Optimizing Defect Detection: A Machine Learning-Ready Data Processing Pipeline

Loredana Cristaldi[1], Emilia Lenzi[1], Davide Martinenghi[1], Luca Martiri[1], Andrea Moschetti[1], Letizia Tanca[1] and Marco Zanoni[1]

[1]*Politecnico di Milano, DEIB, 20133 Milano*

Abstract

Printed Circuit Boards (PCBs) are essential in modern electronics, but ensuring their quality is increasingly difficult due to complex designs and common soldering defects. While AOI and X-ray inspections automate detection, manual data handling still introduces inconsistencies.

This study proposes a data processing pipeline that standardizes AOI and X-ray outputs to improve consistency and support machine learning. It integrates preprocessing, harmonizes defect labels, and leverages a relational database for automated storage and analysis. Tests on real-world data show reduced false positives and less need for manual verification.

Keywords

Data Pipeline, Anomaly Detection, Machine Learning, PCB,

1. Introduction

In automated quality control for printed circuit boards (PCBs), Automated Optical Inspection (AOI) and X-ray Inspection (AXI) generate vast amounts of heterogeneous data, including images, structured test results, and metadata. However, making effective use of this data requires a robust processing pipeline that handles data integration, cleaning, transformation, and storage for subsequent analysis. Traditional AOI systems, while efficient in identifying potential defects, suffer from high false-positive rates due to rule-based heuristics that fail to generalize across different manufacturing conditions. Similarly, AXI data is often stored in unstructured formats, making integration with AOI results complex and prone to inconsistencies [1].

The primary challenge is constructing a comprehensive data pipeline that consolidates information from multiple sources, aligns defect detection outputs, and ensures consistency across inspection methods. This requires structured database design, schema optimization, and automated entity resolution techniques to match test results, component references, and board identifiers [2].

Furthermore, an additional problem for defect detection is caused by the imbalanced nature of the data. Non-defective components vastly outnumber defective ones, skewing machine learning models trained on raw data. Addressing this issue necessitates a preprocessing step that balances the dataset while preserving manufacturing variability [3, 4].

SEBD 2025: 33rd Symposium on Advanced Database System, June 16-19, 2025 - Ischia, Italy

loredana.cristaldi@polimi.it (L. Cristaldi); emilia.lenzi@polimi.it (E. Lenzi); davide.martinenghi@polimi.it (D. Martinenghi); luca.martiri@polimi.it (L. Martiri); andrea.moschetti@polimi.it (A. Moschetti); letizia.tanca@polimi.it (L. Tanca); marco2.zanoni@mail.polimi.it (M. Zanoni)

This paper presents a data pipeline that efficiently processes, integrates, and stores AOI and AXI data, creating a structured foundation for defect detection using experts' analysis input as ground truth for our systems. We detail the steps involved in data transformation, storage, and retrieval, providing a scalable framework applicable to large-scale PCB manufacturing environments. We demonstrate the effectiveness of our process by showcasing how the data preparation pipeline enables the successful application of Random Forest (RF) and Convolutional Neural Networks (CNNs) models in various configurations for defect detection. This approach significantly boosts precision, improving it from 3.03% to over 80% in the binary classification task. Furthermore, our approach introduces monitoring of false negatives, a crucial aspect currently overlooked in the AOI system.

2. Related Work

Automated defect detection in Printed Circuit Boards (PCBs) has transitioned from rule-based methods to data-driven approaches. Early systems relied on handcrafted feature extraction and rule-based classification, but these methods struggled with variations in lighting, alignment, and defect complexity [5]. Machine learning techniques, such as Support Vector Machines (SVMs) and RF, improved classification accuracy by leveraging structured defect data [6], but they remained limited in handling real-world manufacturing variability.

Deep learning further advanced PCB defect detection, with CNNs excelling in feature extraction from AOI images [7]. While CNN-based models outperform traditional methods, they require extensive labeled datasets, making data preprocessing and integration crucial [8]. Alternative approaches, such as object detection models like YOLO and Faster R-CNN, have been applied for precise defect localization, balancing accuracy and computational efficiency [9, 10]. However, challenges persist due to class imbalance and dataset limitations, often addressed through data augmentation and transfer learning [11].

Beyond algorithmic improvements, the effectiveness of defect detection hinges on structured data pipelines that integrate AOI, X-ray, and manual inspection data. Recent studies emphasize the importance of data alignment across heterogeneous sources to reduce false positives and enhance defect classification reliability [12]. This work builds on these findings by proposing a data pipeline solution that ensures consistency in defect detection across multiple inspection methods.

3. Data Processing Pipeline

The proposed methodology addresses the integration and preprocessing of heterogeneous data sources used in Printed Circuit Board (PCB) defect detection. The pipeline consists of three main phases:

- Data transformation: Preprocessing to standardize formats and correct inconsistencies.
- Exploratory Data Analysis (EDA): Understanding data distribution and identifying key challenges.

- Data integration: Merging information from different inspection sources into a unified dataset.

Each phase presents distinct challenges, which are addressed through strategies that improve the overall reliability of defect classification models by ensuring high-quality input data, reducing inconsistencies, and providing a structured foundation for further analysis and defect detection algorithms.

Given the increasing complexity of PCB designs and the higher accuracy demanded in defect detection, traditional single-source inspection approaches are insufficient. By integrating AOI, X-ray, and manual inspection data, our pipeline establishes a high-resolution defect detection framework that minimizes false positives and maximizes classification accuracy. However, achieving this level of integration requires addressing multiple issues related to data format inconsistencies, identifier mismatches, and varying defect classification criteria.

In the following subsections, we will provide a detailed description of the pipeline's phases.

Data Transformation Data was collected from multiple sources, including Automated Optical Inspection (AOI) images, tabular test results, X-ray scans, PCB layout files, and operator manual evaluations, which served as the ground truth for supervised learning models. These data sources exhibited inconsistencies in structure, identifier formats, and missing information, which made direct integration impossible and required a comprehensive transformation strategy.

To standardize AOI images, each image—representing a portion of the whole board—was systematically renamed following a unified naming convention to enable board sub-region-based analysis. However, inconsistencies in naming conventions made automated alignment difficult. To address this, custom scripts were developed to match images with component identifiers based on spatial metadata and inferred relationships. PCB layout diagrams, available as PDF files, lacked coordinate-based information, requiring manual mapping of component names to their corresponding AOI and test results. Similarly, X-ray inspection data, originally stored in HTML format with embedded images, required custom parsing scripts to extract structured defect classifications and unique identifiers for the inspected PCB. Each entry was enriched with inspection dates, retrieved from the timestamp of the X-ray inspection, and details of the imaging equipment used. Since many entries had inconsistent labeling, an adaptive text-matching approach was necessary to unify defect categories. As far as identifier standardization, Locality-Sensitive Hashing (LSH) was used to group similar Board IDs and identify formatting patterns, and regular expressions were then applied to extract relevant components based on these patterns to resolve discrepancies between AOI and X-ray test results. This significantly improved cross-referencing, although some cases persisted where ambiguous abbreviations were used.

Despite these standardization efforts, several challenges remained. AOI image resolution varied across inspections, complicating defect localization. Additionally, missing test result entries and unexpected values without semantic meaning introduced biases, which required mitigation through a semi-supervised labeling approach based on past inspection data. Furthermore, bounding boxes were manually placed around components in AOI images to provide

training data for supervised models. However, this process was labor-intensive, highlighting the need for automated annotation in future iterations.

Exploratory Data Analysis EDA was performed to evaluate the effectiveness of data transformation and identify key issues within the dataset.

First of all, discrepancies between AOI and manual inspections exposed systematic differences in defect assessment criteria. While AOI tended to over-detect minor imperfections, human inspectors focused on severe issues, necessitating a harmonization strategy to align defect classifications across inspection methods.

The analysis revealed that AOI systems detected significantly more defects than human inspectors, indicating a high false-positive rate. These false positives often resulted from minor solder irregularities flagged by pre-set AOI thresholds, which required a statistical analysis to refine these thresholds dynamically. Moreover, Table 1 highlights the limitations of the AOI system when considering the operator's inspection as the ground truth. Notably, the AOI system exhibits a high false positive rate, incorrectly identifying a significant number of components as defective when they are not (29481 cases). This results in particularly low precision ($\approx 3.03\%$), potentially leading to wasted resources spent on inspecting components that were wrongly flagged as defective. On the other hand, the zero associated with false negatives — meaning no instances where the AOI system missed a defect identified by the operator — is misleading. This does not indicate that the system is foolproof in detecting all defects; rather, it reflects a gap in the verification process: false negatives are not systematically checked in the current workflow. Nevertheless, such errors, although not directly reflected in the table, have a significant negative impact on business processes, as undetected defects may result in additional costs and potential non-conformities in production.

Furthermore, the analysis of the X-ray dataset revealed an extreme class imbalance, with defective components accounting for less than 1% of inspected samples. This imbalance presented challenges for machine learning models, prompting the consideration of oversampling, synthetic defect generation, and augmentation strategies as mitigation techniques.

	Operator: No Defect (0)	**Operator: Defect (1)**
AOI: No Defect (0)	2,830,061	0
AOI: Defect (1)	29,481	920

Table 1
Confusion Matrix: AOI vs. Operator Outputs

In this context, a key challenge was the lack of ground truth labels for many defect cases. Operator classifications were prioritized as the most reliable, yet inconsistencies in their evaluation introduced subjectivity. Additionally, defects identified by X-ray were not always detected in AOI, revealing limitations in purely visual inspection techniques. This discrepancy arises because X-ray inspection is typically used for components with solder joints located beneath them, which are not visible through AOI alone. Consequently, combining AOI and X-ray analysis ensures full defect coverage, overcoming the limitations of individual inspection methods. Cross-validation of AOI and X-ray results was explored to enhance defect detection reliability.

Data Integration The final stage of the pipeline involved merging multiple inspection data sources into a structured dataset. To achieve this, identifier matching was carried out, where AOI and X-ray test results were aligned using probabilistic matching techniques and heuristics, improving cross-system consistency. Subsequently, the defect labels from AOI and manual inspection were standardized by mapping the descriptions to a unified defect taxonomy.

Despite these efforts, identifier mismatches persisted due to variations in board numbering conventions.

In the following sections, we will explore the additional steps undertaken to prepare the final dataset for the machine learning task and describe the final structure of the designed database.

3.1. Additional preprocessing actions

As deeply discussed in the previous sections, the initial statistical assessment of the dataset revealed a severe class imbalance, with defective components accounting for less than 1% of all inspected samples. Some defect categories were significantly underrepresented, posing a challenge for machine learning models prone to favoring majority classes. To address this, two dataset configurations were developed: (i) Full Defect Set: Preserved all defect categories, maintaining the original imbalance. While this configuration provided the most complete defect representation, it required additional rebalancing techniques to prevent model bias; (ii) Selected Subset: Focused only on the four most frequently occurring defect types, ensuring a more even class distribution and improving classification reliability for high-occurrence defects.

Moreover, as shown before, the available data included structured test results and unstructured AOI/X-ray images, each requiring different preprocessing steps. To ensure high-quality input for machine learning models, AOI and X-ray images were normalized and standardized to address variations in lighting and exposure across different inspection conditions, in addition, bounding box coordinates were adjusted based on statistical outlier detection, reducing annotation errors in defect localizations.

More precisely, AOI images were systematically linked to defect annotations stored in the structured database, ensuring consistency across multiple inspection sources. Labels were assigned based on operator-confirmed defects, reducing misclassification risks. To address the issue of class imbalance, a controlled sampling strategy was used to pair defective components with comparable non-defective ones, maintaining representativeness while preventing model bias and ensuring balanced training data.

Image preprocessing standardized resolution and intensity levels, addressing exposure variations. Data augmentation techniques—including brightness adjustments and Gaussian noise—were applied to improve model robustness under real-world conditions.

These preprocessing steps ensured that defect localization models could handle imaging inconsistencies and generalize across different manufacturing conditions.

For what concerned the structured test results, the raw inspection data were transformed into feature vectors that represented several key aspects. These included component-level attributes such as board type, component type, test frequency, and past defect history.

Inspection metadata was incorporated, which included details like the test method (whether it was AOI, manual, or X-ray, when available), the inspection timestamp, and defect confidence scores.

Finally, defect classification features were added, which consisted of encoded defect labels, failure severity, and the likelihood of misclassification.

Feature selection was performed using mutual information ranking, ensuring that only the most relevant predictors were retained for defect classification. Additionally, dimensionality reduction via Principal component analysis (PCA) was tested to improve computational efficiency without sacrificing classification accuracy

3.2. Final Dataset Structure

The final dataset was designed to support both structured queries and machine learning-based defect detection. It comprises of a tabular dataset, which functions as a structured database storing the inspection results, linking each component to its corresponding defect classification, metadata, and historical inspection outcomes, and an image dataset, consisting of cropped and preprocessed AOI/X-ray images that were paired with the corresponding defect annotations, formatted to serve as input for deep learning models.

To facilitate scalability and rapid access, the database schema was optimized specifically for machine learning applications. One key optimization was indexing, which was applied to frequently queried attributes such as component IDs and defect categories, allowing for quicker data retrieval. Additionally, partitioning was implemented for tables storing historical inspection data, with partitions organized by time. This approach improved query speed and enabled real-time defect monitoring. To further optimize storage, hybrid storage strategies were employed; rather than storing the images directly in the database, file path references were used. This not only reduced storage overhead but also ensured seamless access to the image datasets.

This structured approach ensures efficient defect retrieval, supporting both traditional statistical analyses and modern AI-driven defect classification methods.

4. Machine Learning for Defect Detection

The generated dataset was used to train and evaluate various machine learning models for classifying defective and non-defective components. Following an initial exploratory phase that tested different model architectures, the most promising candidates advanced to the final training stage. The models selected for evaluation were:

- **Random Forest (RF)**: Trained on numerical features extracted from AOI images and structured test results, RF provided strong interpretability and robustness against class imbalance.
- **Convolutional Neural Networks (CNNs)**: Designed to process AOI images directly, CNNs captured spatial defect patterns to enhance classification granularity.

The models were trained using a 70/30 train-test split on two dataset configurations: (i) the full defect set and (ii) a more balanced subset containing the most frequent defect types. Table 2 summarizes the classification performance. The results indicate that the RF model achieved high accuracy, precision, and recall across both subsets, whereas the CNN's performance declined

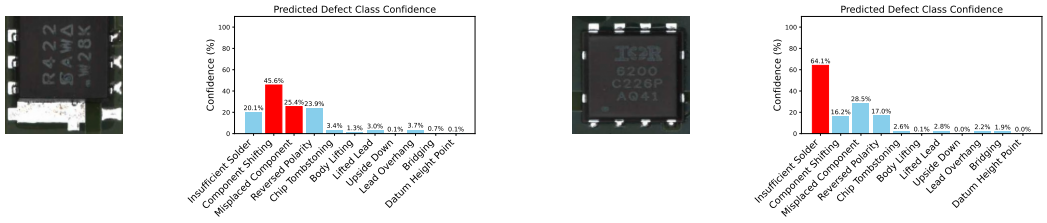

Figure 1: Examples of outputs for the multi-label approach. In red, are shown the expected defect classes.

on the *Full Defect* set due to significant dataset imbalance, particularly in defect classes with fewer than 10 samples.

Approach	RF			CNN		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
Selected Subset	88.9%	89.0%	89.0%	86.2%	87.3%	85.1%
Full Defect Set	89.0%	89.0%	89.2%	82.7%	82.7%	82.8%

Table 2
Comparison of Accuracy, Precision, and Recall for RF and CNN in binary classification.

After testing the binary classification, a multi-label approach was applied using a pre-trained ResNet model with additional convolutional layers to enhance defect classification. This approach provides valuable insights into the probability of each defect occurring in different components, which is crucial for improving maintainability and facilitating faster fault recovery. The proposed method was tested on two different subsets of defective components: the first containing 12 defect classes and the second with 11, where the “Missing Component" fault was excluded, as it represents a fundamentally different issue compared to placement or welding defects, two examples of the output of this step are shown in Figure 1. The results, presented in Table 3, demonstrate high precision in identifying the most probable defect. However, the low recall—particularly when the “Missing Component" class is considered—indicates that the model sometimes predicts defects that are not actually present in the component. These results could be improved by increasing the amount of available training data and adjusting fault detection thresholds based on historical information.

Approach	AUC	Precision	Recall
All Defects	95.4%	86.5%	58.8%
Without “Missing Component"	88.3%	22.0%	70.5%

Table 3
Comparison of AUC (Area Under the Curve), Precision, and Recall for RF in multi-label classification.

To evaluate real-world applicability, models were tested under varying brightness conditions and noise levels (Table 4). CNNs demonstrated higher robustness to brightness fluctuations, while both RF and CNNs showed performance degradation with increasing image noise.

Noise Level	RF			CNN		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
10% black pixels	84.0%	84.1%	83.9%	83.6%	83.0%	83.1%
30% black pixels	77.2%	76.8%	77.1%	77.4%	81.5%	81.0%

Table 4
Impact of noise levels on RF and CNN classification performance.

4.1. Discussion on models performances

Both RF and CNN models contribute significantly to improving AOI-based defect detection, each excelling in different aspects. RF demonstrated strong accuracy, precision, and recall across both dataset configurations, leveraging numerical features for structured classification while effectively handling class imbalance. CNNs, on the other hand, excelled in capturing spatial defect patterns directly from AOI images, offering finer classification granularity. However, its performance declined on the highly imbalanced *Full Defect* set, highlighting the need for data augmentation and rebalancing techniques.

Moreover, both methods significantly improve precision compared to the AOI system in all configurations, effectively addressing the high false positive rate while maintaining strong accuracy and recall and mitigating false negatives.

To further enhance defect classification, both a multi-label approach and a noise-robustness test were conducted. The multi-label approach provided insights into the probability of each defect occurring in different components, while the robustness test evaluated the models' resilience to image distortions caused by machinery faults. These analyses highlighted the importance of establishing performance thresholds to monitor the overall health of the inspection process.

5. Conclusion and Future Works

This study presents a comprehensive approach to PCB defect analysis by combining a robust data pipeline with machine learning techniques. The proposed pipeline effectively integrates AOI, X-ray, and manual inspection data, reducing inconsistencies through identifier standardization and enhancing data alignment. Overall, integrating ML models into a structured defect detection pipeline enhances classification by combining numerical inspection data with image-based defect patterns.

To further improve performance, future research should focus on automating defect annotation using object detection models and exploring active learning strategies to iteratively refine defect classifications. Additionally, multimodal approaches that combine structured tabular data with visual defect patterns hold promise for enhancing robustness across inspection conditions. To further improve detection performance and robustness, dataset expansion, adaptive thresholding, and model fusion strategies can also be considered in future works. These advancements are expected to further improve defect classification accuracy, reduce manual intervention, and strengthen the scalability of the proposed system in real-world manufacturing environments.

Acknowledgment

This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-Generation EU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). CUP MICS D43C22003120001.

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI to check grammar and spelling. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. E. See, Visual inspection: a review of the literature. (2012).
[2] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering 21 (2009) 1263–1284.
[3] Q. Ling, N. A. M. Isa, Printed circuit board defect detection methods based on image processing, machine learning and deep learning: A survey, IEEE Access 11 (2023) 15921–15944.
[4] P. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, F. Khomh, Data cleaning and machine learning: a systematic literature review, Autom. Softw. Eng. 31 (2024) 54.
[5] A.-A. I. Hassanin, F. E. Abd El-Samie, G. M. El Banby, A real-time approach for automatic defect detection from pcbs based on surf features and morphological operations, Multimedia Tools and Applications 78 (2019) 34437–34457.
[6] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[7] M. B. Akhtar, The use of a convolutional neural network in detecting soldering faults from a printed circuit board assembly, HighTech and Innovation Journal 3 (2022) 1–14.
[8] Q. Ling, N. A. M. Isa, Printed circuit board defect detection methods based on image processing, machine learning and deep learning: A survey, IEEE Access 11 (2023) 15921–15944.
[9] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[10] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[11] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.
[12] L. Martiri, A. Moschetti, E. Lenzi, M. Zanoni, L. Cristaldi, L. Tanca, D. Martinenghi, A data pipeline to classify pcb welding defects on noisy data, Accepted in IEEE I2MTC (2025).