National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Alleviating Data Scarcity In Industrial Machine Learning Applications

PHD Thesis

of

SPYROS-CHRISTOFOROS THEODOROPOULOS

Athens, June 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Alleviating Data Scarcity In Industrial Machine Learning Applications

PHD Thesis

of

SPYROS-CHRISTOFOROS THEODOROPOULOS

Advisory Committee : Panayiotis Tsanakas
Dimosthenis Kyriazis
Angelos Amditis

Approved by the examination committee on 5th June 2025.

.....	.....	.....
Panayiotis Tsanakas	Dimosthenis Kyriazis	Angelos Amditis
Professor, NTUA	Professor, Univ. of Piraeus	Reasearcher A', ICCS NTUA

.....	.....
Andreas-Georgios Stafylopatis	Athanasios Voulodimos
Professor, NTUA	Asst. Professor, NTUA

.....	.....
Vangelis Marinakis	Sotirios Xydis
Asst. Professor, NTUA	Asst. Professor, NTUA

Athens, June 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this PhD thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....

Spyros-Christoforos Theodoropoulos, PhD,
School of Electrical and Computer
Engineering,
National Technical University of Athens

5th June 2025

Abstract

Visual defect recognition and its manufacturing applications have been an upcoming topic in recent AI research as an integral part of the manufacturing process that is becoming increasingly automated with the advent of Industry 4.0 and Industry 5.0. While being a very beneficial solution to this problem, AI-driven Computer Vision Algorithms and Deep Neural Networks face several issues that may impede their adoption in practical real-life settings such as a manufacturing shop floor. For instance, defect datasets are often severely imbalanced and can be additionally burdened with separating classes of high visual similarity. Another issue arising during an AI classifier's continuous operation is the frequent lack of robustness to novel defects appearing for the first time. The aim of this thesis is to deal with such challenges by providing augmentations to AI solutions, either on the data or the model level, addressing real-life and benchmark scenarios from the domain of manufacturing.

The initial focus is Imbalanced Learning. Although various methods of data augmentation have been proposed to mitigate class imbalances, they often fail to cope with tinier minority classes or have fidelity issues with smaller defects while, at the same time, needing significant computational resources to train. Also, augmentation based on vector-based oversampling struggles to produce high-fidelity inputs and is hard to apply on custom CNN architectures, which often perform better for this type of problem. Our work presents an image-level oversampling method based on an instance-based image generator that can be applied to any CNN directly during the training process without increasing the order of training time required. It is based on identifying a small number of the most uncertain base samples close to the estimated class boundaries and using them as seeds for augmentation. The resulting images are of high visual quality preserving small class differences, and they also improve the classifier boundary leading to higher recall scores than other state-of-the-art approaches.

Aside from class imbalance, lack of real-world data as well as the strict safety constrains that need to be imposed to manufacturing AI deployments dictate the need for handling novel inputs. Such unanticipated inputs can pose a significant risk to cyber-physical applications as a resulting out-of-context decision could compromise the integrity of the production process. While recent Machine Learning methods can theoretically tackle this problem from different angles (e.g., open-set recognition, semi-supervised learning, intelligent data augmentation), applying them to a real-life setting with a small, imbalanced dataset and high inter-class similarity can be challenging. This work confronts such a use case aiming at the automation of the visual quality inspection of shaver shell brand prints from the electronics industry, which is characterized by data scarcity

and the existence of small local defects. To that end, we introduce a novel data augmentation approach based on the latent space manipulation of StyleGAN, where defect data is intentionally synthesized to simulate novel inputs that can help form a boundary of the model's knowledge. Our approach shows promising results compared to well-established open-set recognition and semi-supervised methods applied to the same problem, while its consistent performance across classifier embeddings indicates lower coupling to the final classifier.

The above mentioned method still requires enough data to train a GAN, which might not always be possible or cost-effective. Collecting more and more defect data is also often not a solution as defects occur rarely in production and the ramp-up time of the AI-driven quality inspector becomes significantly slower. To cope with smaller datasets we apply an innovative approach based on Neurosymbolic AI. Specifically, we use a Logic Tensor Network that expresses the outputs of an unsupervised out-of-distribution detector as symbolic rules and uses them to drive the training of a neural network classifier. The resulting algorithm shows improved results in comparison to other related methods, especially in terms of defect recall, meaning that few defects remain undetected even if completely novel. More specifically, it achieves similar or better recall scores than semi-supervised and unsupervised methods when handling novel defects, but significantly outperforms them in defects that were seen during training. Similarly, when compared to supervised methods, it maintains high performance on known defects but significantly improves on novel ones. These best-of-both-worlds results are illustrated through higher F1-scores in the majority of the test datasets of manufacturing products.

Keywords

Artificial Intelligence, Visual Quality Inspection, Smart Manufacturing, Deep Learning, Defect Recognition, Imbalanced Learning, Data Augmentation, Oversampling, Generative Adversarial Networks, Open-set Recognition, Neurosymbolic AI

Περίληψη

Η αναγνώριση οπτικών ελαττωμάτων όπως εφαρμόζεται στον κατασκευαστικό τομέα είναι ένα θέμα που απασχολεί την τρέχουσα έρευνα στο πεδίο της τεχνητής νοημοσύνης, καθώς αυτός αποτελεί αναπόσπαστο μέρος της διαδικασίας παραγωγής που αυτοματοποιείται ολοένα και περισσότερο με την εμφάνιση της Βιομηχανίας 4.0 και της Βιομηχανίας 5.0. Αν και είναι μια πολύ ευεργετική λύση, οι αλγόριθμοι όρασης υπολογιστών που βασίζονται στη Μηχανική Μάθηση και τα Βαθιά Νευρωνικά Δίκτυα αντιμετωπίζουν πολλά προβλήματα που μπορεί να εμποδίσουν την υιοθέτησή τους σε πρακτικές εφαρμογές, όπως σε μια γραμμή παραγωγής. Τα σύνολα δεδομένων που περιέχουν ελαττώματα δεν έχουν συνήθως ισορροπημένες κλάσεις και πάσχουν κατά τον διαχωρισμό μεταξύ κλάσεων υψηλής οπτικής ομοιότητας. Ένα άλλο ζήτημα που προκύπτει κατά τη συνεχή λειτουργία ενός ταξινομητή μηχανικής μάθησης είναι η έλλειψη ανθεκτικότητας σε νέα ελαττώματα που εμφανίζονται για πρώτη φορά. Ο στόχος αυτής της εργασίας είναι να αντιμετωπίσει τέτοιες προκλήσεις παρέχοντας επαυξήσεις στις λύσεις τεχνητής νοημοσύνης, είτε σε επίπεδο δεδομένων είτε σε επίπεδο μοντέλου, ώστε να μπορούν να ανταποκριθούν σε πραγματικές συνθήκες στον κατασκευαστικό τομέα.

Η αρχική εστίαση είναι στη Μη Ισορροπημένη Μάθηση. Παρόλο που έχουν προταθεί διάφορες μέθοδοι επαύξησης δεδομένων για τον μετριασμό των ανισορροπιών κλάσεων, συχνά αποτυγχάνουν σε ιδιαίτερα ολιγοπληθείς κατηγορίες ενώ, ταυτόχρονα, χρειάζονται σημαντικούς υπολογιστικούς πόρους για εκπαίδευση. Επίσης, η επαύξηση που βασίζεται σε υπερδειγματοληψία βάσει διανυσμάτων δυσκολεύεται να παράγει εισόδους υψηλής ευκρίνειας και είναι δύσκολο να εφαρμοστεί σε προσαρμοσμένες αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ), οι οποίες συχνά αποδίδουν καλύτερα για αυτόν τον τύπο προβλήματος. Η εργασία μας παρουσιάζει μια μέθοδο υπερδειγματοληψίας στο επίπεδο της εικόνας που μπορεί να εφαρμοστεί σε οποιοδήποτε ΣΝΔ απευθείας κατά τη διάρκεια της εκπαιδευτικής διαδικασίας χωρίς μεγάλη επιβάρυνση του απαιτούμενου χρόνου εκπαίδευσης. Ξεκινά με τον εντοπισμό ενός μικρού αριθμού αβέβαιων δειγμάτων κοντά στα εκτιμώμενα όρια μεταξύ δύο κλάσεων και βασίζει τη σύνθεση νέων δεδομένων σε αυτά. Οι εικόνες που προκύπτουν είναι υψηλής οπτικής ποιότητας διατηρώντας μικρές διαφορές μεταξύ των κατηγορίων και χρησιμεύουν στο να βελτιώσουν τα όρια του ταξινομητή, οδηγώντας σε υψηλότερη ανάκληση σε σχέση με άλλες προσεγγίσεις.

Εκτός από την ανισορροπία κλάσεων, η αδυναμία συλλογής πολλών δεδομένων, καθώς και οι αυστηροί περιορισμοί ασφαλείας για τα κυβερνο-φυσικά συστήματα, υπαγορεύουν τον αποτελεσματικό χειρισμό καινοφανών εισόδων. Τέτοιες απρόσμενες είσοδοι μπορεί να αποτελέσουν σημαντικό κίνδυνο, καθώς μια λανθασμένη απόκριση σε αυτές θα μπορούσε να βλάψει την ακεραιότητα της διαδικασίας παραγωγής. Ενώ οι πρόσφατες μέθοδοι Μηχανικής

Μάθησης μπορούν θεωρητικά να αντιμετωπίσουν αυτό το πρόβλημα από διαφορετικές οπτικές γωνίες (π.χ. αναγνώριση ανοιχτού συνόλου, ημι-εποπτευόμενη μάθηση, έξυπνη επαύξηση δεδομένων), εφαρμόζοντάς τες σε ένα πραγματικό περιβάλλον με ένα μικρό, μη ισορροπημένο σύνολο δεδομένων και υψηλή ομοιότητα μεταξύ των κλάσεων αποτελεί πρόκληση. Η παρούσα εργασία αντιμετωπίζει μια τέτοια περίπτωση που αφορά στην αυτοματοποίηση της οπτικής ποιοτικής επιθεώρησης εκτυπώσεων λογοτύπων σε κελύφη ξυριστικών μηχανών από τη βιομηχανία ηλεκτρονικών και χαρακτηρίζεται από σπανιότητα δεδομένων και ύπαρξη μικρών τοπικών ελαττωμάτων. Για το σκοπό αυτό, εισάγεται μια νέα προσέγγιση επαύξησης δεδομένων που βασίζεται στον χειρισμό του λανθάνοντος χώρου του StyleGAN, με αποτέλεσμα τα δεδομένα ελαττωμάτων να συντίθενται σκόπιμα για την προσομοίωση νέων εισόδων με στόχο τον σχηματισμό ενός ορίου γύρω από την γνωστή κατανομή εκπαίδευσης του μοντέλου. Η προσέγγισή μας δείχνει υποσχόμενα αποτελέσματα σε σύγκριση με τις καθιερωμένες μεθόδους αναγνώρισης ανοιχτού συνόλου και τις ημι-εποπτευόμενες μεθόδους που εφαρμόζονται στο ίδιο πρόβλημα, ενώ η σταθερή απόδοσή της σε διαφορετικούς χώρους χαρακτηριστικών υποδεικνύει χαμηλότερη σύζευξη με τη διαδικασία εξαγωγής τους.

Η παραπάνω μέθοδος εξακολουθεί να απαιτεί αρκετά δεδομένα για την εκπαίδευση του StyleGAN, κάτι που μπορεί να μην είναι πάντα δυνατό ή οικονομικά αποδοτικό. Η συλλογή ολοένα και περισσότερων δεδομένων ελαττωμάτων επίσης συχνά δεν είναι λύση, καθώς τα ελαττώματα εμφανίζονται σπάνια στην παραγωγή και ο χρόνος εγκατάστασης του ευφυούς επιθεωρητή ποιότητας γίνεται σημαντικά πιο αργός. Για να αντιμετωπίσουμε μικρότερα σύνολα δεδομένων εφαρμόζουμε μια καινοτόμο προσέγγιση που βασίζεται στη Νευροσυμβολική Τεχνητή Νοημοσύνη. Συγκεκριμένα, χρησιμοποιούμε ένα Δίκτυο Λογικού Τανυστή που εκφράζει τις εξόδους ενός μη-επιβλεπόμενου ανιχνευτή ανωμαλιών ως συμβολικούς κανόνες με στόχο στη συνέχεια να καθοδηγήσει την εκπαίδευση ενός νευρωνικού δικτύου. Ο αλγόριθμος που προκύπτει δείχνει βελτιωμένα αποτελέσματα σε σύγκριση με άλλες σχετικές μεθόδους, ειδικά όσον αφορά στην ανάκληση ελαττωμάτων, πράγμα που σημαίνει ότι λίγα ελαττώματα παραμένουν μη ανιχνεύσιμα ακόμη και αν είναι εντελώς καινοφανή. Πιο συγκεκριμένα, επιτυγχάνει παρόμοια ή καλύτερα αποτελέσματα ανάκλησης από τις ημι-εποπτευόμενες και μη εποπτευόμενες μεθόδους κατά τον χειρισμό νέων ελαττωμάτων, αλλά παράλληλα υπερέχει σημαντικά σε ελαττώματα που παρατηρήθηκαν κατά τη διάρκεια της εκπαίδευσης. Ομοίως, σε σύγκριση με τις εποπτευόμενες μεθόδους, διατηρεί υψηλή απόδοση σε γνωστά ελαττώματα, ενώ ταυτόχρονα δείχνει μεγάλη βελτίωση στα καινοφανή. Τα αποτελέσματα αυτά γίνονται ορατά μέσω των υψηλότερων βαθμολογιών F1 στην πλειονότητα των συνόλων δεδομένων αξιολόγησης.

Λέξεις Κλειδιά

Τεχνητή Νοημοσύνη, Οπτικός Έλεγχος Ποιότητας, Ευφυής Βιομηχανοποίηση, Βαθιά Μάθηση, Αναγνώριση Ελαττωμάτων, Μη Ισορροπημένη Μάθηση, Επαύξηση Δεδομένων, Υπερδειγματοληψία, Παραγωγικά Αντιπαραθετικά Δίκτυα, Αναγνώριση Ανοιχτού Συνόλου, Νευροσυμβολική Τεχνητή Νοημοσύνη

Acknowledgements

Upon reaching the end of this journey, it is important to reflect on all the difficulties and successes, smaller and greater, that made this invaluable experience possible for me. I can say with little doubt that the knowledge and skills acquired during this four-and-a-half-year effort have significantly expanded my outlook not only academically and professionally, but regarding life in general. Of course, none of this would have been possible without the significant support of several people around me.

First and foremost, I would like to thank my supervisor, Prof. Panayiotis Tsanakas, for his confidence in me and for providing me with the academic freedom, continuous guidance, and prompt support that played a fundamental role in motivating and sustaining all my research efforts. I would also like to thank Prof. Dimosthenis Kyriazis, who had an active part as member of my advisory committee, for his continuous support, his academic and professional advice, as well as for exposing me to different challenges, opportunities and environments that made my journey richer and more fruitful. I am also grateful to Dr. Angelos Amditis for his valuable feedback as a member of my advisory committee.

The results of this effort are largely owed to the ideas and influences derived from a lively, collaborative, and demanding research environment, for which I would like to express my graduate to all my collaborators and co-authors on different projects and papers and especially Dr. George Makridis.

Lastly and most importantly, I am deeply grateful to my family and friends who supported me in my choice and coped alongside me through every difficulty.

Athens, June 2025

Spyros-Christoforos Theodoropoulos

Εκτεταμένη Περίληψη

Η παρούσα διατριβή επικεντρώνεται στις εφαρμογές μηχανικής μάθησης σε βιομηχανικά περιβάλλοντα και σε συγκεκριμένα πρακτικά προβλήματα που προκύπτουν λόγω της δυσκολίας συλλογής δεδομένων εκπαίδευσης όπως η Ανισορροπία Κλάσεων και η Εμφάνιση Καινοφανών Δεδομένων. Το τεχνολογικό και ερευνητικό περιβάλλον στο οποίο εξετάζονται αυτά τα προβλήματα είναι αυτό της Βιομηχανίας 5.0. Ο όρος αυτός προκύπτει ως μια επέκταση των μέχρι τώρα τεσσάρων Βιομηχανικών Επαναστάσεων και συγκεκριμένα της 4ης Βιομηχανικής Επανάστασης (Industry 4.0) η οποία χαρακτηρίζεται από τεχνολογίες όπως το Διαδίκτυο των Αντικειμένων, τα Κυβερνο-φυσικά Συστήματα, τα Ψηφιακά Δίδυμα, τα Μεγάλα Δεδομένα η Τεχνητή Νοημοσύνη κ.α. Σε αυτό το υπόβαθρο η Βιομηχανία 5.0 στοχεύει στον συνδυασμό των ανθρώπινων δυνατοτήτων με αυτών των ευφυών μηχανών μέσω συστημάτων προσομοίωσης και συνεργασίας Ανθρώπου-Υπολογιστή. [1]

Πιο συγκεκριμένα, η παρούσα έρευνα επικεντρώνεται στον Αυτόματο Έλεγψο Ποιότητας Βιομηχανικών προϊόντων μέσω τεχνικών Μηχανικής Μάθησης για Υπολογιστική Όραση. Στα πλαίσια της Ποιότητας 4.0 (μέρους της Βιομηχανίας 4.0) στόχος είναι η δημιουργία αυτοελεγχόμενων συστημάτων που μπορούν να μετρήσουν αυτόματα την ποιότητα της εξόδου τους και να αποφασίζουν αυτόνομα για την αποδοχή ή απόρριψή της. Η Βαθιά Μάθηση λόγω της προσαρμοστικότητάς της (π.χ. σε οπτικές αλλαγές στην κλίμακα ή την περιστροφή της εικόνας) έχει βοηθήσει πολύ σε αυτό, αλλά ταυτόχρονα απαιτεί μεγάλο όγκο δεδομένων εκπαίδευσης και δεν είναι ευσταθής σε δείγματα εκτός της κατανομής εκπαίδευσης. Μια λύση που διερευνάται στα πλάισια της Βιομηχανίας 5.0 είναι η ανάπτυξη συστημάτων συνεργασίας Ανθρώπου-Μηχανής όπου η ανθρώπινη νοημοσύνη και εμπειρία θα αναπληρώνει τα μειονεκτήματα της τεχνητής.

Κατά την πορεία της παρούσας έρευνας στον Αυτόματο Έλεγχο Ποιότητας βιομηχανικών προϊόντων μέσω τεχνικών Βαθιάς Μάθησης διαπιστώθηκαν τρείς κύριες προκλήσεις, οι οποίες αποτελούν και το επίκεντρο αυτής της εργασίας:

1. ***Η ανεπάρκεια δεδομένων εκπαίδευσης***, η οποία γίνεται ιδιαίτερα αισθητή σε προϊόντα με σφάλματα. Αυτό συμβαίνει διότι τα σφάλματα εμφανίζονται σπάνια στις γραμμές παραγωγής σε σχέση με τα άρτια προϊόντα οδηγώντας σε *ανισορροπία μεταξύ των δύο κλάσεων*.

2. ***Η μεγάλη οπτική ομοιότητα μεταξύ άρτιων και ελαττωματικών προϊόντων*** η οποία δυσχεραίνει σημαντικά την ικανότητα διάκρισης των ταξινομητών.

3. ***Η εμφάνιση καινοφανών ελαττωμάτων*** κατά τη συνεχή λειτουργία ενός ήδη εκπαιδευμένου αλγορίθμου μπορεί να οδηγήσει σε λανθασμένη ταξινόμηση των προϊόντων ώς άρτια.

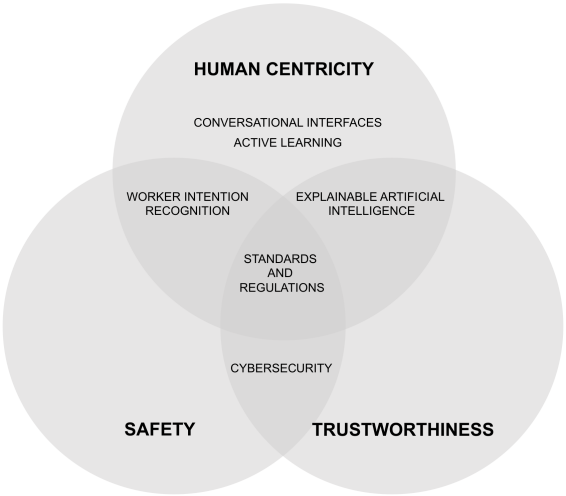


Figure 1. *Βασικές Αρχές και Τεχνολογίες της Βιομηχανίας 5.0 [2]*

Για την αντιμετώπιση της ανισορροπίας κλάσεων αναπτύχθηκε μέθοδος για την επαύξηση των δεδομένων εκπαίδευσης που ανήκουν σε μειονοτικές κλάσεις. Η σύνθεση των δεδομένων έγινε με τεχνικές κατεύθυνσης Παραγωγικών Αντιπαραθετικών Δικτύων (ΠΑΔ), με στόχο να γίνει υπερδειγματοληψία παραδειγμάτων στα οποία οι προβλέψεις του ταξινομητή παρουσιάζουν χαμηλή αξιοπιστία. Η επαύξηση τέτοιων δεδομένων δύναται να παρέχει μεγαλύτερο όφελος στη διαδικασία εκπαίδευσης. [3]

Για τον χειρισμό των καινοφανών εισόδων διερευνήθηκαν παρόμοιες τεχνικές, αυτήν την φορά με στόχο την σύνθεση οριακών παραδειγμάτων με χρήση StyleGAN. Παρότι η διαδικασία παραγωγής δεδομένων που αναπτύχθηκε ξεκινά από τις κατανομές εκπαίδευσης, τα οριακά δεδομένα, χάρη στη γενικευσιμότητα του StyleGAN, παράγονται στα άκρα των κατανομών αυτών και δημιουργούν ένα όριο μεταξύ γνωστών και καινοφανών εισόδων. [4] Σαν επέκταση χρησιμοποιήθηκαν τεχνικές Νευροσυμβολικής τεχνητής νοημοσύνης με στόχο την αύξηση της ανθεκτικότητας όταν βρίσκονται διαθέσιμα ακόμη λιγότερα δεδομένα εκπαίδευσης [5].

Όσον αφορά στην ομοιότητα μεταξύ άρτιων και ελαττωματικών προϊόντων, αυτή συνυπολογίστηκε σε όλες τις παραπάνω μεθόδους. Συγκεκριμένα, χρησιμοποιήθηκαν ΠΑΔ με δυνατότητες πολύ λεπτομερούς σύνθεσης εικόνων, ενώ, όπου ο όγκος των δεδομένων το επέτρεπε, έγινε εκπαίδευση των τελικών ταξινομητών απευθείας στο πρόβλημα χωρίς χρήση μεταφοράς μάθησης από προεκπαιδευμένα δίκτυα.

Διάταξη της Γραμμής Παραγωγής και Δεδομένα

Το στάδιο του ελέγχου ποιότητας της γραμμής παραγωγής τροποποιείται με την τοποθέτηση κάμερας η οποία φωτογραφίζει τα προϊόντα με τη βοήθεια συστήματος που στοχεύει στην τοπική ομογενοποίηση της φωτεινότητας, για να αποφευχθούν σκιές ή θάμπωμα. Πολλοί τέτοιοι σταθμοί μπορούν να τοποθετηθούν σε κοντινή απόσταση με έναν άνθρωποχειριστή υπαύθυνο για όλους. Ενώ ο αλγόριθμος μηχανικής μάθησης έχει ευθύνη για την αρχική ταξινόμηση των προϊόντων, σε περίπτωση που ανιχνέυσει πιθανότητα ελαττώματος, το προϊόν καταλήγει στον υπεύθυνο χειριστή για την τελική απόφαση - αν όντως είναι ελατ-

(a) *Good* **(b)** *Double Print* **(c)** *Interrupted*

Figure 2. *Δείγματα από τα δεδομένα της PCL BV*

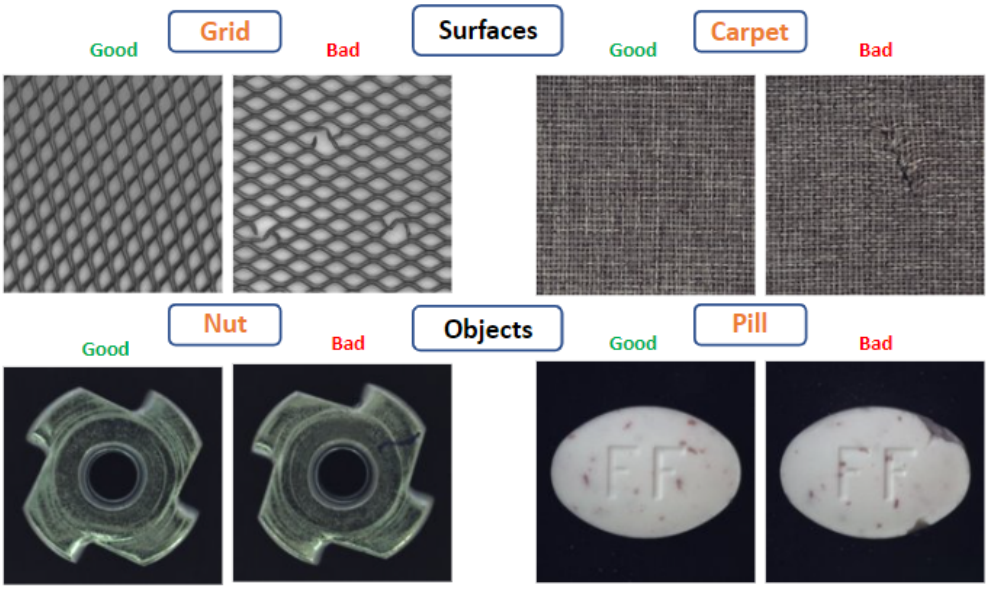


Figure 3. *Δείγματα από τα δεδομένα του MVTEC-AD*

τωματικό ή μπορεί να συνεχίσει στη διαδικασία παραγωγής ως άρτιο. Δεδομένου πως ένα προϊόν έχει χαρακτηριστεί ως άρτιο αυτό εξέρχεται από το σύστημα χωρίς ανθρώπινο έλεγχο (πέρα από τυχαία δειγματοληψία). Για αυτόν τον λόγο πρέπει το σύστημα να είναι ιδιαίτερα αυστηρό με την ταξινόμηση στην άρτια κλάση. Παράλληλα δεν πρέπει να είναι τόσο αυστηρό ώστε να υπερφορτώνει τον χειριστή με άρτια προϊόντα που λανθασμένα έχουν χαρακτηρισθεί ελαττωματικά.

Για την ανάπτυξη και αξιολόγηση των μεθόδων που ακολουθούν χρησιμοποιήθηκαν δύο σύνολα δεδομένων: το πρώτο προέρχεται από την Philips Consumer Lifestyle BV και απεικονίζει εκτυπωμένα λογότυπα της εταιρείας σε βραχίονες ξυριστικών μηχανών και το δεύτερο είναι το ευρύτερα χρησιμοποιούμενο στην ερευνητική κοινότητα MVTEC, το οποίο απεικονίζει διαφορετικά προϊόντα σε ξεχωριστά υποσύνολα δεδομένων. Από αυτά επιλέχθηκαν όσα είχαν ανισορροπίες κλάσεων και μεγάλη ομοιότητα άρτιων και ελαττωματικών προϊόντων.

Αντιμετώπιση Ανισορροπίας Κλάσεων

Πλέον τα βαθιά συνελικτικά δίκτυα (ΒΣΔ) είναι η επικρατέστερη μέθοδος στη βιβλιογραφία για αυτόματο ποιοτικό ελέγχο καθότι συνδυάζουν τα εξής πλεονεκτήματα:

1. Επιτυγχάνουν υψηλή ακρίβεια καθώς μαθαίνουν τα χαρακτηριστικά εκπαίδευσης δυναμικά με βάση τα δεδομένα.

2. Δεν απαιτούν ειδικές γνώσεις του προβλήματος με αποτέλεσμα να απλοποιούν τον σχεδιασμό των διαδικασιών (προ-)επεξεργασίας των δεδομένων.

3. Μπορούν πιο εύκολα να αναπροσαρμοστούν σε παρόμοια προβλήματα με το αρχικό (π.χ. ένα προϊόν με διαφορετική διακόσμηση).

4. Είναι ανθεκτικά σε οπτικούς μετασχηματισμούς όπως οι αλλαγές θέσης και κλίμακας.

5. Παρέχουν δυνατότητα μεταφοράς γνώσης από πολυπληθή σε μικρότερα σύνολα δεδομένων.

Παρόλα αυτά, αν εξαιρέσει κανείς την μεταφορά δεδομένων, η εκπαίδευση αυτών των δικτύων απαιτεί συνήθως 10^3 έως 10^4 παραδείγματα, ενώ είναι ιδιαίτερα ευαίσθητη στις ανισορροπίες του αριθμού παραδειγμάτων μεταξύ των κλάσεων.

Για την αντιμετώπιση της ανισορροπίας δεδομένων έχουν αναπτυχθεί τόσο πιο παραδοσιακές τεχνικές για διανυσματικά δεδομένα (SMOTE, Borderline-SMOTE, ADASYN), όσο και πιο σύγχρονες τεχνικές επαύξησης μέσω σύνθεσης ολόκληρων εικόνων από ΠΑΔ. Η επαύξηση δεδομένων μέσω ΠΑΔ μπορεί να πραγματοποιηθεί είτε απευθείας (π.χ. με χρήση των Wasserstein GAN, DCGAN ή StyleGAN) είτε μέσω προσπαθειών καθοδήγησης των εξόδων του ΠΑΔ (π.χ. μέσω ενισχυτικής μάθησης - Actor-Critic GAN) για να παράγει πιο χρήσιμες εξόδους. Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει μια νέα μέθοδος, η DeepSMOTE, βασισμένη σε αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή που αναπαράγει την διαδικασία SMOTE, αλλά στο επίπεδο της εικόνας, παράγοντας εικόνες από γραμμικές παρεμβολές μεταξύ των γνωστών εικόνων εισόδου [6].

Κατά τη διαρκεία της έρευνας διαπιστώθηκε ότι η επίδοση δικτύων εκπαιδευμένων αποκλειστικά στο πρόβλημα της αναγνώρισης ελαττωμάτων (χωρίς μεταφορά μάθησης) είναι πιο αποτελεσματική, κατά συνέπεια θα βοηθούσε περισσότερο μια τεχνική υπερδειγματοληψίας στο επίπεδο των εικόνων. Λαμβάνοντας υπόψιν τις ιδιαιτερότητες του προβλήματος αναπτύχθηκε μια βελτίωση του DeepSMOTE για το συγκεκριμένο πρόβλημα που στοχέυει στην υπερδειγματοληψία δεδομένων για την ταξινόμηση των οποίων ο αλγόριθμος είναι αβέβαιος με στόχο την βελτιστοποίηση της ανάκλησης του τελικού ταξινομητή. Η σύνθεση των δεδομένων γίνεται εφικτή μέσω της προασαρμογής του BigGAN για λειτουργία σε μικρά σύνολα δεδομένων [7].

Μέθοδος

Σύνθεση Δεδομένων

Λόγω του μικρού σε μέγεθος και άνισα κατανεμημένου σε κλάσεις συνόλου δεδομένων, αποφεύχθηκε η εκπαίδευση ΠΑΔ εξαρχής και χρησιμοποιήθηκε η μέθοδος των Noguchi et al. [7] η οποία προσαρμόζει ένα μοντέλο BigGAN προεκπαιδευμένο στο ImageNet. Συγκεκριμένα για κάθε είσοδο I συγκεκριμένες παράμετροι του ΠΑΔ προσαρμόζονται ώστε να παραχθεί μια παραλλαγμένη μορφή I_z δεδομένου τυχαίου διανύσματος εισόδου z. Οι

παράμετροι του BigGAN που προσαρμόζονται είναι μόνο αυτές του παράγοντα κλίμακας (scale) και της μετατόπισης (shift) στα στρώματα νευρώνων κανονικοποίησης (batch normalization layers). Διαισθητικά αυτό ισοδυναμεί με την επιλογή χαρακτηριστικών χρήσιμων για τα δεδομένα εκπαίδευσης από ένα υπερσύνολο χαρακτηριστικών που έχει μαθευτεί από το Imagenet.

Καθότι τα δεδομένα που παράγονται με αυτόν τον τρόπο μπορεί να μην απεικονίζουν ικανοποιητικά μικροσκόπικα ελαττώματα στις εικόνες, χρησιμοποιούνται δύο επιπρόσθετοι μηχανισμοί. Αρχικά μια συνάρτηση *TilePermutations*, εμπνευσμένη από τους Satoshi et al.[8] παράγει υβριδικές εικόνες, χωρίζοντας τις αρχικές και συνθετικές εικόνες σε μη επικαλυπτόμενα τμήματα και ανασυνδυάζοντάς τα τυχαία. Τέλος ακολουθεί ένα βήμα φιλτραρίσματος όπου μόνο ένα υποσύνολο εικόνων μεγέθους n_{aug} επιλέγεται, αποτελούμενο από τις υβριδικές εικόνες που είναι κοντινότερες στις αρχικές.

Υπολογισμός Αξιοπιστίας Προβλέψεων

Με σκοπό να βελτιστοποιηθεί η υπαρδειγματοληψία, επιλέχθηκε η παραγωγή συνθετικών δεδομένων να είναι βασισμένη σε εικόνες για την ταξινόμηση των οποίων δεν υπάρχει μεγάλος βαθμός σιγουριάς με βάση τον υποκείμενο αλγοριθμο. Η υπόθεση είναι ότι αυτή η πιο στοχευμένη υπερδειγματοληψία, εμπνευσμένη από την βελτιστοποίηση Borderline-SMOTE του SMOTE αλλά στο επίπεδο των εικόνων, θα προσφέρει μεγαλύτερες βελτιώσεις στην τελική διαδικασία μάθησης.

Παρότι στη βιβλιογραφία υπάρχουν πολλοί τρόποι για να ποσοτικοποιηθεί η αξιοπιστία μιας πρόβλεψης που αντιστοιχεί σε μία είσοδο, χρησιμοποιήθηκε η μέθοδος των Elsayed et al. [9] ή οποία δεν απαιτεί μεταβολές ούτε στη διαδικασία μάθησης, ούτε στην αρχιτεκτονική του δικτύου.

Σύμφωνα με αυτήν το όριο απόφασης μεταξύ δύο κλάσεων i και j ορίζεται ως το σύνολο εισόδων για το οποίο ο (ψεύδο-)βαθμός αξιοπιστίας προβλέψεων για ταξινόμηση σε καθεμία κλάση (δηλαδή η έξοδος του στρώματος softmax του δικτύου f) είναι ίσος:

$$D_{\{i,j\}} = \{x \mid f_i(x) = f_j(x)\} \tag{1}$$

Η απόσταση ενός σημείου x από το όριο απόφασής ορίζεται τότε ως η l_p νόρμα της μικρότερης μετατόπισης που πρέπει να υποστεί το σημείο ώστε να υπάρχει ισότητα των (ψεύδο-)βαθμών αξιοπιστίας:

$$d_{f,x,\{i,j\}} = \min_{\delta} \|\delta\|_p \mid f_i(x+\delta) = f_j(x+\delta) \tag{2}$$

Καθότι το πρόβλημα βελτιστοποίησης είναι μη αναλυτικά επιλύσιμο για μη γραμμική f, χρησιμόποιείται το ανάπτυγμα Taylor πρώτου βαθμού για να γραμμικοποιηθεί η f οδηγώντας στην ακόλουθη προσέγγιση της απόστασης από το όριο:

$$\hat{d}_{f,x,\{i,j\}} = \frac{|\, f_i(x) - f_j(x) \,|}{\|\nabla_x f_i(x) - \nabla_x f_j(x)\|_q} \tag{3}$$

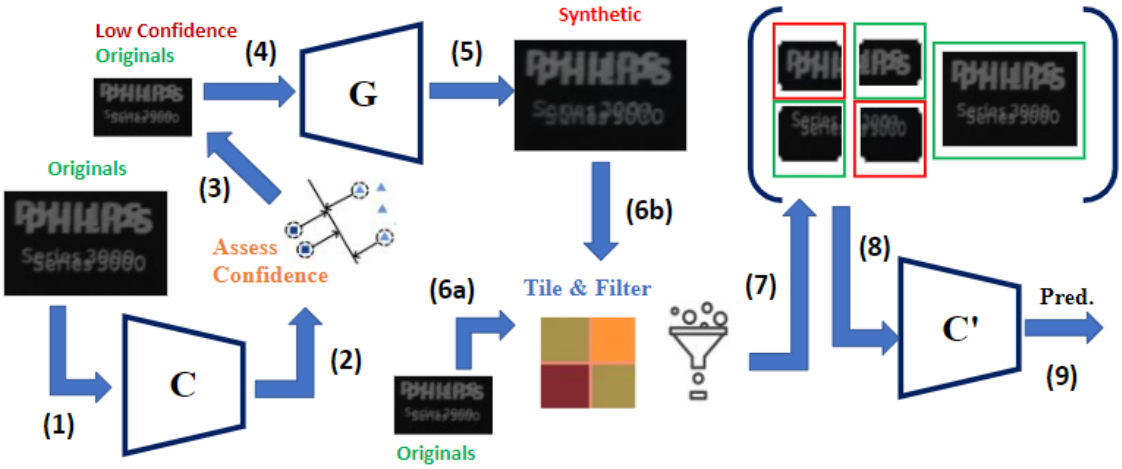


Figure 4. *Αρχιτεκτονικό διάγραμμα της μεθόδου επαύξησης δεδομένων και της διαδικασίας εκπαίδευσης του τελικού ταξινομητή.*

Επαύξηση Δεδομένων σε Πραγματικό Χρόνο

Συνοψίζοντας τα παραπάνω βήματα προκύπτει η τελική μέθοδος επαύξησης δεδομένων σε πραγματικό χρόνο που φαίνεται και στην Εικόνα 4.

Βήμα#1: Ο ταξινομητής C εκπαιδεύεται για καθορισμένο αριθμό n_p εποχών και η απόσταση από το όριο διαχωρισμού άρτιων και μη άρτιων εικόνων προσεγγίζεται για κάθε εικόνα εκπαίδευσης σύμφωνα με την Εξ.2.2.

Βήμα#2: Οι k_{top} εικόνες εισόδου με την μικρότερη απόσταση από το όριο διαχωρισμού επιλέγονται και χρησιμοποιούνται ως γεννήτορες για την σύνθεση δεδομένων.

Βήμα#3: Υπολογίζεται ο αριθμός των παραγόμενων εικόνων που αντιστοιχεί σε κάθε εικόνα γεννήτορα n_{aug} έτσι ώστε τα δεδομένα για την τελική εκπαίδευση να είναι ισορροπημένα ανάμεσα σε άρτια και μη άρτια και εκτελείται η διαδικασία για την παραγωγή των δεδομένων.

Βήμα#4: Ο προεκπαιδευμένος ταξινομητής C εκπαιδεύεται επιπλέον για n εποχές στο επαυξημένο σύνολο δεδομένων με σκοπό να μάθει ένα βελτιωμένο όριο διαχωρισμού.

Αποτελέσματα

Στόχος της προτεινόμενης προσέγγισης είναι να περιοριστεί όσο το δυνατόν ο αριθμός ελαττωματικών προϊόντων που εσφαλμένα ταξινομούνται ως άρτια. Για τον λόγο αυτό η σημαντικότερη μετρική είναι αυτή της ανάκλησης από την πλευρά των ελαττωματικών δεδομένων. Δεδομένου ενός ταξινομητή C, δεδομένων αξιολόγησης X, με ετικέτα εκπαίδευσης $l(x)$, όπου οι ετικέτες των ελαττωματικών κλάσεων δίνονται στο $L_d = \{double\ print, interrupted\}$, αυτή ορίζεται ως εξής:

$$BinaryRecall = \frac{|x \in X : C(x) \in L_d \wedge l(x) \in L_d|}{|x \in X : l(x) \in L_d|}$$

Η συγκεκριμένη μετρική επηρεάζεται επίσης λιγότερο από την ανισορροπία δεδομένων σε σχέση π.χ. με την ακρίβεια (accuracy), παρόλα αυτά για να υπάρχει και έλεγχος των ψευδώς θετικών προβλέψεων, εξιολογούνται και οι μετρικές AUROC, Precision και F1.

Συνολικά, για κάθε συγκρινόμενη μέθοδο εκτελέστηκαν 30 μετρήσεις με χρήση 5-πλης διασταυρωμένης επικύρωσης, ενώ παράλληλα έγινε και βελτιστοποίηση των υπερ-παραμέτρων κάθε μεθόδου. Οι μετρήσεις που παρουσιάζονται είναι ο μέσος όρος των παραπάνω μαζί με τα 95% διαστήματα εμπιστοσύνης.

Method	**Bin. Recall %**	**AUROC %**	**Precision %**	**F1 %**
Resnet50	85.85 ± 1.50	98.85 ± 0.12	94.41 ± 3.27	89.59 ± 1.27
Resnet50+SMOTE	95.84 ± 0.52	98.87 ± 0.13	84.53 ± 3.01	89.61 ± 1.57
Resnet50+ADASYN	95.49 ± 0.99	99.07 ± 0.11	85.14 ± 3.45	89.67 ± 1.69
Custom CNN	95.84 ± 0.39	99.20 ± 0.19	97.53 ± 0.81	96.67 ± 0.56
Custom CNN+LW	96.07 ± 0.39	99.09 ± 0.19	98.34 ± 0.33	**97.19 ± 0.43**
StyleGAN	91.20 ± 2.20	99.01 ± 0.14	**99.17 ± 0.41**	94.95 ± 1.38
DeepSMOTE	93.58 ± 1.07	99.23 ± 0.15	96.93 ± 0.80	95.22 ± 0.87
Ours	**97.27 ± 0.76**	**99.34 ± 0.07**	96.82 ± 1.27	97.03 ± 0.98

Table 1. *Αξιολόγηση στα δεδομένα εικόνων ξυριστικών μηχανών της PCL BV*

Μια αρχική ενδιαφέρουσα παρατήρηση σχετικά με τον Πίνακα 1 είναι ότι το ρηχό συνελικτικό δίκτυο που εκπαιδεύεται εξαρχής στο πρόβλημα έχει καλύτερη επίδοση απέναντι σε δίκτυα με μεταφορά μάθησης ακόμα και όταν χρησιμοποιούν τεχνικές υπαρδειγματοληψίας διανυσμάτων (SMOTE, ADASYN κλπ.). Τεχνικές επαύξησης στο επίπεδο της εικόνας όπως το StyleGAN και το DeepSMOTE δεν κατάφεραν να βελτιώσουν την επίδοση του ρηχού συνελικτικού δικτύου, πιθανότατα, όπως φαίνεται και στην Εικόνα 5, λόγω της αδυναμίας τους να παράγουν αρκετά λεπτομερείς εικόνες ελαττωματικών προϊόντων. Η προτεινόμενη μέθοδος, χάρη στην εισαγωγή των επιπλέον βημάτων σχετικά με την επιλογή σημαντικών εικόνων και την χρήση τους στη σύνθεση εικόνων επαύξησης, κατάφερε τόσο να παράγει εικόνες υψηλής ευκρίνειας, όσο και να βελτιώσει σημαντικά την ανάκληση του τελικού ταξινομητή. Παρότι η προτεινόμενη μέθοδος υπολείπεται σε άλλες μετρικές όπως η F1, αυτό δεν είναι πρόβλημα καθότι η θυσία είναι αρκετά μικρή για την επίτευξη υψηλής ανάκλησης.

Και στο MVTEC-AD (Πίνακας 2) η προτεινόμενη μέθοδος επέτυχε την υψηλότερη ανάκληση σε όλες της περιπτώσεις. Δυστυχώς, λόγω και του μικρού πληθυσμού αυτών των συνόλων δεδομένων, η στατιστική σημαντικότητα του αποτελέσματος δεν ήταν δυνατόν να επιτευχθεί σε όλες τις περιπτώσεις. Ιδιαίτερα στα δεδομένα τύπου Grid οι διακυμάνσεις μεταξύ κάθε πειραματικής εκτέλεσης ήταν σημαντικές. Παρόλα αυτά, στα σύνολα δεδομένων Metal Nut, Pill και Carpet η διαφορά ήταν είτε στατιστικά σημαντική είτε πολύ κοντά σε αυτό. Σε αντίθεση με τα πειράματα του Πίνακα 1 εδώ επιτυγχάνεται βελτίωση και στις μετρικές Precision και F1. Πιθανότατα και πάλι λόγω του μικρού πληθυσμού των δεδομένων, υπάρχουν μεγαλύτερα οφέλη από την επαύξηση των δεδομένων. Το μικρό μέγεθος των συνόλων δεδομένων επίσης δεν επέτρεψε την ικανοποιητική εκπαίδευση των μεθόδων StyleGAN και DeepSMOTE που χρειάζονται περισσότερα δεδομένα για να μπορούν να συνθέσουν νέες εικόνες.

Η βασική υπόθεση της προτεινόμενης μεθόδου είναι ότι χάρη στην επαύξηση δεδομένων βασισμένη στις εικόνες που βρίσκονται πιο κοντά στο όριο ταξινόμησης θα μετατοπίσει αυτό το όριο ώστε να επιτυγχάνεται καλύτερη ανάκληση. Αυτό φαίνεται τόσο από τα αποτελέσ-

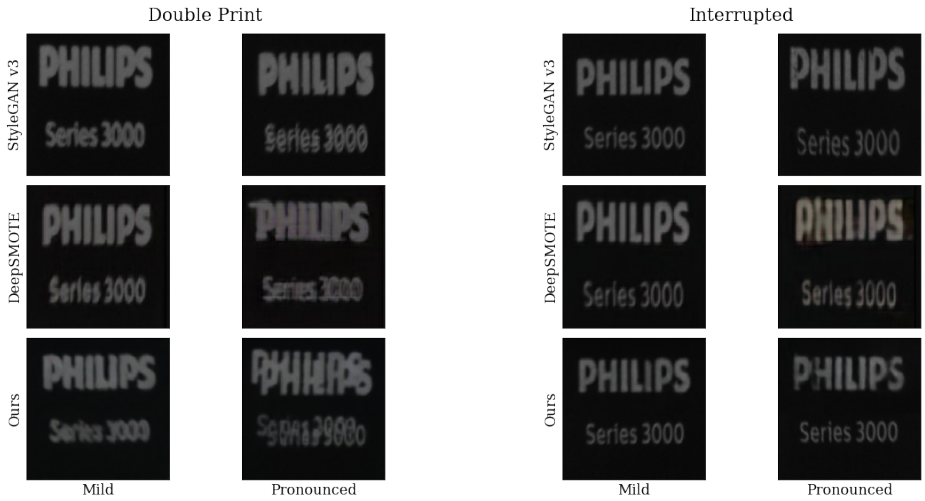


Figure 5. *Συνθετικές εικόνες που παράχθηκαν από τις υπο εξέταση μεθόδους*

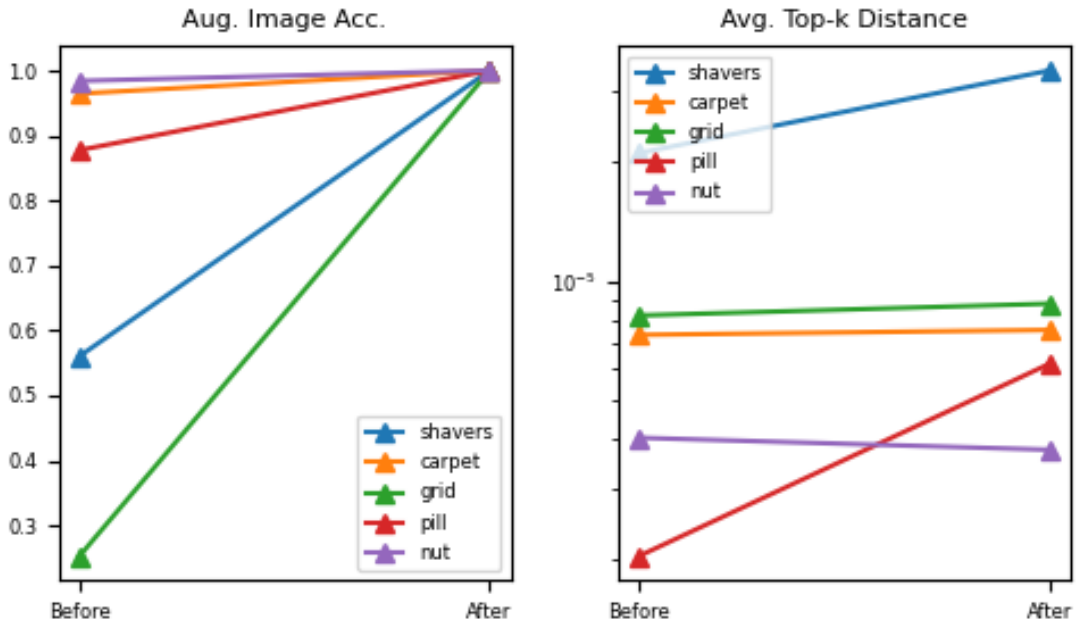


Figure 6. *Ακρίβεια ταξινομητή στα οριακά δεδομένα που επιλέγονται ως γεννήτορες για την σύνθεση των δεδομένων επαύξησης (Αριστερά). Οι k μικρότερες αποστάσεις προς το όριο διαχωρισμού των κλάσεων πριν και μετά την εκπαίδευση στο επαυξημένο σύνολο δεδομένων για k=15 (Δεξιά)*

Dataset	Method	Binary Recall %	AUROC %	Precision %	F1 %
Grid	Resnet50	30.30 ± 5.59	73.29 ± 3.48	42.4 ± 5.57	34.36 ± 5.95
	Resnet50 + SMOTE	35.60 ± 5.71	74.67 ± 4.34	48.27 ± 3.38	43.55 ± 3.48
	Resnet50 + ADASYN	42.57 ± 8.38	74.26 ± 4.04	42.93 ± 4.59	35.93 ± 6.68
	Custom CNN	70.90 ± 10.93	90.71 ± 5.49	80.23 ± 9.97	74.50 ± 10.50
	Custom CNN + LW	69.24 ± 12.25	89.80 ± 6.09	75.38 ± 12.3	71.55 ± 12.15
	Ours	**71.21 ± 9.92**	**91.22 ± 5.12**	**91.43 ± 6.86**	**78.45 ± 8.36**
Carpet	Resnet50	81.89 ± 3.70	97.07 ± 0.41	87.80 ± 3.33	84.20 ± 2.27
	Resnet50 + SMOTE	88.69 ± 1.53	97.21 ± 0.46	79.56 ± 2.11	83.66 ± 0.94
	Resnet50 + ADASYN	84.18 ± 2.71	97.25 ± 0.45	83.96 ± 3.78	83.42 ± 1.28
	Custom CNN	87.77 ± 7.62	98.94 ± 0.49	89.73 ± 1.23	87.48 ± 4.75
	Custom CNN + LW	91.11 ± 6.06	98.90 ± 0.51	88.02 ± 1.78	88.92 ± 3.72
	Ours	**92.22 ± 3.32**	**99.86 ± 0.11**	**92 ± 1.60**	**91.9 ± 1.97**
Metal Nut	Resnet50	84.03 ± 3.46	96.90 ± 0.79	95.33 ± 1.70	88.99 ± 1.97
	Resnet50 + SMOTE	88.30 ± 3.71	97.32 ± 0.51	90.32 ± 1.33	89.07 ± 1.62
	Resnet50 + ADASYN	84.09 ± 3.71	97.01 ± 0.72	95.38 ± 1.63	89.02 ± 2.09
	Custom CNN	82.92 ± 5.36	97.49 ± 1.15	98.33 ± 1.33	89.55 ±3.87
	Custom CNN + LW	82.92 ± 5.36	97.49 ± 1.15	98.33 ± 1.33	89.55 ± 3.87
	Ours	**92.63 ± 3.15**	**98.32 ± 1.22**	**98.75 ± 1.00**	**95.49 ± 2.12**
Pill	Resnet50	71.52 ± 6.29	92.70 ± 1.63	84.84 ± 1.63	76.65 ± 4.29
	Resnet50 + SMOTE	90.02 ± 2.62	91.76 ± 1.82	60.7 ± 1.57	72.34 ± 1.41
	Resnet50 + ADASYN	78.62 ± 4.16	91.87 ± 1.70	82.29 ± 1.54	80.08 ± 2.41
	Custom CNN	88.71 ± 2.18	98.35 ± 0.60	93.48 ± 2.03	90.94 ± 1.76
	Custom CNN + LW	88.71 ± 2.18	98.35 ± 0.60	93.48 ± 2.03	90.94 ± 1.76
	Ours	**92.29 ± 3.79**	**98.80 ± 0.58**	**96.25 ± 1.63**	**94.11 ± 2.68**

Table 2. *Αξιολόγηση στα σύνολα εικόνων από το MVTEC-AD*

ματα των Πινάκων 1 και 2, όσο και από το γράφημα σύγκρισης των αποστάσεων από το όριο στην Εικόνα 6. Ενδιαφέρον παρουσιάζει το γεγονός ότι οι ελάχιστες αποστάσεις είναι μικρές σχετικά με την μεγάλη διαστατικότητα του χώρου, κάτι που πιθανόν να υποδεικνύει πυκνή συγκέντρωση των απεικονίσεων του δικτύου κοντά στα όρια. Φυσικά η μετατόπιση που αναφέρουμε γίνεται από τη μη άρτια κλάση προς την άρτια, με αποτέλεσμα συχνά να συνεπάγεται κάποια θυσία στην ακρίβεια της άρτιας κλάσης. Τέλος, ένας περιορισμός της μεθόδου είναι ότι σε περίπτωση που οι κλάσεις είναι εύκολα διαχωρίσιμες και τα όρια μεταξύ των κλάσεων πιο αραιοκατοικημένα, δεν θα παρέχει κάποια βελτίωση.

Χειρισμός Καινοφανών Δεδομένων

Όπως έχουν δείξει οι Hendrycks et al. [10], συχνά τα Βαθιά Συνελικτικά Δίκτυα είναι ιδιαίτερα ευαίσθητα σε σφάλματα πρόβλεψης όταν πρόκειται για καινοφανή δεδομένα ακόμα και όταν αυτά προέρχονται από μικρής έκτασης αλλαγές σε δεδομένα που υπάρχουν στο σύνολο εκπαίδευσης. Ένα σχετικό παράδειγμα βλέπουμε στην Εικόνα 7. Ακόμα και εξειδικευμένες στην ανίχνευση οπτικά ορατών ελαττωμάτων ημι-επιβλεπόμενες ή μη επιβλεπόμενες τεχνικές αντιμετωπίζουν προβλήματα όταν π.χ. ένα γνωστό ελάττωμα βρίσκεται σε μια εντελώς καινούργια θέση σε μια εικόνα εισόδου [11].

Στην βιβλιογραφία υπάρχουν διάφορες κατηγορίες μεθόδων για τον χειρισμό καινοφανών δεδομένων:

- *Ταξινομητές μίας κλάσης* **(One-class SVMs, Isolation Forests, Local Outlier Factor)** οι οποίοι προσπαθούν να σχηματίσουν ένα όριο που να ξεχωρίζει τα δεδομένα μίας

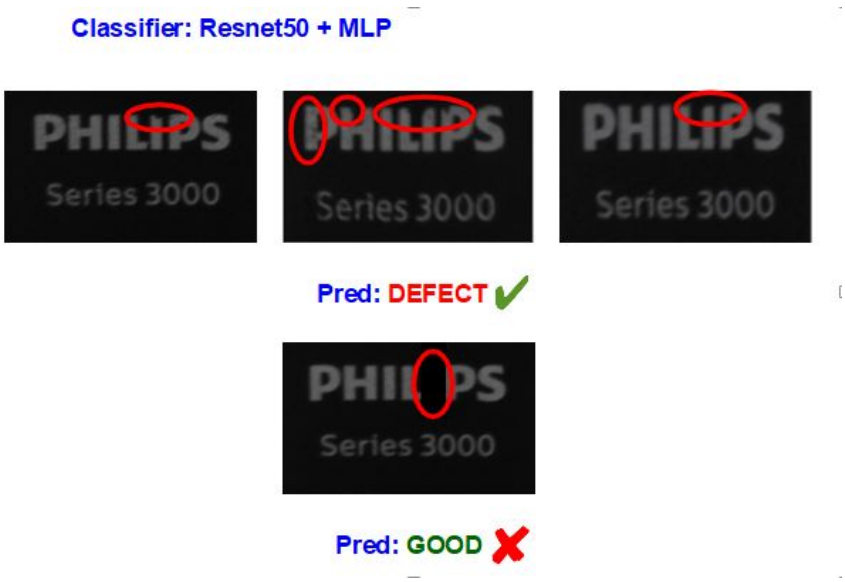


Figure 7. *Προβλήματα ανθεκτικότητας σε νέες εισόδους. Η πάνω σειρά περιέχει διακοπτόμενες σε μικρό βαθμό εικόνες που αντίστοιχές τους υπάρχουν στα δεδομένα εκπαίδευσης και ταξινομούνται σωστά. Αντίθετα εικόνες με μεγαλύτερης έκτασης ελαττώματα που είναι καινοφανείς δεν ταξινομούνται σωστά (κάτω σειρά).*

κλασης (στην περίπτωσή μας αυτήν με τα άρτια προϊόντα) από όλα τα άλλα πιθανά δεδομένα εισόδου.

- *Ημι-επιβλεπόμενες μέθοδοι* **(GANomaly, DFKDE, DFM)** οι οποίες εκπαιδεύονται μόνο με δεδομένα της άρτιας κλάσης και μαθαίνουν να μοντελοποιούν μόνο αυτή και χωρίζουν τα δεδομένα σε άρτια και μη ανάλογα με κάποια μετρική απόστασης από την κλάση που έχουν μάθει.

- *Μέθοδοι επαύξησης δεδομένων* **(OSRCI, OpenGAN)** οι οποίες προσπαθούν να συνθέσουν (συνήθως με χρήση κάποιου ΠΑΔ ή κωδικοποιητή-αποκωδικοποιητή) καινοφανή δεδομένα τα οποία παρέχονται στην διαδικασία εκπαίδευσης μέσω μιας επιπρόσθετης κλάσης.

- *Μέθοδοι αναγνώρισης "ανοιχτού συνόλου"* **(W-SVM, PISVM)** χρησιμοποιούν την Θεωρία Ακραίων Τιμών για να μοντελοποιήσουν ακριβέστερα τα άκρα των κατανομών των γνωστών κλάσεων και να τις ξεχωρίσουν από το "ανοιχτό σύνολο" που αποτελείται από εισόδους άγνωστες κατά την εκπαίδευση.

Μέθοδος

Η μέθοδος που αναπτύχθηκε είναι και αυτή στα πλαίσια της επαύξησης δεδομένων. Σε σύγκριση με μεθόδους όπως οι OSRCI και OpenGAN, χρησιμοποιείται μία νεότερη, μεγαλύτερης ευκρίνειας και πιο γενική αρχιτεκτονική, το StyleGAN v3, το οποίο επίσης παρέχει αυξημένες δυνατότητες καθοδήγησης και ελέγχου της σύνθεσης εικόνων. Αυτή η δυνατότητα καθίσταται εκμεταλλεύσιμη μέσω της Σημασιολογικής Παραγοντοποίησης (Semantic Factorization - SeFa) η οποία ανακαλύπτει κατευθύνσεις στους λανθάνοντες χώρους

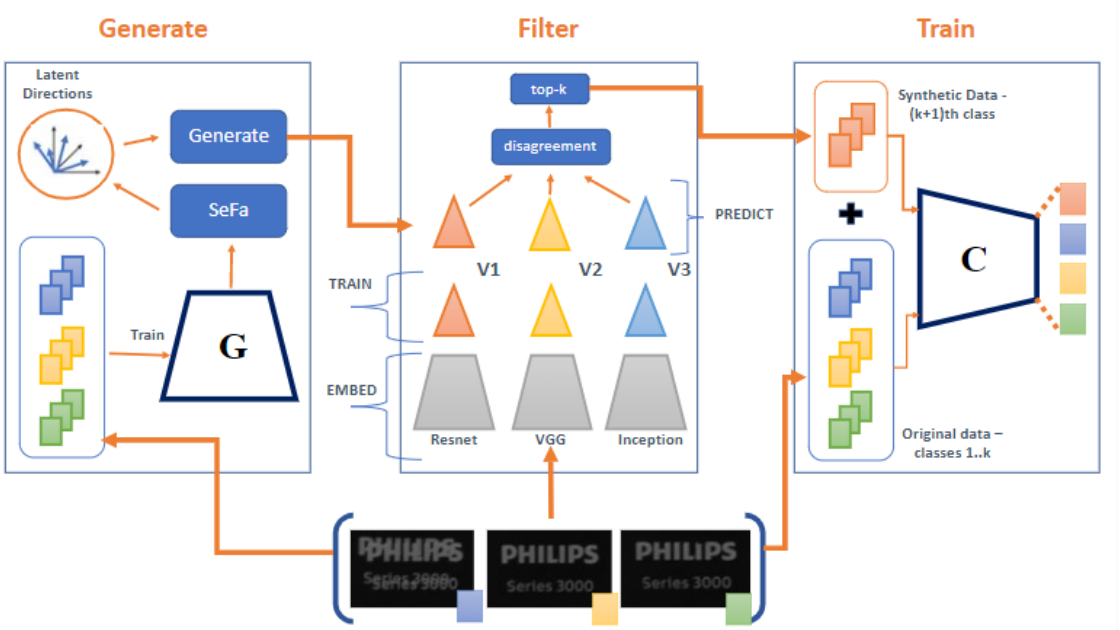


Figure 8. *Αρχιτεκτονικό διάγραμμα της διαδικασίας εκπαίδευσης με παραγωγή συνθετικών δεδομένων μέσω Semantic Factorization και φιλτράρισμά τους μέσω ψηφοφορίας*

εισόδου του StyleGAN κατά μήκος των οποίων μεταβάλλονται σημασιολογικά πλούσια στοιχεία της εικόνας. Με τη βοήθεια αυτής και ενός σταδίου φιλτραρίσματος βασισμένο σε ψηφοφορία, στόχος είναι να παραχθούν εικόνες που να οριοθετούν τις κατανομές εισόδου και να τις ξεχωρίζουν από το "ανοιχτό σύνολο". Αυτές οι οριακές εικόνες προστίθενται στα δεδομένα εκπαίδευσης ως μια επιπρόσθετη κλάση (βλ. Εικόνα 8).

Σημασιολογική Παραγοντοποίηση στο StyleGAN

Η Σημασιολογική Παραγοντοποίηση (SeFa) [12] προσπαθεί να επιτύχει μια αναλυτική λύση του προβλήματος ανακάλυψης σημασιολογικά πλούσιων κατευθύνσεων στον λανθάνοντα χώρο εισόδων του StyleGAN. Το κύριο πλεονέκτημά της σε σύγκριση με παρόμοιες μεθόδους που απαιτούν διαδικασία μάθησης με δεδομένα είναι ότι είναι υπολογιστικά πολύ γρηγορότερη αφού βασίζεται αποκλειστικά σε παραγοντοποίηση πινάκων (Singular Value Decomposition - SVD).

Δεδομένου ενός γεννήτορα ΠΑΔ G που απεικονίζει σημεία z του R^d σε εικόνες του συνόλου $\mathcal{I}$: $I = G(z)$, το αρχικό του στρώμα νευρώνων $G_1(z)$ μπορεί να γραφτεί ως $G_1(z) = Az+b$ όπου ο A περιέχει τα βάρη των νευρώνων. Ύστερα από αλγεβρικούς μετασχηματισμούς οι Shen et al. [12] καταλήγουν στο εξής πρόβλημα βελτιστοποίησης για την εύρεση των k σημασιολογικά πλουσιότερων κατευθύνσεων $N^* = \{n_1, ..., n_k\}$:

$$N^* = \arg\max_{n_1,...,n_k} \sum_{i=1}^{k} ||An_i||^2$$

Το παραπάνω μπορεί να λυθεί μέσω της μεθόδου των πολαπλασιαστών Lagrange και συνεπώς της εύρεσης των ιδιοδιανυσμάτων που αντιστοιχούν στις k μεγαλύτερες ιδιοτιμές του A^TA.

Στην υπό εξέταση περίπτωση εφαρμόζουμε την τεχνική SeFa σε διαφορετικά στρώματα του StyleGAN που ελέγχουν χαρακτηριστικά όπως η υφή, η περιστροφή κλπ. Ξεφεύγοντας ολοένα και περισσότερο από τα δεδομένα εισόδου και ακολουθώντας αυτές τις κατευθύνσεις είναι δυνατό να παράξουμε οριακές εικόνες αναφορικά με τις δυνατότητες σύνθεσης του ΠΑΔ.

Φιλτράρισμα Συνθετικών Δεδομένων μέσω Μηχανισμού Ψηφοφορίας

Έχοντας τη δυνατότητα για σύνθεση εικόνων που προέρχονται από τις σημασιολογικά πλούσιες κατευθύνσεις, χρειάζεται και ένα κριτήριο ανομοιότητας με τα δεδομένα εισόδου που να επιτρέπει την επιλογή των οριακών εικόνων. Για τον σκοπό αυτό χρησιμοποιήθηκε ένα σύνολο ταξινομητών-ψηφοφόρων V_1, V_2, V_3 που χρησιμοποιούν χαρακτηριστικά που έχουν εξαχθεί από τις εικόνες μέσω των δικτύων Resnet50, VGG '16, και Inception v3 αντίστοιχα και ακολουθούνται από ένα ρηχό νευρωνικό δίκτυο μαζί με το οποίο έχουν εκπαιδευτεί ακριβώς στα ίδια δεδομένα με τον τελικό ταξινομητή (βλ. Εικόνα 8). Κάθε ψηφοφόρος εξάγει μια πρόβλεψη για κάθε συνθετική εικόνα. Η διαφωνία μεταξύ των ψηφοφόρων υπολογίζεται στη συνέχεια ως ο αριθμός των διαφορετικών μεταξύ τους προβλέψεων. Η υπόθεση είναι ότι στις γνωστές εικόνες οι ψηφοφόροι θα τείνουν να συμφωνούν, ενώ στο "ανοιχτό σύνολο" όπου θα τραβούν τυχαία όρια διαχωρισμού των κλάσεων θα διαφωνούν. Για αυτόν τον λόγο επιλέγονται τελικά οι συνθετικές εικόνες που προκαλούν τις μεγαλύτερες τιμές διαφωνίας.

Αποτελέσματα

Για την αξιολόγηση της μεθόδου χρησιμοποιήθηκε το σύνολο δεδομένων της PCL. Λόγω του μικρού αριθμού κατηγοριών ελαττωμάτων, δημιουργήθηκαν νέες τεχνητές κατηγορίες που προσομοιώνουν πιθανά σφάλματα στη γραμμή παραγωγής όπως γραμμικά χαράγματα, ελλείψεις γραμμάτων στην επιγραφή, λεκέδες διαφορετικών χρωμάτων, περιστροφές αριστερά/δεξιά 90 μοιρών και περιστροφές 180 μοιρών.

Για την σύγκριση των μεθόδων χρησιμοποιήθηκαν οι μετρικές AUROC, F1, η ανάκληση εικόνων από τις γνωστές (Closed-set Recall) και τις καινοφανείς (Open-set Recall) κατηγορίες. Όμοια με προηγουμένως, ο σημαντικότερος στόχος είναι η ελαχιστοποίηση των σφαλμάτων που ταξινομούνται ως άρτια ενώ στην πραγματικότητα είναι ελαττωματικά και μετρούνται μέσω των δύο ανακλήσεων. Οι μετρικές AUROC και F1 αποσκοπούν στο να επιβεβαιώσουν ότι ο αλγόριθμος έχει ικανοποιητική επίδοση στην ταξινόμηση πραγματικά άρτιων προϊόντων ως άρτια και κατά συνέπεια δεν επιβαρύνει τον χειριστή με πολλούς περιττούς χειροκίνητους ελέγχους. Για ταξινομητές που απαιτούν διανυσματικά δεδομένα χρησιμοποιούμε τρεις διαφορετικούς προεκπαιδευμένους εξαγωγείς χαρακτηριστικών με αρχιτεκτονικές Resnet50, VGG16 και Inception v3. Στον Πίνακα 3 φαίνονται οι πιο υποσχόμενες μέθοδοι από όσες ελέγχθηκαν με τις μετρικές F1 και την μέση ανάκληση (μέσος όρος ανάκλησης κλειστού και ανοιχτού συνόλου).

Σε ότι αφορά τους εξαγωγείς χαρακτηριστικών, ο Resnet50 για όλες σχεδόν τις μεθόδους είχε χαμηλή ανάκληση στο ανοιχτό σύνολο, ενώ ο VGG επέτυχε 0.9208 ανάκληση ανοιχτού συνόλου απλά με ένα ρηχό νευρωνικό δίκτυο (VGG + MLP) χωρίς να χρειαστεί κάποιον ειδικό μηχανισμό για το ανοιχτό σύνολο. Αποδίδουμε αυτήν την διαφορά στο μικρότερο πεδίο

Method	F1-score	p-value	R_{avg}	p-value
DFM	0,8347	0,0011	0,9150	0,0187
OpenMax + VGG	0,9389	0,0005	0,9320	0,0004
PISVM + VGG	0,9533	0,0011	0,9556	0,0158
PISVM + Inception	0,9577	0,0024	0,9448	0,0006
MLP + VGG	0,9633	0,0029	0,9264	0,0144
Proposed + VGG	**0,9796**	—	**0,9756**	—

Table 3. *Σύγκριση των μετρικών F1 και Ανάκλησης για τις πιο υποσχόμενες μεθόδους χειρισμού καινοφανών εισόδων στις εικόνες από βραχίονες ξυριστικών μηχανών της PCL BV.*

υποδοχής (receptive field) του VGG που τον καθιστά πιο ευαίσθητο σε μικρές λεπτομέρειες στην εικόνα [13]. Παρότι στις περισσότερες μεθόδους ανίχνευσης ανοιχτού συνόλου τα εξαγώμενα χαρακτηριστικά παίζουν σημαντικό ρόλο στην τελική ανάκληση, η προτεινόμενη μέθοδος κατάφερε να πετύχει υψηλά ποσοστά ανεξαρτήτως εξαγωγέα χαρακτηριστικών, με λιγο καλύτερο τον VGG ’16.

Οι μέθοδοι ημι-επιβλεπόμενης μάθησης, ενώ έχουν καλές επιδόσεις στο ανοιχτό σύνολο, δεν καταφέρνουν να αναγνωρίσουν εύκολα εικόνες του κλειστού συνόλου λόγω τόσο του ότι έχουν εκπαιδευτεί μόνο με άρτιες εικόνες, όσο και λόγω της μεγαλύτερης ομοιότητας μεταξύ ελαττωματικών και άρτιων προϊόντων στα αρχικά (μη-συνθετικά) δεδομένα. Για τον τελευταίο λόγο υπολείπονται και οι μέθοδοι επαύξησης δεδομένων οι οποίες δεν καταφένουν να συνθέσουν δεδομένα αρκετά όμοια με τα άρτια με αποτέλεσμα να μην βοηθούν στην διαδικασία μάθησης. Το μειονέκτημα αυτό αντιμετωπίζεται από την προτεινόμενη μέθοδο χάρη στη μεγαλύτερη εκφραστικότητα και γενικευσιμότητα του StyleGAN σε σχέση με παλιότερες αρχιτεκτονικες ΠΑΔ. Από τις μεθόδους σύγκρισης ξεχωρίζουν κυρίως η PISVM με χαρακτηριστικά που έχουν εξαχθεί με το Inception v3, καθώς και η ημι-επιβλεπόμενη DFM, οι οποίες έχουν σταθερές επιδόσεις σε όλες τις καινοφανείς κλάσεις κοντά στην προτεινόμενη μέθοδο.

Από την άλλη πλευρά, στα μειονεκτήματα της προτεινόμενης μεθόδου συγκαταλέγονται τόσο η μεγάλη σε διάρκεια διαδικασία εκπαίδευσης του StyleGAN, όσο και η παραγωγή μεγάλων ποσοτήτων συνθετικών δεδομένων που φιλτράρονται ως μη χρήσιμα από την διαδικασία ψηφοφορίας. Περισσότερη έρευνα σε πιο ελαφριές αρχιτεκτονικές ΠΑΔ που προσφέρουν όμως δυνατότητα ελέγχου, όσο και στην αποτελεσματικότερη παραγωγή μειωμένου όγκου συνθετικών δεδομένων θα ήταν σκόπιμες για τη βελτίωση της μεθόδου και πιθανώς τη γενίκευσή της σε μικρότερα σύνολα δεδομένων.

Ενίσχυση της Ανθεκτικότητας του Ταξινομητή με Νευροσυμβολική Τεχνητή Νοημοσύνη

Όπως υπογραμμίστηκε η επαύξηση δεδομένων βασισμένη σε StyleGAN απαιτεί τόσο αυξημένους υπολογιστικούς πόρους όσο και επαρκή δεδομένα εκπαίδευσης. Για να ανταποκριθούμε σε σύνολα όπως το MVTEC-AD, που έχουν μόνο λίγες εκατοντάδες πα-

ραδειγμάτων, μερικές φορές και δεκάδες για κάποιες κλάσεις, προτείνουμε μια Νευροσυμβολική προσέγγιση, η οποία και αποδεικνύεται αρκετά ανθεκτική σε νέα ελαττώματα. Ο στόχος της Νευροσυμβολικής ΤΝ [14] είναι να συγχωνεύσει δύο υπάρχοντες κλάδους της τεχνητής νοημοσύνης, συγκεκριμένα την συμβολική τεχνητή νοημοσύνη και την στατιστική μηχανική μάθηση, ελπίζοντας να συνδυάσει τα οφέλη και των δύο προσεγγίσεων σε μια νέα γενιά μεθόδων ΤΝ [15]. Η συμβολική ΤΝ βασίζεται σε χειροποίητους κανόνες που εκφράζονται μέσω λογικών τύπων και οντολογιών, ενώ η Στατιστική Μηχανική Μάθηση χαρακτηρίζεται από μεθόδους όπως τα νευρωνικά δίκτυα που μαθαίνουν απευθείας από δεδομένα. Ενώ η συμβολική ΤΝ μπορεί να λάβει αυτοματοποιημένες αποφάσεις γρήγορες και επεξηγήσιμες, απαιτεί σημαντική προσπάθεια από εμπειρογνώμονες του τομέα που αφορά (π.χ. ιατρική διάγνωση), οι οποίοι καλούνται να συγκεντρώσουν και να κωδικοποιήσουν τη συμβολική γνώση σε οντότητες, σχέσεις μεταξύ οντοτήτων και τους κανόνες που διέπουν αυτές τις σχέσεις. Επιπλέον, τα προκύπτοντα συστήματα χειρίζονται αμφίσημα ή θορυβώδη δεδομένα, όπως αυτά προκύπτουν σε πραγματικές συνθήκες λειτουργίας, με άκαμπτο τρόπο. Αντιθέτως δεδομένο-κεντρικές και στατιστικές προσεγγίσεις, όπως τα Βαθιά Νευρωνικά Δίκτυα, χειρίζονται επιτυχώς τέτοια δεδομένα με αποτέλεσμα να έχουν βρει ουσιαστική εφαρμογή σε τομείς όπως η όραση υπολογιστών και η επεξεργασία φυσικής γλώσσας. Ωστόσο, αντιμετωπίζουν άλλα προβλήματα όπως η αδιαφάνεια αναφορικά με τις εσωτερικές τους λειτουργίες και ως εκ τούτου η έλλειψη αξιοπιστίας, η έλλειψη ευρωστίας σε κυβερνο-επιθέσεις και άγνωστες εισόδους [16][17], καθώς και η απαίτηση πολλών δεδομένων εκπαίδευσης και η ευαισθησία σε ανισορροπίες δεδομένων [4]. Σε αυτή την εργασία χρησιμοποιούμε την Νευροσυμβολική ΤΝ για την αύξηση της γενίκευσης ενός στατιστικού ταξινομητή, έτσι ώστε να καθίσταται πιο ανθεκτικός σε καινοφανείς εισόδους, δηλαδή νέους τύπους ελαττωμάτων παραγωγής. Ειδικότερα εκμεταλλευόμαστε την έγχυση συμβολικών κανόνων μέσω των Δικτύων Λογικού Τανυστή, οι οποίοι συμβαδίζουν με τις αποφάσεις ενός γενικότερου μη επιβλεπόμενου ανιχνευτή καινοφανών δεδομένων, στη συνάρτηση απώλειας ενός επιβλεπόμενου ταξινομητή προσαρμοσμένου στο συγκεκριμένο υπό εξέταση πρόβλημα. Ενώ από μόνος του ο ταξινομητής χωρίς επίβλεψη παράγει πολλά ψευδώς θετικά στοιχεία, ο συνδυασμός του με τον μη επιβλεπόμενο ταξινομητή μέσω της Νευροσυμβολικής ΤΝ έχει ως αποτέλεσμα αυξημένες δυνατότητες αναγνώρισης καινοφανών εισόδων.

Μέθοδος

Χρησιμοποιώντας τη Νευροσυμβολική τεχνητή νοημοσύνη, και συγκεκριμένα τα Δίκτυα Λογικού Τανυστή (LTN), φιλοδοξούμε να συνδυάσουμε τα οφέλη των μεθόδων μάθησης χωρίς επίβλεψη με αυτά των εποπτευόμενων μεθόδων. Ενώ οι πρώτες αποδίδουν καλά στο γενικότερο πρόβλημα της ανίχνευσης καινοφανών εισόδων, οι επιβλεπόμενες μέθοδοι μπορούν να μάθουν πολύ καλά πώς να αναγνωρίζουν τα συγκεκριμένα ελαττώματα που εμφανίζονται στο σύνολο δεδομένων εκπαίδευσης. Παράλληλα με τις προαναφερθείσες προκλήσεις κατά την αυτοματοποίηση του οπτίκου έλεγχου ποιότητας, η γνώση των ειδικών σχετικά με το τι συνιστά ελάττωμα δεν μπορεί να κωδικοποιηθεί πλήρως σε ξεκάθαρους κανόνες, κάτι που αποτελεί άλλο ένα εμπόδιο στις συμβολικές και νευροσυμβολικές προσεγγίσεις. Ωστόσο, μια Νευροσυμβολική προσέγγιση μπορεί ακόμα να επωφεληθεί από σαφείς, αλλά

μη καθολικές περιπτώσεις (π.χ., όταν υπάρχουν σαφείς ενδείξεις ελαττώματος, αλλά η έκφραση αυτών των ενδείξεων μέσω κανόνων δεν μπορεί να είναι καθολικά εφαρμόσιμη για κάθε παράδειγμα του συνόλου δεδομένων λόγω των πολλών και διαφορετικών υποπεριπτώσεων). Αυτές οι προκλήσεις μας οδήγησαν να επιλέξουμε τα LTN καθώς αυτά δεν επιβάλλουν αυστηρά τους συμβολικούς περιορισμούς τους, επιτρέποντας έτσι μεγαλύτερη ευελιξία στη διατύπωση των συμβολικών κανόνων. Επιπλέον, η γνώση των ξεκάθαρων ελαττωμάτων αποτυπωμένη σε συμβολικούς κανόνες μπορεί ακόμα να αξιοποιηθεί για την επιτάχυνση της εκπαίδευσης σε σύγκριση με έναν κλασικό αλγόριθμο εποπτευόμενης μάθησης.

Μια σημαντική πτυχή των Δικτύων Λογικού τανυστή είναι ο τρόπος με τον οποίο οι περιορισμοί μετασχηματίζονται ώστε να είναι διαφορίσιμοι και να αποτελούν μέρος της διαδικασίας εκπαίδευσης. Αυτό επιτυγχάνεται μέσω μιας τεχνικής που ονομάζεται «γείωση» η οποία είναι πολύ κοντά στις ασαφείς λογικές. Πιο συγκεκριμένα, κάθε μεμονωμένη πρόταση ή γεγονός κωδικοποιείται μέσω ενός πολυδιάστατου τανυστή, ο οποίος στην περίπτωσή μας αντιστοιχεί σε διανυσματικές απεικονίσεις που εξάγονται από τις εικόνες εισόδου. Τα κατηγορήματα μπορούν να εφαρμοστούν σε αυτούς τους τανυστές με τη μορφή διαφορίσιμων μαθηματικών συναρτήσεων που μπορούν επίσης να έχουν προσαρμόσιμες παραμέτρους μέσω μάθησης όπως τα τεχνητά νευρωνικά δίκτυα. Η εφαρμογή αυτών των κατηγορημάτων θα πρέπει να αποδίδει μια πραγματική τιμή μεταξύ 0 και 1 που αντιστοιχεί στον βαθμό αλήθειας του κατηγορήματος που εφαρμόζεται σε μία ή πολλαπλές προτάσεις. Με βάση αυτό, οι λογικοί τελεστές μπορούν να χρησιμοποιηθούν για να συνδυάσουν διαφορετικά αποτελέσματα κατηγορημάτων. Για παράδειγμα, ένα λογικό $a \wedge b$ μπορεί τώρα να υπολογιστεί ως $a\dot{b}$ και το $a \implies b$ υπολογίζεται ως $\frac{b}{a}$ εάν $b < a$ ή 1 αν $b > a$. Φυσικά, υπάρχουν πολλές διαφορετικές αντιστοιχίσεις από τη λογική πρώτης τάξης προς τους πραγματικούς τελεστές, πολλές από τις οποίες περιγράφονται λεπτομερώς στο [18]. Αφού γίνει η λογική πρόταση διαφορίσιμη, ο βαθμός ικανοποίησής της μπορεί να προστεθεί ως όρος της συνάρτησης απώλειας που θα βελτιστοποιηθεί κατά τη διάρκεια της εκπαίδευσης.

Η «γείωση» των συμβολικών κανόνων του LTN σε διαφορίσιμες πραγματικές συναρτήσεις του επιτρέπει να περιορίσει έναν αλγόριθμο στατιστικής μηχανικής μάθησης έτσι ώστε να πληροί τους προκαθορισμένους συμβολικούς κανόνες κατά τη φάση εκπαίδευσής του. Ταυτόχρονα, η αξιοποίηση αυτών των κανόνων προϋποθέτει την κωδικοποίηση της γνώσης ενός ειδικού σε αντίστοιχη μορφή που, στην περίπτωσή μας, είναι δύσκολο να επιτευχθεί. Το σενάριο παραγωγής που αντιμετωπίζουμε αφορά σε μια ευέλικτη γραμμή παραγωγής με συχνές αλλαγές στις προδιαγραφές του προϊόντος. Οι αλλεπάλληλες αλλαγές καθιστούν δύσκολο για τους φορείς παραγωγής να αναπτύξουν αρκετή τεχνογνωσία έτσι ώστε να καταλήξουν σε ένα πλήρες σύνολο κανόνων για τον εντοπισμό ελαττωμάτων. Επιπλέον, η φύση των δεδομένων εικόνας καθιστά δύσκολη τη σύνδεση αυτών των κανόνων με τις ιδιότητες των εικόνων. Μια ιδιότητα όπως, για παράδειγμα, η ομαλότητα της επιφάνειας δεν είναι εύκολο να οριστεί ως συνάρτηση-κατηγόρημα επεξεργασίας εικόνας που θα χρησιμοποιηθεί από το LTN. Για αυτούς τους λόγους χρησιμοποιούμε έναν ταξινομητή χωρίς επίβλεψη στον ρόλο του ειδικού.

Το κριτήριο για την επιλογή ενός ταξινομητή χωρίς επίβλεψη είναι να έχει καλές ιδιότητες αναγνώρισης καινοφανών δεδομένων και μια απλή προσαρμόσιμη υλοποίηση. Ακολουθώντας τα αποτελέσματά μας από την προηγούμενη εργασία [17], επιλέξαμε το Isolation Forest

(IF), καθώς προσφέρει μια τέτοια υλοποίηση, χρειάζεται περιορισμένη προσαρμογή και έχει αποδειχθεί ότι αποδίδει καλά σε μια ποικιλία συνόλων δεδομένων [19]. Παρά την υψηλή απόδοση σε άγνωστες εικόνες, το IF δεν είναι τόσο αποτελεσματικό στις γνωστές κατηγορίες των δεδομένων εκπαίδευσης. Για να ξεπεράσουμε αυτό το μειονέκτημα δημιουργήσαμε τους κανόνες που περιγράφονται παρακάτω, όπου ο A είναι ο βασικός ταξινομητής Multi-Layer Perceptron (MLP) και ο U ο μη εποπτευόμενος ταξινομητής Isolation Forest. Αυτοί οι κανόνες επιβάλλουν στο MLP έναν ήπιο λογικό περιορισμό για να ακολουθεί την έξοδο U όταν προβλέπει ένα ελάττωμα.

$$
\begin{aligned}
SatAgg\Big([\forall x(l_S(x) = 1 \implies A(x) = 1)]\wedge \\
[\forall x(l_S(x) = 0 \implies A(x) = 0)]\wedge \\
[\forall x(U(x) = 0 \implies A(x) = 0)]\Big)
\end{aligned}
$$

Ο παραπάνω τύπος περιέχει δύο πρόσθετους περιορισμούς που απαιτούνται για την ταξινόμηση και διασφαλίζουν ότι η πρόβλεψη $A(x)$ είναι σύμφωνη με την ετικέτα εποπτείας $l_S(x)$. Έτσι, ο βασικός ταξινομητής A εκπαιδεύεται μόνο για να ικανοποιεί το σύνολο κανόνων που περιγράφεται. Η πλήρης διαδικασία εκπαίδευσης απεικονίζεται επίσης ως διάγραμμα στο Σχ.8

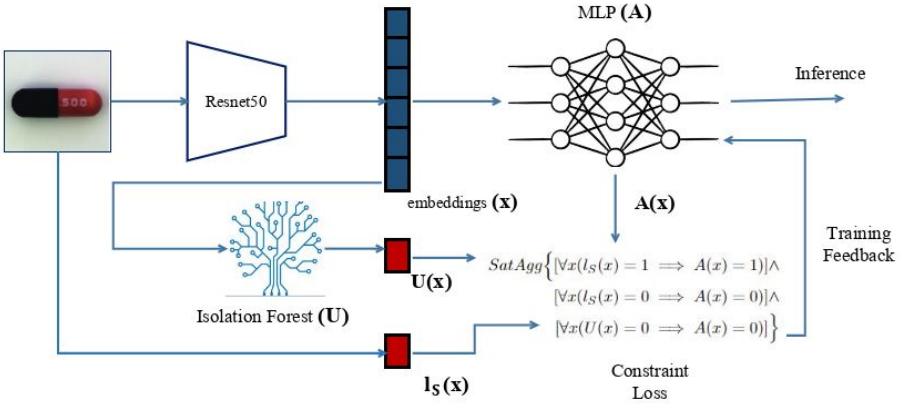


Figure 9. *Διαδικασία εκπαίδευσης του LTN*

Αποτελέσματα

Αναφορικά με τα αποτελέσματα στα σύνολα δεδομένων προϊόντων του MVTEC-AD βλέπουμε διάφορα κοινά μοτίβα. Πρώτον, δεν αποτελεί έκπληξη το γεγονός ότι το Deep Feature Modelling (DFM) επιτυγχάνει τα υψηλότερα αποτελέσματα όσον αφορά την AUROC και την ακρίβεια, καθώς είναι μια ημι-εποπτευόμενη μέθοδος που εκπαιδεύεται μόνο στην "άρτια" κατηγορία και επομένως είναι καλύτερη στην αναγνώριση της. Στις δύο μετρήσεις ανάκλησης, ωστόσο, βλέπουμε ότι το LTN ξεπερνά το DFM σχεδόν σε όλες τις περιπτώσεις, με εξαίρεση την ανάκληση ανοιχτού συνόλου για τα σύνολα δεδομένων "leather" και "grid". Στα περισσότερα σύνολα δεδομένων επιτυγχάνει επίσης υψηλότερη βαθμολογία F1 η οποία

αποτελέι μια πιο σφαιρική μέτρηση της επίδοσης στο συνολικό πρόβλημα, εξισορροπώντας την απόδοση μεταξύ των κατηγοριών άρτιων και ελαττωματικών προϊόντων, ενώ επηρεάζεται λιγότερο από τις ανισορροπίες κλάσεων.

Dataset	Method	AUROC	Prec.	F1-score	R_open	R_closed
Carpet	MLP	92,89 ± 1,08	**95,04 ± 2,58**	83,47 ± 2,45	48,13 ± 10,67	74,80 ± 13,30
	OCSVM	74,23 ± 2,86	63,01 ± 2,79	72,08 ± 1,70	59,64 ± 6,17	59,26 ± 7,11
	IF	86,80 ± 1,58	70,56 ± 2,58	79,39 ± 2,42	**86,00** ± 3,01	81,06 ± 4,60
	DFM	**98,44 ± 0,27**	**99,45 ± 0,47**	**84,08 ± 1,26**	79,37 ± 4,87	79,59 ± 4,26
	WSVM	72,63 ± 2,00	58,81 ± 8,74	58,48 ± 10,11	63,86 ± 15,12	**84,20 ± 8,85**
	LTN	**97,47 ± 0,98**	88,66 ± 3,66	**91,74 ± 2,04**	**89,68 ± 4,00**	**99,53 ± 0,72**
Capsule	MLP	**93,87 ± 1,11**	**98,66 ± 0,81**	**78,99 ± 1,64**	51,28 ± 5,31	**94,20 ± 2,81**
	OCSVM	71,20 ± 2,71	72,82 ± 2,88	66,18 ± 2,19	57,33 ± 5,34	57,93 ± 5,68
	IF	81,25 ± 2,44	75,66 ± 3,03	71,72 ± 2,93	73,46 ± 5,39	67,46 ± 4,31
	DFM	**98,55 ± 0,61**	**98,72 ± 0,64**	**83,02 ± 3,78**	**83,77 ± 5,03**	82,80 ± 8,48
	WSVM	72,79 ± 6,09	56.18 ± 2,57	42,96 ± 3,08	67,46 ± 8.32	87,13 ± 6,02
	LTN	85,92 ± 11,19	83,19 ± 10,88	66,79 ± 23,35	**91,28 ± 8,57**	**99,80 ± 0,31**
Grid	MLP	72,98 ± 2,80	**76,20 ± 4,98**	81,02 ± 1,01	17,46 ± 3,99	**72,53 ± 8,75**
	OCSVM	41,32 ± 2,87	30,52 ± 2,50	66,82 ± 1,79	26,13 ± 4,88	24,13 ± 9,17
	IF	47,65 ± 2,72	33,17 ± 1,84	63,64 ± 1,98	36,80 ± 6,02	35,66 ± 8,97
	DFM	**93,60 ± 1,20**	**91,53 ± 2,55**	**81,23 ± 1,84**	**68,57 ± 5,97**	69,13 ± 5,85
	WSVM	40,18 ± 2,25	38,68 ± 5,10	58,52 ± 7,22	47,51 ± 12,15	63,80 ± 12,29
	LTN	**81,42 ± 6,80**	74,28 ± 11,82	**84,47 ± 4,78**	**62,22 ± 13,98**	**86,13 ± 7,09**
Pill	MLP	86,62 ± 1,60	95,82 ± 3,23	65,88 ± 3,60	32,72 ± 13,68	66,06 ± 12,28
	OCSVM	58,90 ± 2,01	68,79 ± 1,71	56,98 ± 1,45	53,57 ± 4,82	58,80 ± 11,84
	IF	68,28 ± 1,89	72,19 ± 1,78	60,16 ± 2,26	64,13 ± 4,37	58,53 ± 9,86
	DFM	**98,21 ± 0,35**	**99,84 ± 0,22**	**67,84 ± 3,52**	**70,10 ± 4,16**	**70,86 ± 9,62**
	WSVM	62,05 ± 5,00	70,44 ± 3,56	56,28 ± 5,87	54,82 ± 11,92	74,40 ± 8,62
	LTN	**95,43 ± 2,66**	**95,91 ± 1,73**	**88,36 ± 3,43**	**87,92 ± 6,05**	**95,33 ± 2,73**
Tile	MLP	97,74 ± 0,87	**99,38 ± 0,75**	**87,37 ± 2,86**	60,88 ± 10,75	**96,20 ± 3,53**
	OCSVM	66,48 ± 3,35	62,44 ± 2,31	68,22 ± 2,02	55,86 ± 6,56	57,73 ± 7,79
	IF	87,77 ± 2,25	71,91 ± 1,69	77,88 ± 1,69	**84,40 ± 5,94**	82,00 ± 8,21
	DFM	**99,34 ± 0,17**	**99,69 ± 0,34**	83,65 ± 0,26	73,06 ± 9,41	79,40 ± 13,42
	WSVM	65,36 ± 6,65	57,08 ± 5,94	56,64 ± 9,17	54,84 ± 10,17	85,46 ± 5,14
	LTN	**97,92 ± 1,60**	91,13 ± 3,00	**93,02 ± 2,68**	**90,97 ± 7,18**	**96,86 ± 2,61**
Leather	MLP	97,54 ± 0,95	**97,97 ± 1,01**	86,48 ± 2,32	62,93 ± 8,59	93,93 ± 3,02
	OCSVM	70,56 ± 3,60	68,39 ± 2,94	71,16 ± 2,65	64,97 ± 5,99	49,80 ± 6,52
	IF	92,93 ± 1,07	79,35 ± 1,91	85,62 ± 1,65	**96,35 ± 1,70**	95,26 ± 3,14
	DFM	**99,97 ± 0,01**	**99,92 ± 0,01**	**97,60 ± 0,77**	**97,91 ± 1,07**	**95,73 ± 1,47**
	WSVM	60,39 ± 4,61	70,56 ± 7,60	68,98 ± 7,79	49,06 ± 15,13	81,26 ± 4,18
	LTN	**99,00 ± 1,09**	96,42 ± 1,96	**95,30 ± 3,44**	89,73 ± 12,65	**99,66 ± 0,71**

Table 4. *Comparison of methods on the various MVTEC-AD product datasets*

Συνολικά, τα πειραματικά μας αποτελέσματα υποδεικνύουν ότι τα δίκτυα λογικού τανυστή (LTN) και η μοντελοποίηση βαθιών χαρακτηριστικών (DFM) υπερέχουν σταθερά σε σχέση με άλλες μεθόδους σε πολλές μετρήσεις. Τα LTN υπερέχουν τόσο στην ανάκληση ανοιχτού όσο και σε κλειστού συνόλου λόγω της ικανότητάς τους να ενσωματώνουν συμβολικούς κανόνες στη διαδικασία μάθησης, παρέχοντας ένα δομημένο πλαίσιο που ενισχύει την ικανότητα του μοντέλου να γενικεύει σε νέα ελαττώματα. Αυτό το πλεονέκτημα είναι σημαντικό σε περιβάλλοντα παραγωγής όπου τα ελαττώματα είναι σπάνια και ποικίλα, καθιστώντας τις παραδοσιακές μεθόδους λιγότερο αξιόπιστες. Ο συμβολικός συλλογισμός στα LTN επιτρέπει στο μοντέλο να χειρίζεται πιο αποτελεσματικά διφορούμενα δεδομένα αξιοποιώντας τη γνώση του τομέα που κωδικοποιείται σε λογικούς κανόνες. Αντίθετα, το DFM έχει εξαιρετικά καλή απόδοση όσον αφορά το AUROC και την ακρίβεια, καθώς μαθαίνει πολύ καλά πώς πρέπει να μοιάζει ένα «άρτιο» προϊόν, επιτρέποντας στο μοντέλο να κατανοήσει καλύτερα και να ταξινομήσει τα άρτια έναντι των ελαττωματικών δειγμάτων. Ωστόσο, η αντιμετώπιση των ελαττωμάτων από το DFM με αγνωστικιστικό τρόπο, που δεν βασίζεται σε συγκεκριμένα δείγματα εκπαίδευσης, συχνά οδηγεί σε χαμηλή απόδοση στον εντοπισμό ελαττωμάτων κλειστού συνόλου σε σύγκριση με άλλες μεθόδους που περιλαμβάνουν ελαττώματα κλειστού συνόλου στο σύνολο δεδομένων εκπαίδευσής τους. Όσον αφορά το MLP, αναμενόμενα αποδίδει αρκετά καλά στην αναγνώριση των κλάσεων στις οποίες έχει εκπαιδευτεί, αλλά η απόδοσή του

επιδεινώνεται σημαντικά στις καινοφανείς κλάσεις (ανοιχτό σύνολο). Μέθοδοι όπως το One-Class SVM (OCSVM) και το Isolation Forest (IF) εμφάνισαν περιορισμούς κυρίως λόγω των υψηλών ψευδώς θετικών ποσοστών τους όταν αντιμετώπιζαν πολύπλοκα και πολύ παρόμοια οπτικά δεδομένα κλάσης όπως στο παρουσιαζόμενο περιβάλλον παραγωγής.

Οι βελτιωμένες και πιο ισορροπημένες βαθμολογίες ανάκλησης ανοιχτών και κλειστών συνόλων της προσέγγισής μας που βασίζεται σε LTN είναι αποτέλεσμα της ικανότητας του LTN, μέσω της εισαγωγής συμβολικών κανόνων, να συνδυάζει την ικανότητα του μη εποπτευόμενου ταξινομητή να ανιχνεύει εισόδους εκτός κατανομής (υψηλή ανάκληση στο ανοιχτό σύνολο) και την ικανότητα προσαρμογής στα δεδομένα του προβλήματος του βασικού στατιστικού ταξινομητή (υψηλή ανάκληση στο κλειστό σύνολο). Είναι σημαντικό να σημειωθεί ότι τα LTN επιτρέπουν στους συμβολικούς κανόνες να επηρεάζουν το μοντέλο συνεχώς κατά τη διάρκεια της εκπαίδευσης και έτσι έχουν μεγαλύτερη επίδραση στη συμπεριφορά του. Αυτή η ικανότητα καθιστά την προσέγγιση LTN ιδανική για ένα σενάριο σπανιότητας δεδομένων όπου προκλήσεις όπως χαμηλού πληθυσμού ή εντελώς νέες κατηγορίες ελαττωμάτων μετριάζονται μέσω του συμβολικού σκέλους του LTN, ενώ οι υπάρχουσες κλάσεις με αρκετά δεδομένα αλλά ίσως μεγαλύτερη ομοιότητα με την άρτια κατηγορία αναγνωρίζονται καλύτερα από το στατιστικό σκέλος.

Συνεισφορές

- Για την αντιμετώπιση της ανισορροπίας των κλάσεων αναπτύχθηκε μια νέα μέθοδος παραγωγής συνθετικών δεδομένων βασισμένων σε παραδείγματα που βρίσκονται κοντά στο όριο μεταξύ “άρτιας” και “μη-άρτιων” κλάσεων. Η μέθοδος αυτή, συνδυάζοντας την ακρίβεια των τεχνικών υπερδειγματοληψίας και τις συνθετικές δυνατότητες του BigGAN κατάφερε να επιτύχει βελτίωση στην ανάκληση του νευρωνικού δικτύου, μειώνοντας ταυτόχρονα τον χρόνο παραγωγής δεδομένων σε σχέση με τις άλλες τεχνικές βασισμένες σε ΠΑΔ.

- Για τον χειρισμό καινοφανών εισόδων, αναπτύχθηκε μια νέα μέθοδος βασισμένη στην επαύξηση δεδομένων με χρήση του StyleGAN, ιδιαίτερα προσαρμοσμένη σε σύνολα δεδομένων με μεγάλη ομοιότητα μεταξύ των κλάσεων, όπως αυτά που συναντιούνται στον βιομηχανικό έλεγχο. Η νέα μέθοδος βασίζεται τόσο στην ευκρίνεια του StyleGAN όσο και στην δυνατότητα ακριβέστερου και με λογική σημασία χειρισμού της παραγωγής των συνθετικών δεδομένων. Επίσης σημαντικό ρόλο έπαιξε το φιλτράρισμα των παραγόμενων δεδομένων μέσω της ποσοτικοποίησης του βαθμού διαφωνίας διαφορετικών ταξινομητών που έχουν εκπαιδευτεί στα αρχικά δεδομένα. Έτσι διασφαλίζεται ότι τα τεχνητά δεδομένα αντιπροσωπεύουν το “ανοιχτό σύνολο” και μπορούν να επαυξήσουν επαρκώς τα αρχικά ώστε να καθιστούν τον τελικό ταξινομητή πιο ανθεκτικό σε νέο-εμφανιζόμενα δεδομένα κατά την περίοδο συνεχούς λειτουργίας. Η νέα μέθοδος συγκρίθηκε με τις υπόλοιπες και εμφάνισε βελτιωμένα αποτελέσματα σε πραγματικό σύνολο δεδομένων από την βιομηχανία.

- Τέλος ως συνέχεια της προηγούμενης μεθόδου για επέκταση σε μικρότερα σύνολα δεδομένων στα οποία δεν είναι εφικτό να εκπαιδευτεί το StyleGAN έγινε χρήση τεχνικών

Νευροσυμβολικής Τεχνητής Νοημοσύνης. Συγκεκριμένα, χρησιμοποιήθηκε ένα Δίκτυο Λογικού τανυστή που εκφράζει τις εξόδους ενός ανιχνευτή καινοφανών εισόδων με περιορισμένη επίβλεψη ως συμβολικούς κανόνες και τους χρησιμοποιεί για να οδηγήσει την εκπαίδευση ενός νευρωνικού δικτύου. Ο αλγόριθμος που προκύπτει δείχνει βελτιωμένα αποτελέσματα σε σύγκριση με άλλες σχετικές μεθόδους, ιδίως όσον αφορά την ανάκληση ελαττωμάτων, με την έννοια ότι λίγα ελαττώματα παραμένουν απαρατήρητα ακόμα και αν είναι εντελώς καινοφανή. Επιπροσθέτως, επιτυγχάνει παρόμοια ή καλύτερα αποτελέσματα ανάκλησης από ημι-επιβλεπόμενες μεθόδους κατά τον χειρισμό νέων ελαττωμάτων, ξεπερνώντας τες όμως σε ελαττώματα που ανήκουν στις κατανομές των κλάσεων εκπαίδευσης (κλειστό σύνολο). Σε σύγκριση με άλλες επιβλεπόμενες μεθόδους, διατηρεί υψηλή απόδοση σε γνωστά ελαττώματα, αλλά βελτιώνει σημαντικά σε νέα. Ο συνδυασμός των πλεονεκτημάτων αυτών των δύο τύπων μεθόδων απεικονίζεται μέσω υψηλότερων βαθμολογιών F1 στα περισσότερα από τα σύνολα δεδομένων δοκιμής.

Table of Contents

Abstract 1

Περίληψη 3

Acknowledgements 5

Εκτεταμένη Περίληψη 7

1 Introduction 33
- 1.1 Machine Learning and AI applications in real-life industrial Systems 34
  - 1.1.1 From Industry 4.0 to Industry 5.0 35
  - 1.1.2 Data Lifecycle in Industrial ML Applications 37
  - 1.1.3 Challenges of Real-life Industrial ML Applications 41
- 1.2 Automated Visual Quality Inspection in Manufacturing 42
  - 1.2.1 Data Scarcity 45
  - 1.2.2 Robustness and Trustworthiness of AI Visual Inspection Systems . . 50
- 1.3 Contributions and Structure of the Thesis 52

2 On-the-fly Image-level Oversampling for Imbalanced Datasets of Manufacturing Defects 55
- 2.1 Background 55
- 2.2 Related Work 57
  - 2.2.1 GANs in Defect Generation 57
  - 2.2.2 Prediction Confidence in Deep Neural Networks 58
- 2.3 Methods 59
  - 2.3.1 Synthetic Image Generation 60
  - 2.3.2 Confidence Assessment 62
  - 2.3.3 On-the-fly Image-Level Oversampling 62
- 2.4 Results 63
  - 2.4.1 Dataset Information 63
  - 2.4.2 Experimental Setup 65
  - 2.4.3 Experimental Results 66
- 2.5 Summary 72

3 Enhancing Robustness to Novel Visual Defects through StyleGAN Latent Space Navigation: A Manufacturing Use Case 75
- 3.1 Background 75

3.2 Use Case and Dataset . . . 76
3.3 Related Work . . . 77
3.3.1 Open-set Recognition . . . 78
3.3.2 Semi-supervised Defect Detection . . . 80
3.3.3 OSR in Manufacturing Defect Detection . . . 81
3.3.4 GAN Inversion and Latent Space Traversal . . . 81
3.4 Proposed Method . . . 82
3.4.1 Semantic Factorization for Latent Space Traversal . . . 84
3.4.2 Method Description . . . 84
3.5 Results . . . 87
3.5.1 Experimental Setup . . . 87
3.5.2 Examined Methods . . . 88
3.5.3 GAN Training . . . 90
3.5.4 Hyperparameter Tuning . . . 91
3.5.5 Experimental Results . . . 92
3.6 Summary . . . 97

4 Robust Novel Defect Detection with Neurosymbolic AI 99
4.1 Background . . . 99
4.2 Related Work . . . 100
4.2.1 Neurosymbolic AI . . . 100
4.2.2 Open-set Recognition . . . 101
4.3 Methods . . . 102
4.3.1 Problem Setting . . . 102
4.3.2 Why Logic Tensor Networks? . . . 103
4.3.3 Our approach . . . 104
4.3.4 Datasets . . . 105
4.4 Results . . . 107
4.4.1 Experimental Setup . . . 107
4.4.2 Experimental Results . . . 107
4.5 Summary . . . 109

5 Conclusions 111

Bibliography 132

English to Greek Glossary of Terms 133

List of Publications 137

List of Figures

1 Βασικές Αρχές και Τεχνολογίες της Βιομηχανίας 5.0 [2] 8
2 Δείγματα από τα δεδομένα της PCL BV 9
3 Δείγματα από τα δεδομένα του MVTEC-AD 9
4 Αρχιτεκτονικό διάγραμμα της μεθόδου επαύξησης δεδομένων και της διαδικασίας εκπαίδευσης του τελικού ταξινομητή. 12
5 Συνθετικές εικόνες που παράχθηκαν από τις υπο εξέταση μεθόδους 14
6 Ακρίβεια ταξινομητή στα οριακά δεδομένα που επιλέγονται ως γεννήτορες για την σύνθεση των δεδομένων επαύξησης (Αριστερά). Οι k μικρότερες αποστάσεις προς το όριο διαχωρισμού των κλάσεων πριν και μετά την εκπαίδευση στο επαυξημένο σύνολο δεδομένων για k=15 (Δεξιά) 14
7 Προβλήματα ανθεκτικότητας σε νέες εισόδους. Η πάνω σειρά περιέχει διακοπτόμενες σε μικρό βαθμό εικόνες που αντίστοιχές τους υπάρχουν στα δεδομένα εκπαίδευσης και ταξινομούνται σωστά. Αντίθετα εικόνες με μεγαλύτερης έκτασης ελαττώματα που είναι καινοφανείς δεν ταξινομούνται σωστά (κάτω σειρά). 16
8 Αρχιτεκτονικό διάγραμμα της διαδικασίας εκπαίδευσης με παραγωγή συνθετικών δεδομένων μέσω Semantic Factorization και φιλτράρισμά τους μέσω ψηφοφορίας 17
9 Διαδικασία εκπαίδευσης του LTN 22
1.1 Fundamental Principles and corresponding Technologies of Industry 5.0 [2] 34
1.2 Main categories of few-shot learning methods [20] 47
2.1 Basic components and dataflows for the proposed oversampling approach. The sequence of processing steps is outlined with numbers from (1) to (9). 60
2.2 Original Shaver Shell Prints 64
2.3 Samples from the MVTec AD datasets 64
2.4 Artificially generated defect images 69
2.5 Label accuracy of augmented images, before and after augmented training (Left). Top-k distances to classification boundary before and after augmented training for k=15 (Right) 71
2.6 Comparison between simple augmentation and confidence-based oversampling - 6 different instances of 5-fold CV on the shavers dataset 73
3.1 Original Shaver Shell Prints 77
3.2 Synthetic “Unexpected" Defects 77
3.3 Basic components and dataflows for the proposed approach 83
3.4 Images generated from SeFa traversal at given distances. The circled images are retained as out-of-distribution after filtering. 83

3.5 Progression of FID while training candidate generator models. “all_classes" is the class-conditional model. The double and interrupted class models start from a pre-trained model of the “flawless" class for k_img=80. 91
3.6 Box plot of open-set class-specific accuracy scores for highest performing methods per type. 96
4.1 Figure 1(a) is a high-level depiction of the visual quality assessment workflow. Potential defects identified by the AI are also examined by a human before being discarded, while products labelled "GOOD" by the AI pass QA. Figure 1(b) shows how the AI system using supervised learning based on Resnet50 runs into issues when encountering novel defects that, despite looking more severe, are incorrectly labelled. 103
4.2 Training Workflow with LTN using embeddings for empirical learning and symbolic rules derived from an Isolation Forest’s predictions. The symbolic rules are "grounded" and embedded into the loss function to guide training. 105
4.3 Original ((a)-(c)) and Synthetic Test ((d)-(f)) Samples from the Shavers Dataset 106
4.4 Product categories’ samples from the MVTEC-AD datasets 106

List of Tables

1 Αξιολόγηση στα δεδομένα εικόνων ξυριστικών μηχανών της PCL BV 13
2 Αξιολόγηση στα σύνολα εικόνων από το MVTEC-AD 15
3 Σύγκριση των μετρικών F1 και Ανάκλησης για τις πιο υποσχόμενες μεθόδους χειρισμού καινοφανών εισόδων στις εικόνες από βραχίονες ξυριστικών μηχανών της PCL BV. 19
4 Comparison of methods on the various MVTEC-AD product datasets . . . 23
2.1 Number of class instances for the Shavers and MVTec AD product datasets including train and test sets 65
2.2 Comparison of oversampling methods on the shaver-shell prints dataset . 68
2.3 Comparison of oversampling methods on the MVTec AD product datasets . 70
2.4 Table of searched and recommended final hyperparameters per examined method for the shavers dataset 72
3.1 Qualitative comparison summary of the characteristics of the examined methods, according to their provided functionality, implementation and computational infrastructure requirements. 90
3.2 Hyperparameter selection intervals. 92
3.3 Evaluation of OSR methods over pre-extracted Resnet50 features, including AUROC, F1-score, Binary Recall on the closed set classes (R_c), and Binary Recall on the open set classes (R_o) and lastly $R_{avg} = \frac{R_c+R_o}{2}$. 93
3.4 Evaluation of OSR methods over pre-extracted deep VGG '16 features. . . . 94
3.5 Evaluation of OSR methods over pre-extracted Inception v3 features. . . . 94
3.6 Evaluation of semi-supervised and data-augmentation-based methods . . 95
3.7 Comparison against the best performing OSR methods over their F1-score and Recall averaged from both open- and closed-set samples with statistical significance scores. 97
4.1 Comparison of methods on the Shavers dataset 107
4.2 Comparison of methods on the various MVTEC-AD product datasets . . . 108

Chapter 1

Introduction

The main focus of this thesis is on machine learning applications in industrial environments and on specific practical problems that arise due to the difficulty of collecting training data such as Class Imbalance and lack of Resilience to Novel Input Data. The technological and research environment in which these problems are examined is that of Industry 5.0, a term arises as an extension of the 4th Industrial Revolution (Industry 4.0), characterized by technologies such as the Internet of Things, Cyber-physical Systems, Digital Twins, Big Data and Artificial Intelligence. Against this background Industry 5.0 aims to combine human capabilities with that of intelligent machines through simulation systems and Human-AI collaboration. [1]

More specifically, the present work of research focuses on the Automatic Quality Control of Industrial Products through Machine Learning techniques for Computer Vision. In the context of Quality 4.0 (part of Industry 4.0) the goal is to create self-evaluating systems that can automatically measure the quality of their output and decide autonomously on its acceptance or rejection. Deep Learning, due to its adaptability (e.g. to visual changes in scale or rotation of the image), has helped a lot in this, but at the same time it requires a large amount of training data and is not stable to samples outside the training distribution. One solution being explored in the context of Industry 5.0 is the development of Human-Machine collaboration systems where human intelligence and experience will compensate for the shortcomings of AI algorithms.

While researching the application of Deep Learning techniques into the Automatic Quality Control of manufacturing products, three main challenges were identified, which serve as the focus of this work:

1. ***The scarcity of training data***, which is particularly noticeable in products with defects. This is because defects occur less often on production lines than good products, leading to an *imbalance between the two classes*.

2. ***The high visual similarity between good and defective products*** which significantly hampers the ability of classifiers to distinguish between them.

3. ***The appearance of novel defects***, that during the continuous operation of an already trained algorithm can lead to incorrect classification of products as flawless.

To deal with class imbalance, a method was developed to increase training data be-

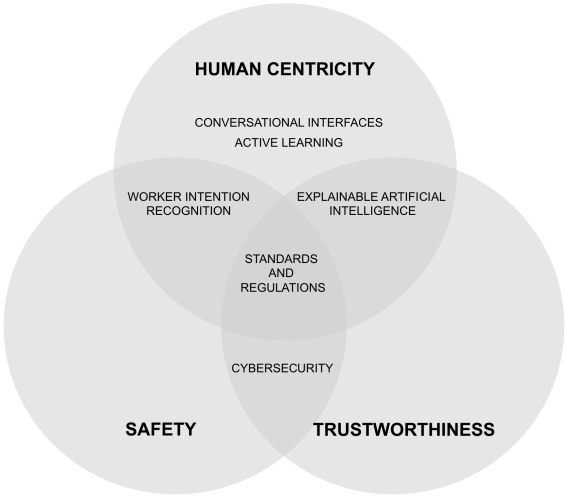


Figure 1.1. *Fundamental Principles and corresponding Technologies of Industry 5.0 [2]*

longing to minority classes. The synthesis of the data was done with techniques aimed at guiding the output of Generative Adversarial Networks (GANs), with the aim of oversampling examples in which the predictions of the classifier show low reliability. The augmentation of such data may provide greater benefit to the training process. [3]

Similar techniques were explored for handling novel inputs, this time with the aim of synthesizing boundary examples using StyleGAN. Although the data generation process developed starts from the training distributions, the boundary data, thanks to the generalizability of StyleGAN, is generated at the edges of the distributions known at training time and creates a boundary between known and novel inputs. [4]. As StyleGAN requires a significant amount of training data, we applied the concept of NeuroSymbolic AI to address smaller datasets. The proposed NeuroSymbolic method combines, using symbolic rules, an unsupervised classifier specialized at detecting novel defects with a supervised one specialized in performing optimally in the known training distribution.

As for the similarity between good and defective products, this was taken into account in both of the above methods. In particular, GANs with very detailed image synthesis capabilities were used, while, where the amount of data allowed, the final classifiers were trained directly on the problem without using transfer learning from pre-trained networks.

1.1 Machine Learning and AI applications in real-life industrial Systems

The holy grail of modern AI research is the achievement of Aritficial General Intelligence (AGI). The first steps towards AGI include systems that can perform a variety of tasks including open-ended learning, innovation and human-like reasoning [21]. Advances in computer vision including various top-scoring methods on the ImageNet [22] benchmark and more recent advances such as GPT-3 and GPT-4 [23][24] by OpenAI or Deepmind's Gato [25] have inspired a significant wave of progress, especially in the domain of general-purpose Large Language Models with multi-modal outputs.

While the above innovations are definitely exciting, care must be taken when applying them to real-life domains such as an Industrial Plant, mainly associated with valid concerns over safety and reliability of such systems. For instance, computer vision systems are known for being sensitive to small differences in input [10], something that might lead to unexpected wrong decisions during the continuous operation of said systems as part of a large cyber-physical deployment. Application of AI in industrial environments, as well as the present thesis, focus therefore more on so-called "Narrow AI" applications. Narrow AI refers to AI systems that focus on performing well, sometimes achieving even super-human performance, on a very narrow task such as visual defect classification, speech recognition, domain specific recommendations or demand forecasting [21]. Narrow AI, has undergone widespread adoption in manufacturing, transforming work design, the allocation of responsibilities, and the socio-economic dynamics of the manufacturing workplace. While AI systems in manufacturing can provide valuable insights, automate repetitive tasks, and assist in decision-making processes, human input still remains crucial, especially in scenarios requiring complex judgments or ethical considerations [26]. A logical consequence is a shift towards the development of synergistic Narrow AI systems that combine the respective strengths of humans and algorithms. The following subsections contain a short review of the main research and industrial trends and challenges that brought human-AI to the forefront of industrial AI research.

1.1.1 From Industry 4.0 to Industry 5.0

Stepping into the "Information Age", the rapid development of ICT together with their democratization through open-source initiatives, have had a significant impact on the manufacturing domain on a worldwide scale. Ranging from the United States with the "Advanced Manufacturing Partnership" to China with "Made in China 2025", government initiatives have sprung up to encourage and facilitate the digitization of industry. The ultimate aim of theses initiatives are manifold, with societal goals such as coping with aging populations and a diminishing workforce to making the industrial sector more competitive, efficient and most importantly environmentally and economically sustainable. [27].

The concept of Industry 4.0, first introduced at the Hannover Industrial Fair in 2011, represents a significant shift in manufacturing. At its core, Industry 4.0 focuses on enhancing operational efficiency and productivity in manufacturing by utilizing intelligent systems that can automate processes, analyze data in real-time, and make informed decisions. It emphasizes the integration of advanced technologies such as the Internet of Things (IoT), cloud computing, artificial intelligence (AI), and Cyber-Physical Systems (CPS). This paradigm shift has led to the development of smart factories where machines, systems, and humans are interconnected, opening up the field for seamless communication and collaboration. However, while Industry 4.0 has primarily focused on technological advancements, there has been a growing recognition of the need to complement these technologies with human-centric solutions. This realization has given rise to the concept of Industry 5.0, which emphasizes the synergistic collaboration between humans and machines, ensuring that the advancements in technology do not overshadow the

importance of human input and well-being. [2]

The integration of software devices associated with the collection of data and its processing towards decision making has given Industry 4.0 adopters a competitive edge but together with that also a set of challenges. For instance, as AI algorithms get complicated, to ensure safety and trustworthiness in the system, system designers, maintainers and operators need to "peer through the black box", i.e. AI systems need to be transparent and understandable [28]. Sometimes, these characteristics can be achieved by including the "human-in-the-loop", so called HIL methods. Such methods could include active or mutual learning [29]. In the former highly uncertain samples (according to the algorithm) are sent to a human operator to label and are given increased importance in subsequent model training rounds, so that the AIs behaviour is improved. For these ideas to be applied successfully in real-life production contexts, the gap between expert and non-expert users needs to be bridged. This can be achieved through human-friendly intuitive interfaces, such as for example spoken dialog systems that help the users interact with intelligent machines easily while carrying out their task without additional burdens [30].

These challenges and their solutions tie in to the broader concept of Industry 5.0 aiming at adapting the efficiency gains of Industry 4.0 to advance the sustainability and human-centricity of the industrial process. A prime example is the concept of Operator 5.0 [31], where operators are envisioned to work alongside intelligent systems that help them complete their tasks, while guarding them from mental and physical stress. Such a goal can be achieved through the combination of Industry 4.0 technologies, such as wearable IoT devices measuring physical stress, together with Industry 5.0 AI systems that formulate production planning collaboratively with humans, taking into account human mental and physical fatigue.

As a further step towards worker well-being, Industry 5.0 aims to address the needs defined in the Industrial Human Needs Pyramid [32], which among others includes the building of trust between humans and machines together on top of workplace safety, leading eventually to worker self-actualization in a supportive environment. Special robots named cobots , or collaborative robots, have been built for this purpose and represent a tangible example of human-machine collaboration. These robots share physical space with human workers, sense and understand their presence, and can perform tasks independently, simultaneously, sequentially, or in a supportive manner [33].

Such semi-autonomous machines can take over physically frustrating and repetitive tasks, while humans can move to more open-ended tasks - and more conducive to human fulfilment - that tap into their critical thinking, creativity and interdisciplinary problem-solving skills. Physical and mental frustration can very often be sources of human error, thus leaving highly repetitive tasks to intelligent machines such as robots and cobots can also reduce waste and cost making the manufacturing process more sustainable [34].

Viewing Industry 5.0 from a systemic viewpoint, around the technologies needed to facilitate, three pillars main pillars can be identified as the main driving forces behind the design and development of these technologies, namely: *Safety*, *Trustworthiness* and *Human-centricity* [2].

Safety: Refers to both mental and physical well-being in the workplace. This includes technical challenges such as making sure AI decisions in cyber-physical systems are reliable and controllable, without unexpected responses due to out-of-distribution or maliciously targeted inputs (cyber-attacks) to endanger worker well-being as well as the correct functioning of the production process. Under this umbrella fall also the enhancements to worker well-being through AI systems mentioned previously.

Trustworthiness: Trustworthiness, though closely connected with safety, is relevant to the controlability as well as the perception of controlability by the AI systems' users. Techniques such as explainable AI (XAI) and Active Learning (AL) are key in fostering trustworthiness and encouraging the widespread adoption of AI in manufacturing production lines. This means visualizing the AI's decision process to human operators so that they can understand the reasoning behind a decision (e.g., through feature importance scores) and also providing interfaces where users can help improve AI decisions (e.g., by providing the correct label for mislabelled inputs).

Human-centricity: The aim here is to design intelligent systems with human needs, competencies and desires at the center in order to promote a healthy work environment where workers can be productive and thrive at the same time [35].

In summary, the transition from Industry 4.0 to Industry 5.0 marks a significant shift in manufacturing, focusing on a more human-centric approach. A key challenge is ensuring that advanced technologies, such as AI and robotics, enhance rather than replace human capabilities, maintaining a balance between automation and human input. Developing intuitive, user-friendly interfaces is crucial to enable seamless human-machine interactions, ensuring accessibility for all workers, regardless of age, gender, or education. Through systematic planning and implementation, Industry 5.0 can leverage the strengths of both humans and technology, offering numerous opportunities to revolutionize manufacturing and create a more sustainable and people-focused industry.

1.1.2 Data Lifecycle in Industrial ML Applications

From its artisanal origins in the Pre-Industrial era, manufacturing has evolved and adopted many forms through continuous technological innovations both in the physical and lately also in the digital domain. The increasing importance of data in the production process has followed the same trend, making modern industrial processes more optimized, tightly controlled and sophisticated than ever. In the pre-industrial years manufacturing products were handcrafted and produced on demand, which was usually small, with knowledge being passed from one generation of artisans to the next or shared inside guilds. It was not until industrialization and mass production arose, that manufacturing processes came under closer and more scientific scrutiny. Starting from the monitoring of a large workforce and the need to predict and meet mass demand, historical data began being recorded on paper [36]. In the mid-20th century, the need to adapt to a more complex, competitive and globalized economic environment intensified data-oriented efforts

with statistics-based production management and operations research (e.g. demand prediction, inventory management, intelligent sampling for quality assurance, floor layout and process optimization, machine failure rates, supply chain optimization). A further amplification of the importance and amount of data came about with the information age, where computers were used to systematize the above processes through different systems and paradigms such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Supply Chain Management (SCM). Additionally, production and design simulation tools (e.g., CAD) and increasing machine automation together with the gradual introduction of industrial robots gave manufacturers the capability to meet customer demand with higher quality and speed at a lower cost [37].

Reaching today's age, the wide proliferation of technologies such as Big Data and Artificial Intelligence in the so-called "Smart Factories" have become evident [38]. In a manufacturing context, "Big Data" refers to large amounts of heterogeneous data produced from multiple sources throughout the lifecycle of a manufacturing product. It can also be characterized by the 5Vs [39]: Volume, Velocity (how close to real-time is data acuisition and processing), Variety (multitude of sources), Veracity (of how good quality is the data) and most importantly Value, which reflects the impact of data utilization for desired business outcomes. This data typically originates from a variety of different sources and can be classified accordingly into categories. Management Data is usually collected by information systems such ERP or CRM and is mostly related with areas such as inventory management, demand forecasting etc. This data is usually stored inside the individual databases for these systems. Equipment Data on the contrary gets collected by IoT devices and is used for monitoring operating conditions or production equipment performance. IoT technologies are also sources for Product Data, which can include context of usage, environmental conditions of operation and biometric information of the user. Finally, User Data and Public Data can be gathered from a variety of widely available APIs and datasets. The first relates to user preferences and can be found in various well-known e-commerce and social media websites, while the second exists in public (e.g. government) datasets and can contain information such as industrial regulations and standards [36].

The sudden availability of such vast and diverse data presents unique opportunities as well as challenges for manufacturing businesses aiming to adopt big data technologies into their business model. Before analyzing these it is worth diving into detail about the different processing and transformation phases that need to be applied to manufacturing data to derive the most business value out of it, namely the Data Lifecycle [40]. Typically a manufacturing data lifecycle consists of the following phases: data collection, transmission, storage, processing, visualization, and application. Data Collection occurs mainly through different IoT devices (e.g. smart sensors, RFID) placed either on the product or the equipment and set to monitor their health status and performance. Additionally, wearables can be used to monitor employees' bodily and mental health status. This phase also includes the collection of user and public data through different APIs or web crawling, as well as management data provided by ERP or SCM systems. After collection, the data is typically stored either as structured, relational (DB tables), semistructured (XML, JSON,

graphs) or unstructured (multimedia, documents) data. Cloud computing technologies play a big role here, making it easier to provision cost-effective, scalable and elastic storage to meet heightened requirements for data velocity and heterogeneity. Above the Data Storage layer lies the Data Processing layer, which aims at the extraction of knowledge for successful business utilization of the data. Initially the data is preprocessed, cleaned and reduced, meaning that duplicate or redundant data is removed, missing values are removed or set to a default, low quality data is filtered out etc. After the data is ready it is processed by the analytics algorithms which usually include Machine Learning (regression, SVMs, Neural Networks, time-series methods) and Data Mining (clustering, association rules, anomaly detection) techniques, often applied at scale in a distributed to system to make the most out of the available data. The clear communication of processing results to the end users is carried out in the Data Visualization phase with the assistance of various graphs and charts as well as virtual reality technologies and smart terminals for real-time data. As most of the processing, preprocessing and storage of data is performed on top of large scale distributed systems that could entail substantial complexity such as federated clouds or fog architectures, Data Transmission can be identified as an important and distinct phase of the data lifecycle. It includes reliable and efficient techniques for transferring large amounts of data with different formats and characteristics across diverse network and computational components. At the very end of the data lifecycle are the Data Applications which in addition to providing insights into the data and the results of its processing, also drive a great deal of automated decision making. These applications can be useful during different manufacturing processes, such as data-driven product design, forecasting and deman analysis, quality control, equipment supervision, failure detection and predictive maintenance.

The recent explosive growth of Big Data and the complexities and challenges in efficiently utilizing it throughout its manufacturing-specific life cycle have led to a number of initiatives that promote the proliferation of Big Data and IoT technologies in the industrial sector such as Industry 4.0 in Europe, Industrial Internet of Things in the US, and the Made in China 2025 [41]. These initiatives aim at providing guidelines for encouraging the easy adoption of these technologies, especially by SMEs, and creating frameworks for interoperability and cooperation across companies and related industry sectors. A notable example is the creation of reference architectures such as IIRA and RAMI 4.0 [42] that serve as common abstracted templates for building problem specific architectures with a strong focus on easy integration and interoperability. There have also been efforts to outline the future research directions and challenges in smart manufacturing, the most characteristic such effort in the EU being carried out by the BDVA. In their 2018 [43] and 2020 [44] reports they identified several key research directions relevant to Smart Product Lifecycle, Smart Supply Change Management and the Smart Factory.

Smart Factory research challenges are further split into Data Management and Lifecycle, Data Processing Architectures, Data Analytics, Data Protection and Security and Data Visualization challenges [43]. Regarding the management and lifecycle of data, the integration of diverse cyber-physical systems and the availability of hererogeneous data produced at different rates are primary concerns. So is the semantic interoperability of

automation systems, usually achieved by the use of ontologies, aiming at the creation of a collaborative information sharing environment. Data annotation is also a relevant direction, it can be performed either on-the-fly or as a seperate processing step and it is worth investigating different ways in which data (e.g., from sensors) can be put in the right context (e.g., mapped to a specific product). Handling missing data is another common issue in the data lifecycle, since for example sensors might be off or fail for a certain period of time. The challenges of data processing architectures are mainly focused on where and how data-intensive computations will be performed. Available choices could be edge servers, HPC infrastructures, clouds or federated clouds depending on requirements such as performance, data confidentiality and a company's limitation of affording or getting value out of computational equipment. Data analytics is probably the richest category in research directions to pursue. Prescriptive maintenance is the enhancement of predictive maintenance in that it tries to discover the causes of failure in a data-driven manner, instead of just predicting them. Such methods can also be used to assist decision making at the management level through parametric analysis of business KPIs and their corresponding risks. Modern ML techniques such as Deep Learning also play a central role with applications in anomaly/fault detection and classification and quality inspection. These can be further enhanced by investigating new patterns of data-human interactions and also by moving some of the processing into embedded systems close to were the data is produced (e.g., to gain more specialized insights into fault occurences for one specific machine). Finally a large chapter of data analytics is simulation and digital twins. Data-driven simulation models can create better opportunities for experimentation and optimization of different production line/machine/cell configurations, which can be made even more accurate through the provision of real time data by digital twins. Of course in a complicated data-rich environment it is only natural for security and privacy concerns to arise. The variety of communication protocols in IoT systems as well as the cyber-physical aspect make smart manufacturing infrastructures not only vulnerable to attack, but also enhance the impact of the attack which can now have impact in the physical production line. Additionally the increasing reliance on data opens up avenues of data corruption or malicious manipulation aiming to derail AI and data-driven models and processes. Another important direction is to establish firm guidelines and protocols about access and privacy of sensitive data and also apply anonymization in a reliable but non-instrusive manner. Last but not least data should be clearly and intuitively presented to all interesting stakeholder, a concern of the Data Visualization domain. The first category of stakeholders, the workers, should be able to obtain context specific visualizations (e.g, when performing remote maintenance with the help of virtual reality) and could also be helped in their work by natural language interaction interfaces powered by NLP. Simulation and smart training environment can also aid their training and familiarization with the modernized smart factory processes. Managers and decision makers also need better visualization tools to understand the data produced in the smart factory as well as the AI-based decision making process to spot patterns and gain deeper insights into the factory's processes. The integration of heterogeneous data and its presentation through common visual interfaces can be combined with data navigation and annotation

techniques, to achieve bi-directional learning between stakeholders and smart processes through continuous feedback.

1.1.3 Challenges of Real-life Industrial ML Applications

Machine Learning results have been impressive in different research scenarios and have been successfully used in commercial applications, most notably recommendation systems and chatbot assistants. However, when it comes to model decisions influencing happenings in the physical world the requirements become much stricter as the margins for failure are required to be minuscule. These new constraints placed upon industrial ML applications create a new set of research challenges mainly focused around the following areas: *Adaptability and Scalability*, *Data Availability*, *Data Privacy and Security*, *Safety*, *Human-AI collaboration* and *Ethics and Compliance* [45] [46].

Adaptability and Scalability: ML applications need to be successfully deployed and integrated into larger cyber-physical systems. For instance Industrial Internet of Things (IIoT) systems often have a requirement for data to be processed in real-time as it is gathered from sensors placed on different parts of the production line. This means that models with fast inference times should be chosen and sometimes that these models need to be deployed at the "edge", closer to where the data is produced. Accommodating such pipelines is especially hard in industries that rely on legacy systems and lack the infrastructure to host real-time data-processing frameworks and state-of-the-art Deep Learning models with high memory and GPU requirements.

Data Availability: While most research results feature clear-cut benchmarks and well-defined datasets, collecting enough and high-quality data in real-life environments remains a challenge. A typical application where this becomes an issue is visual quality inspection, where the collection of images is an expensive process requiring a precise setup that adjusts for lighting differences and keeps precise distances and angles to produce a homogeneous dataset. Additionally the collection of images of product defects is hard when defects are rare, needing many production cycles to complete until enough data is collected. Another typical example is the collection of real-time data for production planning where concept drift and catastrophic forgetting can lead to erroneous AI decisions [47].

Data Privacy and Security: Especially when it comes to worker wellbeing monitoring, industrial ML application will need to collect and process potentially sensitive data, it is important therefore for the appropriate data anonymization and privacy safeguarding techniques to be employed as well as to monitor adherence with the multiple regulatory frameworks. Security also becomes a risk, not only due to sensitive data being at risk of being stolen or tampered with, but also the deployment of ML models risks exposure to a variety of novel attacks such poisoning or inversion attacks. In these examples a malicious adversary tries either to tamper with input data to lead the model to wrong decisions (that translate to the selection of wrong

actions in the cyber-physical system) or to infer sensitive data from a model's output by "inverting" its inference process. [48]

Safety: As previously mentioned, AI models whose decisions influence the physical world have the potential to put human well-being at risk as well as the physical assets of the factory and the production process. Therefore it is important, apart from accuracy metrics, to also consider robustness to inputs that originate from a dynamic real-life environment and which might not always align with inputs viewed by the AI during training. A useful technique that can be employed here is simulation, either through virtual environments for reinforcement learning [49], or through the targeted production of synthetic data that reproduces realistic out-of-training-distribution scenarios.

Human-AI collaboration: Here the aim is twofold: the first part is to help human decision makers and operators trust AI decisions (when it is beneficial to trust them of course) to achieve wider AI adoption. The second part is to create AI systems that do not aim to replace humans (something that can often be risky, or even impossible), but that work synergistically with humans and can combine machine precision and consistency with human open-ended thinking and common sense.

Ethics and Compliance: As AI systems become widespread in modern industry, a number of regulatory frameworks have been created addressing ethical issues such as bias, fairness and accountability. Prime examples are the European Union's AI Act, which categorizes AI systems according to their risk and accordingly imposes different levels of human supervision. [50] Similar regulations have been introduced in Canada, the United Kingdom and China. [51]

While all of the above challenges touch upon this thesis, our main focus will be on the (lack of) *Data Availability* as well as the *Safety* of AI systems used in visual defect recognition.

1.2 Automated Visual Quality Inspection in Manufacturing

Quality evaluation is an arduous and repetitive task that would greatly benefit from automation. Here we focus on visual inspection as it is a common usecase and a good example for showcasing the effectiveness of machine learning models. The benefits of an automated approach are that it can be a scalable and elastic form of non-destructive testing, able to adapt to production volume fluctuation more easily than a fixed number of human workers. Additionally human error phenomena such as inspector-to-inspector inconsistency are largely reduced, providing a more objective and consistent criterion for product quality [52]. Especially given the fact that several modern Deep Learning techniques already achieve higher performance in computer vision tasks [53].

Automated visual quality inspection has been achieved both by supervised and unsupervised learning methods. The latter are quite common since their independence from

labelled training data makes them an attractive choice. Often, however, at least the class of non-defective products needs to be labelled i.e. a semi-supervised scenario. The ideal would be, of course, to use supervised learning with a fully labelled dataset. Due to the arduousness of the labelling process and the low volume of defects in production, such labelled datasets are often small in size and suffer from class imbalance. Supervised methods, nevertheless are better able to discriminate between small differences between defective and non-defective products and can also discriminate between different types of defects [54]. Ultimately the choice will be influence by an organization's capability to collect and label the required data and the types of defects, as for example functional vs. structural vs. cosmetic defect might require different types of ML methods.

Initial attempts at automated defect detection focused on the "hand-crafted" extraction of features. These were mainly computer vision methods, usually coupled with a simple classifier like Random Forests or an unsupervised method such as clustering. For instance [55] used a variation of Otsu's method for adaptive threshholding to detect abnormalities on surface areas (e.g scratches, cracks etc.), achieving quite high accuracy for ceramic and metallic surfaces. In another case [56], phone screen defects were searched for - a harder problem due to the high gloss of screen surfaces. A full pipeline was utilized including image alignment, normalization across noise and lighting conditions and fuzzy c-means based anomaly detection as a final step. A third alternative is edge detection which can be achieved by using the wavelet transform [57] to sharpen edges on wood surfaces and extract defects that disrupt the edge continuity. There is a long list of these methods in the literature, of which the aforementioned ones are only examples. However we can already see a common pattern, namely that these algorithms require a lot of effort to design as they have to normalize the image under different transformations to be able to effectively compare defective and non-defective images. They are also specialized to specific types of surfaces. For example, [55] shows a much lower accuracy in detecting liquid surface contamination, as its method is not designed to cope with intense light reflection.

To satisfy recent industry trends such as part customization and agile manufacturing, quality inspection methods need to be much more versatile and adaptable. This is the reason why the interest of the research community has shifted towards the use of Convolutional Neural Networks (CNNs). These circumvent the need for "hand-crafted" feature extraction algorithms as their convolutional layers can be trained directly from data to extract distinguishing features, often managing to achieve invariance against various conditions such as rotation, translation, lighting and noise. Techniques such as Transfer Learning also offer the opportunity to reuse knowledge from large pretrained networks on smaller never seen before datasets. The literature on CNNs for defect detection includes various different approaches roughly divided into two categories: segmentation-based and one-off classification.

The aim of image segmentation methods is to first extract simpler candidate defect areas from a complex image to be used as features for a classification layer. What the classification layer consists of could range from a dense neural network to a simpler random forest classifier or even to a semi supervised method such as S4VMs. A common

network architecture for segmentation – given the availability of pre-segmented training data - is the U-Net used in [58] to detect small defects in radiographs of aerospace welds. In that study, instead of using the default dense layer for classification, a Random Forest classifier is chosen to convert the segmented image areas to pixelwise probabilities of a defect. A final step includes filtering of the candidate regions using Maximally Stable External Regions (MSER) [59] and thresholding. This substitution of the last layer with the simpler (and less overfitting prone) Random Forest was necessary due to the small amount of defect images in the data – a recurring problem in this defect detection datasets. Another approach in the automotive parts domain tries to create a single DNN similar to One-off methods, but stacking segmentation before classification layers [60]. The aim now is not a binary choice between defect and non-defect but the identification of specific defect types observed from different views (top, bottom, side) of the image. The topmost custom segmentation CNN is followed by a "refinement" network, which performs density slicing, filters the candidate areas and produces the classification output. The resulting network manages to achieve good results on all defect classes with >95% accuracy and F1-scores close to 50%.

A dilemma faced in such computer vision tasks is whether to use transfer learning, i.e. a pretrained model that is fine-tuned on the dataset at hand or a custom model architecture specific to the problem and trained from scratch. In a printing industry usecase described in [61], developing a custom shallow CNN model gave better results, however, significant effort was required to produce a homogeneous training set, especially in terms of lighting conditions. Transfer learning can also be useful for complex products such as vehicle parts, where quick retraining and - to some extent - independence from inputs are highlighted. For instance, [62] achieved best performance on a dataset of vehicle parts by utilizing a pre-trained VGG16 model further fine-tuned on the quality inspection dataset.

Still, one-off classification methods tend to avoid any preprocessing overhead and produce an assessment from just a single image. As a trade-off a larger training dataset is needed for achieving acceptable results. A good example is [63] where a custom CNN is used together with data-augmentation to predict different defect classes that appear on steel strips. The CNN consists of 6 convolutional layers with max pooling along with 2 dense layers leading to 7 output categories, 1 for "non-defect" and 6 defect categories. Although some initial preprocessing was included in order to isolate the part of the image containing the examined surface, this can be viewed as a data quality adjustment. In another usecase, rail defect detection was tackled by an one-off classification approach in [64]. Different custom architectures were compared, with the best performing one consisting of 3 convoluational layers with max pooling and 2 dense layers, mapping to 6 defect categories including "non-defect". More recent approaches have tried sophisticated combinations of methods such Long Short-Term Memory Networks (LSTMs) over pre-extracted CNN features to detect debris in avionic component ducts [65]. The object detection model, YOLO v7, was used for example to detect defective packages to be extracted from transfer pipelines in shipping [66].

Among the various metrics used to measure classification performance for defect

detection the two most important ones are recall and precision. Recall should be given the most attention since false negatives will lead to undetected production defects. Precision, while secondary, is also important since false positives will require unnecessary human inspections, a large number of which defeats the purpose of the automated QA approach. Common metrics based on these two criteria and used in the literature are the F1-score and the Area Under the Curve (AUC). These metrics are most meaningful when separating the normal from the anomalous categories, as the differences between defect categories are usually less consequential.

Choosing the right CNN-based pipeline for defect detection depends on numerous factors. First and foremost is the availability of labeled data. Segmentation-based methods might seem to need less data at the cost of a more sophisticated pipeline, however labeled segmented data is harder to find and more expensive to create. On the other hand one-off methods need more training examples and suffer more from data imbalance (e.g. in the case of a rare defect). Techniques such as data augmentation either through applying predetermined transformation on existing training data or creating new synthetic data can help ease the disadvantages for both cases.

1.2.1 Data Scarcity

While CNNs are indeed performant and flexible they do require large training sets, ideally with many samples both from the "good" or "flawless" products and the defective products as well. In reality, however, collecting this data is often prohibitive due to various reasons, such as high cost, lack of time or manpower or lack of a scalable automated setup. A large portion of modern quality inspection research focuses on mitigating this issue through techniques such as transfer learning, active learning, few-shot learning, oversampling and the generation of synthetic data for data augmentation.

Transfer Learning

Transfer learning as mentioned in the previous section is a technique for using large models, that have been trained in large generic datasets such as Imagenet [67] in environments with enough computational resources (i.e. clusters with multiple CPUs and GPUs). These models are then reapplied to smaller datasets in different ways, that usually involve targeted readjustments of the base model's weights. The simplest way to use transfer learning is to utilize a part of the original model's architecture, with frozen weights, as a feature extractor to avoid costly feature engineering [68]. One can also unfreeze the base model's weights, all or from selected layers such batch normalization layers, and perform end-to-end learning with a ready-made architecture and starting from a "good" weight initialization [69]. Other flavours of transfer learning are domain adaptation and domain randomization, both of which have had successful applications in vision-based reinforcement learning.

Domain adaptation is a set of techniques that help a learning model generalize to a target domain while trained with samples from a different source domain. In the case of robotic grasping, simulation is the source domain and the real production line is the

target. Domain adaptation is widely used in computer vision and can be roughly distinguished into two categories: feature-level and pixel-level. Feature-level is usually based on adaptive feature extraction methods such as CNNs, which already have some degree of transferability between the simulation and reality domains. Also including a domain-level similarity metric such as maximum mean discrepancy in the loss function when retraining in the new domain can help enforce domain invariance [70]. Pixel-level domain adaptation is mainly based on using GANs to restyle simulation images so that they look more similar to real ones [71]. Both of the above techniques can work well on Deep Reinforcement Learning algorithms that base their perception and action planning on CNNs. A good example is GraspGAN [72] which uses simulation with a hybrid adaptation method, combining Domain Adaptation Neural Networks (DANNs) with a novel batch-normalization technique. The proposed method achieved comparable or better performance to vanilla DRL with 50 times fewer real-world samples.

Domain randomization methods have also shown good results for vision-based tasks such as robotic grasping, making simulation-only training feasible. The goal is to train the agent in a wider set of environmental conditions by introducing randomization in the simulated environment at training time. Given that the variability of the conditions is sufficient, the model trained in simulation will be able to generalize in the real world. For instance [73] uses randomization on the following types of features: addition of distractor objects of different shapes and sizes, object position and texture, texture of background objects, camera position, orientation and field of view, number and position of lights and addition of different types of random noise. The trained model produced comparable results to real-world training, even though no real-world data was used

Active Learning

Active learning is applicable under the precondition that some labelled samples exist and that there is a human operator that can help with the learning process by manually labelling pre-selected instances, which the model is highly uncertain of [74]. This leads to a training process consisting of training-labelling-retraining cycles that is very dependent on the quantification of the model's uncertainty over a specific data instance. The role of this quantification is to reduce the amount of manual labelling as mush as possible. Different strategies of selecting the most informative instances for manual labelling have been suggested including uncertainty sampling, representativeness sampling and sampling of adversarial instances [75]. There are a few successful use cases of active learning in manufacturing such as [76], where a training database of samples was continuously enhanced through selective manual labelling during the visual inspection of printed circuit boards and in the prediction of displacements between chip layers [77], a highly sensitive process where manual measurements can be disruptive and should be minimized. Active learning can also be combined with other techniques against data scarcity such as data augmentation. Synthetic data generation was used to reduce the expenses associated with data collection, combined with feedback from active learning regarding the desired characteristics of data that benefits the model the most, to detect defects in a

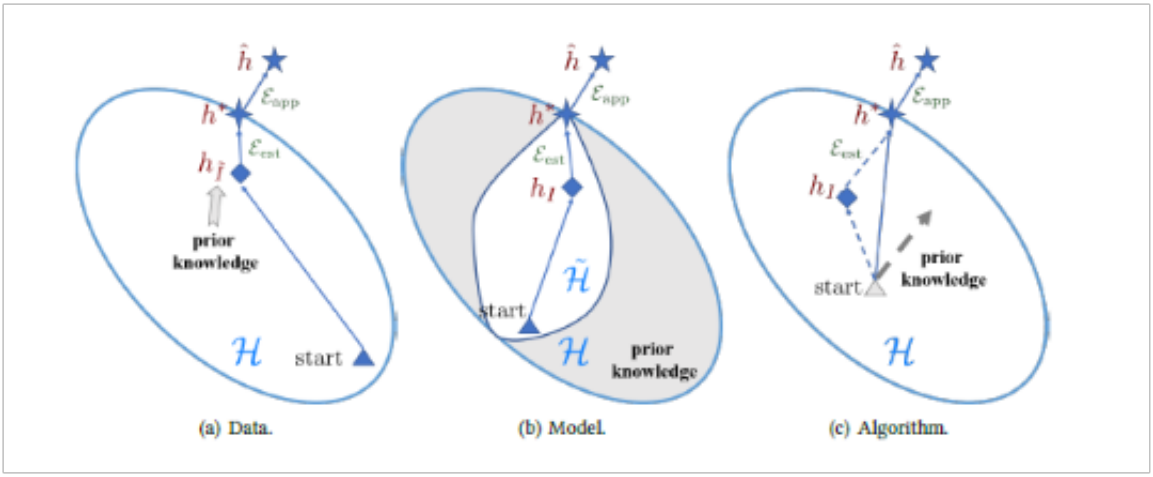


Figure 1.2. *Main categories of few-shot learning methods [20]*

dataset of shaver shell prints [26].

Few-shot Learning

While transfer learning manages to reduce the high data requirements for Deep Learning Models such as CNNs, typically to hundreds of examples, there are cases where the data available is even less. This is especially prevalent in automated quality inspection where production defects are rare and some times minority classes remains at the tens of samples. Few-shot learning is a set of machine learning techniques for dealing with these low data scenarios [20].

Few-shot learning works on three levels, namely the data level, model and algorithm levels. Data-level methods aim to augment the data, usually through synthesizing novel samples with GANs or through applying graphical transformations to existing samples. Model-based approaches are very similar to transfer learning in that they try to constrain the hypothesis space (tunable model parameters) by using prior knowledge (e.g. pre-training the model on a similar but more general problem and freezing some of its weights, while leaving some to be fine-tuned on the small dataset). Finally, algorithm-based methods concentrate on incorporating prior knowledge to the search strategy for optimal parameters. The above categories are more formally illustrated in Fig.1.2.

$\mathcal{H}$, as depicted in Fig.1.2 is the hypothesis space, or space of the family of models (e.g. all CNNs of a specific architecture). The optimization algorithm moves through this space by learning better and better parameters moving from "start" to h_l (note that h_l is dependent on the training dataset), which represents the final learned parameters. ϵ_{est} is the estimation error due to learning inefficiency (e.g. overfitting) and ϵ_{app} the approximation error, due to the limited capacity of the hypothesis space. What FSL is trying to do is bring "start" closer to h^* faster than full model training. For example model-based techniques such as transfer learning try to constrict $\mathcal{H}$ to $\mathcal{H}'$, a smaller hypothesis space learned from another similar problem with a high chance of including h^*. On the other hand the "algorithm" category tries to use prior knowledge over the learning rate and direction of the optimizer so as to decrease the number of model updates. With data

augmentation, which is our main focus, we are trying to improve the accuracy gained by the model by adding additional samples and bringing the final stage h'_l of the training closer to h^*.

An approach of few shot learning that differentiates it from transfer learning is meta-learning [78]. Usually meta-learning approaches are based on learning a distance metric between classes making them a hybrid between supervised and unsupervised methods. More specifically, metric-based approaches learn a mapping to an embedding space where instances of the same class are mapped close together while instances from different classes are further away, thus roughly forming a cluster for each of the classes. After the mapping a simple nearest neighbors classifier can be used to determine the class of a new instance. A typical example of this method are prototypical networks [79].

FSL also includes optimization based methods where gradient step sizes for example are imported from another problem where training data is abundant. These methods usually work in two stages, the meta-learning and the task-specific learning. The meta-learner model could be trained on a different task, or a set of tasks that cumulatively have enough samples. Thereafter it updates the parameters of the task-specific learner, which is then fine-tuned on its task of focus. Typical examples are model-agnostic meta-learning (MAML) [80] and Reptile [81].

Finally, there is another category of methods that are worth mentioning which do not attempt at all to rely on prior knowledge. Instead, they attempt to build models with architectures specific for fast learning, such as memory-based architectures [82] and rapid-adaptation architectures [83].

Various of the above types of methods have been used in the context of Visual Quality Inspection. For instance, relation networks, a method similar to prototypical networks, was used on top of pre-extracted CNN features to detect defects on bar surfaces [84]. For the pre-extracted features, attention modules were employed to make surface defects more salient before passing on to the relation networks. MAML was utilized in [85] to detect bearings defects, by treating the detection of different defect categories as different tasks for the meta-learner. A low-parameter model - Resnet-10 - geared towards faster learning was investigated in [86] to detect defects in lithium batteries. To boost the size of the inputs and ensure enough diversity in the data various data augmentation techniques were also included in their pipeline. Protypical networks were used both in [87] and [88] for fabric and auto-part defect detection respectively. In the former they were combined with class activation mapping to enhance the contrast of defect locations, while in the latter they formed part of a custom network with attention mechanisms used for the same purpose. Attention mechanisms were also part of the approach suggested in [89] for manufacturing defect detection using the MVTEC-AD dataset [11]. This time the classifier was a Siamese Network with pair-balanced contrastive loss to account for the class imbalance between defects and non-defects. [26] compares prototypical networks with and without data augmentation vs. supervised approaches in low data scenarios with manufacturing components from Philips Consumer Lifestyle BV and Iber-Oleff - Componentes Tecnicos Em Plástico, S.A. and finds them competitive. In addition, few shot learning was enhanced with different sampling strategies for creating and labelling

the initial support set.

Oversampling

Contrary to few-shot learning where the whole dataset is small in size, what is more often the case in visual quality inspection in manufacturing is that the defect classes have very few instances. This is a natural byproduct of a production process working mostly as expected and rarely outputting defective parts. On the other hand, this makes the automation of the process harder by complicating the data labelling process and making it more costly in terms of time, i.e. too many products will need to pass through the production line before sufficient defective products have been collected. This can be mitigated through oversampling and its extension, data augmentation. We differentiate between the two as oversampling happens on the feature level, while data augmentation on the input data/image level. For oversampling to be applicable, dedicated feature selection methods should be in place or instead pre-extracted features using transfer learning can be utilized to obtain lower dimensionality feature vectors.

Imbalanced data is a common issue in many other domains such as tumor classification or security attack detection, that are closely related to anomaly detection and where the minority class is usually of much higher importance to predict correctly [90]. Due to its ubiquity, a number of methods have been developed to cope with the imbalance at the feature level. The two most well known ones are SMOTE [91] and ADASYN [92]. In SMOTE pairs of minority class instances are connected with line segments and over these segments new instances are sampled so as to hit a target that will make the dataset balanced. A few variations of SMOTE have been then introduced such as Borderline-SMOTE [93] that try to focus sampling near the classification boundary, sometimes also oversampling edge instances from the majority class to create more refined boundaries. ADASYN is a more sophisticated extension of the same idea, where high-uncertainty samples are those near the boundary or in sparsely populated regions. These instances are then perturbed to produce synthetic instances between them and their closest neighbours.

While oversampling methods have existed for a long time and have proven themselves in different applications, when it comes to image classification with very high-dimensional inputs they are not effective and require the use of feature-extractors. Another idea examined in the coming section is to produce synthetic instances at the input level using modern techniques such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs).

Data Augmentation

Moving oversampling to the image level can take many forms, the simplest of which is to use simple graphical transformations. The emergence of sophisticated deep generative methods, however, such as GANs and VAEs can bring enhanced capabilities by approximating the true distribution of input images and therefore managing to generate high-fidelity outputs [94]. The major issue here is that GANs and VAEs are even more data-hungry than traditional deep learning methods. Nevertheless, even if training from

scratch is not possible, transfer learning can come to the rescue, this time in the context of generation rather than classification. This solution has been explored in current research and although traditional weight fine-tuning was not enough for small datasets [95], the fine-tuning of batch normalization layers only produced promising results on BigGAN [7]. While not perfect, the resulting images retained many useful features for classification. Further fusing with original raw images also showed some usefulness in a few-shot learning scenario [8].

Applying data augmentation to automated visual quality inspection can be tricky as defective images are very much alike to non-defective ones making high-fidelity generation challenging. Despite that there have been promising applications in the manufacturing domain. For instance, in a dataset of shaver shell prints, [52] utilized Lightweight GAN [96], a low resource GAN, to generate high-fidelity augmentation images and improve the AUROC score of the final classifier. In [97] data augmentation was utilized in a different way, namely samples were augmented by outputs from an unsupervised anomaly localization classifier in the form of heatmaps highlighting potential defects. These augmented samples were classified with higher accuracy without the need for additional rebalancing methods.

1.2.2 Robustness and Trustworthiness of AI Visual Inspection Systems

A further issue that the adoption of AI systems in manufacturing settings faces is the real and/or perceived lack of robustness. Industry decision makers and regulators are often sceptical of adopting AI systems in physical environments a they appear as black boxes, which nobody know if they suddenly come up with a very unexpected and potentially dangerous decision. This scepticism is of course not without merit and it is important for the research community to come up with techniques that i) shed light into the inner workings of AI systems and ii) attempt to make these systems more robust. Although the subject of this thesis is more related to data scarcity and the robustness issues that might result from this particular cause, ideas such as Explainable AI (XAI) and techniques to mitigate cyber-security attacks against Deep Learning algorithms offer many insights and similarities with techniques aiming at making AI systems more robust in general.

eXplainable AI (XAI)

As state-of-the-art Deep Neural Networks are starting to surpass human ability in various specific tasks, their complexity (i.e., number of layers, parameters, complexity of the loss function) increases to such an extent that they, especially to non-experts, become black boxes. Interpretability or explainability of Deep Neural networks is the ability to provide insight into the inner workings of a DNN in a human understandable form [98]. A variety of XAI methods have been developed for computer vision that could be applied to an Industrial Visual Quality Inspection setting. Usually these methods involve Post-Hoc explanations, focusing on the reasoning behind the decisions made by the model for specific instances rather than trying to explain the whole model [99]. One way to achieve

this is through perturbing certain input feature to gauge their impact on the final output.

Prediction Difference Analysis (PDA) [100] and Meaningful Perturbations [101] are prime example of such methods, for instance replacing parts of an image with constant values, noising or blurring the image or regions thereof to measure changes in activations and/or classification scores. Local Interpretable Model-Agnostic Explanation (LIME) [102] is a popular model-agnostic method that has also been extended to images. For a particular instance it generates local perturbations and trains a simple self-interpretable local model (e.g. a decision tree) on the local perturbations and the original model outputs. Results from these explainability methods on images are often represented through salience maps, where the brightness of a pixel is dependent on its importance for the production of the specific output. Saliency maps can vary from method to method on how they localize high-salience reqions and measuring the quality of saliency maps is an open research question [103].

One might attempt to gather more insight from a model by considering knowledge of its architecture as a given. This is what techniques such as Deep Learning Important FeaTures (DeepLIFT) [104] and Class Activation Mapping (CAM) [105] try to achieve. In DeepLIFT, activation differences from a network's final layers are backpropagated similar to gradients to match class activations to important parts of inputs or input features. CAM used the fact that convolutional layers very often make objects present in the image more salient and their local outputs can be pooled together to extract regions of the image that highly contribute to the class prediction. Layer activations are combined with gradient information in GradCAM [106] to produce better localized explanation regions.

In the context of data augmentation saliency maps and heatmaps generated from the above methods can prove useful. In [97], heatmaps from an explainable semi-supervised defect localizer were combined with raw inputs to improve classification performance in a setting with data imbalances. However, the extracted information from explainability methods can also be used to improve robustness, as for instance in [16] where images with masked super-pixels from the LIME method made the network more robust to poisoning attacks.

Sensitivity to Small Differences between Inputs, Safety and Security

The fact that Deep Convolutional Neural Networks (DCNNs) are considered black boxes is most clearly illustrated in [10], where small corruptions in the image inputs can lead to mistaken predictions. This fact gains in importance when adversarial attacks against neural networks are also considered. Adversarial corrupted images that are almost identical to real ones can be created to induce wrong model decisions - also known as poisoning attack [107]. While several techniques have been suggested to mitigate this kind of attack, such as Gradient Masking [108], Robust Optimization [109] and Adversary Detection [110], XAI also plays a significant role in identifying these attacks and defending against them. For example, Similarity Difference And Uniqueness method (SIDU) [111] aims to provide visual explanations tailored to the detection of poisoning attacks and showed promising performance against fixation maps on datasets with noisy inputs [112]. Shap-

ley Additive Explanation Values [113] have also proven useful in filtering out adversarial inputs in conjunction with traditional anomaly detection methods. Finally an interesting approach are evasion attacks, which try to identify which features are non-robust in producing predictions and manipulating those to introduce errors [114]. Data augmentation was shown to be effective against this attack by using masked-superpixel inputs generated by LIME as a from of synthetic data [16].

Vision-based Deep Reinforcement Learning has been an area of particular interest as the stakes related to these sensitivities are high when AI agents interact with a physical environment. Most often a simulation environment is used to robustly train Deep RL agents and reduce their interactions with the real-world as much as possible [70][73]. But even for less risky scenarios such as Visual Quality Inspection these errors can be quite costly, potentially derailing the production process and causing significant material waste. For this reason the use of synthetic data can be of high value. Especially given the fact that defects occur rarely, it is very possible that some defect categories will have not appeared during the data gathering process for the training set. It is therefore important to produce simulated inputs that will train Visual Quality Inspection systems in a way that is as robust as possibly and prepare the algorithm as much as possible for the occurrence of novel unanticipated defects before it is deployed in the manufacturing environment.

1.3 Contributions and Structure of the Thesis

The structure of this thesis is built around its three main axes of contributions. Firstly, the problem of class imbalance is tackled, as it is the most common Data Scarcity issue in Visual Quality Inspection Scenarios. This is followed by an attempt to handle defect classes for which various constraints (e.g., ramp up time) did not allow the collection of any samples, and therefore the system used needs to be ready for unexpected inputs that are not in its training set. As the method developed for this is quite data-hungry and computationally intensive, to extend into smaller datasets, techniques from the emerging field of NeuroSymbolic AI were employed. The progress and contributions made along these three axes are the following:

- To deal with class imbalance a new method was developed to generate synthetic data based on examples that are close to the boundary between the "good" and "defect" classes. This method, combining the precision of oversampling techniques and the synthetic capabilities of BigGAN, managed to achieve an improvement in the recall of the neural network, while reducing the data generation time compared to other GAN-based techniques.

- To handle novel inputs, a new method based on data augmentation using StyleGAN was developed, particularly adapted to datasets with high similarity between classes, such as those encountered in industrial quality control. The new method relies on both the high-fidelity generation of StyleGAN and the ability to more accurately and meaningfully guide the synthetic data output. Also an important role was played by the filtering of generated data by quantifying the degree of disagreement

of different classifiers trained on the original data. This ensures that the artificial data represent the "open set" and can augment the initial training set sufficiently to make the final classifier more robust to novel defects during the continuous operation period. The new method showed improved results on a real dataset from manufacturing.

- Finally, as a continuation of the previous method to extend to smaller data sets in which it is not possible to train StyleGAN, NeuroSymbolic Artificial Intelligence techniques were used. Specifically, a Logic Tensor Network was used that expresses the outputs of a supervised novel input detector as symbolic rules and uses them to drive the training of a neural network. The resulting algorithm shows improved results compared to other related methods, especially in defect recall, in the sense that few defects remain undetected even if they are completely novel. Additionally, it achieves similar or better recall results than semi-supervised methods when handling new defects, but outperforms them on defects belonging to the training class distributions (closed set). Compared to other supervised methods, it maintains high performance on known defects but improves significantly on novel ones. The combination of advantages of these two types of methods is illustrated by higher F1 scores on most of the test datasets.

Chapter 2

On-the-fly Image-level Oversampling for Imbalanced Datasets of Manufacturing Defects

2.1 Background

Automatically detecting and classifying object defects is an important application of modern manufacturing AI systems that presents unique challenges, such as severe class imbalance, high inter-class similarity, and a requirement for high classification performance in real-life settings. Addressing these challenges can provide novel insights and improvements in the general context of imbalanced learning. Class imbalance is an inherent and very frequent issue in datasets of defects used for automated visual quality inspection owing to the rarity of defect occurrences in real-life processes [115]. For instance, in many modern manufacturing processes, a defect may occur in one per thousand manufactured objects making the collection of sufficient data for a balanced dataset either too costly or in the worst case nearly impossible. Even though defects are rare, the ability to detect them automatically or in a synergistic way between human and AI algorithms is of great value, since, it not only reduces costs and worker fatigue but also frees up human resources to perform more challenging, less repetitive, and more creative work [116].

Early approaches in automated visual inspection did not run into the problem of class imbalance as they mainly relied on traditional computer vision methods using pre-extracted features [117]. These methods were custom-designed using rules derived from an expert's domain knowledge and were completely unsupervised, both regarding feature extraction and rule-based decision-making, with no requirement for collecting training data. Even later, more flexible methods such as Histogram of Gradients [118] and Viola-Jones [119] relied on the extraction of custom features tailored to the problem at hand. However, since the introduction of Deep Convolutional Neural Networks (DCNNs) [22] it was made possible to achieve good accuracy scores by deriving extracted features directly from the training data. Despite the new approach requiring the collection of large amounts of data, in some cases even 10^3 to 10^4 samples, and being very sensitive to class imbalance [120], it offers several distinct advantages that have made it very popular in the current research:

1. There is little need for expert domain knowledge during feature extraction or decision-

making as DCNNs learn mainly from the data. This avoids the development of complex and error-prone data pipelines.

2. Due to this independence from domain expertise DCNNs can be more easily adapted to tackle similar problems (e.g., defect inspection of a similar but different product produced by the same organization), and can also easily accommodate new defect types, given enough training data, without change to the recognition algorithm.

3. DCNNs can easily adapt to differences in simple visual conditions such as translation and scale [121].

4. Knowledge extracted from large datasets can be adapted to smaller datasets through transfer learning, thus, coping to a certain extent with high data requirements.

In visual quality inspection, which is the focus of our work, the most frequent approach in the current literature, aimed at mitigating class imbalance, is data augmentation [122] [123]. Traditionally, image defect datasets are augmented via various graphical transformations, such as scaling, rotation, translation, shearing, blur, illumination, etc. However, those image-level transformations do not contribute sufficiently to the clearer separation between different classes, especially when the separation depends on higher-level features [124]. To overcome the limitations of traditional image processing methods, Convolutional Variational Autoencoders (CVAEs) [125] have been proposed and used successfully in a dataset of metal surfaces [126]. Generative Adversarial Networks (GANs) [127] is another important tool, which can efficiently address different kinds of imbalances such as inter-class, intra-class (e.g., person reidentification), and object and pixel level imbalances for segmentation tasks [128]. A third family of methods is based on Neural Style Transfer attempting to fuse a "style" image (defect) and a "content" image. Defects can be generated through global [129] and local [130] style transfer, using extracted defect patches and suitably placing them on the target object. However many of the above methods still require a significant amount of data (being Deep Learning methods) and may not be suitable for all datasets depending on their degree of imbalance as well as the similarity between classes that makes the generation of high-fidelity images difficult. Such methods are also usually computationally intensive requiring long training times. Nevertheless, many modern GAN architectures can be controlled through manipulation of their latent space and therefore can be suitably adapted to specific problems and potentially also made to work with smaller datasets as described in more detail in the Related Work section.

In our work, we applied data augmentation to mitigate class imbalance in a dataset of logo print images on top of manufactured shaver shells. Following our early experiments we noticed that custom shallow CNN architectures that are trained end-to-end on the dataset at hand achieved the most promising performance, therefore we introduced a data augmentation method compatible with end-to-end training. Our approach's novelty lies in using a small sample GAN introduced in [7] in a confidence-aware manner. This leads the generator to produce synthetic images based on highly uncertain training samples that lie near the classification boundary. The resulting method achieved promising

results against recent and established methods based on deep data generation or vector-based oversampling, while also retaining good computational performance by generating synthetic images on the fly.

2.2 Related Work

The current work builds upon two areas of research. The first is on using GANs for generating defects. GANs have proven very reliable in producing high-quality images and many works have managed to apply them to imbalanced and smaller datasets. We also build upon advances in assessing the reliability of neural network predictions. This line of research focuses on ways to obtain confidence estimates of the network's predictions, which we aim to utilize to bias our generation process towards low-confidence samples.

2.2.1 GANs in Defect Generation

GANs have been successfully used in many different industrial, biomedical, and other scenarios to tackle the class imbalance found in defect detection problems. The most straightforward way to use them is by training them on the same set of data as the final detector/classifier and then generating data to augment the initial dataset. A step further is to introduce customizations to control a GAN's output either through manipulation of its latent space or the influence of its loss function. A common example of the latter that is very popular in defect detection is encoder/decoder-based architectures.

Direct Data Augmentation A variety of architectures have been tried for direct augmentation, for instance, TransGAN [131], a transformer-based GAN was used in an agricultural setting for detecting fruit surface defects [132], as well as CGAN, a class-conditioned architecture, able to more precisely synthesize classes of defects [133]. A very popular architecture for these scenarios is Deep Convolutional GAN (DCGAN). In a comparative study of steel strip defect detection [134], it outperformed models such as the information-theoretic InfoGAN [135], and has improved accuracy metrics in imbalanced datasets from a variety of domains such as fiber layup inspection [136], liver lesion classification [137] and defect generation [138], often trained with the help of additional data augmentation via geometric or stylistic transformations. An improved version of DCGAN, capable of producing more diverse data, Wasserstein GAN (WGAN), was applied to the detection of weld [139] and decorative sheet [140] defects, however, complicated defects such as "burn-through" and "crack" welding defects still needed to be synthesized graphically using human prior knowledge. Finally, a more recent and sophisticated architecture, StyleGAN2-ADA [141], was capable of high-fidelity generation of structural adhesive defects trained over a small input dataset of fewer than two hundred images [142], with limited additional augmentation and manual labeling.

Customized Architectures Apart from the direct application of GAN architectures, several customized architectures have been developed to specifically tackle defect synthesis.

For instance, AC-PG GAN is a combination of the Progressive Growing GAN (PGGAN) [143] and Auxilliary Classifier GAN (ACGAN) [144] aimed at the quality assessment of photovoltaic modules through electroluminescence images [145]. In the biomedical field a similar modification towards a conditional PGGAN has yielded improvements for brain metastases detection in magnetic resonance images [146]. One Class GAN (OCGAN) and Multi-modal One Class GAN (MMOCGAN) presented in [147] are an attempt to cope with statistically non-meaningful defect classes by generating samples from the complementary distribution of the "good" class. Reinforcement Learning (RL) methods have also been used to guide data generation and increase the intra-class variability of the generated data, An example is the Actor-Critic GAN (AC-GAN) [148], which aims to identify sub-classes from a given class in a preprocessing step and then use Actor-Critic RL on top of the GAN to adjust loss weighting so that augmentation of each sub-class is either encouraged or inhibited. Finally, enabling generation for even smaller datasets is the BigGAN [149] adaptation method described in [7], which proved useful for few-shot learning in [8] and, though still untried in defect detection, served as a major inspiration for our work.

Encoder/Decoder Architectures A common type of customized architecture is one based on encoder/decoder approaches to generation. For example, [150] uses an improved combination of similar encoder/decoder-based generators, namely BEGAN [151] and Skip-GANomaly [152]. Defect-GAN [153] copes with the lack of defect data by synthesizing defects through unpaired image-to-image translation, thus creating additional defects using good images. Its encoder/decoder architecture corresponds to a defacement and restoration process and makes use of a spatial and categorical control map as well as the injection of adaptive noise to increase image diversity. A similar image-to-image translation idea is implemented in the surface defect-generation adversarial network (SDGAN) [154] and in [155] which is built around CycleGAN. A recent and well-performing approach, DeepSMOTE [6], tries to mimic vector-based oversampling approaches but on the level of raw images. It uses an encoder-decoder architecture to produce linear interpolations in the image space similar to SMOTE [91]. Although the above method was not used for defect classification, it served as inspiration for our approach of performing oversampling on the image level, which we further adapted to the defect classification problem.

While the proposed methods facilitate both high-fidelity image generation from limited data and targeted oversampling of important inputs, the approach introduced aims to combine the two leading to a more efficient and less computationally intensive oversampling method performed at the image level.

2.2.2 Prediction Confidence in Deep Neural Networks

As we saw in the previous sections, DNNs and especially convolutional ones, are a powerful learning model. This has come at a cost, however, as the growing model complexity of neural networks - which is also the cause of their better test accuracy -

introduces more overconfidence in their predictions [156]. In this section, we focus on the problem of classification. One way to define confidence in a classification setting is as the maximal value of the last softmax layer, which determines the class of a given input. Comparing this with the validation accuracy of the network for a given class using a reliability diagram for different confidence ranges and their corresponding accuracy scores, [157] found a significant difference in the 110-layer ResNet compared to the better calibrated but more primitive five-layer LeNet model on the CIFAR-100 dataset.

The main reason for this increasing miscalibration due to increasing model complexity is that DNNs additionally suffer from a more subtle case of overfitting. Namely, they tend to overfit the negative log-likelihood loss invisibly. In contrast, their visible generalization accuracy measured by a 0/1 loss seems to remain stable. This is a sign of unreliability that has limited DNN use in real-world safety-critical applications.

Many methods have been proposed to counter prediction overconfidence. The first category of calibration methods tries to adjust softmax outputs as a post-processing step to resemble the actual confidence probabilities or follow an ordering where a higher value will correspond to higher true confidence. Histogram Binning [158], Isotonic Regression, and Bayesian Binning Quantiles (BBQ) [159] are example methods that solve optimization problems after the model training to bring softmax output close to their confidence values as estimated on a validation set. Platt Scaling [160] and its generalizations Matrix Weighting [156], and Temperature Scaling [161] are applied on the logit layer just before the softmax aiming to calibrate the weights of the final layer so that outputs are close to the validation set confidence probabilities. Temperature scaling is the most popular approach, as it has the benefit of not influencing the ordering of the class predictions and therefore guaranteeing the exact class prediction as before.

A further category of confidence assessment methods tries to make changes to the learning algorithm so that the training process is constrained to output reasonable measures of the model's true confidence. Most notable is the addition of a penalty term to the loss function that discourages ordering inconsistencies in the output pseudoprobabilities [162]. Finally, regularization techniques such as dropout, weight decay, label smoothing [163] and mixup [164] have also been shown to improve confidence estimates.

Accurately quantifying the prediction confidence of Deep Neural Networks plays an important role in our approach since it helps us determine which samples need to be reinforced through data augmentation. As we are less interested in obtaining probabilistic estimates of confidence and also want to avoid risking a deterioration of the classifier's performance by treating the model as a black box, we focus on a less invasive method introduced for approximating the distance to the classification boundary [9], which does not require any changes in the network's architecture or the way it is trained.

2.3 Methods

This work introduces an oversampling method that is applied directly to raw images. The rationale for our approach is that we want to perform oversampling in a way that is decoupled from deep feature extraction, making it possible to train the final classifier

end-to-end on the augmented dataset. It can be seen as a method similar to Borderline-SMOTE [93] focusing on samples close to the classification boundary, but on the level of raw images. Aiming to generate images that are most informative for the way the classifier separates between classes, we rank images according to initial classifier confidence and use low-confidence ones to guide our generation process.

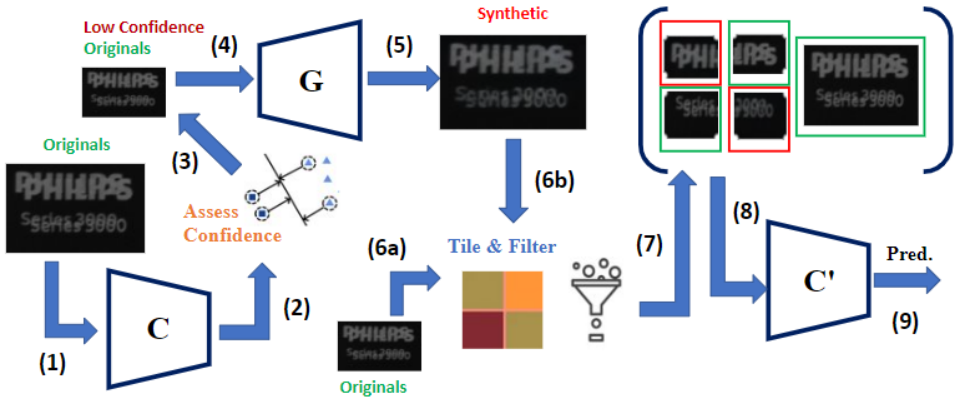


Figure 2.1. *Basic components and dataflows for the proposed oversampling approach. The sequence of processing steps is outlined with numbers from (1) to (9).*

Fig. 4.2 depicts an overview diagram for our proposed approach. It consists of an initial pre-training stage performed on the original imbalanced dataset. The resulting weights are used for the estimation of the boundary between classes and the ranking of instances according to model confidence. After the most informative instances have been selected from the original dataset they are used as seeds for an instance-based generator, which produces similar images introducing small variations. After post-processing (tiling and fusion with original images) and filtering of sub-standard quality images, we use the generated data to augment the original dataset. The training of the classifier is completed by fine-tuning the weights of the pre-trained classifier using the newly augmented, balanced dataset.

In the following subsections, we provide more details on the Synthetic Image Generation and Confidence Assessment components before fitting everything together to the final oversampling process.

2.3.1 Synthetic Image Generation

Producing high-fidelity images for fine-grained classification is challenging, however, state-of-the-art networks such as BigGAN [149] or StyleGAN [165] have been able to achieve it. Of course, both consist of millions of learnable parameters and require vast training datasets along with the corresponding computational resources. Instead of training such a model from scratch, we make use of a technique inspired by [7] and [8], which aims to perform transfer learning on a pre-trained BigGAN on ImageNet. BigGAN's generator G is isolated from the discriminator and its weights are initialized to the values obtained from ImageNet. Then for each input image I in the dataset, it is fine-tuned to produce an image I_z, as similar as possible to the original given a random noise vector z as

input. This fine-tuning includes only the relearning of the scale and shift parameters of the batch normalization layers. Intuitively this corresponds to selecting only the features relevant to the target dataset from a super-set of features learned through pre-training on ImageNet. The loss function for the fine-tuning is as follows:

$$\mathcal{L}_G(G, I_z, z) = \mathcal{L}_1(G(z), I_z) + \lambda_p \mathcal{L}_{perc}(G(z), I_z) + \lambda_z \mathcal{L}_{EM}(z, r) \tag{2.1}$$

$\mathcal{L}_1$ is the L1 distance and $\mathcal{L}_{EM}$ the earth mover distance, which tries to regularize z as a Gaussian sample ($r \sim N(0, 1)$); $\mathcal{L}_{perc}$ is the perceptual loss and λ_p, λ_z are regularization coefficients. Finally, to generate multiple images from input I, some random noise is added to the input so that $I_z = G(z + \epsilon)$.

Αλγοριθμος 2.1: *Generate Synthetic Data*

Input: $G, \mathcal{L}_G$ image generator and loss, I_b base images, n_{gen} aug. target
Output: I_{out} set of $|I_b| \cdot n_{gen}$ generated images

$I_{out} \Leftarrow \{\}$
$M \Leftarrow MinHeap()$
for $i \in I_b$ **do**
  for $n \in range(n_{gen})$ **do**
    $z \Leftarrow \mathcal{U}^{n_z}(0, 1)$
    $i_g \Leftarrow G(i, z, \mathcal{L}_G)$
    $I_p \Leftarrow TilePermutations(i, i_g, \{2, 4\})$
    for $p \in I_p$ **do**
      $M \Leftarrow M \cup \{(p, mse(p, i))\}$
    end for
  end for
  for $k \in range(n_{gen})$ **do**
    $m_k, l_k \Leftarrow \underset{m \in M}{argmin}[mse(m, i)]$
    $M \Leftarrow M \setminus \{(m_k, l_k)\}$
    $I_{out} \Leftarrow I_{out} \cup \{m_k\}$
  end for
end for

In Algorithm 2.1 we use the aforementioned generator as an instance-based generator that allows us to produce small variations of an input image. In practice, we observed that it usually produced high-quality defect images. To address the cases where it didn't we added additional quality enhancement measures. The most important of those is provided by the *TilePermutations* function, whose aim is to produce hybrid images by splitting its inputs into halves and quadrants and producing all possible combinations of the split parts (without of course changing their position in the original images). The resulting hybrid images together with the synthetic images are more populous than the n_{aug} images we need per base image. For this reason, we store all synthetic and hybrid images in a min-heap M from which we pick the top n_{aug} images with the lowest mean squared error (MSE) compared to the originals.

2.3.2 Confidence Assessment

To determine which defect instances the classifier is most uncertain of, and can thus benefit from seeing more similar examples of, the approximation of the distance to the classification boundary is the most straightforward approach. Of course, confidence cannot be viewed as a probability, but the relative ordering between distances together with a threshold can give us a limit of the model's knowledge boundaries. Contrary to SVMs, determination of the margin in deep neural networks is a challenging problem, nevertheless [9] suggests the following approximation, which is used in their calculation of the Large Margin Loss.

The decision boundary between classes i and j is defined as the set of inputs for which the confidence for two classes is equal, f being the (confidence) output of the NN:

$$D_{\{i,j\}} \triangleq \{x \mid f_i(x) = f_j(x)\}$$

The distance of a point x to the decision boundary is then defined under an l_p norm as the smallest displacement of the point that results in confidence equality:

$$d_{f,x,\{i,j\}} \triangleq \min_{\delta} \|\delta\|_p \text{ s.t } f_i(x+\delta) = f_j(x+\delta)$$

The above optimization problem is intractable for a non-linear f, therefore using the 1st order Taylor approximation to linearize f they obtain the following final approximation for the distance to the margin:

$$\hat{d}_{f,x,\{i,j\}} = \frac{|\, f_i(x) - f_j(x) \,|}{\|\nabla_x f_i(x) - \nabla_x f_j(x)\|_q} \tag{2.2}$$

2.3.3 On-the-fly Image-Level Oversampling

Algorithm 2.2 incorporates the outcomes as per the previous sections into the training process. The inputs are a CNN architecture C and the training data (X, Y), as well as the instance-based GAN G, adapted from BigGAN according to [7] with a loss $\mathcal{L}_G$. Further parameters include n_p which is the number of pre-training epochs to get sufficiently updated weight values to assess model confidence and n the number of epochs to train on the full augmented dataset. k_{top} indicates the number of most informative images selected to serve as seeds for the generation process.

After pre-training for n_p epochs, the distance to the boundary for each training image is computed according to Eq.2.2. As expected, this approximation does not provide good results for all images but it works well for images close to the class boundary assigning them smaller values than clearly classified images that are away from the boundary. Data instances and their confidence scores are stored in a min-heap out of which the k_{top} lowest distance images are extracted to form the base set for the generation. After determining the number of synthetic images to be generated per base image, needed for rebalancing the dataset, we pass the base images to the generation process described in Algorithm2.1. The parameter n_{aug} defines the number of images to be generated per selected base image so that the final dataset is balanced between defects and non-defects

Αλγοριθμος 2.2: *On-the-fly Image-Level Oversampling*

Input: C the CNN, X, Y train data, $G, \mathcal{L}_G$ the GAN and its loss function, k_{top} size of generation base, n_p, n pre-train and train epochs, l_g the label for the good class, L_d the set of defect class labels
Output: C'' the trained classifier after oversampling

$M \Leftarrow MinHeap()$
$C' \Leftarrow train(C, X, Y, n_p)$
$X_{good}, Y_{good} \Leftarrow \{(x_i \in X, y_i \in Y) : y_i = l_g\}$
$X_{defect}, Y_{defect} \Leftarrow \{(x_i \in X, y_i \in Y) : y_i \in L_d\}$
for $x \in X_{defect}$ **do**
    $\hat{d} = \frac{|C'_{good}(x) - C'_{defect}(x)|}{\|\nabla_x C'_{good}(x) - \nabla_x C'_{defect}(x)\|_\infty}$
    $M \Leftarrow M \cup \{(x, \hat{d})\}$
end for
$I_b \Leftarrow \{\}$
$n_{aug} \Leftarrow \lfloor \frac{|Y_{good}| - |Y_{defect}|}{k_{top}} \rfloor$
for $k \in range(k_{top})$ **do**
    $x_k, d_k \Leftarrow \underset{x \in M}{argmin}(\hat{d})$
    $M \Leftarrow M \setminus \{(x_k, d_k)\}$
    $I_b \Leftarrow I_b \cup \{m_k\}$
end for
$X_{aug}, Y_{aug} \Leftarrow generate(G, \mathcal{L}_G, I_b, n_{aug})$
$C'' \Leftarrow train(C', X \cup X_{aug}, Y \cup Y_{aug}, n)$

and is determined by the integer division of the difference between the number of images in the good class $|Y_{good}|$ and the number of total defective images $|Y_{defect}|$ over the number of base-images k_{top}. Note that the number of generated images per individual defect class might differ; there is only a constraint that the total defects are balanced with the good images. Depending on how many low-confidence images a defect class has, the more it needs to be augmented according to our approach. Following the augmentation step, the pre-trained classifier is trained for a further n epochs to produce a better classification boundary.

2.4 Results

Throughout our experiments, we show how the presented oversampling method benefits the general defect classification problem, by comparing it both with state-of-the-art approaches used in defect datasets and image- and vector-level oversampling approaches. We performed our experiments using a dataset of shaver shell logo print images from a real production line presented in the section below.

2.4.1 Dataset Information

The dataset used was provided by Philips Consumer Lifestyle B.V. and was collected from their pad printing process to serve the need for building an automated quality inspection system. As described earlier, owing to the infrequency of defects in their process,

it was hard to gather many defect images leading to an imbalanced dataset.

(a) *Good* **(b)** *Double Print* **(c)** *Interrupted*

Figure 2.2. *Original Shaver Shell Prints*

The dataset consists of JPEG RBG images with dimensions 220×360. They are divided into three classes, one good and two defect classes, namely double prints and interrupted prints. Representative examples of each class are presented in Figure 4.3. The number of correctly printed images is 2684, of double prints 244, and of interrupted prints 598. One important feature to note is that interrupted prints can be very similar to good prints, making their distinction difficult, as well as the generation of sufficiently differentiated images from these two classes.

Moreover, to verify the robustness of our method we used four additional datasets of product defects from the MVTec AD collection [11]. This is a collection of datasets consisting of surface and object defects. For our evaluation, we chose two products from each category that exhibited similar defects to the shavers dataset leading to the high similarity between classes. From the surfaces, we used the carpet and grid datasets and from the objects the pill and metal nut datasets, samples of which are shown in Fig. 2.3.

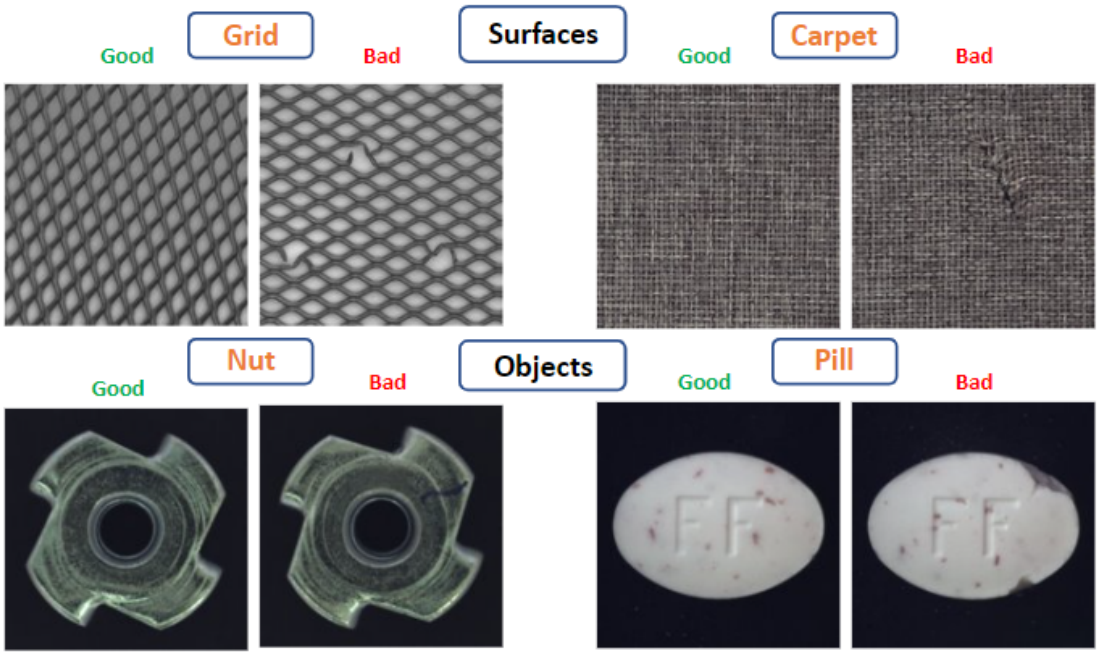


Figure 2.3. *Samples from the MVTec AD datasets*

Table 2.1 shows the number of instances belonging to each class for all datasets used

as well as their train and test set sizes as determined by the 5-fold cross-validation scheme described in the next section.

Table 2.1. *Number of class instances for the Shavers and MVTec AD product datasets including train and test sets*

	Train		Test		Total	
Datasets	Good	Defects	Good	Defects	Good	Defects
Shavers	2147	674	537	168	2684	842
Grid	211	46	53	11	264	57
Carpet	224	71	56	18	280	89
Metal Nut	176	74	44	19	220	93
Pill	214	113	53	28	267	141

2.4.2 Experimental Setup

Our experimental process was designed to compare our approach with three other families of approaches that have been common in the literature. The first is the attempt to directly generate data of the highest fidelity possible using a powerful generation method. We use StyleGAN as a comparison which achieved good results in [166]. The second type is the use of transfer learning and namely Resnet50 used in many works such as [167] and [168] for transfer learning, also viewing it in combination with vector-based oversampling. Thirdly we compare against DeepSMOTE [6], a state-of-the-art approach of performing SMOTE-like oversampling on the image level. For the non-transfer learning scenarios, we used as a classifier a customized shallow CNN for this dataset consisting of a convolutional layer with two parallel filters of $(3 \times 3) \times 16$ and $(1 \times 1) \times 16$ followed by a dense and a softmax layer.

The metric monitored was the binary recall from the perspective of the defect classes (Table 2.2), i.e. the defect class is considered the positive class for measuring recall. We found this metric most appropriate for a defect classification example as it better suits the way automated visual inspection is envisioned to work on a real production line. More specifically, positive predictions (good) usually receive the green light with no or little manual checking, while negative ones (defects) are put aside to be further examined by a human operator. Our aim is to minimize the number of defects that are mistakenly labeled as high-quality products. A more precise definition of Binary Recall, as used in the current context, can be formulated given the classifier C, test data X, the labeling function l, and the set of defect labels $L_d = \{double\ print, interrupted\}$, as follows:

$$BinaryRecall = \frac{|x \in X : C(x) \in L_d \wedge l(x) \in L_d|}{|x \in X : l(x) \in L_d|}$$

Another benefit of this metric is that it does not suffer from the dataset skew as, for example, accuracy which is dominated by the accuracy in the majority class. One must also be careful while maximizing binary recall so that not every product image is classified as a defect. For this reason, we also evaluated the ROC-AUC score which measures class

separability, though with a relative skew towards the majority class. The ROC-AUC score was satisfying for all experiments with values greater than 98%.

To complete the picture of the final classifiers' performance, we include the binary Precision and F1-scores, again, measured from the perspective of the defect class. Precision performance will be determined by the percentage of good images that get mistakenly classified as defects, while the F1-score will attempt to give a balanced account of the methods' effects on precision and recall. More precisely, these metrics are calculated using the same notation as for Binary Recall as follows:

$$Precision = \frac{|x \in X : C(x) \in L_d \wedge l(x) \in L_d|}{|x \in X : C(x) \in L_d|}$$

$$F1 - score = \frac{2 \times Precision \times BinaryRecall}{Precision + BinaryRecall}$$

Additionally, we compare a simplistic augmentation using our generation method in an untargeted fashion against our targeted oversampling approach based on the selection of the most informational examples (Fig. 2.6). This helps us gain further insight into how and in which cases targeted oversampling is helpful. We also monitor additional metrics such as the number of images generated for each defect class and the class-specific recalls.

The experiments consist of a total of 30 model runs using 5-fold cross-validation on a single NVidia K80 GPU used for both the training and data generation processes. Binary Recall scores are presented with their 95% confidence intervals.

Hyperparameter Tuning

For most comparison methods, an exhaustive search was carried out over ranges around initial well-performing hyperparameters (HPs) determined through trial and error. The best-performing hyperparameters were chosen over a stratified 5-fold cross-validation scheme similar to that followed for the showcased experiments resulting in an overall nested cross-validation (or double-cross) scheme as described in [169] and specified in pseudo-code in Algorithm 2.3. Specifically, the inner cross-validation produces validation sets for the selection of HPs and the outer cross-validation produces independent test sets for out-of-sample evaluation of the methods with the best-performing HPs. From this scheme, we extract the most frequently selected HP combinations as the recommended set of HPs to use for each method, which could provide the interested reader with insight into the dataset from an oversampling perspective.

2.4.3 Experimental Results

As shown in Table 2.2 our method outperforms all state-of-the-art approaches in terms of binary recall. However, there are several interesting points to note. Firstly, we observe that the custom CNN architecture outperforms transfer learning and transfer learning with oversampling approaches because the features are learned end-to-end specifically for the dataset at hand instead of being adapted from imagenet. Secondly, the impact of

Αλγοριθμος 2.3: *Nested Cross-Validation for Evaluation and Hyper-Parameter Tuning*

Input: C the Classifier (possibly including pre-trained embeddings and/or an oversampling method), X the input data (images), Y the input labels, CV the stratified cross-validation scheme, $\mathcal{H}$ the set of possible hyperparameter combinations, r_s the random seed for the current evaluation run

Output: M the complete final metrics per fold and seed, H_f the selected hyperparameter combinations per fold and seed

$M \Leftarrow \{\}$
$H_f \Leftarrow \{\}$
$fold \Leftarrow 0$
for $Train, Test \in CV.split(X, Y, folds = 5, random = r_s)$ **do**
  $H_m = MinHeap()$
  for $H \in \mathcal{H}$ **do**
    $fold$ += 1
    $R_{avg} \Leftarrow 0$
    for $Train_{HP}, Test_{HP} \in CV.split(Train, 5, r_s)$ **do**
      $C_H \Leftarrow train(C, Train_{HP}, H)$
      $m_H \Leftarrow evaluate(C_H, Test_{HP})$
      $R_{avg} \Leftarrow R_{avg} + m_H.recall$
    end for
    $R_{avg} \Leftarrow R_{avg}/5$
    $H_m \Leftarrow H_m \cup \{(H, R_{avg})\}$
  end for
  $h_{top}, r_{top} \Leftarrow \underset{(h,R)\in H_m}{argmax}[R]$
  $C_f \Leftarrow train(C, Train, h_{top})$
  $m \Leftarrow evaluate(C_f, Test)$
  $H_f \Leftarrow M \cup \{(r_s, fold, h_{top})\}$
  $M \Leftarrow M \cup \{(r_s, fold, m)\}$
end for

oversampling on the vanilla Resnet50 approach is much larger, than the effect of both Loss Weighting and our approach on the custom CNN. This can be attributed to greater margins for improvement in lower recalls, but also to the imperfection of the generation methods at the image level, which is a much more complicated, high-dimensional process than generating simple vectors.

Table 2.2. *Comparison of oversampling methods on the shaver-shell prints dataset*

Method	**Bin. Recall %**	**AUROC %**	**Precision %**	**F1 %**
Resnet50	85.85 ± 1.50	98.85 ± 0.12	94.41 ± 3.27	89.59 ± 1.27
Resnet50+SMOTE	95.84 ± 0.52	98.87 ± 0.13	84.53 ± 3.01	89.61 ± 1.57
Resnet50+ADASYN	95.49 ± 0.99	99.07 ± 0.11	85.14 ± 3.45	89.67 ± 1.69
Custom CNN	95.84 ± 0.39	99.20 ± 0.19	97.53 ± 0.81	96.67 ± 0.56
Custom CNN+LW	96.07 ± 0.39	99.09 ± 0.19	98.34 ± 0.33	**97.19 ± 0.43**
StyleGAN	91.20 ± 2.20	99.01 ± 0.14	**99.17 ± 0.41**	94.95 ± 1.38
DeepSMOTE	93.58 ± 1.07	99.23 ± 0.15	96.93 ± 0.80	95.22 ± 0.87
Ours	**97.27 ± 0.76**	**99.34 ± 0.07**	96.82 ± 1.27	97.03 ± 0.98

Most interestingly, we observed that the augmentation approaches based on StyleGAN and DeepSMOTE had an adverse effect on the custom CNN's performance. This is mainly attributed to their inability to produce realistic defect images that are close but not identical to the high-quality images and can also be hinted at by the samples of generated images shown in Fig 2.4. In fact, on the MVTec AD datasets, which are one order of magnitude smaller in size, these generative methods failed to produce plausible defect images, most probably due to the documented early overfitting approach of GAN architectures on small datasets [170]. Therefore they are also not included in Table 4.2. Our generator, thanks to the additional processing steps introduced manages to usually depict these kinds of small defects, which occur mostly in the interrupted class of the shaver dataset. Nevertheless, confusing synthetic images were still occasionally produced in some of the dataset's splits leading to a small deterioration in performance, highlighting a possible limitation of the proposed method.

It is important to note that in terms of AUROC, our method does not provide a significant improvement as it does with binary recall. The purpose of monitoring the AUROC metric, as mentioned in the experimental results section, is to ensure that while our method improves recall in the defect classes, it does not, at the same time, significantly sacrifice performance in the good class. Let us also note that since AUROC considers the dataset as a whole it makes it difficult for improvements in recall to be reflected since they are overshadowed by the performance in the majority class, which is more similar across the different methods.

Of particular interest is the effect of the proposed method on the Precision and F1 metrics in this dataset of high inter-class similarity. As explained in the experimental results section, our on-the-fly oversampling method was designed to optimize recall, which in the case of datasets with high inter-class similarity might come at the expense of precision i.e. mistakenly classifying more good images than before as defects. This is also illustrated by the method's performance in the precision metric which is lower than

all other methods utilizing a custom CNN classifier. Consequently, its F1-score, while second highest, is overcome by the Custom CNN with Loss Weighting. The sacrifice of the F1 metric, however, is only 0.16, with largely overlapping confidence intervals between the two methods, showing a small sacrifice in the overall problem performance.

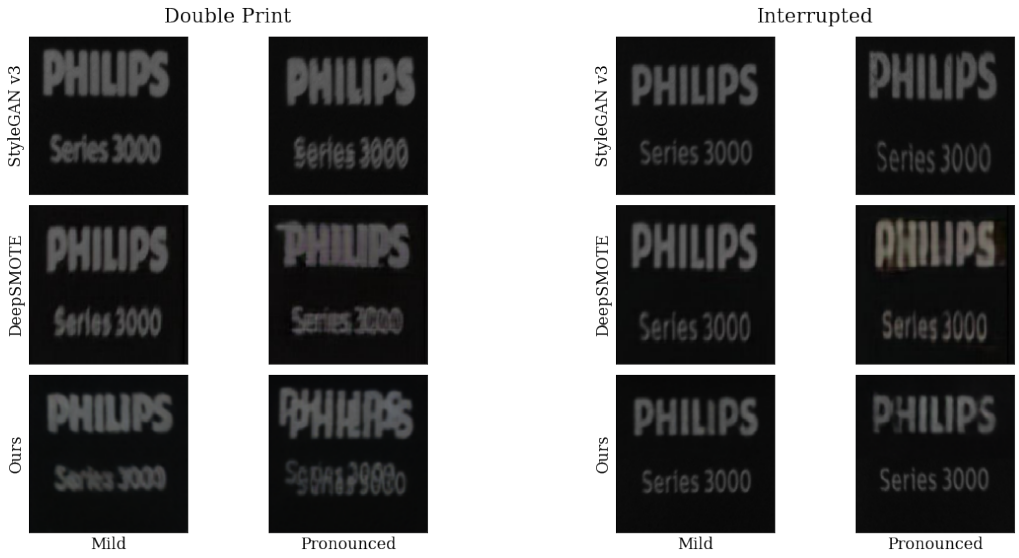


Figure 2.4. *Artificially generated defect images*

On the MVTec AD datasets, confidence-aware oversampling managed to provide the biggest improvements upon the end-to-end trained network, achieving the best recall scores in all cases. However, establishing statistical significance through confidence intervals was harder in this case, due to the very low number of defects in the test sets (see Table 2.3). As a consequence of the low number of defect samples, mispredicting just a few images has a pronounced impact on the overall binary recall score, which unfortunately presents a limitation of the evaluation scheme of our method when faced with smaller minority classes. Still, the improvement in the Metal Nut dataset was significant in comparison with most methods, while in the pill and carpet datasets, there are some indications of improvement. The grid dataset was harder for all methods producing results with very high variability between individual run scores.

Contrary to the shaver's dataset performance in the precision and F1 scores is consistently the highest across the four MVTec AD products - in the Metal Nut and Pill datasets being also statistically significant. We attribute this difference in the corresponding performances as measured on the original Shavers dataset, again, to the smaller amount of data which benefits significantly from the addition of the augmented images resulting in more precise boundaries from the perspective of both classes. For this reason, our image-level oversampling method has a more global effect on the classifiers' performance, not suffering from the trade-offs appearing in the more populous shavers dataset.

To understand the proposed method in more depth, Fig. 2.5 shows the changes in classifying augmented images and in the top-15 minority instance distances to the boundary before and after augmentation. There is an indication that boundaries shift

Table 2.3. *Comparison of oversampling methods on the MVTec AD product datasets*

Dataset	Method	Binary Recall %	AUROC %	Precision %	F1 %
Grid	Resnet50	30.30 ± 5.59	73.29 ± 3.48	42.4 ± 5.57	34.36 ± 5.95
	Resnet50 + SMOTE	35.60 ± 5.71	74.67 ± 4.34	48.27 ± 3.38	43.55 ± 3.48
	Resnet50 + ADASYN	42.57 ± 8.38	74.26 ± 4.04	42.93 ± 4.59	35.93 ± 6.68
	Custom CNN	70.90 ± 10.93	90.71 ± 5.49	80.23 ± 9.97	74.50 ± 10.50
	Custom CNN + LW	69.24 ± 12.25	89.80 ± 6.09	75.38 ± 12.3	71.55 ± 12.15
	Ours	**71.21 ± 9.92**	**91.22 ± 5.12**	**91.43 ± 6.86**	**78.45 ± 8.36**
Carpet	Resnet50	81.89 ± 3.70	97.07 ± 0.41	87.80 ± 3.33	84.20 ± 2.27
	Resnet50 + SMOTE	88.69 ± 1.53	97.21 ± 0.46	79.56 ± 2.11	83.66 ± 0.94
	Resnet50 + ADASYN	84.18 ± 2.71	97.25 ± 0.45	83.96 ± 3.78	83.42 ± 1.28
	Custom CNN	87.77 ± 7.62	98.94 ± 0.49	89.73 ± 1.23	87.48 ± 4.75
	Custom CNN + LW	91.11 ± 6.06	98.90 ± 0.51	88.02 ± 1.78	88.92 ± 3.72
	Ours	**92.22 ± 3.32**	**99.86 ± 0.11**	**92 ± 1.60**	**91.9 ± 1.97**
Metal Nut	Resnet50	84.03 ± 3.46	96.90 ± 0.79	95.33 ± 1.70	88.99 ± 1.97
	Resnet50 + SMOTE	88.30 ± 3.71	97.32 ± 0.51	90.32 ± 1.33	89.07 ± 1.62
	Resnet50 + ADASYN	84.09 ± 3.71	97.01 ± 0.72	95.38 ± 1.63	89.02 ± 2.09
	Custom CNN	82.92 ± 5.36	97.49 ± 1.15	98.33 ± 1.33	89.55 ±3.87
	Custom CNN + LW	82.92 ± 5.36	97.49 ± 1.15	98.33 ± 1.33	89.55 ± 3.87
	Ours	**92.63 ± 3.15**	**98.32 ± 1.22**	**98.75 ± 1.00**	**95.49 ± 2.12**
Pill	Resnet50	71.52 ± 6.29	92.70 ± 1.63	84.84 ± 1.63	76.65 ± 4.29
	Resnet50 + SMOTE	90.02 ± 2.62	91.76 ± 1.82	60.7 ± 1.57	72.34 ± 1.41
	Resnet50 + ADASYN	78.62 ± 4.16	91.87 ± 1.70	82.29 ± 1.54	80.08 ± 2.41
	Custom CNN	88.71 ± 2.18	98.35 ± 0.60	93.48 ± 2.03	90.94 ± 1.76
	Custom CNN + LW	88.71 ± 2.18	98.35 ± 0.60	93.48 ± 2.03	90.94 ± 1.76
	Ours	**92.29 ± 3.79**	**98.80 ± 0.58**	**96.25 ± 1.63**	**94.11 ± 2.68**

from the minority classes closer to the majority classes so that generated images that were misclassified before augmented training are now learned by the model. Of course, this shift in the distances is varied and cannot easily be correlated with performance increases, due to the complexity of the deep learning process and the approximate nature of the distance calculation method. It is important to note that distances both before and after augmentation are low in magnitude considering the high dimensionality of the feature space, hinting at the existence of highly populated boundaries. The goal of our method is to push those boundaries slightly so that they are biased toward the minority class - whose recall is more important - while perhaps, as in the case of the Shavers dataset, sacrificing prediction accuracy over the majority class - which is desirable given that the performance sacrifice is limited. This small shift could be significant exactly because the boundaries are densely populated due to high-class similarity. In the case of the MVTec AD datasets, this process leads to an overall improvement of class separability as highlighted by the increases in both precision and recall.

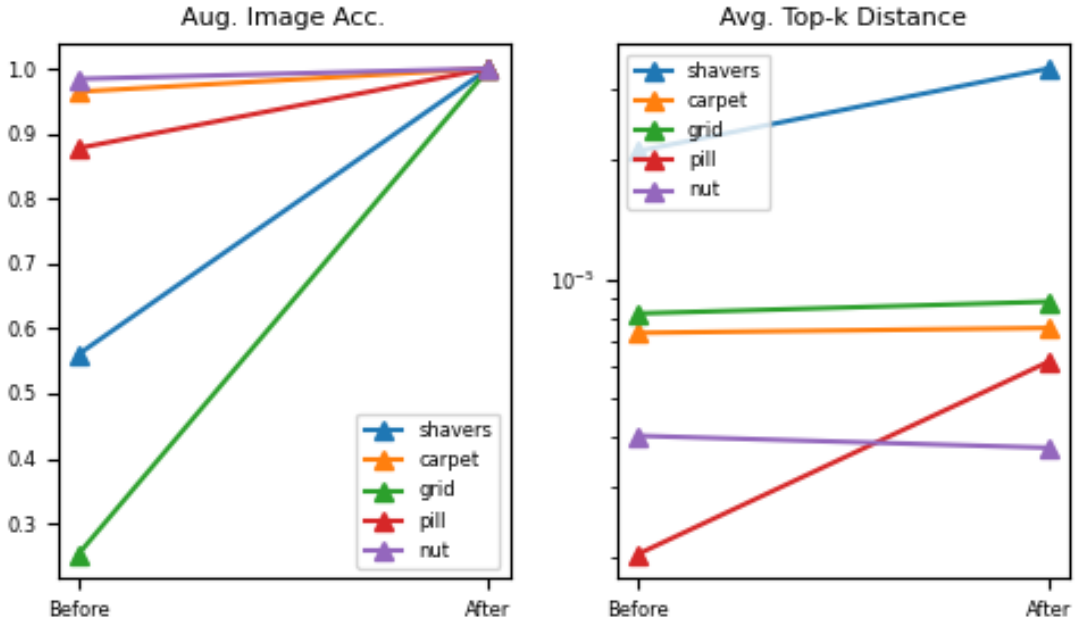


Figure 2.5. *Label accuracy of augmented images, before and after augmented training (Left). Top-k distances to classification boundary before and after augmented training for k=15 (Right)*

In terms of the specific hyperparameters (HPs) of our approach defined as inputs to Algorithm 2.2, after comparing the final classification performance of different combinations we chose 20 epochs for pretraining and 30 epochs with early stopping for training on the augmented dataset as the best-performing way to split the 50 total epochs needed to reach a stable loss plateau. We also determined the best value for k_{top} to be 15 images. In all other approaches, the training epochs for the classifier were 50 with early stopping, so that all comparison classifiers have time to reach their loss plateaus and equal to the total amount of training epochs used in our approach. The number of augmentation examples produced for the comparison methods was always the required amount for every class to

have as many instances as the good class, resulting in a balanced dataset. The ranges of HPs examined and the final recommended HPs for the Shavers dataset are shown in Table 2.4. For the image generation of StyleGAN and DeepSMOTE, we used the settings suggested for small datasets in the respective papers ([165], [6]).

Table 2.4. *Table of searched and recommended final hyperparameters per examined method for the shavers dataset*

Method	**Searched HP**	**Recommended HP**
SMOTE	*type* $\in$ {*None*, *borderline*1, *borderline*2}, $k \in [2, 20]$, $m \in [0, 22]$	*type* = *borderline*2 $k = 2$, $m = 20$
ADASYN	$k \in [2, 20]$	$k = 5$
Custom CNN	*batch-size* $\in \{4, 8, ..., 64\}$, $l_r \in \{10^{-5}, 10^{-4}, ..., 10^{-2}\}$, *dropout* $\in \{0.2, 0.3, ..., 0.8\}$	*batch-size* = 4 $l_r = 10^{-4}$ *dropout* = 0.4
Ours	*top-k* $\in [5, 50]$, *pre-eps* $\in \{5, 10, .., 45\}$	*top-k* = 15 *pre-eps* = 20

Finally, in terms of computation time, the introduced method was much quicker by approximately 3× the training time without augmentation (~ 30 minutes for a full run), while other image-level approaches such as StyleGAN and DeepSMOTE took more than 20h to train. This is because our method uses a small base set of images for generation and the time taken is linearly proportional to the number of base images. It is also built on top of a lightweight transfer learning method for GANs, while DeepSMOTE needs to be trained from scratch and StyleGAN's fine-tuning is more time-consuming due to its vast number of parameters.

Fig. 2.6 shows more closely how our oversampling method helps the classifier's learning process in the shavers dataset. We compare binary and class-specific recalls by using our generation method in a uniform way with the whole training set as seeds and selecting the seed set based on a distance-to-boundary confidence measure. What stands out is that the majority of the images close to the decision boundary belong to the interrupted class which is most similar to the good class. Basing the augmentation off of those images is also what brings the largest gains in recall performance. In the double print category, such gains are not visible, in fact, performance slightly deteriorates. This hints at a limitation of our method consistently producing performance gains over a range of imbalanced learning scenarios as it has been primarily designed for problems with high inter-class similarity.

2.5 Summary

In this chapter, we introduced a novel method for performing oversampling at the image level in the context of defect detection. Data generation is now performed more efficiently based on images that are estimated to be close to the classification boundary.

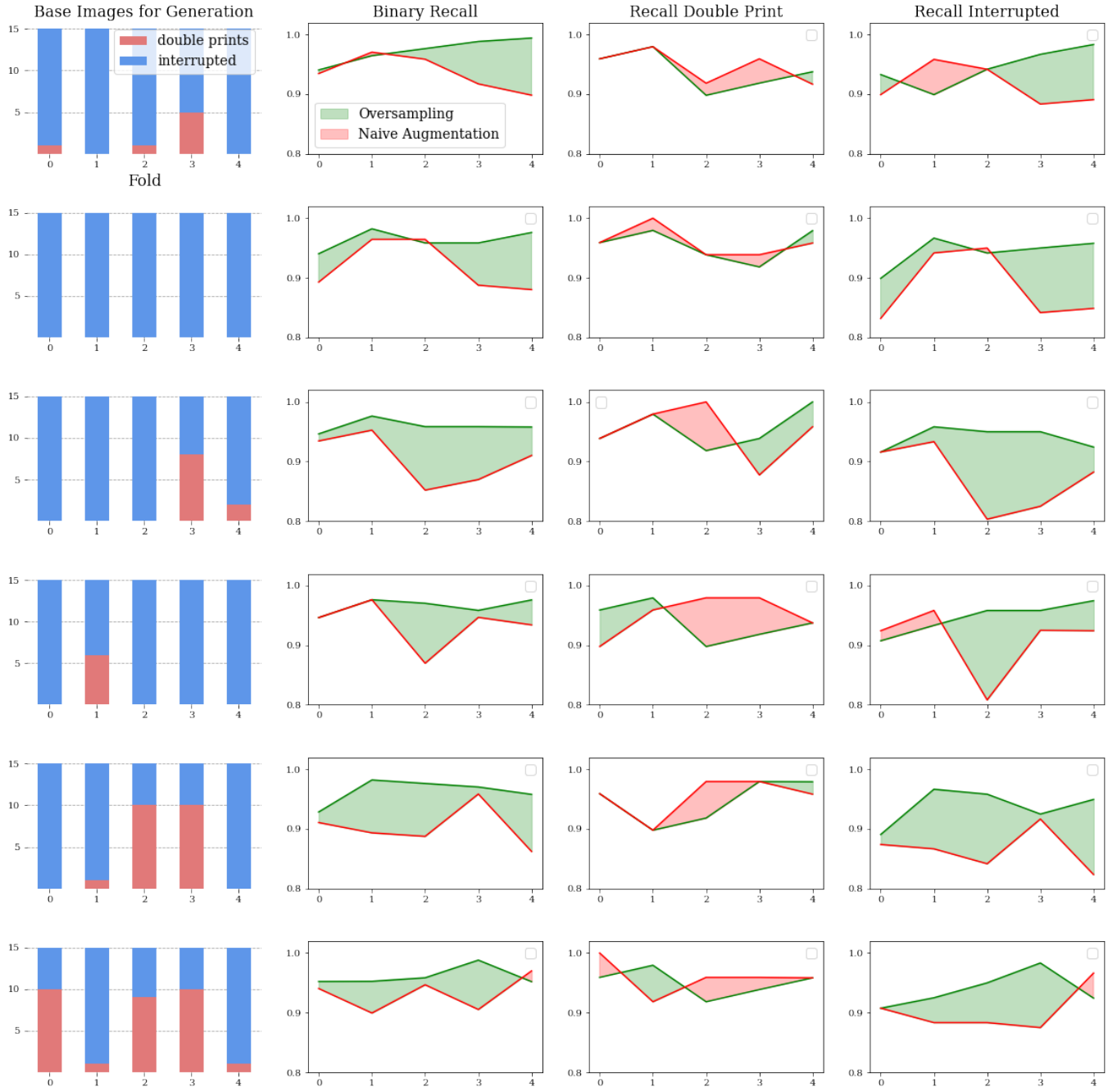


Figure 2.6. *Comparison between simple augmentation and confidence-based oversampling - 6 different instances of 5-fold CV on the shavers dataset*

The high-fidelity images generated helped improve the classification results over a dataset containing defects of varying perceptibility. The runtime and computational costs of generating synthetic data were also greatly reduced compared to other state-of-the-art approaches.

We believe that future advances in instance-based or few-shot image generation can greatly help improve our work by producing images of higher fidelity and variability from a small selected seed set of low-confidence images. Further opportunities for improvement lie in the way original and synthetic images are fused, which could potentially be performed in a smoother way than tiling using a few-shot learning-based fusion method. Finally, it is worth investigating how to produce linear interpolations between low-confidence samples through a suitable encoder/decoder architecture.

Chapter 3

Enhancing Robustness to Novel Visual Defects through StyleGAN Latent Space Navigation: A Manufacturing Use Case

3.1 Background

Quality Inspection, a key component of all production systems, has been following the trend of digitalization introduced by Industry 4.0 through the connection of digital sensors on the shop floor to state-of-the-art statistical and Artificial Intelligence (AI) algorithms running on the cloud and edge infrastructure [171]. The capabilities to collect information through sensors in real-time, in a non-destructive manner as well as the capabilities to store and process the large volumes of complexly interrelated data generated by the continuous operation of the shop floor through Machine Learning and Deep Learning has made it possible to develop sophisticated platforms that provide global view and control of quality in the factory [172]. Our work focuses on the specific case of visual inspection of the finished part, which is necessary when it comes to painting and decorating products. To that end, current AI research has provided many Image Processing, Computer Vision, and, lately, Deep Learning techniques that can meaningfully process rich image signals [173]. However, full automation of the Visual Quality Inspection can still be improved. In this work we examine a real-life manufacturing use case of automatically assessing the quality of brand prints on shavers produced by Philips Consumer Lifestyle B.V. While working on this use case we identified three significant challenges:

1. Typical *insufficiency of training data*, especially regarding rarely-occurring production defects.
2. *High visual similarity between flawless and defective products* might not be easily recognizable by an AI algorithm.
3. *Occurrence of unanticipated defects* during the continuous operation phase, which can lead to wrong AI decisions since they lie outside of the algorithm's training domain.

While the main focus in this section is on the last issue, the first two challenges, and especially high inter-class similarity, are not taken into account by most existing deep

learning robustness methods. However, they play an essential role in selecting suitable algorithms and evaluating the results in the problem of defect classification.

After carefully examining the suitability of different methods for identifying defects that were not anticipated during the AI model's training, a novel approach to open-set recognition for defect detection is proposed which relies on data augmentation and is more tailored to the defect classification problem and its aforementioned challenges. The presented method is based on the state-of-the-art GAN architecture of StyleGAN v3 [174], chosen due to its high fidelity, degree of generalization, and advanced manipulation capabilities. The latter are then leveraged through a computationally efficient closed-form factorization method [12] that discovers the most impactful directions for image generation in the GAN's latent space. After generating images along these directions, a novel criterion is applied for deciding if a synthetic image can be considered "unknown" relative to the used classifier and thus added to the augmented training set. The intended effect of this method is to introduce images that lie at the edge of the known classes and can define our classifier's area of competence. Consequently, any image that occurs at test time and is mapped outside this area can be considered unknown. The proposed method could potentially also be utilized in other areas where small visual anomalies need to be detected, such as civil infrastructure inspection or biomedical image processing.

3.2 Use Case and Dataset

The examined use case features a human-AI collaboration scenario, where products are first examined by the AI system to identify en masse potentially defective products that are then examined by a human operator who makes the final decision whether to discard or keep the product. The same scenario is also presented with an enhanced role for the operator in the form of active learning in [26]. The associated product image dataset provided by Philips Consumer Lifestyle B.V. was collected from the factory's pad printing process to serve the need for building an automated quality inspection system. The images in the dataset have been collected from the real-life production process before automation was introduced and have been manually labeled by multiple quality inspectors working in the factory to ensure correct labeling before their use in AI training. As is often the case, manufacturing defects are rare and this resulted in an imbalanced dataset. In the current context, we do not focus on solving the imbalance issue, nevertheless, we take it into account during the evaluation of our experimental results so that they represent a realistic scenario.

The collected dataset in digital form consists of RGB images in PNG format with dimensions 220×360. They are divided into three classes, one with flawless products, and two defect classes, namely double prints and interrupted prints. Representative examples of each class are presented in Fig. 4.3.

The number of correctly printed images is 2684, of double prints 244, and of interrupted prints 598. One important feature to note is that interrupted prints can be very similar to flawless prints, making their distinction difficult as well as the generation of sufficiently differentiated images from these two classes. The full collected dataset is split

(a) *Flawless* **(b)** *Double Print* **(c)** *Interrupted*

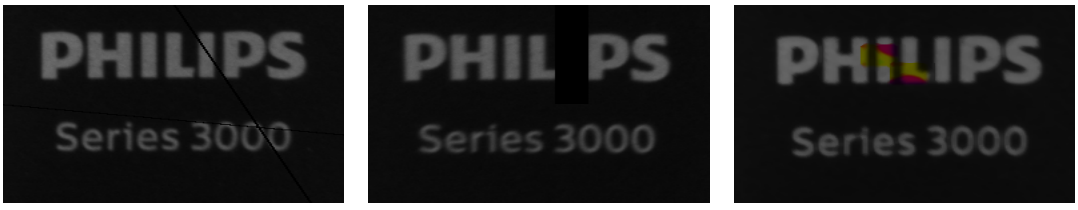
Figure 3.1. *Original Shaver Shell Prints*

into approx. 70% of the images to be used for training and 30% of images to be used for performance evaluation of the trained algorithm.

To additionally evaluate robustness, novel defects (unseen in the training set) were created synthetically to simulate possible unexpected defects that might occur during Automated Visual Quality Inspection, namely:

- **Line Interruptions**, which could result from preexisting scratches on the printing pad.
- **Missing Letters**, which could be due to a defect in the printer head
- **Discoloration**, due to the corruption/mixing of the sprayed color
- **Horizontal and Vertical Flips** due to a wrong setting of the printer head

Synthetic images of the first three categories can be found in Fig. 3.2. Images from these categories are merged with the test set in the same proportion to flawless images as the original defects (approx. 3 flawless to 1 defective), to represent a realistic imbalance scenario that could potentially occur in the production line. Therefore, the final test set, over which all methods are evaluated, contains 800 flawless images and 250 images with known defects, augmented with 250 novel defect images randomly and uniformly generated from one of the synthetic classes above.



(a) *Line Scratches* **(b)** *Missing Letter* **(c)** *Discoloured*

Figure 3.2. *Synthetic "Unexpected" Defects*

3.3 Related Work

Two branches of recent AI research aimed at the development of systems robust to out-of-distribution samples and also applicable in Visual Quality Inspection are open-set recognition (OSR) and semi-supervised anomaly detection. Traditionally, classifiers

have been evaluated on closed set problems, where the classes in the test domain are identical to those in the training data. However, practical use cases often require the classification of so-called “unknown unknowns" [175] corresponding to unmodelled aspects of the problem domain, which tend to confuse learning algorithms. OSR attempts to minimize the risks associated with these unknowns while preserving performance in the training classes. We identified three categories of OSR methods that could apply to our use cases, the current literature on which is mentioned below: Statistical OSR, OSR for Deep Learning, and Data Augmentation.

On the other hand, semi-supervised approaches view the problem over a binary lens, trying to model the flawless class and identify visual deviations as anomalies. Despite lacking the granularity of multi-class classification methods [176], they are beneficial to cases where the “closed set" consists only of the flawless class, and all defects belong to the “open set". It was demonstrated in [177], that these methods achieve comparable but lower performance on some metrics to open-set recognition, even though finding appropriate decision thresholds to address the aforementioned class-similarity problem can be tricky.

3.3.1 Open-set Recognition

OSR can be applied in two ways: by separating unknown from known instances in a binary way and then performing the usual multi-class classification task or by maintaining classification accuracy for known instances by grouping unknown ones in a newly added background class [178]. The OSR problem was formally defined by [179] as they attempt to minimize open space risk, where the open space O refers to the space away from the mass of known instances. The risk of labeling such an instance as a member of a known class is defined with the help of an indicator function f as:

$$R_O(f) = \frac{\int_O f(x)dx}{\int_{S_O} f(x)dx}$$

Where $f(x) = 1$ if an open-space instance is defined as known and is 0 otherwise and S_O is the total space including both open and closed-set instances. Subsequently, OSR is posed as an optimization problem of minimizing the open space risk R_O together with the empirical risk R_ϵ depending on the recognition function f from measurable space H with training data V and a regularization coefficient λ_r:

$$\arg\min_{f \in H} \{R_O f + \lambda_r R_\epsilon(f(V))\}$$

Regarding specific implementations, OSR can be further subdivided into three families of methods: Statistical, Deep Learning, and Data Augmentation-based OSR.

Statistical Methods

Most statistical methods for OSR make use of Extreme Value Theory (EVT) [180], a branch of statistics that has been successfully applied to areas such as financial and en-

vironmental risk management and anomaly detection (e.g., intrusion detection in security monitoring [181]. The goal of EVT is to label a sample as extreme through modeling a distribution's tails and subsequent application of appropriate thresholds. In most OSR applications, EVT is used over the distribution of classifier scores.

Scheirer et al. [179], combined Support Vector Machines (SVMs) with EVT by defining and adjusting an extra hyperplane to divide known classes and open space, thus bounding open-space risk. One of the most significant contributions of this research was the Compact Abating Probability (CAP), which requires that the probability of an instance belonging to a specific class decreases in all directions leading from the training space to the open space. Their algorithm was named the "1 vs. set machine".

Weibull SVMs (W-SVM) is an attempt by [182] to extend the "1 vs. set machine" using score calibration based on the Weibull distribution (a common choice in EVT for modeling distribution tails) and nonlinear boundaries. The scores to be calibrated are a combination of a One-class SVM [183] using an RBF kernel for differentiating between open/closed set instances and a multiclass SVM to classify amongst known classes. This approach has been widely used in fingerprint recognition [184] and intrusion detection [181].

A further development in this direction was the Probability of Inclusion SVM (PI-SVM) by [185], which models the posterior probability of inclusion for each class and rejects unknown samples based on an appropriate threshold value. This modeling happens via an RBF kernel SVM using a "1-vs-all" approach, where classification scores from instances close to the positive class limit are used to fit a Weibull distribution. Instances are assigned to whichever class their probability of inclusion after Weibull calibration is highest and above a certain threshold. They are marked as unknown if they are below the threshold for all candidate classes.

OSR and Deep Learning

Similarly to the calibration of SVM scores, EVT has been used over Deep Neural Network (DNN) scores to minimize open space risk. Initial approaches for OSR on DNNs focused on thresholding the softmax output of a network [186], which squeezes the activations of the last layer between 0 and 1, providing a pseudo-probabilistic output. However, due to its steep form, the softmax function will not only misclassify an out-of-distribution sample but also likely assign its prediction a high confidence score, making thresholding on the softmax score problematic. A possible solution is using a background or garbage class [187]. Even though this worked well in the benchmarked pedestrian datasets, it was insufficient for other real-world use cases with practically infinite open-space risk. A further step is creating a softmax-like layer with an extra class output introduced in Openmax [188] and aiming to redistribute scores between closed and open-set classes while retaining the benefits of softmax. OpenMax operates on distances to mean activation vectors (MAV) exported from the final layer of the DNN before the softmax. After determining the highest per-class distances, it uses EVT to fit a Weibull distribution on top of them. After thresholding, it leads to a CAP adhering distribution with rejected samples

being assigned to the unknown class after the overall final scores are normalized. Unfortunately, OpenMax is constrained by the underlying feature representation of the original network architecture, which might not necessarily drive toward better representations for the differentiation of unknown instances.

Data Augmentation

Attempts to directly learn open-set feature representations have been largely based on data augmentation techniques, based on the notion that training the model on synthetic open-set instances will produce representations that will remain robust at test time. To that end, generative adversarial networks (GANs) have been utilized in works such as [189] and [190]. GANs consist of two antagonistic networks: the generator (which produces images similar to the training data from a small noise input vector) and the discriminator (that tries to differentiate between real and synthetic samples). During adversarial training, the generator becomes increasingly better at fooling the discriminator.

G-Openmax [189] is a GAN-augmented extension of Openmax working under the assumption that open-set classes are usually closely related to the original training classes. The synthetic instances created as additional input to the Openmax augmented network training data result from linear latent space interpolations between samples belonging to different classes. While this technique improves upon OpenMax in handwritten digits and characters datasets, it does not make a difference in more realistic use cases.

Aiming to improve upon G-Openmax [190], propose Open-Set Recognition with Counterfactual Images (OSRCI), augmenting the training set with counterfactual images. These are generated by posing GAN latent space traversal as an optimization problem where the nearest noise vector to a class's latent representations for whom the generator output is misclassified, serves as the seed for generating a counterfactual image. This idea is closely related to the CAP notion. [191] similarly propose an adversarial sample generation (ASG) method. Finally, OpenGAN [192] uses a vector encoding semantic information together with inter-class interpolation in the latent space to drive the generation of novel images.

3.3.2 Semi-supervised Defect Detection

Semi-supervised anomaly detection is a close relative to OSR. It is most useful when images for the normal class are available with none or very few anomalous samples. This is very often the case in manufacturing production lines. Therefore, we use various models to compare their performance to open-set recognition techniques. Methods that apply to visual inspection most commonly rely on image reconstruction pipelines and generative models and are usually based on encoder-decoder, GAN, and, more recently, normalizing flow architectures [193].

The idea behind image reconstruction methods is to train an encoder-decoder-like architecture to reconstruct only normal images. When defects are seen at test time, their reconstruction will not be as accurate, leading to a measurable (e.g., using the Structural Similarity Index) difference between flawless and anomalous image reconstructions.

Variants of this approach have been applied in use cases such as the inspection of civil infrastructure [194], the production of hot-rolled strips [195], and railway rail insulator patches [196]. GANs have also been widely used in this scheme. For instance, [197] introduced GANomaly, which adds another encoder on top of its encoder-decoder GAN-based reconstruction pipeline. The output of this new encoder is compared to the latent space representation of the original encoder to determine whether the image contains a defect.

An additional approach to semi-supervised visual defect detection is based on calculating appropriate distance-based distributions and thresholds on top of pre-extracted embeddings from large datasets such as Imagenet. Deep Feature Kernel Density Estimation (DFKDE) [198] follows the pre-trained backbone network with Principal Component Analysis and Gaussian Kernel Density Estimation, while Deep Feature Modelling (DFM) [199] also applies PCA followed by fitting a mixture of Gaussian on the features with lowered dimensionality, as extracted from the flawless class images.

3.3.3 OSR in Manufacturing Defect Detection

Although semi-supervised learning has been widely applied to manufacturing quality inspection problems, few research works study open-set recognition settings. This is also partly due to the proliferation of datasets fitting the semi-supervised setting, such as MV-TEC and Kollektor SDD [198]. Despite the defect detection problem being a binary classification task in many use cases, open-set recognition can offer more flexibility with the ability to distinguish between different defect classes and open-set instances. One such work is presented in [200], which applied a CNN with a distance or clustering-based approach in an embedded space to a wafer map inspection scenario.

3.3.4 GAN Inversion and Latent Space Traversal

Attempts to control the output of GANs are directly related to OSR data augmentation methods as they can be leveraged to produce realistic novel image data. This is the domain of GAN Inversion Research which also enables the targeted traversal of a GAN's input space in a meaningful and sometimes interpretable way. The initial goal of GAN inversion is to map an image that is relevant to the GAN's training domain, backwards to a latent space noise vector that when provided as input to the GAN produces an accurately reconstructed version of the image [201]. Of course, such a task is made more difficult with modern complicated, but also very realistic GAN architectures such as PGGAN [202], BigGAN [203] and StyleGAN [204]. Towards that end many supervised, unsupervised and optimization-based methods have been introduced, each with its unique trade-offs. Supervised methods attempt to learn a mapping from the generated images to their latent space origin vectors hoping to extend it to non-synthetic images as well, but often introducing bias towards the sampled synthetic images used to form the training set, while optimization-based methods try to minimize a reconstruction loss type constraint by a directed search of the latent space, which however comes at a higher computation cost [205].

An important aspect of GAN inversion is also the latent space that is used. The first choice is the so-called Z-space which is available to every architecture as the space of possible inputs. However, architectures such as StyleGAN provide more degrees of freedom and better disentanglement regarding semantic attributes. One such space is the W-space which is the result of a fully connected MLP applied to the Z-space inputs to map them to a more disentangled space [204], characteristic of StyleGAN architectures. W-space vectors are further processed by the AdaIN layers and are fed at different layer depths to StyleGAN's generator architecture. These processed per-layer inputs bundled together form the W+ space [206]. More elaborate latent spaces such as S-space and P-space have also been introduced [207], [208], but are out of scope for this research.

The end goal of mapping an image inversely to one of the aforementioned latent spaces is to provide the capability to traverse this latent space in a way that modifies or edits semantically meaningful attributes of the image [209], [210]. More specifically, the goal is to find those direction vectors n for which linear traversals with step a in the form $z' = z + an$ produce meaningful changes to the output image. These directions can be discovered both in supervised and unsupervised manners. For instance, [211][212] gather many latent space vector/image pairs and tries discovering directions relating to features like color, rotation, or facial attributes using corresponding pre-trained classifiers, which of course might not be available for all possible required attributes/datasets. On the other hand [213] attempts to discover high-impact traversals by applying Principal Component Analysis (PCA) on the latent space. More recent approaches however try to find closed-form solutions, such as SeFa [12] which makes use of StyleGAN architecture specifics to approximate the problem in an analytically solvable manner, with substantial gains in computation time.

3.4 Proposed Method

Our proposed approach follows the example of the Data Augmentation methods for open-set recognition described in the previous section. Similarly, it adds an "unknown" class to the problem, for which synthetic images are generated using GAN manipulation. Compared to approaches such as OSRCI and OpenGAN, which were described in Section 3.3.1, we use a newer and more expressive GAN architecture namely StyleGAN v3. StyleGAN produces higher-fidelity outputs, which helps when generating images with high inter-class similarity such as the flawless and interrupted images from our use case. It also offers more easily manipulable latent spaces and a variety of different methods for traversing them. Those two aspects are foundational to our approach which is broadly described in Fig.4.2.

The three basic components of our approach are the *Generation of Synthetic Data*, the *Voting-based Filtering*, and the *Training on the Augmented Dataset*. The data generation process uses a pre-trained StyleGAN model (G) fine-tuned on our use-case dataset. The StyleGAN generator is then passed onto the Semantic Latent Factorization model which discovers the semantic directions over which the greatest change in the output occurs. These directions are then traversed and produce synthetic images corresponding to speci-

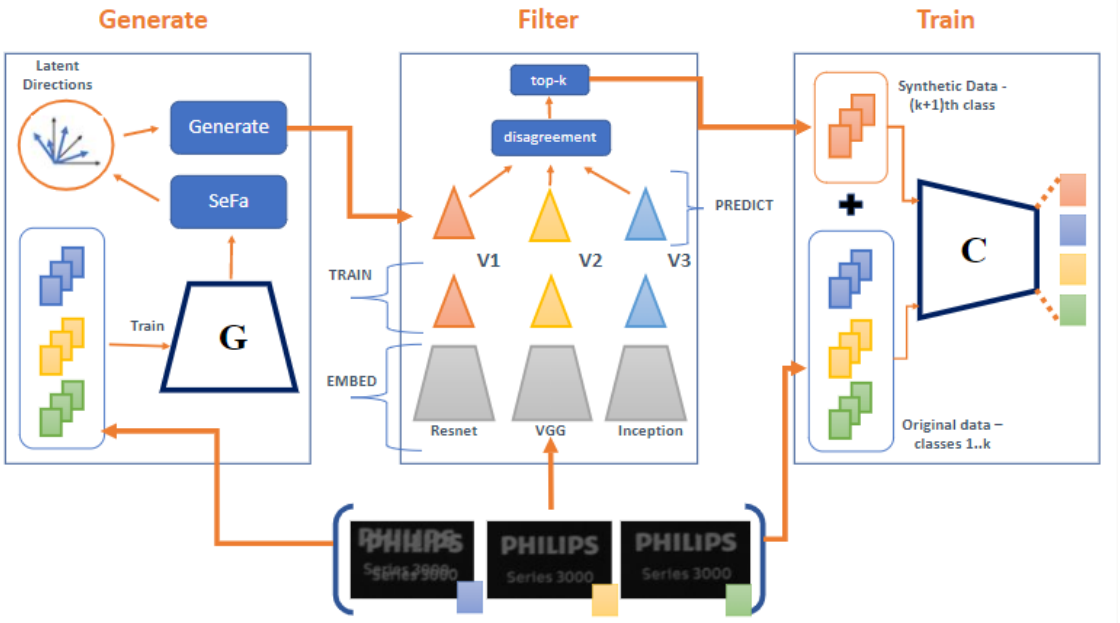


Figure 3.3. *Basic components and dataflows for the proposed approach*

fied distance points on the latent direction lines. Of course, the images produced through this process might not be novel but instead belong to one of the original classes, as shown in Fig. 3.4. However, our rationale is that images originating from points that lie on the edge of the Generator's learned latent distributions could be sufficiently unrecognizable by a classifier to be considered novel or "unknown" and thus be used through augmentation to form a boundary around the distribution that is known during training time.

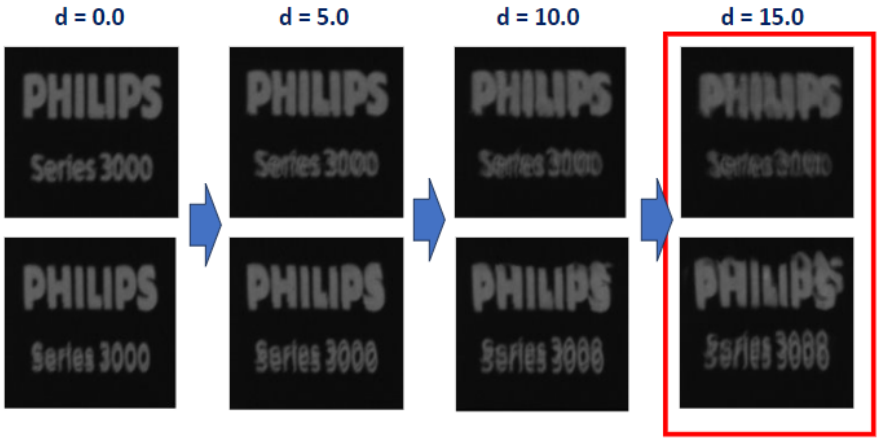


Figure 3.4. *Images generated from SeFa traversal at given distances. The circled images are retained as out-of-distribution after filtering.*

Data Filtering attempts to identify and collect these extreme images, while discarding those that are easily recognizable. This is achieved through a vote gathered from three voter classifiers (V_i). Each of these classifiers is trained on the same data but using transfer learning from different pre-trained embeddings so that the problem is learned from different angles. We measure how novel an image is through the disagreement of

the voter classifiers i.e., the number of distinct different classes predicted for the image. The images that cause high enough disagreement are then grouped into the “unknown" class and added to the original training data. A classifier C is trained on the augmented set with the additional class. We show that by adding these extreme images to the training set we make the classifier more resilient to novel inputs that might occur during testing or continuous real-life operation. The rationale behind this type of augmentation is that “extreme" images form a boundary between the original classes and the open space helping out-of-distribution inputs fall into the added “unknown" class.

3.4.1 Semantic Factorization for Latent Space Traversal

Semantic Factorization (SeFa) [12] is an attempt at a closed-form solution to the problem of discovering semantically meaningful latent space traversals. We leverage this method for our approach since it is a closed-form and therefore computationally light, method that performs on par with previous learning-based methods as described in Section 3.3.4. The method is based on the singular value decomposition of a GAN's first layer weight matrix. Assuming a generator G mapping inputs from R^d to the space of possible images $\mathcal{I}$, i.e. $I = G(z)$, the first layer output $G_1(z)$ can be represented as an affine transformation of the latent space input z: $G_1(z) = y = Az + b$, where A is the matrix of first layer weights of G. Then $G_1(z')$ for a sample in direction n, $z' = z + an$ starting from a randomly selected z and placed at a distance regulated by the constant a, was expressed by the authors in terms of $G_1(z)$ as $G_1(z') = G_1(z) + Aan$, meaning the difference between the two outputs are dependent only on the weight matrix A and therefore reducing the search for k most meaningful semantic directions $N^* = \{n_1, ..., n_k\}$ to the optimization problem:

$$N^* = \arg\max_{n_1,...,n_k} \sum_{i=1}^{k} ||An_i||^2$$

After the use of Lagrange multipliers, the problem is further reduced to finding the eigenvectors corresponding to the k largest eigenvalues of the matrix A^TA.

In our use case, we apply SeFa to different StyleGAN layers that control style attributes such as pose, texture, etc., and collect the directions of highest change from all layers together to later produce synthetic images. The reason for choosing this method, apart from its computational efficiency (closed from - no learning model needed), is its clear way of discovering directions of steep change. These directions make it possible to produce samples that lie at the edge of the generator's capabilities.

3.4.2 Method Description

In this section more details are provided on the proposed approach as specified in Algorithms 3.1, 3.2 and 3.3, loosely corresponding to the main components in Fig. 4.2.

Algorithm 3.1 describes the process for generating synthetic images. A pre-trained StyleGAN generator G is provided, as well as a list of layer IDs ($\mathcal{L}_G$) that correspond to the latent spaces that can be explored e.g., Z-space, W-space, etc. The generation process is

followed per input class belonging to the closed-set classes C. The first step is to use Semantic Factorization (SeFa) to produce the set of most semantically significant directions and the layers to which they belong. The number of directions used is defined by the number of semantics N_{SEM} that is passed as a parameter. For each dimension/semantic a number of N_{SAM} points are sampled from its associated latent space in the StyleGAN generator architecture to serve as a starting point for a traversal. How far away one can go from the starting point is bounded by d_{min} and d_{max}, while the number of steps s defines the number of intermediate images generated and saved across the chosen semantic direction. The outputs of the generation procedure are the total images gathered from the executed traversals.

Αλγοριθμος 3.1: *Generate Synthetic Open-set Data*

Input: G StyleGAN image generator, $C = \{c_1, c_2, ..., c_k\}$ closed-set classes, N_{SEM} # semantic directions, N_{SAM} # samples per direction, t truncation factor, (d_{max}, d_{min}, s) direction bounds and step, $\mathcal{L}_G$ list of layer IDs
Output: I_S synthetic traversal images

$I_{out} \Leftarrow \{\}$
for $c \in C$ **do**
  $\{l, \mathbf{n}\}_{i=1:N_{SEM}} \Leftarrow SeFa(G^c, \mathcal{L}_G)$, where $\forall i, l \in \mathcal{L}_G$
  for $j \in range(N_{SEM})$ **do**
    $n_z \Leftarrow dim(\mathbf{n}_j)$
    $G' \Leftarrow G^c_{l_j}$
    for $k \in range(N_{SAM})$ **do**
      $z \Leftarrow \mathcal{U}^{n_z}(0, 1)$
      for $d \in range(d_{min}, d_{max}, s)$ **do**
        $z' \Leftarrow z + d \cdot \mathbf{n}_j$
        $img \Leftarrow G'(c, z', t)$
        $I_S \Leftarrow I_S \cup \{img\}$
      end for
    end for
  end for
end for

Following Fig. 4.2 the next step in the process is the filtering of synthetically generated images to keep potential candidates for populating the “unknown" class. The filtering is based on a set of voter classifiers (in our case V_1, V_2, V_3) which are trained on the same training set as the final classifier C in Fig. 4.2. Each of these classifiers uses different pretrained embeddings, namely Resnet50, VGG ’16, and Inception v3 embeddings. For each sample in the set of synthetic images X_S the predicted classes C_{pred} from each classifier are gathered and their disagreement is measured as the number of distinct elements in the set of the predicted classes, namely its cardinality. All images are ranked by being inserted into a min-heap according to their disagreement score and the top n_{gen} images are kept for data augmentation. The reasoning behind using voting for filtering is that all trained classifiers will be able to agree in areas near the training samples but might draw arbitrary boundaries in the so-called open space, away from the training samples. This means that

if we filter for images that are embedded in areas where the classifiers disagree, we will get the open-space images that are needed for populating the newly created “unknown" class. Of course, the optimal choice of the number and architectures of the voter classifiers may play a role and are a fruitful direction for future research. As a minimum for introducing the idea, we chose three significantly different architectures, as many as the training classes in our problem, so that we can get informative disagreements scores. For example, had we used 10 voters, it would still be impossible to get a disagreement score higher than 3.

Αλγοριθμος 3.2: *Filter Synthetic Open-set Data*

Input: $D_C = \{X_C, y_C\}$ the input closed set dataset, X_S the synthetic open set images, $\mathcal{V} = \{V_1, V_2, ..., V_n\}$ voting classifiers, n_{gen} augmentation target
Output: $I_{\mathcal{F}}$ the filtered synthetic open-set images

$I_{\mathcal{F}} \Leftarrow \{\}$
$M \Leftarrow MinHeap()$
for $V_i \in \mathcal{V}$ **do**
    $V_i \Leftarrow train(V_i, D_C)$
end for
for $x_i \in X_S$ **do**
    $C_{pred} \Leftarrow \{\}$
    for $V_j \in \mathcal{V}$ **do**
        $c \Leftarrow V_j(x_i)$
        $C_{pred} \Leftarrow C_{pred} \cup \{c\}$
    end for
    $D_i \Leftarrow \mathbf{card}(C_{pred})$
    $M \Leftarrow M \cup \{i, D_i\}$
end for
for $k \in range(n_{gen})$ **do**
    $m_k, d_k \Leftarrow \underset{m \in M}{argmin}[D(m)]$
    $M \Leftarrow M \setminus \{(m_k, d_k)\}$
    $I_{\mathcal{F}} \Leftarrow I_{\mathcal{F}} \cup \{m_k\}$
end for

Finally, Algorithm 3.3 puts together all steps into a process of assembling an augmented training set that will render the learned classifier robust to novel defect types that have not yet occurred in its training set. Apart from the parameter used to call the generation and filtering procedures outlined previously, several parameters have to be set to ensure the right number of generated images is produced. We have already seen N_{SEM} as the number of top semantics extracted by SeFa. This parameter is the first one needing to be fixed and is at the moment determined empirically by visual inspection of how many directions produce images that differ substantially as the distance from the seed sample increases. The maximum traversal distance d is also chosen in the same manner. Afterwards, parameters M_{R_1} and M_{R_2} should be chosen. These “redundancy multipliers" ensure that a sufficient number of images are generated so that after filtering the augmentation target can be reached and the augmented data is sufficiently novel concerning

the training data. Thus, data generation is seen as a lossy process that will not generate many truly novel images. In our use case, we observed around 2-3% of generated images as having the highest disagreement score, so we generated 40-50 times the size of the original data to finally achieve a balanced training set of closed- vs. open-set samples. M_{R_1} is used as a multiplier for the number of samples per direction and M_{R_2} for the number of images generated from each per-sample traversal. The aim is to have $M_{R_1} \cdot M_{R_2} \cdot \mathbf{card}(D_C)$ images after generation so that we end up with approximately $\mathbf{card}(D_C)$ (the size of the closed-set training set) synthetic images after filtering. These final synthetic images will be grouped into a new “unknown" class and added to the original training set to perform the training of the robust classifier.

Αλγοριθμος 3.3: *OSR method based on Data Augmentation*

Input: C the CNN used for classification, $G, \mathcal{L}_G$ the pre-trained StyleGAN generator and its layer IDs, $\mathcal{C} = \{c_1, c_2, ..., c_k\}$ closed-set classes, $\mathcal{O} = \{c_{k+1}\}$ the open-set class, $D_C = \{X_C, y_C\}$ the input closed set dataset, $\mathcal{V} = \{V_1, V_2, ..., V_n\}$ voting classifiers, M_{R_1}, M_{R_2} the redundancy multipliers, N_{SEM} # semantic directions for generation, d maximum traversal distance, t data gen. truncation factor
Output: C' the CNN trained on the augmented dataset

1: $G' \Leftarrow train(G, D_C)$
2: $num\_samples \Leftarrow \frac{M_{R_1} \cdot \mathbf{card}(D_C)}{N_{SEM}}$
3: $step \Leftarrow \frac{2d}{M_{R_2}}$
4: $I_S \Leftarrow \mathbf{generate}(G', C, N_{SEM}, num\_samples, t, (d, -d, step), \mathcal{L}_G)$
5: $I_{\mathcal{F}} \Leftarrow \mathbf{filter}(D_C, I_S, \mathcal{V}, \mathbf{card}(D_C))$
6: $X_{aug}, Y_{aug} \Leftarrow (I_{\mathcal{F}}, \{c_{k+1}\}^{\mathbf{card}(D_C)})$
7: $C' \Leftarrow train(C, \{X \cup X_{aug}, Y \cup Y_{aug}\})$

3.5 Results

3.5.1 Experimental Setup

To compare our methods against some of the most promising ones from the related work in Section 3.3, we looked across four metrics: the Area Under the Receiving Operating Characteristic (AUROC) curve, the F1-Score, and the Binary Recalls from the perspective of the defect class for closed-set and open-set defects, as well as their average. For evaluation, we chose binary metrics to have a uniform comparison between OSR and Semi-supervised methods, the latter not distinguishing between specific defect classes. This also aligns with our use case, where samples marked as “defects" or “unknown" will both be examined by human operators before being discarded, so the actual decision is whether a given sample is OK or needs a human check. The recall metric for defect classes is particularly important since it indicates what percentage of defects move through the system unnoticed by being marked as flawless. Regarding this metric, we also distinguish between open and closed set classes to allow us to discover potential trade-offs between the two types of classes. Finally, the F1-score and AUROC metrics showcase whether the models have a reasonable performance in the flawless class and

overall problem. For instance, some methods might achieve perfect recall by marking too many flawless images as defects, making them inefficient for a practical setting, since they will substantially increase the work of human operators, especially given that the majority of products passed through the system are not expected to be defective.

Moreover, for OSR classifiers operating on pre-extracted features, we compare across three of the most prevalent CNN architectures, namely Resnet50, VGG '16, and Inception v3, since the characteristics of the feature space greatly influence the models' discrimination capabilities. The baseline for our approach is a Multi-Layer Perceptron (MLP) operating on one of the above embeddings in a one-vs-all fashion. Additional baselines are provided through well-established anomaly detectors, namely One-Class SVM (OCSVM) [183], Isolation Forest (IF) [214] and Local Outlier Factor (LOF) [215].

Finally, we present a more fine-grained view of the performance of the most promising classifiers from each method category in Fig. 3.6 to examine the influence of each new class's features on the underlying algorithm's uncertainty profiles. We also choose the F1 metric as the ultimate performance indicator, containing both open- and closed-set performance information, and compare our approach against the most promising approaches including a statistical significance test.

The results are the average outcomes of 30 independently seeded runs for each measurement. They were performed in an environment with 4 CPU cores of 2.3GHz, 16GB of RAM, and access to an NVidia K80 GPU.

3.5.2 Examined Methods

As presented in the results section, we divided our compared methods into three categories according to their implementation requirements. The first group (**I-VII**) is methods operating on vector data for which we used pre-extracted features from Convolutional Neural Networks (CNN) trained initially on Imagenet (Resnet50, VGG16, and Inception v3). The second group consists of semi-supervised methods that learn only from the non-defective (**VIII-X**), followed by data augmentation techniques (**XI-XII**). Next, we describe how each of these methods has been applied to our use case:

I. MLP We used a single hidden layer architecture with 100 neurons leading to a 3-class classification head, both for the open and closed-set cases, and the 'adam' optimizer. We assessed the performance in both open and closed set cases based on whether a defective instance was assigned to any defect class.

II. SVM Despite being categorized as an unsupervised one-class classification method, OCSVM allows a small proportion of outlier instances in training, corresponding to the parameter ν. We fill out this proportion using the known defect instances in the training set. Otherwise, OCSVM works like a usual SVM but only forms a boundary for separating the good class from the rest of the instances. In our experiments, we used $\nu = 0.3$ and an RBF kernel.

III. Isolation Forest The idea behind isolation forests is the linear splitting of the feature space by individual trees until a point is "isolated" in a tree leaf. The anomaly

score assigned by the forest is an accumulation of how quickly each tree manages to separate the anomaly from the rest of the dataset. For the training of the IF we additionally use closed-set defect samples similar to OCSVM, setting the contamination factor (which again corresponds to the proportion of total defects to good images equal to 0.3). We also set the number of isolation tree estimators to 100.

IV. Local Outlier Factor A density-based anomaly detection method that is again trained on both the good and defect classes using a contamination factor of 0.3. The main idea behind LOF is that it compares the local point density of a given point to that of k of its neighbors and labels those with lower relative densities as anomalies. We chose $k = 20$ neighbors based on the Euclidean distance.

V. W-SVM As mentioned in the Related Work section, W-SVM is an ensemble of one-class and multi-class SVMs, whose scores are combined and calibrated using the Weibull distribution according to EVT. We used an RBF kernel and a 0.1 probability threshold for rejecting samples as an open set for the experiments.

VI. PI-SVM A more sophisticated extension of W-SVM is trying to model the probability of inclusion for each class using only in-class samples and EVT. The model was parameterized in a similar way to W-SVM.

VII. OpenMax OpenMax operates on the penultimate layer of a DNN to accommodate an “unknown" class and recalibrates scores using EVT. We used a tail size of ten samples to fit the Weibull distribution and an $a = 3$ corresponding to the total number of classes whose scores are recalibrated. In our case, we have very few (three) original classes, so we recalibrate all of them. For the DNN, we use the same MLP on top of pre-trained embeddings as above.

VIII. GANomaly A short description of functionality is provided in the Related Work. We used a latent vector size of 100 dimensions along with $w_{adv} = 1$, $w_{con} = 50$, and $w_{enc} = 1$ for the coefficients of the adversarial, contextual and encoder loss coefficient defined in [197].

IX. DFKDE This method consists of a backbone network to extract deep features followed by Principal Component Analysis (PCA) and Gaussian Kernel Density Estimation ([198]). In our use case, we use the 16 principal components explaining the most variance along with the euclidean distance and a 0.5 score threshold for anomaly classification.

X. DFM This approach tries to fit a Gaussian distribution or mixture of Gaussians to a DNN’s features after a DNN has been trained on a specific classification task and PCA has been applied to the feature vectors to reduce their dimensionality and thus improve computational speed. ([199]). In our use case, we train the model only on good images and use a Resnet50 backbone with a 0.97 variance retaining threshold for the PCA and the feature reconstruction score to rank anomalies.

XI. OSRCI Data augmentation through counterfactual images described in the Related Work. In terms of parameters, we followed [190] using a 20-dimensional latent space and a classifier architecture of two convolutional layers followed by two fully connected layers.

XII. OpenGAN Similar technique to OSRCI (see Related Work), using latent space interpolations between classes. A Resnet18 backbone is used for the feature extractor and Gaussian Kernel Density estimation for the final classifier.

Characteristics Methods	Pre-Extracted Features	Multi-class	Synthetic Data	#Hyper-parameters	Memory Usage	Requires GPU	Execution Time
MLP	✓	✓		Low	Low		Low
One-class SVM	✓	✓		Low	Medium		Low
Isolation Forest	✓	✓		Low	Low		Low
Local Outlier Factor	✓	✓		Low	Low		Low
WSVM	✓	✓		Medium	High		Low
PI-SVM	✓	✓		Medium	High		Medium
OpenMax	✓	✓		Medium	Medium		Low
Ganomaly				High	Medium	✓	High
DFKDE				Medium	Medium		Low
DFM				Medium	Medium		Low
OSRCI		✓	✓	High	Medium	✓	High
OpenGAN		✓	✓	High	Medium	✓	High

Table 3.1. *Qualitative comparison summary of the characteristics of the examined methods, according to their provided functionality, implementation and computational infrastructure requirements.*

A comparison summary of the implemented methods is shown in Table 1. In short, methods requiring pre-extracted features are low in terms of computational demands and hyperparameters, with the exceptions of SVM-based methods that need to load the whole dataset in memory. Semi-supervised methods are slightly more intensive computationally and in terms of hyperparameters but offer no multi-class functionality. An exception is Ganomaly which is closer to the data augmentation methods with high computational and hyperparameter requirements due to the adversarial training. Finally for pure data augmentation methods the creation and storage of the synthetic data should be taken into account.

3.5.3 GAN Training

To train the StyleGAN v3 generator used in our approach we had to choose between training a conditional model that would be able to generate images from all three classes and training three individual, unconditional models for each class. As shown in Fig. 3.5 the latter option yielded lower Frechet Inception Distance (FID) scores, at the cost,

however, of training three models instead of one. In the training process for each network we followed the guidelines provided in [174] for smaller datasets, based on finetuning the pre-trained network on FFHQ with a 256x256 image size.

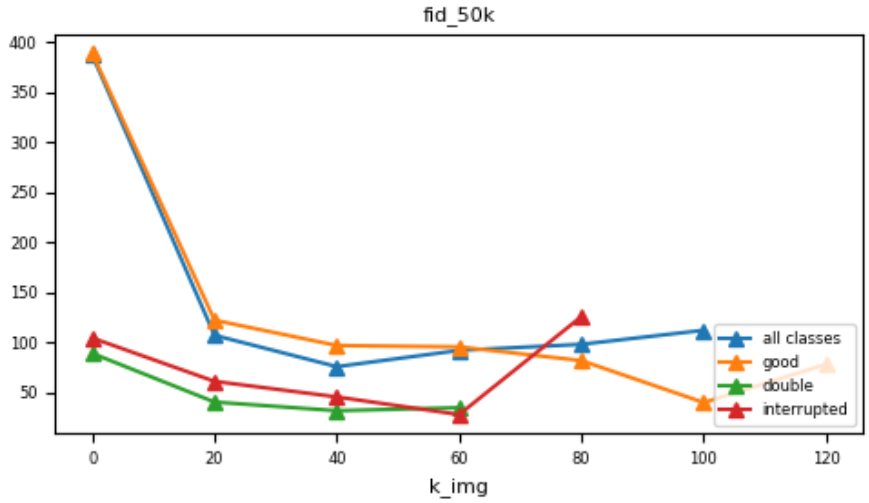


Figure 3.5. *Progression of FID while training candidate generator models. "all_classes" is the class-conditional model. The double and interrupted class models start from a pre-trained model of the "flawless" class for k_img=80.*

Moreover, due to the class imbalance, we first trained a model on the majority class (flawless) for 80k images. This model was then fine-tuned for all three classes until reaching a minimum (dip) in the FID score. While the conditional model achieved a minimum FID of 75.54, the individually trained models were able to reach FID scores below 40, by retaining FID improvements over more iterations. Training three models instead of a conditional one is of course more computationally expensive and could be prohibitive for problems with many classes. However, it was particularly beneficial in our use case where the high inter-class similarity requires lower FID scores.

3.5.4 Hyperparameter Tuning

To ensure that all comparisons and proposed methods were sufficiently tailored to the presented use case, hyperparameter tuning using grid search was performed to select the best hyperparameters from empirically sensible intervals as shown in Table B1.

For the more computationally expensive Deep Learning methods such as OSRCI and OpenGAN, we parameterized them following the given guidelines for parameterization for the datasets CIFAR-10 and Flowers102 respectively. For the proposed approach using StyleGAN v3 and Semantic Factorization, most hyperparameter tuning was focused on the number of images generated and the length of the semantic directions (max_dist). The number of semantics and the images generated per semantic traversal were both empirically set to 10. The chosen number of semantics represents the most important semantics as outlined by SeFa, which we threshold to include those that are visually meaningful. The number of images per traversal is empirically not that important since usually the

Method	**Parameter**	**Values**	**Chosen**
MLP	hidden layer size	[50, 100, 250, 500]	100
One-class SVM	kernel	[rbf, poly-3, poly-5]	rbf
	nu	[0.1, 0.2, 0.3, 0.4, 0.5]	0.3
Isolation Forest	estimators	[50, 100, 250, 500]	100
	contamination	[0.1, 0.2, 0.3, 0.4, 0.5]	0.3
Local Outlier Factor	neighbors	[10, 20, 50, 100]	20
	contamination	[0.1, 0.2, 0.3, 0.4, 0.5]	0.3
PI-SVM	kernel	[rbf, poly-3, poly-5]	rbf
	P-threshold	[0.05, 0.1, 0.15, 0.2]	0.1
WSVM	kernel	[rbf, poly-3, poly-5]	rbf
	P-threshold	[0.05, 0.1, 0.15, 0.2]	0.15
OpenMax	tail size	[10, 25, 30, 50]	10
	alpha	[1, 3, 5, 10]	3
Ganomaly	latent dim.	[50, 100, 250, 500]	100
	w_bec	[0.5, 1, 10, 50]	1
	w_rec	[0.5, 1, 10, 50]	50
	w_enc	[0.5, 1, 10, 50]	1
DFKDE	principal components	[8, 16, 32, 64]	16
	anomaly threshold	[0.4, 0.5, 0.6]	0.5
DFM	PCA threshold	[0.9, 0.91, ..., 0.99]	0.97
	anomaly threshold	[0.4, 0.5, 0.6]	0.5
OSRCI	latent dim.	[10, 20, 50, 100]	20
OpenGAN	num_images	[250, 2500]	250
	iters	[60000]	60000
	norm	[batch, instance]	instance
Proposed	mult_coeff	[1.5, 2, 3, 4]	4
	max_dist	[12, 15, 20]	15
	num_voters	[3, 9, 15]	3

Table 3.2. *Hyperparameter selection intervals.*

remaining images after filtering occur at the edges of the traversals and will be generated no matter the number set. We also considered different numbers of voters which did not impact our current voting scheme based on disagreement, however, examining it in conjunction with alternative voting schemes could lead to optimizations regarding the number of required synthetic images before filtering and is a fruitful direction for future research.

3.5.5 Experimental Results

Tables 1 to 3 show the results for pre-extracted feature-dependent methods across the three base networks. An immediate observation is that Resnet50 features achieve lower AUROC and F1 scores for most classifiers, with few exceptions, such as the Local Outlier Factor. On the other hand, it is surprising that the baseline method with VGG embeddings scores over 95% on those metrics and achieves a 0.9208 open-set recall without using open-set mechanisms. Although explaining the differences attributed to different base networks is difficult due to the complexity of their architectures, we spec-

	Resnet				
Method	AUROC	F1	R_c	R_o	R_{avg}
MLP	0,7414	0,8462	0,8800	0,2386	0,5593
One-class SVM	0,8353	0,7924	0,6245	**1,0000**	0,8122
Isolation Forest	0,8764	0,8213	0,6920	**1,0000**	0,8460
Local Outlier Factor	0,9121	0,8451	0,7285	**1,0000**	0,8642
WSVM	0,8596	0,7659	0,7385	0,8274	0,7830
PI-SVM	0,6709	0,8375	0,8628	0,1684	0,5156
OpenMax	0,6973	0,7737	**0,9165**	0,5834	0,7500
Proposed	**0,9952**	**0,9650**	0,8670	0,9772	**0,9221**

Table 3.3. *Evaluation of OSR methods over pre-extracted Resnet50 features, including AUROC, F1-score, Binary Recall on the closed set classes (R_c), and Binary Recall on the open set classes (R_o) and lastly $R_{avg} = \frac{R_c+R_o}{2}$.*

ulate that it could be explained by the difference in their receptive fields [13]. Resnet50 and Inception v3 have larger receptive fields than VGG16 which could make them more efficient in recognizing large objects but could also lead them to miss small details such as those found in interrupted prints or small discolorations. These facts make evident the importance of trying out different types of classifier embeddings when trying to optimize open-set performance. For this reason, we chose the three main, and most common in the literature, approaches for building CNNs namely Resnet, VGG, and Inception, which also have receptive fields of differentiated sizes, that could lead to significantly different results.

In general, some of the best-performing combinations are VGG with the baseline MLP, PI-SVM, W-SVM, and OpenMax and Inception with PI-SVM and OpenMax. Another interesting observation is that one-class classifiers achieve perfect recall R_o on open-set instances irrespective of the underlying embeddings. However, this comes with a significant decrease in closed-set recall R_c. Across all cases the embeddings seem to have a larger influence than the open-set mechanisms, as, for instance, highlighted by PISVM achieving a very high open-set recall on VGG embeddings (95.56%) but performing very low on Resnet50 embeddings (16.48%). In comparison, the proposed method shows consistency across different embeddings and between closed- and open-set recall, while at the same time maintaining high performance in AUROC and F1 scores, meaning that few flawless products will end up falsely marked as potential defects.

Table 4 contains results from semi-supervised and data-augmentation-based methods. Semi-supervised methods are trained on a subset of the flawless class instances and evaluated over a test set with data from all classes. In this case, all classes can be considered open-set since they are unknown at training time. However, we still evaluate it separately in defects of the original dataset and on simulated open-set defects. Between the two, there is a marked difference in performance for all methods, which we attribute to the approximate nature of simulated defects. Across all methods, DFM has the highest values across all metrics. We note that semi-supervised methods achieve a lower closed-set recall (R_c) than the methods presented in Tables 1-3, which is expected since

	VGG				
Method	AUROC	F1	R_c	R_o	R_{avg}
MLP	0,9777	0,9633	0,9320	0,9208	0,9264
One-class SVM	0,8767	0,8256	0,7060	**1,0000**	0,8530
Isolation Forest	0,8731	0,8598	0,8607	0,8664	0,8635
Local Outlier Factor	0,9090	0,8430	0,7095	**1,0000**	0,8548
WSVM	0,9022	0,8088	0,8122	0,9631	0,8877
PI-SVM	0,9902	0,9533	0,9111	**1,0000**	0,9556
OpenMax	0,9630	0,9389	**0,9707**	0,8932	0,9320
Proposed	**0,9965**	**0,9796**	0,9560	0,9952	**0,9756**

Table 3.4. *Evaluation of OSR methods over pre-extracted deep VGG '16 features.*

	Inception				
Method	AUROC	F1	R_c	R_o	R_{avg}
MLP	0,9325	0,9117	0,8695	0,6754	0,7724
One-class SVM	0,8771	0,8258	0,7080	**1,0000**	0,8540
Isolation Forest	0,9051	0,8389	0,7565	**1,0000**	0,8783
Local Outlier Factor	0,9149	0,8498	0,7030	**1,0000**	0,8515
WSVM	0,9106	0,6797	0,9200	0,8856	0,9028
PI-SVM	0,9834	0,9577	0,9040	0,9856	0,9448
OpenMax	0,9409	0,9289	**0,9500**	0,8464	0,8982
Proposed	**0,9954**	**0,9752**	0,9490	0,9884	**0,9687**

Table 3.5. *Evaluation of OSR methods over pre-extracted Inception v3 features.*

Method	AUROC	F1	R_c	R_o	R_{avg}
Ganomaly	0,8242	0,8930	0,6100	0,9460	0,7780
DFKDE	0,9848	0,7401	0,7700	0,9720	0,8710
DFM	0,9909	0,8347	0,8500	0,9800	0,9150
OSRCI	0,7884	0,6813	**0,9900**	0,8540	0,9220
OpenGAN	0,9399	0,8858	0,8483	0,6860	0,7672
Proposed + VGG	**0,9965**	**0,9796**	0,9560	**0,9952**	**0,9756**

Table 3.6. *Evaluation of semi-supervised and data-augmentation-based methods*

closed-set defects are considered by these methods as “unknown" at training time. This could also explain the lower F1 scores. However, their AUROC scores tend to be higher (e.g., 98.48% for DFKDE and 99.09% for DFM), possibly due to their better-calibrated probability outputs.

In regards to methods based on data augmentation, OSRCI, shows high defect recall scores with lower AUROC and F1, hinting at a potential marking of many flawless instances as defects. On the other hand, OpenGAN is more stable across these metrics despite slightly lower recalls. Overall, the inability of these more sophisticated methods to outperform previous ones could be attributed to the difficulty of generating differentiated data for use cases where instances from different classes are very similar to each other such as defect detection. The relative improvement shown by the proposed method could be attributed in part to StyleGAN’s higher expressive and generalization capabilities compared to earlier GAN architectures, but also to our novel voting-based filtering mechanism.

To also shed more light on the open-set performance we present conducted measurements using a fine-grained boxplot of the class-specific accuracies, against 5 selected high-performing methods representing each method type (semi-supervised, SVM-based, data augmentation, etc.) (Fig. 3.6). We note that horizontal and vertical flips and discolorings are well-recognized by all top-performing methods. In line defects and missing letters, we see MLP and Openmax on VGG features and OSRCI having more difficulties as well as more variable results. Generally from the existing OSR approaches, PISVM on Inception features performs more stably across all classes, on par with DFM and the proposed approach.

Finally, Table 5 compares the F1-score and average recall of the most promising methods, also evaluating the statistical significance of their differences to the proposed approach using a paired t-test. We chose the F1-score as the ultimate measure of comparison since it is evaluated on a set containing both “novel" and “known" examples, that could realistically occur in a production environment, and is also less sensitive to class imbalances. In contrast to the recall metrics presented, the F1-score includes information on the methods’ performance on flawless images. As an illustrative example, a hypothetical method marking every image as a defect would have perfect open-set and closed-set recalls, but its F1-score would be low due to every flawless image being misclassified. The F1-score is therefore also evaluated as an attempt to provide a more balanced and

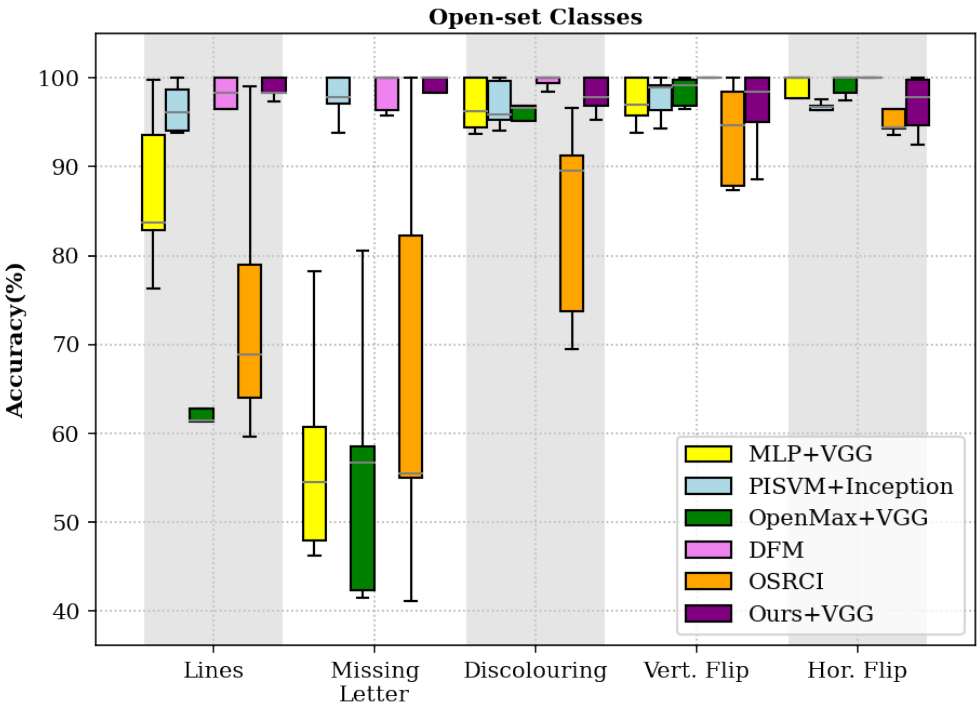


Figure 3.6. *Box plot of open-set class-specific accuracy scores for highest performing methods per type.*

all-around picture of the methods' performance.

In summary, it is important to observe how the three main challenges of automated visual quality inspection, identified in the introductory section, manifest themselves throughout our experimental process and how they are addressed. Firstly, the lack of collectible defect data is evident in the number of interrupted and especially double print images in the examined dataset. Although we do not address the resulting class imbalance directly, we can observe that methods that have been pre-trained on a large and diverse dataset (e.g., StyleGAN trained on a dataset of celebrity faces, Resnet50 trained on Imagenet, etc.) can cope with class imbalance in the "known" classes. Class imbalance has also been taken into account when evaluating, both by generating equally many "unknown" defects as "known" ones and by using the Recall metric which is not affected by the majority of images belonging to the flawless class. Secondly, the high inter-class similarity proved especially problematic for the data augmentation and semi-supervised methods, which most likely had trouble either generating sufficiently differentiated defects or recognizing very small defects without being given training samples. Our choice of StyleGAN as a generator was key in tackling this issue, as due to its more sophisticated architecture it could generate images with small differences with high fidelity. Last but not least, a satisfactory solution to the third challenge, the robustness to novel defects, was achieved, however, the improvement would not have been possible without considering and successfully addressing the two previous challenges.

Method	F1-score	p-value	R_{avg}	p-value
DFM	0,8347	0,0011	0,9150	0,0187
OpenMax + VGG	0,9389	0,0005	0,9320	0,0004
PISVM + VGG	0,9533	0,0011	0,9556	0,0158
PISVM + Inception	0,9577	0,0024	0,9448	0,0006
MLP + VGG	0,9633	0,0029	0,9264	0,0144
Proposed + VGG	**0,9796**	—	**0,9756**	—

Table 3.7. *Comparison against the best performing OSR methods over their F1-score and Recall averaged from both open- and closed-set samples with statistical significance scores.*

3.6 Summary

In this work, we introduced a novel data-augmentation method to make defect recognition classifiers more robust against novel defects unseen in the training set. Applied to a real-life manufacturing use case along with methods from the relevant literature ranging from SVM-based approaches to semi-supervised methods it achieved high performance as well as consistency across different classifier embeddings. This could be attributed to the fidelity and variability of synthetic images that can be generated from StyleGAN as well as to the steerability of its latent space. An important feature is the treatment of novelty data generation through latent space traversals as an imperfect process that needs to undergo a filtering step. To that end, a simple voting scheme was introduced to isolate images that cause high confusion between voting classifiers and add them to the augmented dataset in the form of an “unknown" class. Despite its high performance, the proposed method is still subject to improvements. Its main drawbacks are the high training times required by StyleGAN, even when transferring knowledge from a pre-trained model, as well as the large amount of redundant data generated, which is then discarded by the filtering process. These open up two avenues for future work, namely the investigation of more lightweight but still steerable GAN architectures and the more efficient extraction of confusing samples from the latent space.

Chapter 4

Robust Novel Defect Detection with Neurosymbolic AI

4.1 Background

Quality inspection in production systems is advancing with digitization, integrating sensors and AI algorithms. Visual inspection, crucial for tasks like painting, benefits from AI techniques in image processing and computer vision. However, fully automating visual quality inspection faces significant challenges. In our previous work on assessing brand prints on finished shaver shells [17], we identified three main challenges: insufficient training data, high visual similarity between flawless and defective products, and unanticipated defects during operation. Traditional methods like Convolutional Neural Networks (CNNs) often fail in these applications due to their dependency on extensive labeled data and their limited ability to handle novel defect types. For instance, [11] highlighted that sometimes unsupervised and semi-supervised anomaly detection methods missed defects due to the variability in defect appearance and position, demonstrating the limitations of traditional machine learning models in real-world manufacturing scenarios. Additionally, it has been shown that achieving robustness remains a significant challenge for conventional methods, which are not designed to identify novel samples even when they result from small image corruptions to known samples [10]. In this section, we propose a Neurosymbolic approach to defect detection, which also proves to be quite robust to novel defects.

The aim of Neurosymbolic AI [14] is to fuse two existing branches of AI, namely Symbolic AI (or sybmolism) and Statistical Machine Learning (or connectionism), so as to combine the benefits of both approaches into the next generation of AI [15]. Symbolic AI relies on hand-crafted rules expressed through Logic Formulas and Ontologies, while Statistical Machine Learning is mainly characterized by neural networks that learn from data. While Symbolic AI makes automated decisions fast and explainable, it requires significant effort from domain experts to gather and codify the symbolic knowledge consisting of entities, relationships and rules governing those relationships. Additionally, the resulting systems handle ambiguous or noisy, real-world data inflexibly. On the other hand, bottom-up statistical approaches, such as (Deep) Neural Networks, deal with these problems quite well having found substantial real-world application, most notably in the

domains of Computer Vision and Natural Language Processing. Nevertheless, they come with their own set of issues, such as opaqueness to their inner workings and therefore lack of explainability and trustworthiness, lack of robustness to adversarial attacks and unknown inputs [16][17], as well as data inefficiency and sensitivity to data imbalances [4]. In this work we use Neurosymbolic AI to increase the generalizability of a statistical ML classifier, and make it more robust to novel inputs (i.e., novel production defects). Specifically we take advantage of the infusion of symbolic rules via Logic Tensor Networks to enhance a fine-grained problem-specific supervised classifier with the capabilities of a more general unsupervised classifier. While alone the unsupervised classifier generates many false positives, its combination with the original classifier through Neurosymbolic AI results in increased open-set recognition capabilities.

4.2 Related Work

4.2.1 Neurosymbolic AI

Neurosymbolic AI has been applied in various application scenarios, introducing new learning capabilities in different domains, such as common-sense reasoning [216], visual scene understanding [217][218] and scientific Discovery [219]. While there exist many taxonomies of Neurosymbolic AI methods, the most notable and extensive one being [220], the two categories we consider most fundamental are the ones described in [221], namely Learning for Reasoning and Reasoning for Learning.

The first group of methods are extensions of existing symbolic reasoning methods that utilize empirical machine learning either to make sense of unstructured data or to speed up their reasoning process. For instance, Neuro-Symbolic Concept Learner (NS-CL) [222] uses a CNN-based visual perception module followed by a semantic parsing module and a symbolic reasoning module to make sense of visual scenes. Additionally, there are approaches that use statistical machine learning methods to automate the building of logical rules in a data-driven manner, such as methods extending Markov Logic Networks [223] [224] and differentiable Inductive Logic Programming [225]. In the Natural Language Processing (NLP) domain, IBM toolkit's Neural Unification for Logic Reasoning over Natural Language [226] uses transformers to help detect logical contradictions between a natural language corpus and a natural language query.

The second group of the taxonomy, Reasoning for Learning, uses neural classifiers as the basis for learning, that are assisted through the incorporation of symbolic knowledge, either in the form of knowledge transfer ([227], [228]) or in the form of constraining/regularization. Two important constraining approaches that are very relevant to this work are Logic Tensor Networks (LTN) [18] and the Symbolic Probabilistic Layer (SPL) [229].

The Semantic Probabilistic Layer (SPL) introduces a fully independent layer that can be added on top of an existing network architecture (e.g., Resnet50) enforcing external logical constraints. In this layer simple logical formulas are encoded as Ordered Binary Decision Diagrams (OBDDs) which in turn are transformed to differentiable Probabilistic Circuits (PCs). It is important to note that even though this transformation is calculated quickly in

practice, its worst case can be exponential. The incorporation of this final layer leads to a readjustment of the conversion of logits to probabilities so that, for instance, prohibitive logical constraints output a pseudo-probability of 0, while the rest of the probability mass in readjusted. SPL guarantees strict consistency with the symbolic rules and has low sample complexity. However, it only works with simple logical propositions and does not incorporate first-order logic.

Logic Tensor Networks (LTN) is one of the most established loss-based regularization methods. LTNs use grounding, a technique that maps first-order logic propositions to real-valued tensors and corresponding mathematical operations. These tensors have to be of different sizes depending on the input datatype and their elements are between 0 and 1 corresponding to their truth value (similar to fuzzy logic). The end result of this process is a real-valued equation of tensor variables (these depend on the algorithm inputs or on features of the inputs) whose result is the degree of truth of the initial logical proposition. This new equation is differentiable and can be used as a term in the loss function that will guide weight updates in a Neural Network during back-propagation. LTNs have been used in a variety of real-life domains such as manufacturing [230] and maintain high accuracy also guaranteeing a high degree of satisfiability of the constraints as well as lower sample complexity. However, complete satisfiability of the symbolic constraints is not guaranteed.

Specifically in the domain of defect detection, Neurosymbolic AI has been used to improve transparency and explainability in cantilever beam defect detection [230] and to drive diagnosis of automotive production faults [231]. In [232] convolutional neural networks perform localization and recognition on video inputs gathered from real-life food product labelling production lines. Their predictions are then used by a knowledge-base-aided symbolic component to support decision making over the state of the production system. In our work we will be applying Neurosymbolic AI with a different but complementary purpose, namely to enable neural network classifiers to expand their capabilities to novel defects.

4.2.2 Open-set Recognition

In the proposed approach we will use Neurosymbolic AI as a means towards Open-Set Recognition (OSR). OSR is about classifying instances in the open-set, meaning the set of classes the classifier has not seen any instances of during training. Contrary, the closed-set contains classes the classifier has been trained on. The OSR problem was formally defined by [233] as an attempt to minimize misclassification risk in the open space. There is a variety of OSR implementations, such as Statistical Methods (e.g., WSVMs [182]), Semi-supervised Deep Learning (e.g. Deep Feature Modelling (DFM) [198]), but broadly the OSR problem can also be addressed by general anomaly detection techniques (e.g., Isolation Forest). In the context of defect detection it has been mostly applied in the semiconductor industry. For instance, in [200], a CNN with a distance and clustering-based approach was applied to a wafer map inspection scenario to detect wafer map products deviating from the training set. Also applied on wafer maps, Optimal Bin

Embedding [234] relies on extracting meaningful embeddings that aim to increase cluster quality and differentiation between the open and closed sets. Another approach based on specialized embedding extraction is introduced in [235] which uses a Submanifold Sparse Convolutional Network architecture to extract a latent representation serving as input to a Gaussian Mixture Model (GMM) outlier detector. We address a very similar Open-set recogntition problem, but for data-scarce scenarios that need to leverage OSR techniques over the use of predefined embeddings (e.g., Resnet50) with transfer learning. The advantage of addressing such scenarios is the lower demands on data collection in a domain where collecting sufficient defects is difficult and detrimental to the ramp-up time of a Visual QA system on new products.

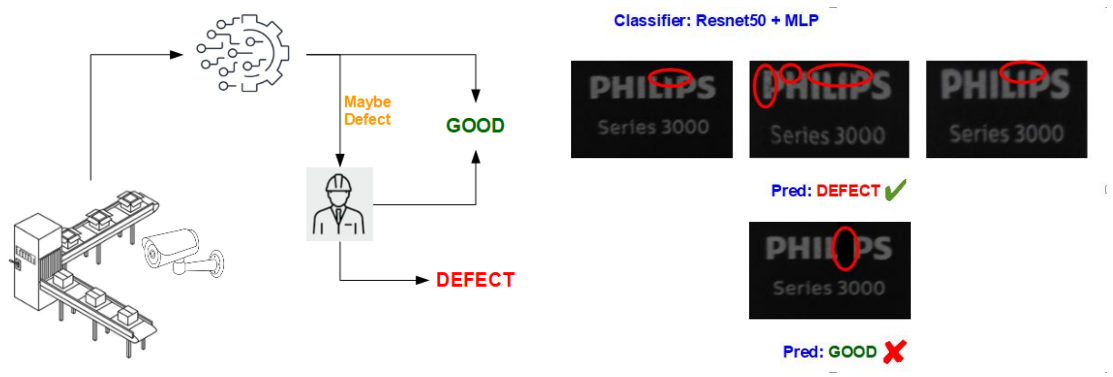
Existing solutions in Neurosymbolic AI and Open-set Recognition (OSR) offer valuable capabilities for defect detection but also exhibit notable limitations(also see the Experimental Results, Section 4.4.2). Neurosymbolic methods such as Logic Tensor Networks (LTNs) [18] and the Symbolic Probabilistic Layer (SPL) [229] introduce symbolic constraints to enhance trustworthiness and safety, yet they rely heavily on well-defined logical rules which, if strictly enforced such as in SPL, would struggle with ambiguous or noisy data common in manufacturing. Similarly, traditional OSR methods, including One-Class SVM (OCSVM) and Weibull-calibrated SVM (WSVM) [182], aim to minimize misclassification risk but often fail in dynamic environments due to high false positive rates when applied to complex visual data. Unsupervised methods like Isolation Forest (IF) [236] are effective in identifying novel defects but tend to produce many false positives [19], which might be exacerbated in cases such as ours due to the lack context-specific knowledge to differentiate benign variations from actual defects. Semi-supervised methods such as Deep Feature Modeling (DFM) [198], although effective in recognizing known defects, often require extensive fine-tuning and are less effective with limited labeled data. These limitations underscore the need for innovative approaches, like the proposed Neurosymbolic AI framework, which combines the strengths of symbolic reasoning and statistical learning to improve defect detection robustness and generalizability.

4.3 Methods

4.3.1 Problem Setting

The specific setting of the real-life problem we are examining regarding the quality assessment of shaver shell prints is as follows. A camera system is placed on the production line and specific measures are taken to enforce uniform lighting conditions to avoid shadowing and gloss. The images taken are saved in a local server running a machine learning defect recognition model. The outputs of this model are "GOOD" and "Maybe Defect". "GOOD" products are moved on to the next production stage, although they can be occasionally sampled for manual Quality Assessment (QA). Potential Defects are sent to human operators to finally determine if the product is indeed defective or just a false positive. As it will be explained in the results section this system is designed to be safe in terms of defect recall, meaning it is very strict in what constitutes a "GOOD" product,

since products marked as such can pass through QA mostly without human supervision. This also leads to an increased number of false positives which are OK, as long as they do not over-burden human operators. A diagram detailing the above is shown in Fig.4.1(a).



(a) *QA workflow* **(b)** *Robustness problem*

Figure 4.1. *Figure 1(a) is a high-level depiction of the visual quality assessment workflow. Potential defects identified by the AI are also examined by a human before being discarded, while products labelled "GOOD" by the AI pass QA. Figure 1(b) shows how the AI system using supervised learning based on Resnet50 runs into issues when encountering novel defects that, despite looking more severe, are incorrectly labelled.*

The challenge in this setting is collecting enough defect images to create the training dataset. As all images come from real-life production and defects are usually rare, not all types of possible defects can show up during collection. Therefore, the system should also be robust to novel defects it has never seen before. However, as we realised using a vanilla Resnet50 Multi-Layer Perceptron (MLP) classifier this is not always the case. As illustrated in Fig.4.1(b), the system learns to recognize small interruptions that have many samples in the training set. Nevertheless, when faced with a much larger and more obvious, but otherwise novel interruption such as a missing letter, it fails to recognize it. This is what led us to investigate augmenting the ML algorithm with techniques such as One-class Learning, Open-set Recognition, Semi-supervised Learning and most importantly Neurosymbolic AI and Logic Tensor Networks.

4.3.2 Why Logic Tensor Networks?

By using Neurosymbolic AI, and specifically LTNs, for the problem setting described in Section 4.3.1, our ambition is to combine the benefits of unsupervised learning methods with the specificity of supervised methods. While the former can generalize to any anomalous output, the latter can learn very well how to recognize the particular defects that occur in the training dataset. As mentioned in Section 4.1, visual defect detection and classification is a problem with very particular challenges, which obstruct its full automation. Additionally, expert knowledge about what constitutes a defect cannot be fully encoded into clear-cut rules, which is another hindrance to symbolic and Neurosymbolic approaches. However, a Neurosymbolic approach can still benefit from clear-cut, but non-universal cases (e.g., when there are clear indications of a defect but the ex-

pression of these indications through rules cannot be universally applicable due to its many edge-cases). These challenges led us to choose LTNs for this problem. LTNs do not strictly enforce their symbolic constraints, thus allowing their user to be more lax with formulating the symbolic rules. Moreover, the knowledge of clear-cut defects can still be leveraged to speed-up training compared to a classical supervised learning algorithm.

An important aspect of Logic Tensor Networks is how constraints are transformed to be differentiable and part of the end-to-end training process. This is achieved through a technique called "grounding" which is very close to fuzzy logics. More specifically each individual proposition or fact is encoded through a multidimensional tensor, which in our case corresponds to vector embeddings extracted from the input images. Predicates can be applied to these tensors in the form of differentiable mathematical functions which can also have learnable parameters such as Artificial Neural Networks. The application of these predicates should yield a real value between 0 and 1 which corresponds to the degree of truth of the predicate applied to one or multiple propositions. Building on top of that, logical operators can be used to combine different predicate results. For example, a logical $a \wedge b$ can now be calculated as $a\dot{b}$ and $a \implies b$ is calculated as $\frac{b}{a}$ if $b < a$ or 1 if $b > a$. Of course there are many different mappings from first-order logic to real operators, many of which are described in detail in the LTN paper [18]. After making the logical propositional differentiable, their degree of satisfaction can be added as a loss function term to be optimized during training.

4.3.3 Our approach

LTN's "grounding" of symbolic rules to their real-valued logic equivalents enable it to constrain a statistical machine learning algorithm to closely adhere to pre-defined symbolic rules during its training phase. At the same time, utilizing these rules requires the encoding of expert knowledge in a corresponding form which, in our case, is difficult to achieve. The production scenario described in Section 4.3.1 is supposed to operate in a flexible and agile manufacturing production line. Such production lines are characterized by a large degree of customization leading to frequent changes in product specifications. This constant flux makes it hard for production operators to develop enough expertise to come up with a complete set of rules for defect detection. Additionally, the nature of the image data makes it hard to link these rules with properties of images. A property such as, for example, surface smoothness is not straightforward to define as an image processing function/predicate to be used by the LTN. For these reasons we use an unsupervised classifier as the "expert".

The criterion for choosing an unsupervised classifier is for it to have good novelty recognition properties and a simple adaptable implementation. Following our results from previous work [17] we chose the Isolation Forest, as it offers a scalable implementation, needs limited fine-tuning and has been shown to perform well in a variety of datasets [19]. Despite its high performance on unknown images, IF is not that effective in the known classes from the training set. To overcome this shortcoming we created the rules outlined below, where A is the base MLP classifier and U the unsupervised Isolation Forest

classifier. These rules enforce upon the MLP a soft logical constraint to follow the output of U when it predicts a defect.

$$\begin{aligned} SatAgg\Big\{ & [\forall x(l_S(x) = 1 \implies A(x) = 1)] \wedge \\ & [\forall x(l_S(x) = 0 \implies A(x) = 0)] \wedge \\ & [\forall x(U(x) = 0 \implies A(x) = 0)]\Big\} \end{aligned}$$

The formula above contains two additional constraints needed for classification that ensure that the prediction $A(x)$ is consistent with the supervision label $l_S(x)$. Thus, the base classifier A is only trained to satisfy the rule-set outlined. The complete training process is also illustrated as a diagram in Fig.4.2

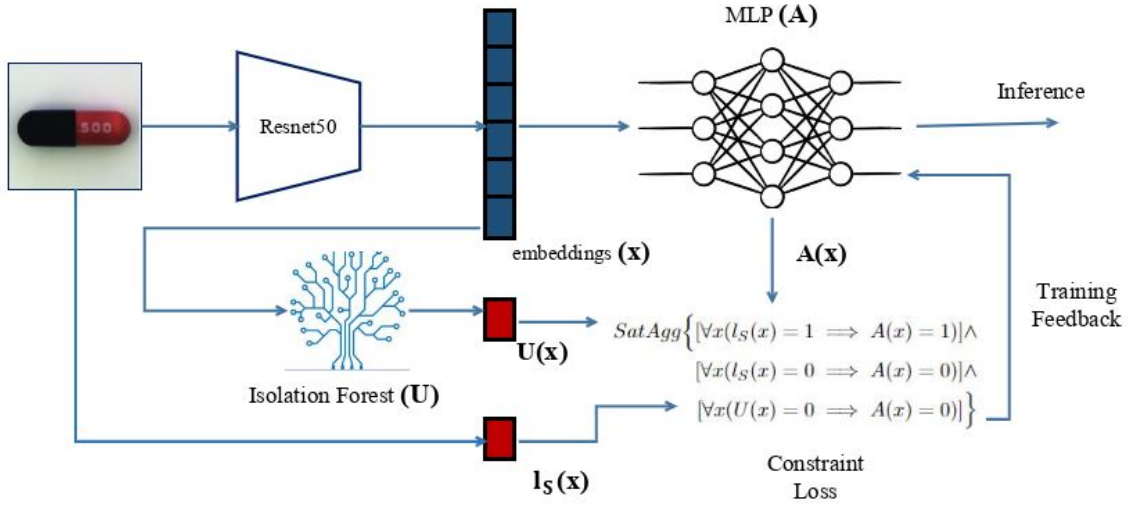


Figure 4.2. *Training Workflow with LTN using embeddings for empirical learning and symbolic rules derived from an Isolation Forest's predictions. The symbolic rules are "grounded" and embedded into the loss function to guide training.*

4.3.4 Datasets

The dataset provided by Philips Consumer Lifestyle B.V. consists of RGB images collected from the factory's pad printing process for building an automated quality inspection system. It contains images of flawless products as well as two types of defects: double prints and interrupted prints. The dataset has been manually labeled by multiple quality inspectors to ensure accuracy. Manufacturing defects are rare, resulting in an imbalanced dataset, which was taken into account during evaluation. The images are 220×360 pixels in size, and the dataset is divided into training and testing sets.

Representative examples of flawless products, double prints, and interrupted prints are shown in Figure 4.3(a)-(c). The training set comprises approximately 70% of the images, while the remaining 30% are used for performance evaluation.

To assess robustness, synthetic images simulating novel defects were created, includ-

(a) *Good* **(b)** *Double Print* **(c)** *Interrupted*

(d) *Line Scratches* **(e)** *Missing Letter* **(f)** *Discoloured*

Figure 4.3. *Original ((a)-(c)) and Synthetic Test ((d)-(f)) Samples from the Shavers Dataset*

ing line interruptions, missing letters, discoloration, and flips. These synthetic defects were merged with the test set in proportion to the original defects, resulting in a realistic imbalance scenario for evaluation.

The final test set contains 800 flawless images and 250 images with known defects, augmented with 250 novel defect images randomly generated from the synthetic classes. Synthetic examples of unexpected defects are depicted in Figure 4.3(d)-(f). This comprehensive dataset allows for the evaluation of machine learning algorithms in a realistic manufacturing defect detection scenario.

Moreover, we additionally assessed our method on six additional datasets of product defects from the MVTec AD collection [11]. This is a collection of datasets consisting of surface and object defects. For our evaluation, we chose products with many different defect classes available, so that in each run we could keep two randomly-chosen defect classes in the training set (the same number as in the shavers dataset) and use the rest as open-set defects. To that end, we used the carpet, capsule, grid, pill, tile and leather datasets, samples of which are shown in Fig. 4.4.

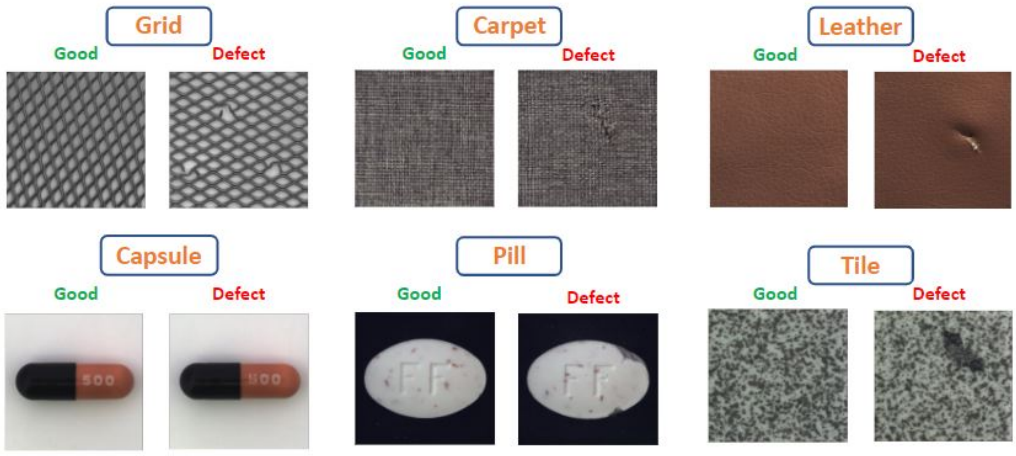


Figure 4.4. *Product categories' samples from the MVTEC-AD datasets*

4.4 Results

4.4.1 Experimental Setup

To evaluate our method, we compared it with promising ones from the areas of Open-set Recognition (One-class SVM - OCSVM, Weibull SVM - WSVM) and Unsupervised (Isolation Forest - IF) and Semi-supervised (Deep Feature Modelling - DFM) anomaly detection, as well as a Multi-Layer Perceptron (MLP) baseline. All methods used pre-extracted Resnet50 embeddings. We focused on four key metrics: Area Under the Receiving Operating Characteristic (AUROC) curve, the overall test-set Defects Precision, F1-Score, and Binary Recalls for closed-set and open-set defects. We chose binary metrics for uniform comparison across supervised and Semi-supervised methods, aligning with our use case where both "defects" and "unknown" samples are examined by human operators. The recall metric for defect classes (open and closed-set) is crucial, its complement indicating the percentage of defects missed by the system. We distinguish between open and closed set classes to uncover potential trade-offs. F1-score and AUROC metrics assess models' performance in the flawless class and the overall problem, ensuring efficient performance without excessive marking of flawless images as defects. Defect precision is also monitored, as a low score in this metric suggests overburdening the human operator with defect false positives. Results represent averages from 30 independently seeded runs, conducted on a system with 4 CPU cores, 16GB RAM, and an NVidia K80 GPU. 95% confidence intervals are also give for each metric.

4.4.2 Experimental Results

According to our experimental setup we first present the results for the shavers dataset in Table 4.1 and then proceed to the MVTEC-AD products which are shown collectively in Table 4.2. The best scores for each metric are highlighted in bold, while second-best scores are shown in gray. It is important to mention again that from our usecase's perspective the most important metric is closed-set recall since this concerns the most common defects, and we want to make sure as few of them as possible pass through the system in Fig. 4.1(a) unnoticed. The second most important metric is open-set recall as this shows our system's robustness to novel defects that are rarer but might still appear in the production line. The purpose of the other metrics (AUROC, Precision, F1-score) is to check that the trade-offs of achieving high closed-set and open-set recall scores are acceptable.

Dataset	Method	AUROC	Prec.	F1-score	R_open	R_closed
Shavers	MLP	74,94 ± 1,27	**94,11 ± 2,36**	87,67 ± 0,50	24,55 ± 1,45	**91,63 ± 2,60**
	OCSVM	83,02 ± 0,44	65,05 ± 0,72	79,55 ± 0,42	**100,00 ± 0,0**	62,66 ± 1,80
	IF	87,10 ± 0,41	69,08 ± 0,68	82,40 ± 0,45	**100,00 ± 0,0**	70,46 ± 1,88
	DFM	**99,13 ± 0,19**	84,22 ± 1,35	**99,64 ± 0,14**	90,33 ± 3,02	84,99 ± 2,25
	WISVM	81,99 ± 2,80	60,91 ± 1,80	75,84 ± 1,31	86,71 ± 2,86	77,89 ± 1,49
	LTN	**97,80 ± 0,40**	**93,34 ± 1,16**	**92,79 ± 1,01**	60,69 ± 8,15	**98,96 ± 0,54**

Table 4.1. *Comparison of methods on the Shavers dataset*

The first method we assess is the MLP on top of Resnet50 embeddings which achieves

high precision, indicating that when it predicts a defect, it is usually correct. However, its recall for open defects is very low, suggesting it may miss some instances of open-set defects. This is the baseline issue we want to address. We see that other dedicated methods such as One Class SVM and Isolation Forest achieve perfect scores in the open-set, however their performance in the closed-set is lacking as they are not explicitly trained on the training set itself. Deep Feature Modelling (DFM), which is trained on good images only, outperforms other methods in terms of AUROC and F1-score. This shows a very good ability to recognize good images, and it also achieves high open-set performance. Our newly introduced LTN approach achieves high AUROC, Precision and F1-scores, being consistently very close to DFM and the MLP, indicating its ability to accurately classify flawless products. It most importantly demonstrates the highest recall for closed defects, being by 8% higher than the second-best MLP. Its open set recall is comparatively low but we still see that the addition of Neurosymbolic AI to the MLP baseline brings a substantial - almost threefold - increase in this metric.

Dataset	Method	AUROC	Prec.	F1-score	R_open	R_closed
Carpet	MLP	92,89 ± 1,08	**95,04 ± 2,58**	83,47 ± 2,45	48,13 ± 10,67	74,80 ± 13,30
	OCSVM	74,23 ± 2,86	63,01 ± 2,79	72,08 ± 1,70	59,64 ± 6,17	59,26 ± 7,11
	IF	86,80 ± 1,58	70,56 ± 2,58	79,39 ± 2,42	**86,00** ± 3,01	81,06 ± 4,60
	DFM	**98,44 ± 0,27**	**99,45 ± 0,47**	**84,08 ± 1,26**	79,37 ± 4,87	79,59 ± 4,26
	WSVM	72,63 ± 2,00	58,81 ± 8,74	58,48 ± 10,11	63,86 ± 15,12	**84,20 ± 8,85**
	LTN	**97,47 ± 0,98**	88,66 ± 3,66	**91,74 ± 2,04**	**89,68 ± 4,00**	**99,53 ± 0,72**
Capsule	MLP	**93,87 ± 1,11**	**98,66 ± 0,81**	**78,99 ± 1,64**	51,28 ± 5,31	**94,20 ± 2,81**
	OCSVM	71,20 ± 2,71	72,82 ± 2,88	66,18 ± 2,19	57,33 ± 5,34	57,93 ± 5,68
	IF	81,25 ± 2,44	75,66 ± 3,03	71,72 ± 2,93	73,46 ± 5,39	67,46 ± 4,31
	DFM	**98,55 ± 0,61**	**98,72 ± 0,64**	**83,02 ± 3,78**	**83,77 ± 5,03**	82,80 ± 8,48
	WSVM	72,79 ± 6,09	56.18 ± 2,57	42,96 ± 3,08	67,46 ± 8.32	87,13 ± 6,02
	LTN	85,92 ± 11,19	83,19 ± 10,88	66,79 ± 23,35	**91,28 ± 8,57**	**99,80 ± 0,31**
Grid	MLP	72,98 ± 2,80	**76,20 ± 4,98**	81,02 ± 1,01	17,46 ± 3,99	**72,53 ± 8,75**
	OCSVM	41,32 ± 2,87	30,52 ± 2,50	66,82 ± 1,79	26,13 ± 4,88	24,13 ± 9,17
	IF	47,65 ± 2,72	33,17 ± 1,84	63,64 ± 1,98	36,80 ± 6,02	35,66 ± 8,97
	DFM	**93,60 ± 1,20**	**91,53 ± 2,55**	**81,23 ± 1,84**	**68,57 ± 5,97**	69,13 ± 5,85
	WSVM	40,18 ± 2,25	38,68 ± 5,10	58,52 ± 7,22	47,51 ± 12,15	63,80 ± 12,29
	LTN	**81,42 ± 6,80**	74,28 ± 11,82	**84,47 ± 4,78**	**62,22 ± 13,98**	**86,13 ± 7,09**
Pill	MLP	86,62 ± 1,60	95,82 ± 3,23	65,88 ± 3,60	32,72 ± 13,68	66,06 ± 12,28
	OCSVM	58,90 ± 2,01	68,79 ± 1,71	56,98 ± 1,45	53,57 ± 4,82	58,80 ± 11,84
	IF	68,28 ± 1,89	72,19 ± 1,78	60,16 ± 2,26	64,13 ± 4,37	58,53 ± 9,86
	DFM	**98,21 ± 0,35**	**99,84 ± 0,22**	**67,84 ± 3,52**	**70,10 ± 4,16**	**70,86 ± 9,62**
	WSVM	62,05 ± 5,00	70,44 ± 3,56	56,28 ± 5,87	54,82 ± 11,92	74,40 ± 8,62
	LTN	**95,43 ± 2,66**	**95,91 ± 1,73**	**88,36 ± 3,43**	**87,92 ± 6,05**	**95,33 ± 2,73**
Tile	MLP	97,74 ± 0,87	**99,38 ± 0,75**	**87,37 ± 2,86**	60,88 ± 10,75	**96,20 ± 3,53**
	OCSVM	66,48 ± 3,35	62,44 ± 2,31	68,22 ± 2,02	55,86 ± 6,56	57,73 ± 7,79
	IF	87,77 ± 2,25	71,91 ± 1,69	77,88 ± 1,69	**84,40 ± 5,94**	82,00 ± 8,21
	DFM	**99,34 ± 0,17**	**99,69 ± 0,34**	83,65 ± 0,26	73,06 ± 9,41	79,40 ± 13,42
	WSVM	65,36 ± 6,65	57,08 ± 5,94	56,64 ± 9,17	54,84 ± 10,17	85,46 ± 5,14
	LTN	**97,92 ± 1,60**	91,13 ± 3,00	**93,02 ± 2,68**	**90,97 ± 7,18**	**96,86 ± 2,61**
Leather	MLP	97,54 ± 0,95	**97,97 ± 1,01**	86,48 ± 2,32	62,93 ± 8,59	93,93 ± 3,02
	OCSVM	70,56 ± 3,60	68,39 ± 2,94	71,16 ± 2,65	64,97 ± 5,99	49,80 ± 6,52
	IF	92,93 ± 1,07	79,35 ± 1,91	85,62 ± 1,65	**96,35 ± 1,70**	95,26 ± 3,14
	DFM	**99,97 ± 0,01**	**99,92 ± 0,01**	**97,60 ± 0,77**	**97,91 ± 1,07**	**95,73 ± 1,47**
	WSVM	60,39 ± 4,61	70,56 ± 7,60	68,98 ± 7,79	49,06 ± 15,13	81,26 ± 4,18
	LTN	**99,00 ± 1,09**	96,42 ± 1,96	**95,30 ± 3,44**	89,73 ± 12,65	**99,66 ± 0,71**

Table 4.2. *Comparison of methods on the various MVTEC-AD product datasets*

Regarding the results in the MVTEC-AD datasets we see various common patterns. Firstly, it is not surprising that DFM achieves the highest results in terms of AUROC and Precision since it is a semi-supervised method trained only in the "good" class and is therefore better at recognizing it. On the two Recall metrics however we see that the LTN outperforms DFM in almost all cases with the exception of open-set recall for leather and grid. In most datasets it also achieves a higher F1-score which more closely approximates a global measure for the overall problem, balancing performance in the

"GOOD" and "DEFECT" classes while being less affected by class imbalances.

Overall, our experimental results indicate that Logic Tensor Networks (LTNs) and Deep Feature Modeling (DFM) consistently outperform other methods across multiple metrics. LTNs excel in both open-set and closed-set recall due to their ability to incorporate symbolic rules into the learning process, providing a structured framework that enhances the model's ability to generalize to novel defects. This advantage is significant in manufacturing environments where defects are rare and diverse, making traditional methods less reliable. The symbolic reasoning in LTNs allows the model to handle ambiguous data more effectively by leveraging domain-specific knowledge encoded in logical rules. In contrast, DFM performs exceptionally well in terms of AUROC and precision since it learns very well what a "good" product should look like, enabling the model to better understand and classify normal versus defective samples. However, DFM's treatment of defects in an agnostic way, not based on concrete training samples, often leads to under-performance in the detection of closed-set defects compared to other methods that include closed-set defects in their training set. Regarding the MLP, it expectedly performs quite well on recognizing the classes it has been trained on, but its performance significantly deteriorates in the unseen open-set classes. Methods like One-Class SVM (OCSVM) and Isolation Forest (IF) showed limitations mainly due to their high false positive rates when faced with complex and highly similar visual class data as in the presented manufacturing setting. OCSVM defines a boundary around known classes [237], which fails to adapt to the often small variability in defect appearance in the high dimensional visual feature space of this particular problem, leading to a high rate of false positives. Similarly, IF's non-parametric nature [19] makes it effective at identifying anomalies but results in many benign variations being misclassified as defects due to its lack of contextual understanding.

The improved and more balanced open and closed-set recall scores of our LTN-based approach are a result of LTN's capability, through the infusion of symbolic rules, to combine the unsupervised classifier's ability to detect out-of-distribution inputs (high R-open) and the problem-specific training of the base statistical classifier (high R-closed). It is important to note that LTNs allow the symbolic rules to influence the model continuously during training and thus have a larger effect on its inference behaviour. This capability makes the LTN approach ideal for a data-scarce scenario where challenges (see also Section 4.1) such as lowly-populated or completely novel defect classes are mitigated via the symbolic part of the AI, while existing classes with enough data but perhaps higher similarity to the good class are better recognized by the neural part.

4.5 Summary

In general we can conclude that the use of Neurosymbolic AI through LTNs can have a significant benefit on the base classifier's recall both in the open and closed cases, leading to fewer defects making it to market. It also maintains competitive scores in the recognition of good images meaning human operators will not be over-burdened by examining lots of false positives. In comparison to semi-supervised methods which are more commonly used in this setting, it maintains comparable overall recognition perfor-

mance, but what we consider most important, is that it consistently holds higher recall scores and is therefore more trustworthy for a real-life system. Of course Neurosymbolic AI is a very young field and there are still many areas to be researched. As the most important next steps to enhance this work we consider the experimentation with different symbolic rules derived from other unsupervised and semi-supervised methods or expressed through image-processing function predicates. Another parameter to vary is the arrangement and way of expressing the symbolic constraints which could possibly lead to different outcomes in the Neurosymbolic learning process. Finally, reasoning-for-learning Neurosymbolic methods that enforce strict constraints such as the Semantic Probabilistic Layer (SPL) can also be considered.

Still this work represents a promising and practical solution that can be readily applied to real-life settings. Additionally, Neurosymbolic AI can be adapted for various other applications beyond defect detection in manufacturing. In healthcare, for instance, these methods could enhance diagnostic systems by combining medical images with patient data for more accurate disease identification. Similarly, in finance, they could improve fraud detection by interpreting complex transaction patterns in conjunction with natural language data. These advancements could significantly impact multiple industries, setting a new standard for AI applications.

Chapter 5

Conclusions

The present thesis has attempted to tackle some core issues in adopting and deploying AI applications in real-life manufacturing production lines. First of all, the introduced novel method for image-level oversampling aimed at alleviating data imbalance in visual quality inspection systems was a significant step in efficiently generating synthetic data by focusing on images that are near the classification boundary. The high-fidelity synthetic images produced by this method have demonstrated promising performance in identifying defects, particularly in datasets where defects vary in perceptibility. Moreover, this was achieved at a significantly lower computational and runtime cost compared to other state-of-the-art methods.

As a next step, another novel data-augmentation method was developed, aimed at making defect classifiers more robust against previously unseen defects. Applied in a real manufacturing setting, the method outperformed existing approaches, thanks to the high fidelity and variability of synthetic images generated using StyleGAN. A key feature is the filtering of novelty data generated through latent space traversals, where a voting scheme identifies highly confusing images to be labeled as "unknown." Despite its effectiveness, the method faces challenges such as relatively high data volume requirements, lengthy training times and redundant data generation, pointing to future research opportunities in developing more efficient GAN architectures and better data extraction techniques.

For these reasons, and most importantly to address smaller datasets, the incorporation of Neurosymbolic AI into our defect detection framework has shown significant benefits when transferring novel defect detection to smaller datasets such as MVTEC-AD. This has been showcased through the improvement of recall rates while competitive recognition performance was maintained. The potential applications of Neurosymbolic AI, especially in outlier and novel input detection in small datasets extend beyond manufacturing, with promising opportunities in healthcare, finance, and other industries where accurate and reliable AI-driven decision-making is essential.

The proposed methods for handling novel defects at operational runtime are also highly relevant in the context of Industry 4.0 and the emerging Industry 5.0 paradigms, where the integration of advanced AI technologies into manufacturing processes is critical for achieving higher levels of automation, precision, and customization. Industry 4.0 emphasizes the use of interconnected systems and smart technologies to create more efficient and flexible production environments. Our work contributes to this by enhancing

the robustness of defect detection systems, which are essential for maintaining high standards of quality in automated manufacturing. By improving the reliability and accuracy of these systems, our method supports the goals of Industry 4.0, enabling manufacturers to detect and address defects more effectively, thus reducing waste and improving overall product quality.

As we move towards Industry 5.0, where the focus shifts to human-centric manufacturing and the collaboration between humans and intelligent machines, the need for more advanced and adaptable AI systems becomes even more critical. Our approaches to image-level oversampling and novel defect detection align with the principles of Industry 5.0 by enabling machines to better assist human operators in identifying and addressing defects. The use of high-fidelity synthetic images helps ensure that defect detection systems are not only accurate but also consistent, reducing the likelihood of false negatives that would lead to unidentified defects while keeping a low level of false positives that could burden human operators.

Our research also highlights the potential for future advancements in AI, particularly in the areas of instance-based or few-shot image generation, which could further enhance the fidelity and variability of synthetic images. These advancements would allow for the generation of more diverse and realistic images from a small set of low-confidence samples, improving the robustness of defect detection systems. Additionally, exploring new methods for fusing original and synthetic images could lead to more seamless integration and better overall performance.

In conclusion, our work presents practical and innovative solutions for enhancing visual quality inspection in manufacturing, with broader implications for various industries as they adopt the principles of Industry 4.0 and Industry 5.0. The advancements made in this research contribute to the ongoing development of more intelligent, adaptable, and human-centric AI systems that are poised to play a key role in the future of manufacturing and beyond.

Bibliography

[1] Julian Müller και others. *Enabling Technologies for Industry 5.0: Results of a workshop with Europe's technology leaders*. *Directorate-General for Research and Innovation*, 2020.

[2] Joze M. Rozanec και others. *Human-centric artificial intelligence architecture for industry 5.0 applications*. *International Journal of Production Research*, 61(20):6847–6872, 2023.

[3] Spyros Theodoropoulos, Patrik Zajec, Joze M. Rozanec, Dimosthenis Kyriazis και Panayiotis Tsanakas. *On-the-fly image-level oversampling for imbalanced datasets of manufacturing defects*. *Machine Learning*, 2024.

[4] Spyros Theodoropoulos, Dimitrios Dardanis, Georgios Makridis, Patrik Zajec, Joze M. Rozanec, Dimosthenis Kyriazis και Panayiotis Tsanakas. *Enhancing robustness to novel visual defects through StyleGAN latent space navigation: a manufacturing use case*. *Journal of Intelligent Manufacturing*, 2024.

[5] Spyros Theodoropoulos, Georgios Makridis, Dimosthenis Kyriazis και Panayiotis Tsanakas. *Robust Novel Defect Detection with Neurosymbolic AI*. *Advances in Production Management Systems. Production Management Systems for Volatile, Uncertain, Complex, and Ambiguous Environments*Matthias Thürer, Ralph Riedel, Gregorvon Cieminski και David Romero, επιμελητές, σελίδες 381–396, Cham, 2024. Springer Nature Switzerland.

[6] Damien Dablain, Bartosz Krawczyk και Nitesh V. Chawla. *DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data*. *IEEE Transactions on Neural Networks and Learning Systems*, σελίδες 1–15, 2022.

[7] Atsuhiro Noguchi και T. Harada. *Image Generation From Small Datasets via Batch Statistics Adaptation*. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, σελίδες 2750–2758, 2019.

[8] David Crandall Satoshi Tsutsui, Yanwei Fu. *Meta-Reinforced Synthetic Data for One-Shot Fine-Grained Visual Recognition*. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[9] Gamaleldin Fathy Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan και Samy Bengio. *Large Margin Deep Networks for Classification*. 2018.

[10] Dan Hendrycks και Thomas Dietterich. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. *Proceedings of the International Conference on Learning Representations*, 2019.

[11] Paul Bergmann, Michael Fauser, David Sattlegger και Carsten Steger. *MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection*. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 9584–9592, 2019.

[12] Y. Shen και B. Zhou. *Closed-Form Factorization of Latent Semantics in GANs*. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1532–1540, Los Alamitos, CA, USA, 2021. IEEE Computer Society.

[13] Andre Araujo, Wade Davenport Norris και Jack Sim. *Computing Receptive Fields of Convolutional Neural Networks*. *Distill*, 2019.

[14] Wenguan Wang, Yi Yang και Fei Wu. *Towards Data-and Knowledge-Driven Artificial Intelligence: A Survey on Neuro-Symbolic Computing*. 2022.

[15] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum και Samuel J. Gershman. *Building Machines That Learn and Think Like People*. *Behavioral and Brain Sciences*, 40, 2017.

[16] Georgios Makridis, Spyros Theodoropoulos, Dimitrios Dardanis, Ioannis Makridis, Maria Margarita Separdani, Georgios Fatouros, Dimosthenis Kyriazis και Panagiotis Koulouris. *XAI enhancing cyber defence against adversarial attacks in industrial applications*. *2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS)*, τόμος Five, σελίδες 1–8, 2022.

[17] Spyros Theodoropoulos, Patrik Zajec, Joze M. Rozanec, Dimitrios Dardanis, Georgios Makridis, Dimosthenis Kyriazis και Panayiotis Tsanakas. *Identifying Novel Defects during AI-driven Visual Quality Inspection*. *IFAC-PapersOnLine*, 56(2):3738–3743, 2023. 22nd IFAC World Congress.

[18] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini και Michael Spranger. *Logic Tensor Networks*. *Artificial Intelligence*, 303:103649, 2022.

[19] Yousra Chabchoub, Maurras Ulbricht Togbe, Aliou Boly και Raja Chiky. *An In-Depth Study and Improvement of Isolation Forest*. *IEEE Access*, 10:10219–10237, 2022.

[20] Yaqing Wang, Quanming Yao, James T Kwok και Lionel M Ni. *Generalizing from a few examples: A survey on few-shot learning*. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[21] Haroon Sheikh, Corien Prins και Erik Schrijvers. *Artificial Intelligence: Definition and Background*, σελίδες 15–41. Springer International Publishing, Cham, 2023.

[22] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks*. *Communications of the ACM*, 60(6):84–90, 2017.

[23] Tom B Brown. *Language models are few-shot learners*. *arXiv preprint ArXiv:2005.14165*, 2020.

[24] OpenAI Josh Achiamet al. *GPT-4 Technical Report*. 2023.

[25] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley D. Edwards, Nicolas Manfred Otto Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar και Nandode Freitas. *A Generalist Agent*. *ArXiv*, abs/2205.06175, 2022.

[26] Joze M. Rozanec, Luka Bizjak, Elena Trajkova, Patrik Zajec, Jelle Keizer, Blaz Fortuna και Dunja Mladenic. *Active learning and novel model calibration measurements for automated visual inspection in manufacturing*. *Journal of Intelligent Manufacturing*, 2023.

[27] Chu Chi Kuo, Joseph Z. Shyu και Kun Ding. *Industrial revitalization via industry 4.0 β€“ A comparative policy analysis among China, Germany and the USA*. *Global Transitions*, 1:3–14, 2019.

[28] Zoe Alexander, Duen Horng Chau και Christopher Saldaña. *An Interrogative Survey of Explainable AI in Manufacturing*. *IEEE Transactions on Industrial Informatics*, 20(5):7069–7081, 2024.

[29] Fazel Ansari, Selim Erol και Wilfried Sihn. *Rethinking human-machine learning in industry 4.0: how does the paradigm shift treat the role of human learning?* *Procedia manufacturing*, 23:117–122, 2018.

[30] Heiner Ludwig, Thorsten Schmidt και Mathias KΓ'Ohn. *Voice user interfaces in manufacturing logistics: a literature review*. *International Journal of Speech Technology*, 26(3):627–639, 2023.

[31] David Romero και Johan Stahre. *Towards The Resilient Operator 5.0: The Future of Work in Smart Resilient Manufacturing Systems*. *Procedia CIRP*, 104:1089–1094, 2021. 54th CIRP CMS 2021 - Towards Digitalized Manufacturing 4.0.

[32] Yuqian Lu, Hao Zheng, Saahil Chand, Wanqing Xia, Zengkun Liu, Xun Xu, Lihui Wang, Zhaojun Qin και Jinsong Bao. *Outlook on human-centric manufacturing towards Industry 5.0*. *Journal of Manufacturing Systems*, 62:612–627, 2022.

[33] Shirine El Zaatari, Mohamed Marei, Weidong Li και Zahid Usman. *Cobot programming for collaborative industrial tasks: An overview*. *Robotics and Autonomous Systems*, 116:162–180, 2019.

[34] Antonio Giallanza, Giada La Scalia, Rosa Micale και Concetta Manuela La Fata. *Occupational health and safety issues in human-robot collaboration: State of the art and open challenges*. *Safety Science*, 169:106313, 2024.

[35] Gökan May, Marco Taisch, Andrea Bettoni, Omid Maghazei, Annarita Matarazzo και Bojan Stahl. *A new human-centric factory model*. *Procedia CIRP*, 26:103–108, 2015.

[36] Fei Tao, Qinglin Qi, Ang Liu και Andrew Kusiak. *Data-driven smart manufacturing*. *Journal of Manufacturing Systems*, 48:157–169, 2018.

[37] George Chryssolouris, Dimitris Mavrikios, Nikolaos Papakostas, Dimitris Mourtzis, George Michalos και Konstantinos Georgoulias. *Digital manufacturing: History, perspectives, and outlook*. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 223, 2009.

[38] *How Big Data Transforms Manufacturing Industry*. *International Journal of Strategic Engineering*, 2019.

[39] A. Gandomi και M. Haider. *Beyond the hype: Big Data concepts, methods, and analytics*. *International Journal of Information Management*, 35(2):137–144, 2015.

[40] A. Siddiqa, I. A. T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband και A. Gani. *A survey of big data management: taxonomy and state-of-the-art*. *Journal of Network and Computer Applications*, 71:151–166, 2016.

[41] M. Speringer και J. Schnelzer. *Differentiation of Industry 4.0 Models. The 4th Industrial Revolution from different Regional Perspectives in the Global North and Global South*. *Innovations for Development: Towards Sustainable, Inclusive, and Peaceful Societies*, Vienna, 2019. Research Academy of the United Nations (RAUN).

[42] Industrial Internet Consortium. *Architecture Alignment and Interoperability*, 2017. [White paper].

[43] Big Data Value Association. *Big Data Challenges in Smart Manufacturing*, 2018. [White paper].

[44] Big Data Value Association. *Big Data Challenges in Smart Manufacturing Industry*, 2020. [White paper].

[45] Lucy Ellen Lwakatare, Aiswarya Raj, Ivica Crnkovic, Jan Bosch και Helena HolmstrΓ¶m Olsson. *Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions*. *Information and Software Technology*, 127:106368, 2020.

[46] Massimo Bertolini, Davide Mezzogori, Mattia Neroni και Francesco Zammori. *Machine Learning for industrial applications: A comprehensive literature review*. *Expert Systems with Applications*, 175:114820, 2021.

[47] Yaoyao Fiona Zhao, Jiarui Xie και Lijun Sun. *On the data quality and imbalance in machine learning-based design and manufacturingβ€”A systematic review*. *Engineering*, 2024.

[48] Christopher Frederickson, Michael Moore, Glenn Dawson και Robi Polikar. *Attack Strength vs. Detectability Dilemma in Adversarial Machine Learning*. *2018 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8, 2018.

[49] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba και Pieter Abbeel. *Domain randomization for transferring deep neural networks from simulation to the real world*. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, σελίδες 23–30, 2017.

[50] Nativi S και De Nigris S. *AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework*. (KJ-NA-30772-EN-N (online)), 2021.

[51] Giusella Finocchiaro. *The regulation of artificial intelligence*. *AI & SOCIETY*, 39(4):1961–1968, 2024.

[52] JoЁYe M. RoЁYanec, Patrik Zajec, Spyros Theodoropoulos, Erik Koehorst, BlaЁY Fortuna και Dunja MladeniΔ‡. *Synthetic Data Augmentation Using GAN For Improved Automated Visual Inspection*. *IFAC-PapersOnLine*, 56(2):11094–11099, 2023. 22nd IFAC World Congress.

[53] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan και Joseph Walsh. *Deep learning vs. traditional computer vision*. *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, σελίδες 128–144. Springer, 2020.

[54] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song και Shu Tao Xia. *Iterative learning with open-set noisy labels*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 8688–8696, 2018.

[55] H F Ng. *Automatic thresholding for defect detection*. *Pattern Recognition Letters*, 27(14):1644–1649, 2004.

[56] C X Jian, J Gao και Y H Ao. *Automatic surface defect detection for mobile phone screen glass based on machine vision*. *Applied Soft Computing*, 52:348–358, 2017.

[57] X Yang, D Qi και X Li. *Multi-scale Edge Detection of Wood Defect Images Based on the Dyadic Wavelet Transform*. *International Conference on Machine Vision and Human-Machine Interface*, σελίδες 120–123. IEEE, 2010.

[58] X. Dong, C.J. Taylor και T.F. Cootes. *Small Defect Detection Using Convolutional Neural Network Features and Random Forests*. *Computer Vision - ECCV 2018 Workshops*Laura Leal-Taixé και Stefan Roth, επιμελητές, τόμος 11132 στο *Lecture Notes in Computer Science*, σελίδες 476–490. Springer, Cham, 2019.

[59] J. Matas, O. Chum, M. Urban και T. Pajdla. *Robust wide baseline stereo from maximally stable extremal regions*. *Proceedings of the British Machine Vision Conference*, σελίδες 384–393, 2002.

[60] Zhenshen Qu, Jianxiong Shen, Ruikun Li, Junyu Liu και Qiuyu Guan. *PartsNet: A Unified Deep Network for Automotive Engine Precision Parts Defect Detection*. *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence (CSAI '18)*, σελίδες 594–599, New York, NY, USA, 2018. Association for Computing Machinery.

[61] Javier Villalba-Diez, Daniel Schmidt, Roman Gevers, Joaquín Ordieres-Meré, Martin Buchwitz και Wanja Wellbrock. *Deep learning for industrial computer vision quality control in the printing industry 4.0*. *Sensors*, 19(18):3987, 2019.

[62] Wang Liqun, Wu Jiansheng και Wu Dingjin. *Research on vehicle parts defect detection based on deep learning*. *Journal of Physics: Conference Series*, τόμος 1437, σελίδα 012004. IOP Publishing, 2020.

[63] Li Yi, Guodong Li και Ming Jiang. *An End-to-End Steel Strip Surface Defects Recognition System Based on Convolutional Neural Networks*. *Steel Research International*, 88(3):1600068, 2017.

[64] S. Faghih-Roohi, S. Hajizadeh, A. Nunez, R. Babuska και B. De Schutter. *Deep Convolutional Neural Networks for Detection of Rail Surface Defects*. *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 2584–2589, Vancouver, Canada, 2016. IEEE.

[65] Carlos Beltrán-González, Matteo Bustreo και Alessio Del Bue. *External and internal quality inspection of aerospace components*. *2020 IEEE 7th International Workshop on Metrology for AeroSpace (MetroAeroSpace)*, σελίδες 351–355. IEEE, 2020.

[66] Mohammad Shahin, F Frank Chen, Ali Hosseinzadeh, Hamed Bouzary και Awni Shahin. *Waste reduction via image classification algorithms: beyond the human eye with an AI-based vision*. *International Journal of Production Research*, σελίδες 1–19, 2023.

[67] Alex Krizhevsky, Ilya Sutskever και Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks*. *Advances in neural information processing systems*, 25, 2012.

[68] Jonathan Long, Evan Shelhamer και Trevor Darrell. *Fully convolutional networks for semantic segmentation*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 3431–3440, 2015.

[69] Tobias Glasmachers. *Limits of end-to-end learning*. *Asian Conference on Machine Learning*, σελίδες 17–32. PMLR, 2017.

[70] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko και Trevor Darrell. *Deep Domain Confusion: Maximizing for Domain Invariance*. *arXiv preprint arXiv:1412.3474*, 2014.

[71] Jun Yan Zhu, Taesung Park, Phillip Isola και Alexei A. Efros. *Unpaired Image-to-Image Translation using Cycle Consistent Adversarial Networks*. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[72] Konstantinos Bousmalis και others. *Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping*. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, σελίδες 4243–4250. IEEE, 2018.

[73] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba και Pieter Abbeel. *Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World*. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[74] Burr Settles. *Active learning literature survey*. 2009.

[75] Samuel Budd, Esther C. Robinson και Bernhard Kainz. *A survey on active learning and human-in-the-loop deep learning for medical image analysis*. *Medical Image Analysis*, 71:102062, 2021.

[76] Weili Dai, Abdul Mujeeb, Marius Erdt και Alexei Sourin. *Towards automatic optical inspection of soldering defects*. *2018 International Conference on Cyberworlds (CW)*, σελίδες 375–382. IEEE, 2018.

[77] Keesvan Garderen. *Active Learning for Overlay Prediction in Semi-conductor Manufacturing*, 2018. PhD thesis.

[78] Archit Parnami και Minwoo Lee. *Learning from few examples: A summary of approaches to few-shot learning*. *arXiv preprint arXiv:2203.04291*, 2022.

[79] Jake Snell, Kevin Swersky και Richard S. Zemel. *Prototypical Networks for Few-shot Learning*. *ArXiv*, abs/1703.05175, 2017.

[80] Chelsea Finn, Pieter Abbeel και Sergey Levine. *Model-agnostic meta-learning for fast adaptation of deep networks*. *International conference on machine learning*, σελίδες 1126–1135. PMLR, 2017.

[81] Alex Nichol και John Schulman. *Reptile: a Scalable Metalearning Algorithm*. *arXiv: Learning*, 2018.

[82] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan και Tao Mei. *Memory matching networks for one-shot image recognition*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 4080–4088, 2018.

[83] Tsendsuren Munkhdalai και Hong Yu. *Meta networks*. *International conference on machine learning*, σελίδες 2554–2563. PMLR, 2017.

[84] Qianwen Lv και Yonghong Song. *Few-shot learning combine attention mechanism-based defect detection in bar surface*. *ISIJ International*, 59(6):1089–1097, 2019.

[85] Shen Zhang, Fei Ye, Bingnan Wang και Thomas G Habetler. *Few-shot bearing anomaly detection via model-agnostic meta-learning*. *2020 23rd International Conference on Electrical Machines and Systems (ICEMS)*, σελίδες 1341–1346. IEEE, 2020.

[86] Ke Wu, Jie Tan, Jingwei Li και Chengbao Liu. *Few-shot learning approach for 3D defect detection in lithium battery*. *Journal of Physics: Conference Series*, τόμος 1884, σελίδα 012024. IOP Publishing, 2021.

[87] Zhu Zhan, Jinfeng Zhou και Bugao Xu. *Fabric defect classification using prototypical network of few-shot learning algorithm*. *Computers in Industry*, 138:103628, 2022.

[88] Jiancheng Xu και Jialei Ma. *Auto Parts Defect Detection Based on Few-shot Learning*. *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, σελίδες 943–946. IEEE, 2022.

[89] Hironori Takimoto, Junya Seki, Sulfayanti F. Situju και Akihiro Kanagawa. *Anomaly detection using siamese network with attention mechanism for few-shot learning*. *Applied Artificial Intelligence*, 36(1):2094885, 2022.

[90] Paula Branco, Luís Torgo και Rita P Ribeiro. *A survey of predictive modeling on imbalanced domains*. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[91] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall και W. Philip Kegelmeyer. *SMOTE: Synthetic Minority over-Sampling Technique*. *J. Artif. Int. Res.*, 16(1):321–357, 2002.

[92] Haibo He, Yang Bai, Edwardo A. Garcia και Shutao Li. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, σελίδες 1322–1328, 2008.

[93] Hui Han, Wenyuan Wang και Binghuan Mao. *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. *ICIC*, 2005.

[94] Vignesh Sampath, Iñaki Maurtua, Juan José Aguilar Martín και Aitor Gutierrez. *A survey on generative adversarial networks for imbalance problems in computer vision tasks*. *Journal of Big Data*, 8(1):27, 2021.

[95] Yaxing Wang, Chenshen Wu, Luis Herranz, Joostvan de Weijer, Abel Gonzalez-Garcia και B. Raducanu. *Transferring GANs: generating images from limited data*. *ECCV*, 2018.

[96] Bingchen Liu, Yizhe Zhu, Kunpeng Song και Ahmed Elgammal. *Towards faster and stabilized gan training for high-fidelity few-shot image synthesis*. *International Conference on Learning Representations*, 2020.

[97] JoEYe M. RoEYanec, Patrik Zajec, Spyros Theodoropoulos, Erik Koehorst, BlaEY Fortunat και Dunja MladeniΔ‡. *Robust Anomaly Map Assisted Multiple Defect Detection with Supervised Classification Techniques*. *IFAC-PapersOnLine*, 56(2):7846–7851, 2023. 22nd IFAC World Congress.

[98] D Gunning. *Explainable artificial intelligence (xai) darpa-baa-16-53*. *Defense Advanced Research Projects Agency*, 2016.

[99] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alejandro Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins και others. *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. *Information Fusion*, 58:82–115, 2020.

[100] Marko Robnik-Šikonja και Igor Kononenko. *Explaining classifications for individual instances*. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600, 2008.

[101] Ruth C. Fong και Andrea Vedaldi. *Interpretable explanations of black boxes by meaningful perturbation*. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, σελίδες 3429–3437. IEEE, 2017.

[102] Marco Tulio Ribeiro, Sameer Singh και Carlos Guestrin. *Anchors: High-precision model-agnostic explanations*. *Proceedings of the AAAI Conference on Artificial Intelligence*, τόμος 32, 2018.

[103] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt και Been Kim. *Sanity checks for saliency maps*. *Advances in Neural Information Processing Systems (NeurIPS)*, τόμος 31, 2018.

[104] Anshul Shrikumar, Peyton Greenside και Anshul Kundaje. *Learning important features through propagating activation differences*. *Proceedings of the International Conference on Machine Learning (ICML)*, σελίδες 3145–3153. PMLR, 2017.

[105] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva και Antonio Torralba. *Learning deep features for discriminative localization*. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 2921–2929. IEEE, 2016.

[106] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh και Dhruv Batra. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. *Proceedings of the IEEE international conference on computer vision*, σελίδες 618–626, 2017.

[107] Han Xu, Yao Ma, Hao Chen Liu, Debayan Deb, Hui Liu, Ji Liang Tang και Anil K Jain. *Adversarial attacks and defenses in images, graphs and text: A review*. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[108] Anish Athalye, Nicholas Carlini και David Wagner. *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples*. *International conference on machine learning*, σελίδες 274–283. PMLR, 2018.

[109] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras και Adrian Vladu. *Towards deep learning models resistant to adversarial attacks*. *arXiv preprint arXiv:1706.06083*, 2017.

[110] Nicholas Carlini και David Wagner. *Adversarial examples are not easily detected: Bypassing ten detection methods*. *Proceedings of the 10th ACM workshop on artificial intelligence and security*, σελίδες 3–14, 2017.

[111] Satya M Muddamsetty, NS Jahromi Mohammad και Thomas B Moeslund. *SIDU: similarity difference and uniqueness method for explainable AI*. *2020 IEEE International Conference on Image Processing (ICIP)*, σελίδες 3269–3273. IEEE, 2020.

[112] LM Fenoy και A Ciontos. *Performance evaluation of Explainable AI methods against adversarial noise*.

[113] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King Wai Low, Shu Fang Newman, Jerry Kim και others. *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery*. *Nature Biomedical Engineering*, 2018.

[114] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran και Aleksander Madry. *Adversarial examples are not bugs, they are features*. *Advances in neural information processing systems*, 32, 2019.

[115] Yasmin Fathy, Mona Jaber και Alexandra Brintrup. *Learning With Imbalanced Data in Smart Manufacturing: A Comparative Analysis*. *IEEE Access*, 9:2734–2757, 2021.

[116] Judi E. See. *Visual inspection : a review of the literature*. *Sandia Report SAND2012-8590, Sandia National Laboratories, Albuquerque, New Mexico*, 2012.

[117] Andrei Alexandru Tulbure, Adrian Alexandru Tulbure και Eva Henrietta Dulf. *A review on modern defect detection models using DCNNs - Deep convolutional neural networks*. *Journal of Advanced Research*, 35:33–48, 2022.

[118] Navneet Dalal και Bill Triggs. *Histograms of oriented gradients for human detection*. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, τόμος 1, σελίδες 886–893. Ieee, 2005.

[119] Paul Viola και Michael Jones. *Rapid object detection using a boosted cascade of simple features*. *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, τόμος 1, σελίδες I–I. Ieee, 2001.

[120] Manuella Kadar και Daniela Onita. *A deep CNN for Image Analytics in Automated Manufacturing Process Control*. *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, σελίδες 1–5, 2019.

[121] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[122] Gongjie Zhang, Kaiwen Cui, Tzu Yi Hung και Shijian Lu. *Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection*. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, σελίδες 2523–2533, 2021.

[123] Fatima A. Saiz, Garazi Alfaro, Inigo Barandiaran και Manuel Grana. *Generative Adversarial Networks to Improve the Robustness of Visual Defect Segmentation by Semantic Networks in Manufacturing Components*. *Applied Sciences*, 11(14), 2021.

[124] Pornntiwa Pawara, Emmanuel Okafor, Lambert Schomaker και Marco Wiering. *Data Augmentation for Plant Classification*. *Advanced Concepts for Intelligent Vision Systems*Jacques Blanc-Talon, Rudi Penne, Wilfried Philips, Dan Popescu και Paul Scheunders, επιμελητές, σελίδες 615–626, Cham, 2017. Springer International Publishing.

[125] Kihyuk Sohn, Honglak Lee και Xinchen Yan. *Learning Structured Output Representation using Deep Conditional Generative Models*. *Advances in Neural Information Processing Systems*C. Cortes, N. Lawrence, D. Lee, M. Sugiyama και R. Garnett, επιμελητές, τόμος 28. Curran Associates, Inc., 2015.

[126] Jong Pil Yun, Woosang Crino Shin, Gyogwon Koo, Min Su Kim, Chungki Lee και Sang Jun Lee. *Automated defect inspection system for metal surfaces based on deep learning and data augmentation*. *Journal of Manufacturing Systems*, 55:317–324, 2020.

[127] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville και Yoshua Bengio. *Generative Adversarial Nets*. *Advances in Neural Information Processing Systems*Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence και K.Q. Weinberger, επιμελητές, τόμος 27. Curran Associates, Inc., 2014.

[128] Vignesh Sampath, Iñaki Maurtua, Juan José Aguilar Martín και Aitor Gutierrez. *A survey on generative adversarial networks for imbalance problems in computer vision tasks*. *Journal of Big Data*, 8:1–2, 2021.

[129] Lizhe Liu, Danhua Cao, Yubin Wu και Taoran Wei. *Defective Samples Simulation through Adversarial Training for Automatic Surface Inspection*. *Neurocomput.*, 360(C):230–245, 2019.

[130] Fujun Luan, Sylvain Paris, Eli Shechtman και Kavita Bala. *Deep Painterly Harmonization*. *Computer Graphics Forum*, 37, 2018.

[131] Yifan Jiang, Shiyu Chang και Zhangyang Wang. *Transgan: Two transformers can make one strong gan*. *arXiv preprint arXiv:2102.07074*, 1(3), 2021.

[132] Chenglong Wang και Zhifeng Xiao. *Lychee Surface Defect Detection Based on Deep Convolutional Neural Networks with GAN-Based Data Augmentation*. *Agronomy*, 11(8), 2021.

[133] Jordan J. Bird, Chloe M. Barnes, Luis J. Manso, Aniko Ekart και Diego R. Faria. *Fruit quality and defect image classification with conditional GAN data augmentation*. *Scientia Horticulturae*, 293:110684, 2022.

[134] Saksham Jain, Gautam Seth, Arpit Paruthi, Umang Soni και Girish Kumar. *Synthetic data augmentation for surface defect detection and classification using deep learning*. *Journal of Intelligent Manufacturing*, σελίδες 1–14, 2020.

[135] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever και P. Abbeel. *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*. *NIPS*, 2016.

[136] Sebastian Meister, Nantwin Mueller, Jan Stoeve και Roger Groves. *Synthetic image data augmentation for fibre layup inspection processes: Techniques to enhance the data set*. *Journal of Intelligent Manufacturing*, 32, 2021.

[137] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger και Hayit Greenspan. *Synthetic data augmentation using GAN for improved liver lesion classification*. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, σελίδες 289–293, 2018.

[138] Yunyan Wang, Shuai Luo και Huaxuan Wu. *Defect detection of solar cell based on data augmentation*. *Journal of Physics: Conference Series*, 1952(2):022010, 2021.

[139] Haodong Zhang, Zuzhi Chen, Chaoqun Zhang, Juntong Xi και Xinyi Le. *Weld Defect Detection Based on Deep Learning Method*. *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, σελίδες 1574–1579, 2019.

[140] Xinyi Le, Junhui Mei, Haodong Zhang, Boyu Zhou και Juntong Xi. *A learning-based approach for surface defect detection using small image datasets*. *Neurocomputing*, 408:112–120, 2020.

[141] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen και Timo Aila. *Training Generative Adversarial Networks with Limited Data*. *Proc. NeurIPS*, 2020.

[142] Ricardo Silva Peres, Miguel Azevedo, Sara Oleiro Araujo, Magno Guedes, Fabio Miranda και Jose Barata. *Generative Adversarial Networks for Data Augmentation in Structural Adhesive Inspection*. *Applied Sciences*, 11(7), 2021.

[143] Tero Karras, Timo Aila, Samuli Laine και Jaakko Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. *ArXiv*, abs/1710.10196, 2018.

[144] Changhee Han, Kohei Murao, Tomoyuki Noguchi, Yusuke Kawata, Fumiya Uchiyama, Leonardo Rundo, Hideki Nakayama και Shin'ichi Satoh. *Learning More with Less: Conditional PGGAN-Based Data Augmentation for Brain Metastases Detection Using Highly-Rough Annotation on MR Images*. CIKM '19, σελίδα 119–127, New York, NY, USA, 2019. Association for Computing Machinery.

[145] Z Luo, S Y Cheng και Q Y Zheng. *GAN-Based Augmentation for Improving CNN Performance of Classification of Defective Photovoltaic Module Cells in Electroluminescence Images*. *IOP Conference Series: Earth and Environmental Science*, 354(1):012106, 2019.

[146] Augustus Odena, Christopher Olah και Jonathon Shlens. *Conditional Image Synthesis with Auxiliary Classifier GANs*. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, σελίδα 2642–2651. JMLR.org, 2017.

[147] Wei Xiong, Janghwan Lee, Shuhui Qu και Wonhyouk Jang. *Data Augmentation for Applying Deep Learning to Display Manufacturing Defect Detection*. *SID Symposium Digest of Technical Papers*, 51:1210–1213, 2020.

[148] Lijyun Huang, Kate Ching Ju Lin και Yu Chee Tseng. *Resolving Intra-Class Imbalance for GAN-Based Image Augmentation*. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, σελίδες 970–975, 2019.

[149] Andrew Brock, Jeff Donahue και Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. *ArXiv*, abs/1809.11096, 2019.

[150] Ta Wei Tang, Wei Han Kuo, Jauh Hsiang Lan, Chien Fang Ding, Hakiem Hsu και Hong Tsu Young. *Anomaly Detection Neural Network with Dual Auto-Encoders GAN and Its Industrial Inspection Applications*. *Sensors*, 20(12), 2020.

[151] David Berthelot, Tom Schumm και Luke Metz. *BEGAN: Boundary Equilibrium Generative Adversarial Networks*. *ArXiv*, abs/1703.10717, 2017.

[152] Samet Akçay, Amir Atapour-Abarghouei και T. Breckon. *Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection*. *2019 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8, 2019.

[153] Gongjie Zhang, Kaiwen Cui, Tzu Yi Hung και Shijian Lu. *Defect-GAN: High-Fidelity Defect Synthesis for Automated Defect Inspection*. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, σελίδες 2523–2533, 2021.

[154] Shuanlong Niu, Bin Li, Xinggang Wang και Hui Lin. *Defect Image Sample Generation With GAN for Improving Defect Recognition*. *IEEE Transactions on Automation Science and Engineering*, 17(3):1611–1622, 2020.

[155] Dongjie Li, Wenbo Xie, Baogang Wang, Weifeng Zhong και Hongmin Wang. *Data Augmentation and Layered Deformable Mask R-CNN-Based Detection of Wood Defects*. *IEEE Access*, 9:108162–108174, 2021.

[156] Chuan Guo, Geoff Pleiss, Yu Sun και Kilian Q. Weinberger. *On Calibration of Modern Neural Networks*. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, σελίδα 1321–1330. JMLR.org, 2017.

[157] Alexandru Niculescu-Mizil και Rich Caruana. *Predicting Good Probabilities with Supervised Learning*. *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, σελίδα 625–632, New York, NY, USA, 2005. Association for Computing Machinery.

[158] Bianca Zadrozny και Charles Elkan. *Transforming Classifier Scores into Accurate Multiclass Probability Estimates*. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, σελίδα 694–699, New York, NY, USA, 2002. Association for Computing Machinery.

[159] Mahdi Pakdaman Naeini, Gregory F. Cooper και Milos Hauskrecht. *Obtaining Well Calibrated Probabilities Using Bayesian Binning*. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, σελίδα 2901–2907. AAAI Press, 2015.

[160] John Platt. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. *Adv. Large Margin Classif.*, 10, 2000.

[161] Geoffrey Hinton, Oriol Vinyals και Jeffrey Dean. *Distilling the Knowledge in a Neural Network*. *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[162] Jooyoung Moon, Jihyo Kim, Younghak Shin και Sangheum Hwang. *Confidence-Aware Learning for Deep Neural Networks*. *ICML*, 2020.

[163] R. Müller, Simon Kornblith και Geoffrey E. Hinton. *When Does Label Smoothing Help?* *NeurIPS*, 2019.

[164] S. Thulasidasan, Gopinath Chennupati, J. Bilmes, Tanmoy Bhattacharya και S. Michalak. *On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks*. *ArXiv*, abs/1905.11001, 2019.

[165] T. Karras, S. Laine και T. Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228, 2021.

[166] Harold Achicanoy, Deisy Chaves και Maria Trujillo. *StyleGANs and Transfer Learning for Generating Synthetic Images in Industrial Applications*. *Symmetry*, 13(8), 2021.

[167] Yan Zhang, Shiyun Wa, Pengshuo Sun και Yaojun Wang. *Pear Defect Detection Method Based on ResNet and DCGAN*. *Information*, 12(10), 2021.

[168] Xinglong Feng, Xianwen Gao και Ling Luo. *A ResNet50-Based Method for Classifying Surface Defects in Hot-Rolled Strip Steel*. *Mathematics*, 9(19), 2021.

[169] M. Stone. *Cross-Validatory Choice and Assessment of Statistical Predictions*. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[170] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen και Timo Aila. *Training Generative Adversarial Networks with Limited Data*. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[171] Virginia Pilloni. *How Data Will Transform Industrial Processes: Crowdsensing, Crowdsourcing and Big Data as Pillars of Industry 4.0*. *Future Internet*, 10(3), 2018.

[172] Hasan Tercan και Tobias Meisen. *Machine learning and deep learning based predictive quality in manufacturing: a systematic review*. *Journal of Intelligent Manufacturing*, 33(7):1879–1905, 2022.

[173] Milica Babic, Mojtaba A. Farahani και Thorsten Wuest. *Image Based Quality Inspection in Smart Manufacturing Systems: A Literature Review*. *Procedia CIRP*, 103:262–267, 2021. 9th CIRP Global Web Conference on Sustainable, resilient, and agile manufacturing and service operations : Lessons from COVID-19.

[174] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen και Timo Aila. *Alias-Free Generative Adversarial Networks*. *Proc. NeurIPS*, 2021.

[175] Thomas G. Dietterich. *Steps Toward Robust Artificial Intelligence*. *AI Magazine*, 38(3):3–24, 2017.

[176] Miryam Elizabeth Villa-Perez, Miguel A. Alvarez-Carmona, Octavio Loyola-Gonzalez, Miguel Angel Medina-Perez, Juan Carlos Velazco-Rossell και Kim Kwang Raymond Choo. *Semi-supervised anomaly detection algorithms: A comparative summary and future research directions*. *Knowledge-Based Systems*, 218:106878, 2021.

[177] Spyros Theodoropoulos, Patrik Zajec, Joze M. Rozanec, Dimitrios Dardanis, Georgios Makridis, Dimosthenis Kyriazis και Panayiotis Tsanakas. *Identifying Novel Defects during AI-driven Visual Quality Inspection*. *IFAC-PapersOnLine*, 56(2):3738–3743, 2023. 22nd IFAC World Congress.

[178] Chuanxing Geng, Sheng Jun Huang και Songcan Chen. *Recent Advances in Open Set Recognition: A Survey*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3614–3631, 2021.

[179] Walter J. Scheirer, Andersonde Rezende Rocha, Archana Sapkota και Terrance E. Boult. *Toward Open Set Recognition*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

[180] Enrique Castillo. *Extreme Value Theory in Engineering*. *Extreme Value Theory in Engineering*Enrique Castillo, επιμελητής, σελίδες 183–209. Academic Press, San Diego, 1988.

[181] Steve Cruz, Cora Coleman, Ethan M. Rudd και Terrance E. Boult. *Open set intrusion recognition for fine-grained attack categorization*. *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, σελίδες 1–6, 2017.

[182] Walter J. Scheirer, Lalit P. Jain και Terrance E. Boult. *Probability Models for Open Set Recognition*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014.

[183] Bernhard Scholkopf, Robert C. Williamson, Alex Smola, John Shawe-Taylor και John C. Platt. *Support Vector Method for Novelty Detection*. *Neural Information Processing Systems*, 1999.

[184] Ajita Rattani, Walter J. Scheirer και Arun Ross. *Open Set Fingerprint Spoof Detection Across Novel Fabrication Materials*. *IEEE Transactions on Information Forensics and Security*, 10(11):2447–2460, 2015.

[185] Lalit P. Jain, Walter J. Scheirer και Terrance E. Boult. *Multi-class Open Set Recognition Using Probability of Inclusion*. *Computer Vision - ECCV 2014*David Fleet, Tomas Pajdla, Bernt Schiele και Tinne Tuytelaars, επιμελητές, σελίδες 393–409, Cham, 2014. Springer International Publishing.

[186] Andras Rozsa, Manuel Günther και Terrance E. Boult. *Adversarial Robustness: Softmax versus Openmax*. *ArXiv*, abs/1708.01697, 2017.

[187] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang και Bernt Schiele. *Towards Reaching Human Performance in Pedestrian Detection*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):973–986, 2018.

[188] A. Bendale και T. E. Boult. *Towards Open Set Deep Networks*. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 1563–1572, Los Alamitos, CA, USA, 2016. IEEE Computer Society.

[189] Sergey Demyanov Zongyuan Ge και Rahil Garnavi. *Generative OpenMax for Multi-Class Open Set Classification*. *Proceedings of the British Machine Vision Conference (BMVC)*Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou και Krystian Mikolajczyk, επιμελητές, σελίδες 42.1–42.12. BMVA Press, 2017.

[190] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng Keen Wong και Fuxin Li. *Open Set Learning with Counterfactual Images*. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[191] Yang Yu, Wei Yang Qu, Nan Li και Zimin Guo. *Open-Category Classification by Adversarial Sample Generation*. IJCAI'17, σελίδα 3357–3363, 2017.

[192] Luke Ditria, Benjamin J. Meyer και Tom Drummond. *OpenGAN: Open Set Generative Adversarial Networks*. *ACCV*, 2020.

[193] Xian Tao, Xinyi Gong, Xin Zhang, Shaohua Yan και Chandranath Adak. *Deep Learning for Unsupervised Anomaly Localization in Industrial Images: A Survey*. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[194] J.K. Chow, Z. Su, J. Wu, P.S. Tan, X. Mao και Y.H. Wang. *Anomaly detection of defects on concrete structures with the convolutional autoencoder*. *Advanced Engineering Informatics*, 45:101105, 2020.

[195] Sanyapong Youkachen, Miti Ruchanurucks, Teera Phatrapomnant και Hirohiko Kaneko. *Defect Segmentation of Hot-rolled Steel Strip Surface by using Convolutional Auto-Encoder and Conventional Image processing*. *2019 10th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, σελίδες 1–5, 2019.

[196] Gaoqiang Kang, Shibin Gao, Long Yu και Dongkai Zhang. *Deep Architecture for High-Speed Railway Insulator Surface Defect Detection: Denoising Autoencoder With Multitask Learning*. *IEEE Transactions on Instrumentation and Measurement*, 68(8):2679–2690, 2019.

[197] Samet Akcay, Amir Atapour-Abarghouei και Toby P. Breckon. *GANomaly: Semi-supervised Anomaly Detection via Adversarial Training*. *Computer Vision – ACCV 2018*C. V. Jawahar, Hongdong Li, Greg Mori και Konrad Schindler, επιμελητές, σελίδες 622–637, Cham, 2019. Springer International Publishing.

[198] Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja και Utku Genc. *Anomalib: A Deep Learning Library for Anomaly Detection*, 2022.

[199] Nilesh A. Ahuja, Ibrahima Ndiour, Trushant Kalyanpur και Omesh Tickoo. *Probabilistic Modeling of Deep Features for Out-of-Distribution and Adversarial Detection*, 2019.

[200] Jaeyeon Jang, Minkyung Seo και Chang Ouk Kim. *Support Weighted Ensemble Model for Open Set Recognition of Wafer Map Defects*. *IEEE Transactions on Semiconductor Manufacturing*, 33(4):635–643, 2020.

[201] W. Xia, Y. Zhang, Y. Yang, J. Xue, B. Zhou και M. Yang. *GAN Inversion: A Survey*. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–17, 5555.

[202] Tero Karras, Timo Aila, Samuli Laine και Jaakko Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. *ArXiv*, abs/1710.10196, 2017.

[203] Andrew Brock, Jeff Donahue και Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[204] Tero Karras, Samuli Laine και Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021.

[205] Rameen Abdal, Yipeng Qin και Peter Wonka. *Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?* *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, σελίδες 4431–4440, 2019.

[206] X. Huang και S. Belongie. *Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization*. *2017 IEEE International Conference on Computer Vision (ICCV)*, σελίδες 1510–1519, Los Alamitos, CA, USA, 2017. IEEE Computer Society.

[207] Zongze Wu, Dani Lischinski και Eli Shechtman. *StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation*. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 12858–12867, 2020.

[208] Peihao Zhu, Rameen Abdal, Yipeng Qin και Peter Wonka. *Improved StyleGAN Embedding: Where are the Good Latents?* *ArXiv*, abs/2012.09036, 2020.

[209] Peiye Zhuang, Oluwasanmi Koyejo και Alexander G. Schwing. *Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigation*. *ArXiv*, abs/2102.01187, 2021.

[210] A. V. Cherepkov, Andrey Voynov και Artem Babenko. *Navigating the GAN Parameter Space for Semantic Image Editing*. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 3670–3679, 2020.

[211] Rameen Abdal, Peihao Zhu, Niloy J. Mitra και Peter Wonka. *StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows*. *ACM Trans. Graph.*, 40(3), 2021.

[212] Ali Jahanian, Lucy Chai και Phillip Isola. *On the "steerability" of generative adversarial networks*. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[213] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen και Sylvain Paris. *GANSpace: Discovering Interpretable GAN Controls*. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[214] Fei Tony Liu, Kai Ming Ting και Zhi Hua Zhou. *Isolation Forest*. *2008 Eighth IEEE International Conference on Data Mining*, σελίδες 413–422, 2008.

[215] Markus M. Breunig, Hans Peter Kriegel, Raymond T. Ng και Jörg Sander. *LOF: Identifying Density-Based Local Outliers*. *SIGMOD Rec.*, 29(2):93–104, 2000.

[216] Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria και Tom M. Mitchell. *Conversational Neuro-Symbolic Commonsense Reasoning*. *ArXiv*, abs/2006.10022, 2020.

[217] Saeed Amizadeh, Hamid Palangi, Oleksandr Polozov, Yichen Huang και Kazuhito Koishida. *Neuro-Symbolic Visual Reasoning: Disentangling "Visual" from "Reasoning"*. *ArXiv*, abs/2006.11524, 2020.

[218] Ivan Donadello, Luciano Serafini και Artur S.d'Avila Garcez. *Logic Tensor Networks for Semantic Image Interpretation*. *International Joint Conference on Artificial Intelligence*, 2017.

[219] M. Cranmer, Alvaro Sanchez-Gonzalez, Peter W. Battaglia, Rui Xu, Kyle Cranmer, David N. Spergel και Shirley Ho. *Discovering Symbolic Models from Deep Learning with Inductive Biases*. *ArXiv*, abs/2006.11287, 2020.

[220] Henry Kautz. *The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture*. *AI Magazine*, 43(1):105–125, 2022.

[221] Dongran Yu, Bo Yang, Dayou Liu, Hui Wang και Shirui Pan. *A survey on neural-symbolic learning systems*. *Neural Networks*, 166:105–126, 2023.

[222] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum και Jiajun Wu. *The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision*. *ArXiv*, abs/1904.12584, 2019.

[223] Giuseppe Marra και Ondřej Kuzelka. *Neural Markov Logic Networks*. *Conference on Uncertainty in Artificial Intelligence*, 2019.

[224] Giuseppe Marra, Michelangelo Diligenti, Francesco Giannini, Marco Gori και Marco Maggini. *Relational Neural Machines*. *European Conference on Artificial Intelligence*, 2020.

[225] Yuan Yang και Le Song. *Learn to Explain Efficiently via Neural Logic Inductive Learning*. *ArXiv*, abs/1910.02481, 2019.

[226] Gabriele Picco, Hoang Thanh Lam, Marco Luca Sbodio και Vanessa Lopez Garcia. *Neural Unification for Logic Reasoning over Natural Language*. *Conference on Empirical Methods in Natural Language Processing*, 2021.

[227] X. Wang, Yufei Ye και Abhinav Kumar Gupta. *Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs*. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 6857–6866, 2018.

[228] Michael C. Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang και Eric P. Xing. *Rethinking Knowledge Graph Propagation for Zero-Shot Learning*. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 11479–11488, 2018.

[229] Kareem Ahmed, Stefano Teso, Kai Wei Chang, GuyVan den Broeck και Antonio Vergari. *Semantic Probabilistic Layers for Neuro-Symbolic Learning*. *Advances in Neural Information Processing Systems*S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave,

K. Cho και A. Oh, επιμελητές, τόμος 35, σελίδες 29944–29959. Curran Associates, Inc., 2022.

[230] Darian M. Onchis, Gilbert Rainer Gillich, Eduard Hogea και Cristian Tufisi. *Neuro-symbolic model for cantilever beams damage detection*. *Computers in Industry*, 151:103991, 2023.

[231] Tim Bohne, Anne Kathrin Patricia Windler και Martin Atzmueller. *A Neuro-Symbolic Approach for Anomaly Detection and Complex Fault Diagnosis Exemplified in the Automotive Domain*. *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, σελίδα 35–43, New York, NY, USA, 2023. Association for Computing Machinery.

[232] Vladimir Golovko, Aliaksandr Kroshchanka, Mikhail Kovalev, Valery Taberko και Dzmitry Ivaniuk. *Neuro-Symbolic Artificial Intelligence: Application for Control the Quality of Product Labeling*. *Open Semantic Technologies for Intelligent System*Vladimir Golenkov, Victor Krasnoproshin, Vladimir Golovko και Elias Azarov, επιμελητές, σελίδες 81–101, Cham, 2020. Springer International Publishing.

[233] Walter J. Scheirer, Anderson Rocha, Archana Sapkota και Terrance E. Boult. *Toward Open Set Recognition*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1757–1772, 2013.

[234] MinSik Chu, Seongmi Park, Jiin Jeong, Kyonghee Joo, Yongyeol Lee και Jihoon Kang. *Recognition of unknown wafer defect via optimal bin embedding technique*. *The International Journal of Advanced Manufacturing Technology*, 121:1–13, 2022.

[235] Luca Frittoli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto και Giacomo Boracchi. *Deep open-set recognition for silicon wafer production monitoring*. *Pattern Recognition*, 124:108488, 2022.

[236] Fei Tony Liu, Kai Ming Ting και Zhi Hua Zhou. *Isolation Forest*. *2008 Eighth IEEE International Conference on Data Mining*, σελίδες 413–422, 2008.

[237] Naeem Seliya, Azadeh Abdollah Zadeh και Taghi M. Khoshgoftaar. *A literature review on one-class classification and its potential applications in big data*. *Journal of Big Data*, 8(1):122, 2021.

English to Greek Glossary of Terms

Machine Learning
Μηχανική Μάθηση

Deep Learning
Βαθιά Μάθηση

Computer Vision
Υπολογιστική Όραση

Artificial Intelligence (AI)
Τεχνητή Νοημοσύνη (ΤΝ)

Industry 4.0
4η Βιομηχανική Επανάσταση

Industry 5.0
Βιομηχανία 5.0

Quality 4.0
Ποιότητα 4.0

Generative Adversarial Networks (GANs)
Παραγωγικά Αντιπαραθετικά Δίκτυα (ΠΑΔ)

Neurosymbolic Artificial Intelligence
Νευροσυμβολική Τεχνητή Νοημοσύνη

Out-of-Distribution (OOD)
Δείγματα Εκτός Κατανομής Εκπαίδευσης

Class Imbalance
Ανισορροπία Κλάσεων

Data Augmentation
Επαύξηση Δεδομένων

Training Data
Δεδομένα Εκπαίδευσης

Oversampling
Υπερδειγματοληψία

Novelty Detection

Ανίχνευση Καινοφανών Εισόδων

Cyber-Physical Systems

Κυβερνο-φυσικά Συστήματα

Digital Twin

Ψηφιακό Δίδυμο

Big Data

Μεγάλα Δεδομένα

Human-Machine Collaboration

Συνεργασία Ανθρώπου-Μηχανής

Human-in-the-Loop (HITL)

Άνθρωπος στο Βρόχο (Human-in-the-Loop)

Deep Convolutional Neural Networks (CNNs)

Βαθιά Συνελικτικά Νευρωνικά Δίκτυα (ΒΣΝΔ)

SMOTE (Synthetic Minority Over-sampling Technique)

Συνθετική Τεχνική Υπερδειγματοληψίας Μειονοτικών Κλάσεων (SMOTE)

Borderline-SMOTE

Παραλλαγή της SMOTE που επικεντρώνεται σε δείγματα κοντά στα όρια ταξινόμησης

ADASYN (Adaptive Synthetic Sampling)

Προσαρμοστική Συνθετική Δειγματοληψία

Wasserstein GAN (WGAN)

ΠΑΔ με Απόσταση Wasserstein

DCGAN (Deep Convolutional GAN)

Βαθύ Συνελικτικό Παραγωγικό Αντιπαραθετικό Δίκτυο

Actor-Critic GAN

Παραγωγικό Αντιπαραθετικό Δίκτυο με αρχιτεκτονική Actor-Critic

DeepSMOTE

Τεχνική επαύξησης εικόνων μέσω γραμμικών παρεμβολών στο επίπεδο χαρακτηριστικών (βασισμένη σε αρχιτεκτονική Κωδικοποιητή-Αποκωδικοποιητή)

BigGAN

Παραγωγικό Αντιπαραθετικό Δίκτυο Μεγάλης Κλίμακας

ImageNet

Διάσημο μεγάλο σύνολο δεδομένων εικόνων για εκπαίδευση αλγορίθμων μηχανικής όρασης

Binary Recall

Ανάκληση Δυαδικής Ταξινόμησης

AUROC (Area Under the Receiver Operating Characteristic Curve)

Εμβαδόν Κάτω από την Καμπύλη ROC

Precision

Ακρίβεια (Θετικών Προβλέψεων)

F1 Score

Μέτρο F1 (Αρμονικός Μέσος Ακρίβειας και Ανάκλησης)

Cross-Validation

Διασταυρωμένη Επικύρωση

Hyperparameter Optimization

Βελτιστοποίηση Υπερπαραμέτρων

MVTEC-AD

Δημόσιο σύνολο δεδομένων ανίχνευσης ελαττωμάτων στη βιομηχανία

Decision Boundary

Όριο Απόφασης (Ταξινόμησης)

Open-set Recognition

Αναγνώριση Ανοιχτού Συνόλου — Η ικανότητα ενός συστήματος ταξινόμησης να εντοπίζει δείγματα που δεν ανήκουν σε καμία από τις γνωστές κατηγορίες εκπαίδευσης.

Latent Space

Λανθάνων Χώρος

Semantic Factorization (SeFa)

Σημασιολογική Παραγοντοποίηση — Μέθοδος ανάλυσης λανθάνουσας αναπαράστασης μέσω παραγοντοποίησης ιδιοτιμών για την ανεύρεση σημασιολογικά πλούσιων κατευθύνσεων μεταβολής εικόνων.

Singular Value Decomposition (SVD)

Παραγοντοποίηση Ιδιαζουσών Τιμών

One-Class Classifier

Ταξινομητής Μίας Κλάσης

Extreme Value Theory (EVT)

Θεωρία Ακραίων Τιμών

Logical Tensor Networks (LTN)

Δίκτυα Λογικού Τανυστή — Πλαίσιο Νευροσυμβολικής ΤΝ που εισάγει λογικούς κανόνες σε μορφή διαφορίσιμων συναρτήσεων, οι οποίες ενσωματώνονται στη διαδικασία εκπαίδευσης ενός στατιστικού μοντέλου.

Grounding

Γείωση — Διαδικασία αντιστοίχισης των λογικών προτάσεων σε πραγματικές (διαφορίσιμες) συναρτήσεις που μπορούν να χρησιμοποιηθούν σε αλγορίθμους βελτιστοποίησης.

Fuzzy Logic

Ασαφής Λογική — Λογική που επιτρέπει βαθμούς αλήθειας μεταξύ 0 και 1 αντί για δυαδικές τιμές (0 ή 1), και χρησιμοποιείται για να εκφράσει αβεβαιότητα ή ασάφεια.

Out-of-Distribution (OOD)

Εκτός Κατανομής (ΕΚ) — Δεδομένα εισόδου που δεν αντιπροσωπεύονται από την κατανομή των δεδομένων εκπαίδευσης, συχνά συνδεδεμένα με το πρόβλημα της γενίκευσης.

Multi-Layer Perceptron (MLP)

Πολυεπίπεδο Perceptron — Αρχιτεκτονική νευρωνικού δικτύου πλήρως συνδεδεμένων στρωμάτων, συχνά χρησιμοποιούμενη για εποπτευόμενες εργασίες ταξινόμησης.

List of Publications

Journals

1. Theodoropoulos, S., Zajec, P., Rožanec, J. M., Kyriazis, D., & Tsanakas, P. (2024). On-the-fly image-level oversampling for imbalanced datasets of manufacturing defects. *Machine Learning, 113*(7), 4013–4035.

2. Theodoropoulos, S., Dardanis, D., Makridis, G., Kyriazis, D., & Tsanakas, P. (2024). Enhancing robustness to novel visual defects through StyleGAN latent space navigation: A manufacturing use case. *Journal of Intelligent Manufacturing*.

3. Zajec, P., Rožanec, J. M., Theodoropoulos, S., Fortuna, B., & Mladenić, D. (2024). Few-shot learning for defect detection in manufacturing. *International Journal of Production Research, 62*(19), 6979–6998.

4. Rožanec, J. M., Novalija, I., Zajec, P., Mladenić, D., & Soldatos, J. (2022). Human-centric artificial intelligence architecture for industry 5.0 applications. *International Journal of Production Research*.

Conferences

1. Theodoropoulos, S., Makridis, G., Kyriazis, D., & Tsanakas, P. (2024). Robust Novel Defect Detection with NeuroSymbolic AI. In *Advances in Production Management Systems. Production Management Systems for Volatile, Uncertain, Complex, and Ambiguous Environments. APMS 2024. IFIP Advances in Information and Communication Technology, vol 732. Springer, Cham.*.

2. Theodoropoulos, S., Zajec, P., Rožanec, J. M., Kyriazis, D., & Tsanakas, P. (2023). Identifying novel defects during AI-driven visual quality inspection. *IFAC-PapersOnLine, 56*(2), 3738–3743.

3. Rožanec, J. M., Zajec, P., Theodoropoulos, S., Fortuna, B., & Mladenic, D. (2023). Robust anomaly map assisted multiple defect detection with supervised classification techniques. *IFAC-PapersOnLine, 56*(2), 7846–7851.

4. Rožanec, J. M., Zajec, P., Theodoropoulos, S., Fortuna, B., & Mladenic, D. (2023). Synthetic data augmentation using GAN for improved automated visual inspection. *IFAC-PapersOnLine, 56*(2), 11094–11099.

5. Makridis, G., Theodoropoulos, S., Dardanis, D., Kyriazis, D., & Koulouris, P. (2022). XAI enhancing cyber defense against adversarial attacks in industrial applications. In *5th IEEE International Image Processing, Applications and Systems Conference, IPAS*.

6. Rožanec, J. M., Zajec, P., Kenda, K., Theodoropoulos, S., & Soldatos, J. (2021). STARdom: An architecture for trusted and secure human-centered manufacturing systems. In *IFIP Advances in Information and Communication Technology, 633*, 199–207.

Workshops

1. Theodoropoulos, S., Makridis, G., Pnevmatikakis, A., Moulos, V., Kyriazis, D., & Tsanakas, P. (2025). A NeuroSymbolic Human-in-the-loop Approach Towards Fusing Medical Expert Knowledge with ANNs. In *SilverTech: Empowering the Future of Ageing Through Advanced AI-based Technologies. AIAI 2025. (Accepted)*.

Book Chapters

1. Theodoropoulos, S., Dardanis, D., Sofianidis, G., Tsanakas, P., & Kyriazis, D. (2021). Confidence assessment of AI models in simulated industrial environments. In *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production* (pp. 114–131).