Chapter 5
Explainable AI for Automated Defect Detection:
Combining ML Algorithms and LLMs

Raj Kishor Verma
https://orcid.org/0009-0009-9645-718X
Galgotias University, India

Gaurav Agarwal
Galgotias University, India

Nitin Pandey
Galgotias University, India

Raj Kishor Verma
https://orcid.org/0009-0005-7216-7752
ABES Institute of Technology, India

ABSTRACT

Defect detection systems that are automated are increasingly mixing machine learning (ML) with big language models (LLMs) to enhance accuracy and interpretability in industrial quality control. Albeit ML methods such as convolutional neural networks (CNN) and image transformers are powerful in abnormality recognition for visual data, the black-box nature of their models often limits transparency and user trust. Combining the LLMs with explainable AI (XAI) techniques such as post-hoc methods like Grad-CAM and SHAP makes it possible for these systems to produce

DOI: 10.4018/979-8-3373-4460-7.ch005

human-understandable stories about reports on defects as well as offer reasons why they happened. LLMs can also be used to not only classify defects but also provide real-time correction advice while avoiding slowing down operation during transmission of data. By this combined method introduction, not only does defect detection performance increase but it also supports regulatory compliance, and boosts user confidence.

INTRODUCTION

Quality and reliability of products have become a major challenge in the current situation for both the manufacturing as well as software industries. Legacy defect detection (for example, manual inspection or rule-based automation) is often unable to handle the number, the complexity, and the subtlety of the defects found in modern high-speed production lines and complex software systems. As industries work to achieve increased standards and efficiency, Artificial Intelligence (AI) has become a solution of transformation, automating and improving the accuracy of defect detection, (Verma, 2025a). Machine Learning (ML) algorithms, especially computer vision and deep learning, have shown impressive results in detecting anomalies, classifying defects, and forecasting failures using large and complex data sets. Such tools can help autonomous systems scan images, sensor data, and even text for tiny irregularities that could elude human inspectors. Due to the use of advanced AI models—namely deep neural networks—we now encounter new challenges with regard to the interpretability and transparency of the models' decisions. In applications such as manufacturing or software testing, high-stakes decisions lead stakeholders to request, not only high-performing classifiers, but also the justification behind the decisions made so it is understandable for being regulatory (e.g. regulatory explanations) compliant and (e.g. operational) trustable, (Verma, 2025b) Explainable Artificial Intelligence (XAI) has been developed to meet this need by providing information on the processes that AI models use to reach their decisions to make their predictions more transparent and useful. Interest has increased recently in the integration of XAI methods with ML-based defect detection systems, to enable practitioners to gain a visual and intuitive understanding of the features and patterns that cause a particular type of defect to be classified. In addition, the emergence of Large Language Models (LLMs) has provided new possibilities for explainability and automation. LLMs are able to understand, summarize, and contextualize findings related to the defect detection results and to process unstructured textual data like inspection reports or maintenance logs, thus enriching classical ML methods with advanced natural language comprehension, (Verma & Kishor, 2024). Through the integration of ML's strength in automated