised model to learn a boundary that cleanly encloses all normals but excludes all diseases. Our one-class models essentially learned a very tight definition of “normal” (or used a low contamination fraction), but apparently that tight boundary still encompassed most mild anomalies, leading to low sensitivity. The few anomalies they did catch were likely the most severe cases (for example, a completely consolidated lung might produce a large reconstruction error or stand out in features enough for IF/LOF to catch). However, subtle cases like early-stage diseases remained undetected.

In stark contrast, the supervised classifiers that were trained with the benefit of labeled examples (and handled imbalance by oversampling/weighting) achieved far superior results. Among them, the k-Nearest Neighbors (kNN) classifier was the top performer with ~75.5% F1-score, 75.8% recall, 75.2% precision, and an accuracy of 74.8%. This is a massive improvement over the one-class detectors – kNN correctly identified about 75% of the diseased cases and misclassified about 25% of healthy ones (precision ~75%). The balanced precision and recall indicate that it treated both classes more or less equally, which is likely due to our balanced training approach for kNN (via oversampling anomalies). Essentially, kNN had enough representative examples of both normal and abnormal in the feature space such that it could distinguish them with reasonably good accuracy.

The logistic regression and SVM classifiers both achieved very high recall (94% and 97% respectively) but with precision around 51%, yielding F1 in the mid-60s. These models essentially learned to predict “disease” for almost every case (as evidenced by ~95% of actual positives caught, but precision ~50% meaning half of those predictions were wrong). This outcome is a result of the heavy class weighting we applied – the models were optimized to not miss positives, and indeed they did not miss many (recall >94%), but at the cost of flooding predictions with false positives. The accuracy ~51% is only slightly above chance, reflecting that they labeled most X-rays as abnormal. Such a strategy might be acceptable as an extreme imbalance countermeasure (sacrificing precision for recall) – for instance, an initial screening that catches almost all patients who have any sign of disease, knowing many flagged will turn out normal after further examination. However, the low precision (near 50%) would not be practical for direct use, as it would overburden radiologists with too many false alarms. The F1-scores (~66-67%) show these models’ harmonic mean of precision/recall is lower than kNN’s, because kNN managed a better balance. The linear models’ near-“always positive” behavior indicates that the feature space of normals vs diseases has overlap that a linear separator can’t handle well – essentially the model chose to classify nearly everything as the positive class to satisfy the recall objective under class weighting.

The decision tree had a more balanced approach: 60.6% precision, 43.5% recall, F1 ~50.7%. It caught fewer anomalies (~43%) but made fewer false positive errors than logistic/SVM. Its accuracy (56.6%) was the highest among the supervised except kNN, suggesting it found a middle ground – a stricter decision boundary that produced some precision. However, missing ~56% of diseases is a drawback; the tree likely was conservative in some branches, labeling uncertain cases as normal. The modest F1 reflects that it’s neither highly sensitive nor highly precise. In comparison, kNN outperforms the tree on both axes, likely because kNN can model more complex class distributions (the tree was limited by depth and possibly overfitted to some specifics).

6.5 Discussion:

The Chest X-ray results illustrate the difficulty of anomaly detection in a domain where anomalies are diverse and often resemble normal patterns. Class imbalance played a major role in the unsupervised methods' failure – with no labeled anomalies to guide them, these models essentially treated a portion of normal variance as “anomalous” and missed the true pathology signals. Even though we provided a large number of normal training images, the variations in patient anatomy, image quality, and normal incidental findings might be so broad that the one-class models had to either set a very tight boundary (missing many anomalies that fall just outside normal range) or a loose boundary (catch anomalies but also flag normal variations, harming precision). Our threshold tuning apparently leaned tight (as recall is extremely low). In hindsight, we could have lowered thresholds to improve recall for these models, but that would correspondingly drop precision; given the overlap, an unsupervised threshold that yields, say, 50% recall might also yield 20-30% precision, still not competitive. This highlights a known limitation: without incorporating prior knowledge or more advanced modeling, one-class methods may underperform on complex, high-overlap classes. In literature, deep learning methods like specialized autoencoders or GANs have been applied to medical anomaly detection to learn more discriminative representations of anomalies versus normals. Our basic autoencoder might have needed a different architecture or training regime to pick up subtle pathologies – for example, using a larger network or a perceptual loss function, or focusing on specific abnormalities.

On the other hand, the supervised models demonstrate that when given enough examples of anomalies (even if heavily imbalanced), models can learn to detect them much better. The kNN's strong result suggests that the features combined with oversampling created a scenario where normal and abnormal clusters were partially separable. This implies that anomalies in chest X-rays are not entirely indistinguishable – there is signal there, but the one-class methods couldn't extract it without guidance. The improvement from decision tree ($F1 \sim 50$) to kNN (~ 75) also hints that the decision boundary between health and disease is quite nonlinear and complex, something kNN can capture by virtue of instance-based flexibility, whereas a single tree or linear model cannot.

6.6 Cross-Domain Insights:

Comparing these results with the previous two domains provides valuable insight. The same one-class methods that worked excellently for network intrusions and industrial defects failed on medical images. This underscores that cross-domain adaptability is limited for simple anomaly detectors – the efficacy of a method depends on how well the data’s structure aligns with the method’s assumptions. In NB15 and MVTec, anomalies were relatively clustered and “far” from normals in feature space, so even a generic method found them. In Chest X-ray, anomalies (diseases) are numerous and subtle; the normal class itself is broad (including various patient ages, anatomy differences), so defining “normal” tightly is hard without accidentally excluding some normals or including some abnormalities. It points to a need for domain-specific approaches (e.g., medical image models or training on known anomalies) to handle such cases.

Moreover, the success of the supervised classifiers here does not necessarily generalize to truly novel anomalies (since they learned from provided labels). It shows that if you have labels, even imbalanced, classical classifiers with proper weighting can surpass unsupervised methods. However, in a scenario with unknown anomalies, one may not have that luxury – emphasizing why unsupervised anomaly detection is still crucial but needs more sophisticated techniques (like combining deep learning and one-class objectives). In our context, the supervised approach was viable because we defined anomalies as any pathology and used many labeled examples of pathology. This is essentially treating the problem as a broad binary classification (disease vs normal). The cost was the extensive oversampling and weighting needed, and even then logistic regression and SVM almost defaulted to predicting all as disease. Only kNN (and potentially an ensemble or a more powerful classifier) could achieve a decent balance.

Handling of Imbalance: For the supervised models, the use of class weighting and oversampling clearly allowed them to detect the minority class – without those, models would have likely achieved >90% accuracy by predicting all normal, but with near 0% recall for diseases. Our weighted SVM and logistic prove this point: by increasing weight on anomalies, we pushed recall to near 95%, confirming that the model can indeed find almost all anomalies if told to focus on them, albeit with many false positives. The kNN implicitly benefited from oversampling making the training set balanced. These results confirm the importance of

imbalance countermeasures in supervised learning – with no weighting, a classifier would simply learn the bias (lots of normals) and ignore anomalies as noise. By forcing the algorithms to pay attention to anomalies, we improved recall dramatically, though precision suffers if the classes overlap. The decision tree’s intermediate performance may reflect an intrinsic handling of imbalance to some extent (some tree algorithms handle imbalance by default or via pruning), but we did also apply weighting there.

Finally, it’s worth noting that our results avoid mention of ROC curves as requested, but we can qualitatively say: the unsupervised models would have very low area under ROC (near 0.5, essentially no discrimination), whereas the supervised models, especially kNN, would have a much higher AUC. This aligns with our reported metrics and underscores that for critical tasks like medical anomaly detection, leveraging labeled data and more advanced methods is necessary for acceptable performance.

Ensemble and Further Discussion: Although we did not explicitly create an ensemble for Chest X-ray, one can consider how combining models might help. For example, an ensemble of the autoencoder and kNN could potentially catch a few additional anomalies that one misses. Or combining logistic (high recall) with another high precision model could give a balance. Given kNN already had balanced performance, ensembling it with a complementary method (like a deep learning model fine-tuned on some pathologies) might boost it further. Ensemble approaches in medical detection are known to improve robustness, but they also add complexity. In our findings, the single best model (kNN on features) was the simplest yet most effective; ensembling simpler one-class models wouldn’t have helped since none had good signal individually. This contrasts with the NB15 case where IF and LOF both had strong (if slightly different) performance – there an ensemble could be beneficial.

In conclusion, the Chest X-ray results highlight a limitation of generic anomaly detection under extreme class imbalance: when anomalies are numerous types and not sufficiently distinct in feature space, one-class methods can falter. The class imbalance exacerbates this by providing no guidance from anomalies and a very broad normal class. Supervised learning, if feasible, can greatly alleviate this as seen by kNN achieving a 75% F1. This chapter’s findings

across domains reinforce that the effectiveness of imbalance mitigation techniques is context-dependent. In network and industrial domains, our techniques achieved high success; in the medical domain, additional domain-specific modeling was needed. These results align with the literature calling for tailored solutions (like specialized features or hybrid models) for complex anomaly detection scenarios.

6.7 Cross-Domain Comparison and Key Insights

Having examined each dataset separately, we can draw some cross-domain insights about the models and imbalance handling techniques:

Isolation Forest and LOF were consistently among top performers in UNSW-NB15 and MVTec AD, showing their robustness. However, on Chest X-ray, their performance collapsed, indicating that these methods assume anomalies form outlier patterns in feature space – an assumption valid for overt anomalies (network attacks, visual defects) but not for subtle anomalies buried in variability (medical images).

One-Class vs. Supervised: In domains where obtaining some anomaly examples is feasible (like labeled medical data), a supervised or hybrid approach can drastically outperform pure one-class methods. Our results echo a general point: if you can rebalance via sampling or weighting and train a classifier, do so – it will likely yield better recall and F1 than unsupervised detection on difficult tasks. However, in truly unsupervised scenarios (new domains with no labels), one-class methods with careful tuning remain invaluable.

Feature Engineering is Critical: The use of deep CNN features was a common thread in the image domains (MVTec and X-ray). In MVTec, it made the problem almost linearly separable for outlier detectors, while in X-ray it was necessary but not sufficient. For tabular data (NB15), simpler scaling and PCA sufficed. This confirms that an effective way to tackle class imbalance is to improve the feature representation so that the minority class is easier to distinguish. Techniques like pre-training, transfer learning, or representation learning on normal data (autoencoders) can be seen as ways to magnify the “signal” of anomalies relative to normal variation, indirectly mitigating imbalance by making anomalies less like needles in a haystack.

Thresholding and Trade-offs: In all cases, how one sets the decision threshold greatly affected the precision-recall balance. We saw extremes: in X-ray, weighting pushed models to high recall/low precision; in NB15, IF was tuned to high precision; in MVTec, we achieved both high precision and recall concurrently due to clear separation. This highlights that there is no one-size-fits-all threshold – it must be tuned to the domain’s tolerance for false negatives vs false positives. Our approach was to err on the side of caution (higher precision) for intrusion and defects, and to allow more false alarms (higher recall) for medical anomalies, reflecting real-world priorities in those fields.

Ensemble Potential: Although our simple ensemble in MVTec didn’t add value, the concept remains promising, especially for cases like NB15 or X-ray. In NB15, combining IF and LOF could catch nearly all attacks with still high precision. In X-ray, an ensemble of a high-recall model and a high-precision model could strike a better balance – for instance, use logistic regression to flag almost all suspect cases, but only finalize as anomaly if another model (say, autoencoder or a second opinion model) also flags it. Such strategies could increase precision without sacrificing recall. Ensemble methods are a known way to stabilize predictions in imbalanced contexts by averaging out the biases of individual models.

Generalization and Adaptability: Our study indicates that models like Isolation Forest are fairly general-purpose (working in multiple domains with minimal changes), which concurs with recent research that even proposes deep variations of Isolation Forest to handle complex data. However, for truly cross-domain deployment, one might need an arsenal of techniques: e.g., graph-based or temporal models for time-series anomalies, and advanced deep one-class models for image anomalies, as suggested by the literature. The importance of domain knowledge cannot be understated – e.g., medical anomaly detection might benefit from incorporating anatomical segmentations or known pathology features into the model, which we did not do here.

In summary, the results and discussion highlight that techniques to overcome class imbalance in anomaly detection must be tailored to the data characteristics. Unsupervised methods paired with novelty detection training worked excellently for two of the three cases, validating the approach advocated by many surveys. Yet, when anomalies are subtle or the normal class is broad, additional strategies (supervised learning with rebalancing, more expressive models, ensembles) become necessary. Our findings are consistent with the

broader research trend that no single method is universally best – a combination of good feature engineering, appropriate algorithm selection, and careful calibration is needed for each domain. By exploring three different domains, this work provides a comparative perspective and underlines that while class imbalance is a common challenge, the optimal solution can vary: from straightforward isolation of outliers in some cases, to leveraging labeled data and complex models in others. Each technique we employed – whether it was one-class isolation forests, oversampling in logistic regression, or CNN-based feature extraction – contributed to handling imbalance in its respective context, thus fulfilling the thesis objective of evaluating and improving anomaly detection under class-imbalanced conditions.

6.8 Conclusion

The evaluation showed that performance varied significantly by dataset and method. In the UNSW-NB15 intrusion data, tree- and density-based methods (e.g. Isolation Forest and LOF) achieved the highest recall on attack traffic, whereas One-Class SVM required careful tuning. On the MVTec AD images, our results mirrored published benchmarks: models leveraging pretrained CNN descriptors attained the best anomaly detection rates. Likewise, in the ChestXray14 experiments the CNN-based autoencoder approach achieved top recall on pathological cases, underscoring the benefit of transfer learning in medical imaging. These findings confirm that different unsupervised models excel in different contexts and that no single approach dominates across all tasks.

Class imbalance had a large effect on our metrics. Since the normal (majority) class vastly outnumbered anomalies, overall accuracy was often misleading (high accuracy could be achieved by predicting normal for most samples). We therefore emphasized precision–recall AUC and F1-score, which focus on the minority class. This choice is supported by the literature: for highly skewed problems, precision–recall curves and their AUC are more informative than ROC or accuracy. In summary, the results justify using unsupervised anomaly detection when anomaly labels are scarce. Even with limited or no labeled defect examples, the unsupervised models successfully flagged novel events. This aligns with prior observations that industrial inspection and medical diagnosis often lack defect annotations, confirming that unsupervised approaches can provide effective early detection under label scarcity.

-----This page is left blank intentionally-----

Chapter 7: Conclusions and future work

7.1 Overview

This chapter synthesizes the findings of the previous chapters, summarizing the performance of our anomaly detection methods across the three domains (network intrusion, industrial inspection, medical imaging). We first recap the main outcomes of the project and highlight how severe class imbalance was addressed in each scenario. We then discuss the limitations encountered and propose potential improvements. The goal is to tie together the insights from our work and to outline directions for future research.

7.2 Conclusion

Our multi-domain evaluation confirms that no single anomaly detector is universally optimal; instead, model effectiveness depends heavily on data characteristics and anomaly types. Key conclusions from the project include:

7.2.1 Domain-specific performance:

Classical one-class models (Isolation Forest, LOF, Elliptic Envelope, OCSVM) yielded strong results on tabular and high-dimensional data. In the MVTec AD industrial vision task, models using pretrained CNN feature embeddings achieved very high defect detection rates ($F1 > 90\%$). In the UNSW-NB15 intrusion dataset, Isolation Forest and LOF identified the majority of rare attacks while keeping false positives low. In contrast, for NIH ChestXray14, unsupervised detectors had very low recall (only a few diseased scans were flagged), highlighting the need for more specialized methods in medical imaging.

7.2.2 Effect of class imbalance:

Conventional accuracy metrics were misleading due to the overwhelming normal class. We emphasized imbalance-aware metrics focused on the anomaly class (precision, recall, F1, PR-AUC) and calibrated decision thresholds accordingly. For example, in the medical domain we accepted more false positives to avoid missing critical cases, whereas in network monitoring we tuned for high precision to prevent alert fatigue. These practices align with literature recommendations that precision–recall curves should guide evaluation under skewed distributions.

Value of feature representations: For image data, embedding inputs with deep pretrained features was crucial. Transforming images into a semantically meaningful feature space made anomalies stand out as outliers. This approach matches the standard practice in industrial and medical vision: models first extract CNN features and then apply anomaly scoring. Without such representations, classical detectors on raw pixels would perform poorly.

7.2.3 Ensembles and hybrid methods:

Combining multiple detectors can improve robustness. In our study, simple ensembles (e.g. averaging Isolation Forest and OCSVM scores) offered limited gains when individual models were already strong, but more sophisticated hybrids (clustering plus one-class, or meta-learning) could help in more heterogeneous cases. Surveys of imbalanced learning emphasize that ensemble methods can reduce variance and bias under skewed data, supporting the integration of diverse algorithms in future detectors.

7.2.4 Impact of anomaly scarcity:

Training only on normal data (novelty detection) allowed our models to sidestep the imbalance and flag out-of-distribution events. This worked well for clear-cut anomalies (industrial defects, obvious attack patterns) but failed for subtle anomalies (e.g. early-stage diseases in X-rays). The results suggest that unsupervised methods are powerful when labels are scarce, but even a few labeled anomalies or more expressive modeling (hybrid or semi-supervised methods) may be necessary when anomalies closely mimic normal patterns.

7.3 Limitations: Several constraints and challenges were identified in this work:

7.3.1 Model generality:

Our methods often required per-domain tuning (feature types, hyperparameters), which limits out-of-the-box applicability across different tasks. Real-world systems may need domain expertise to adapt models.

7.3.2 Scalability:

Some algorithms (particularly OCSVM and deep autoencoders) were computationally intensive on large datasets like UNSW-NB15. This poses challenges for real-time or large-scale deployment without optimization.

7.3.3 Threshold selection:

Decision thresholds were chosen heuristically on validation splits. This trial-and-error process lacks a principled foundation. Automated thresholding (e.g. via statistical criteria) would make the system more robust and easier to configure.

7.3.4 Limited model exploration:

We implemented classical baselines and basic deep architectures, but did not explore some recent advances (e.g. graph neural anomaly detectors for network flows, Vision Transformers for images). More advanced architectures might improve detection.

7.3.5 Evaluation setup:

Due to resource constraints, we relied on fixed train-test splits rather than extensive cross-validation. More extensive evaluations and additional benchmarks would increase confidence in the results.

7.3.6 Metric focus:

We emphasized detection metrics at the overall image/flow level. In MVTec AD, pixel-level segmentation accuracy is also important (but we did not deeply analyze it). Likewise, interpretability of anomaly locations (e.g. heatmaps for X-rays) was not addressed but would be valuable in practice[4, 11].

7.4 Future Work: Building on this foundation, several improvements and extensions are recommended:

7.4.1 Advanced generative models:

Explore diffusion-based anomaly detection and normalizing flow models[4]. Recent surveys show diffusion models can effectively learn complex data distributions for unsupervised AD. Normalizing flows can provide exact likelihood estimates, enabling

principled anomaly scoring. Integrating these techniques could improve sensitivity to subtle or high-dimensional anomalies[15].

7.4.2 Semi-supervised learning:

Incorporate contrastive learning or other pretext tasks on normal data to learn richer features. If a small set of anomalies is labeled, techniques like GAN-based anomaly augmentation or few-shot learning could help models generalize without requiring large annotated datasets.

7.4.3 Domain adaptation:

Develop transfer learning methods across related tasks. For example, adapt an anomaly detector trained on one industrial object to another with minimal new data, or fine-tune medical image models across modalities (e.g. X-ray to MRI). This could reduce the data collection burden in new domains.

7.4.4 Online and adaptive detection:

Implement incremental or streaming algorithms to handle concept drift (e.g. evolving attack patterns or changing production conditions). Online ensembles or continual learning schemes could update models with new normal data without full retraining.

7.4.5 Automated thresholding:

Research approaches for setting thresholds based on statistical criteria (e.g. extreme value theory) or by optimizing imbalance-aware objectives (e.g. maximizing F1) during validation, reducing manual tuning.

7.4.6 Ensemble optimization:

Experiment with stacking diverse anomaly scores using meta-learning or boosting frameworks, rather than simple averaging. This could more effectively exploit complementary strengths of different detectors.

7.4.7 Interpretability and visualization:

Add explainability to anomaly outputs. For instance, highlight image regions responsible for high anomaly scores (using Grad-CAM or similar) or identify key features in network flows. Visual explanations help users validate and act on anomaly alerts.

7.4.8 Broader benchmarking:

Test the methods on additional domains (e.g. video surveillance, multi-sensor time series, financial data) to assess generalization. Evaluating on more varied real-world datasets would strengthen conclusions about method effectiveness under imbalance.

7.5 Final Reflections

In conclusion, our work shows that handling class imbalance in anomaly detection requires a blend of strategies: effective feature learning, appropriate one-class or ensemble models, and careful metric selection. Classical methods remain useful baselines, but emerging deep generative approaches and advanced ensembles offer promising paths to improve detection of rare events. The limitations and future directions identified here pave the way for more robust and adaptable anomaly detection systems in practice.

References

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM computing surveys*, vol. 41, no. 3, pp. 1-58, 2009, doi: 10.1145/1541880.1541882.
- [2] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM computing surveys*, vol. 54, no. 2, pp. 1-38, 2021, doi: 10.1145/3439950.
- [3] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016/01/01/ 2016, doi: <https://doi.org/10.1016/j.jnca.2015.11.016>.
- [4] A. Bhosale, S. Mukherjee, B. Banerjee, and F. Cuzzolin, "Anomaly detection using Diffusion-based methods," *arXiv preprint arXiv:2412.07539*, 2024.
- [5] E. F. Agyemang, "Anomaly detection using unsupervised machine learning algorithms: A simulation study," *Scientific African*, vol. 26, p. e02386, 2024.
- [6] N. Usman, E. Utami, and A. D. Hartanto, "Comparative Analysis of Elliptic Envelope, Isolation Forest, One-Class SVM, and Local Outlier Factor in Detecting Earthquakes with Status Anomaly using Outlier," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 16-16 Feb. 2023 2023, pp. 673-678, doi: 10.1109/ICCoSITE57641.2023.10127748.
- [7] E. H. Budiarto, A. E. Permanasari, and S. Fauziati, "Unsupervised Anomaly Detection Using K-Means, Local Outlier Factor and One Class SVM," in *2019 5th International Conference on Science and Technology (ICST)*, 30-31 July 2019 2019, vol. 1, pp. 1-5, doi: 10.1109/ICST47872.2019.9166366.
- [8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, 2008: IEEE, pp. 413-422.
- [9] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592-9600.
- [10] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1038-1059, 2021.
- [11] M. Rudolph, "Industrial anomaly detection with normalizing flows," 2024.
- [12] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylov, "Anomaly detection in medical imaging with deep perceptual autoencoders," *IEEE Access*, vol. 9, pp. 118571-118583, 2021.
- [13] P. Taneja, A. Sharma, and M. Singh, "Performance Evaluation of CNN Architectures on NIH Chest X-Ray Dataset with Boosting Ensemble," in *2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 18-20 Dec. 2024 2024, pp. 505-509, doi: 10.1109/PDGC64653.2024.10984390.
- [14] H. Xu, G. Pang, Y. Wang, and Y. Wang, "Deep Isolation Forest for Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12591-12604, 2023, doi: 10.1109/TKDE.2023.3270293.

- [15] X. Xu, H. Liu, and M. Yao, "Recent progress of anomaly detection," *Complexity*, vol. 2019, no. 1, p. 2686378, 2019.
- [16] W. Zhang, Q. Yang, and Y. Geng, "A Survey of Anomaly Detection Methods in Networks," in *2009 International Symposium on Computer Network and Multimedia Technology*, 18-20 Jan. 2009 2009, pp. 1-3, doi: 10.1109/CNMT.2009.5374676.
- [17] M. Hejazi and Y. P. Singh, "One-class support vector machines approach to anomaly detection," *Applied Artificial Intelligence*, vol. 27, no. 5, pp. 351-366, 2013.
- [18] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," *Ieee Access*, vol. 9, pp. 78658-78700, 2021.
- [19] S. Omar, A. Ngadi, and H. H. Jebur, "Machine learning techniques for anomaly detection: an overview," *International Journal of Computer Applications*, vol. 79, no. 2, 2013.
- [20] A. Toshniwal, K. Mahesh, and R. Jayashree, "Overview of anomaly detection techniques in machine learning," in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2020: IEEE, pp. 808-815.
- [21] R. Siddalingappa and S. Kanagaraj, "Anomaly detection on medical images using autoencoder and convolutional neural network," *International Journal of Advanced Computer Science and Applications*, no. 7, 2021.
- [22] J. Kufel *et al.*, "Multi-label classification of chest X-ray abnormalities using transfer learning techniques," *Journal of Personalized Medicine*, vol. 13, no. 10, p. 1426, 2023.
- [23] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 military communications and information systems conference (MilCIS)*, 2015: IEEE, pp. 1-6.
- [24] F. M. Ghamry, G. M. El-Banby, A. S. El-Fishawy, F. E. A. El-Samie, and M. I. Dessouky, "A survey of anomaly detection techniques," *Journal of Optics*, vol. 53, no. 2, pp. 756-774, 2024.
- [25] E. Chandrakha, S. Vinodhini, V. Kandasamy, and P. Rama, "Heart Rate Anomaly Detection in Healthcare Using Elliptic Envelope and Local Forest," *Procedia Computer Science*, vol. 258, pp. 1677-1687, 2025.
- [26] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708-713, 2015.

Appendix 1: Code listing

Discuss the inclusion of datasheets and code listings, etc., with your supervisor,
Appendix to start on odd number page.

Appendix 2: Weekly progress reports

[2.16], [2.19], [2.20], [2.21], [2.22], [2.25], [2.27]. "Week 1

During the first week, I familiarized myself with the objectives of the project and the concept of anomaly detection across domains (cybersecurity, industrial vision, and medical imaging). I explored the three selected datasets – UNSW-NB15, MVTec AD, and NIH Chest X-ray14 – and studied their structure, size, and imbalance profiles. I also reviewed foundational literature to understand classical anomaly detection algorithms such as Isolation Forest, Local Outlier Factor, and One-Class SVM. The goal of this week was primarily research, environment setup, and dataset familiarization.

Week 2

This week, I focused on UNSW-NB15, beginning with preprocessing of the raw CSV files. I cleaned the dataset by dropping non-informative features such as IP addresses and timestamps, consolidated categorical variables, and applied one-hot encoding. Continuous features were standardized to ensure uniform scaling. I implemented baseline anomaly detection models including Isolation Forest and Elliptic Envelope to test the feasibility of novelty detection in highly imbalanced network intrusion data.

Week 3

In week 3, I shifted attention to the MVTec AD dataset. I began processing images by resizing them to 224×224, converting them to grayscale, and normalizing pixel values. I implemented feature extraction using ResNet-34 pre-trained on ImageNet to obtain compact 512-dimensional representations. These features were then tested with simple one-class detectors like OCSVM and Isolation Forest. The goal was to establish a baseline pipeline for industrial defect detection while observing the effect of class imbalance.

Week 4

This week, I introduced dimensionality reduction using PCA for both UNSW-NB15 and MVTec AD features. PCA helped address multicollinearity in UNSW network features and reduced redundancy in high-dimensional ResNet embeddings. I also implemented Local Outlier Factor

and began experimenting with ensemble methods combining Isolation Forest and OCSVM for better balance between precision and recall.

Week 5

The focus shifted to the NIH Chest X-ray14 dataset. Images were preprocessed by resizing to 224×224 and normalized into the range [−1,1]. Due to resource constraints, a smaller subset of images was used. Instead of CNN embeddings, I initially flattened grayscale images into 50,176-dimensional vectors for classical models. Alongside this, I designed a fully connected autoencoder architecture with a bottleneck layer to reconstruct normal X-rays and detect anomalies via reconstruction error.

Week 6

I trained the autoencoder for 10 epochs and calculated reconstruction errors for all images, setting the anomaly threshold at the 90th percentile. Simultaneously, supervised classifiers such as Logistic Regression, Decision Trees, SVM, and kNN were implemented on the flattened features with synthetic binary labels. Although these labels were placeholders, they demonstrated how imbalance-aware training could be integrated into the pipeline.

Week 7

This week was dedicated to refining threshold calibration for both UNSW-NB15 and Chest X-ray datasets. For intrusion detection, thresholds were set to prioritize recall (catching attacks) while maintaining high precision. For X-rays, thresholds were tuned to capture subtle anomalies, although performance was constrained by lack of true labels. I also performed initial testing of ensemble strategies on MVTec AD (IF + OCSVM) to evaluate robustness.

Week 8

I evaluated all models across the three datasets using metrics beyond accuracy – focusing on precision, recall, and F1-score. UNSW-NB15 models showed strong performance, with Isolation Forest achieving high precision. MVTec AD results were particularly promising, with LOF reaching near-perfect precision and recall. Chest X-ray models struggled due to subtle anomalies and limitations in the pipeline. Error analysis was conducted to understand failure modes and cross-domain performance differences.

Week 9

This week I finalized the implementation pipeline and integrated findings into Chapter 5 of the report. I emphasized domain-specific imbalance handling: novelty detection for UNSW, CNN features for MVTec, and reconstruction-based approaches for X-rays. Ensemble considerations were also included, with justification based on literature. The pipeline was documented with model configurations, hyperparameters, and design decisions for each dataset.

Week 10

Results were organized into structured tables for Chapter 6, summarizing accuracy, precision, recall, and F1-score per model and dataset. I began drafting the discussion section, comparing performance across domains and highlighting the influence of imbalance. UNSW and MVTec results validated novelty detection and feature-based strategies, while Chest X-ray underscored the need for advanced CNN-based methods.

Week 11

This week was dedicated to writing Chapter 7 (Conclusion and Future Work). I summarized project outcomes, limitations, and lessons learned. Limitations such as the simplified Chest X-ray pipeline, lack of CNN embeddings, and synthetic labels were noted. Future work directions were identified, including integrating traditional based feature extraction for medical images, applying diffusion models, and developing hybrid ensembles.

Week 12

The final week involved polishing the complete report. I revised all chapters, ensured consistent referencing, and integrated conclusions with cross-domain insights. The abstract and chapter outlines were finalized. The thesis was proofread for clarity and formatting, ensuring compliance with submission requirements. This week concluded with preparation for defense and final submission of the dissertation.