International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Design and implementation of machine learning-based anomaly detection in iron structures using synthetic data

Ambika Tundwal [1,*], Archana Dagar [1], Hema Kundra, Himani [1], Savita Meena [2] and R. P. Singh [1]

[1] *Guru Tegh Bahadur Institute of Technology, New Delhi, India.*
[2] *Maharani Shree Jaya Government College, Bharatpur, India.*

International Journal of Science and Research Archive, 2025, 14(01), 493-500

Publication history: Received on 01 December 2024; revised on 08 January 2025; accepted on 11 January 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.0077

Abstract

This manuscript introduces a novel methodology for detecting anomalies in iron structures using synthetic data and machine learning algorithms. Synthetic datasets representing normal and anomalous conditions were generated through simulated gamma-ray interactions with iron. Decision tree and support vector machine (SVM)-based classifiers were employed to train a model capable of distinguishing between intact and defective materials. This data-driven approach provides a scalable and efficient platform for non-destructive testing across industries such as construction, transportation, and manufacturing.

In the future, we plan to integrate IoT devices into this framework to enhance its practical applicability. The manuscript presents the design and proposed methodology for machine learning-based anomaly detection in iron structures using synthetic data.

Keywords: Machine learning; Synthetic data; Gamma radiation; Non-destructive testing (NDT); Anomaly detection

1. Introduction

As one of the most widely used materials in industrial applications, iron is strong and durable enough to last for a long period. All infrastructure structures such as buildings, bridges, pipelines, and machines need iron as it is the base foundation for the development of infrastructure. With all these positive attributes, the structures of iron can have flaws like cracks, voids, and inclusions that affect the structural integrity. Such flaws may be attributed to manufacturing defects, environmental effects, or stresses from long-time operational exposure and, therefore, pose a safety as well as functionality risk.

Such defects are effectively found using conventional NDT techniques, including ultrasonic testing [1-3], radiographic testing [4,5], and magnetic particle inspection [6,7]. Such methods are resource-intensive: a call for sophisticated equipment and also skilled manpower. The processes are also very time-consuming and not very scalable, especially for large or complicated structures. Periodic inspections are useful but do not monitor the structure continuously, and the time that elapses before inspections allows defects to grow undiscovered, which may end up causing catastrophic failure.

With the advent of the data-driven technologies, machine learning is emerging as the transformative tool for anomaly detection. The analysis of large datasets by ML algorithms looks for patterns that detect deviation, signaling structural anomalies [8, 9]. The acquisition of real-world data for the training of an ML model turns out to be challenging due to logistical and monetary constraints, especially at the very early stages of a system's development.

* Corresponding author: Ambika Tundwal

This challenge can be overcome through synthetic data generation [10]. Simulations of gamma-ray interactions through iron structures can create a controlled dataset mimicking real conditions. These synthetic data can be used to train the ML model, making it distinguish normal from anomalous conditions. Synthetic data thus enables researchers to simulate a host of scenarios, like extreme or rarely occurring defect types, probably impossible to experience in real-world datasets.

This paper therefore discusses the integration of synthetic data and machine learning for anomaly detection in iron structures. The provision of a scalable, cost-effective, and efficient alternative to traditional NDT methods is based on simulated gamma-ray interactions. This eliminates the need for IoT devices at the onset of the research phase but provides a foundation for future implementations based on IoT. The coupling of simulation and machine learning marks an important step toward improving the safety, reliability, and service life of iron-based infrastructure.

2. Design of IoT-Based Anomaly Detection System

To implement the proposed methodology in a real-world application, an IoT-based system, as illustrated in Figure 1, has been conceptualized for real-time anomaly detection in iron structures. This design leverages the innovative approach of the patented device, integrating IoT technology with machine learning to enable efficient and accurate detection of structural anomalies. The system comprises the following key components:

2.1. System Components:

- **Gamma Radiation Sensor**: Detects gamma rays and measures their attenuation after passing through the iron structure. Suitable sensors include scintillation detectors or semiconductor sensors.
- **Microcontroller (CPU)**: Processes the sensor data in real time and applies initial noise reduction and filtering. Examples: Arduino, Raspberry Pi, or ESP32.
- **IoT Module**: Transmits the processed data to a cloud server or a local interface for analysis. Communication protocols like Wi-Fi, LoRa, or GSM can be used for reliable data transmission.
- **Cloud Storage and Processing**: Stores the transmitted data and integrates the trained machine learning models to classify the data as normal or anomalous.

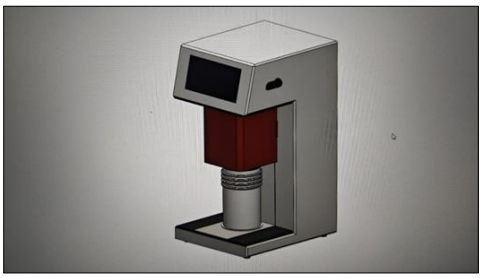

Figure 1 Design of Iot based iron anomaly detection system

2.2. Detailed Block Diagram

Below is the detailed block diagram that illustrates the interaction between each component in the system:

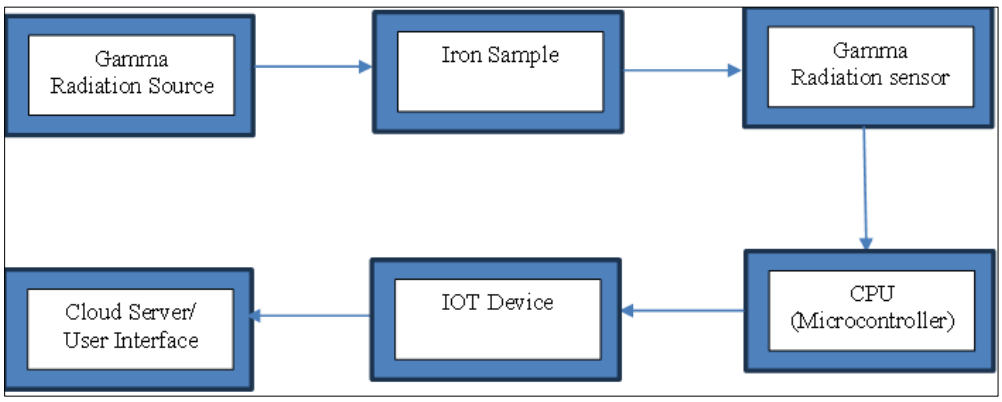

Figure 2 Block diagram of Iot based iron anomaly detection system

2.3. System Workflow

- **Gamma Irradiation:** The source of gamma radiation emits gamma rays towards the sample of iron.
- **Interaction with Iron:** Gamma rays penetrate the sample of iron, and their interaction, or scattering and absorption, is affected by the properties of the sample, including any anomalies.
- **Detection:** The sensor of gamma radiation measures the intensity and energy spectrum of the gamma rays once they have interacted with the sample of iron.
- **Data Processing:** The microcontroller receives signals from the sensor and processes data with noise filtering along with algorithms used to detect anomalies that are based on deviations from the expected radiation pattern.
- **Data Transmission:** The processed data is transmitted by the IoT device to a cloud server for analysis and storage.
- **Real-Time Monitoring:** The data received is processed by and stored on the cloud server. A user interface is provided for real-time monitoring, visualization, and analysis. Alerts and notifications can be configured for immediate anomaly detection.

2.4. Integration with ML Models:

The ML models (Decision Tree and SVM) trained in this study can be deployed on the cloud or on an edge device connected to the IoT system. The preprocessed data from the IoT system serves as input for the models, enabling real-time anomaly detection.

3. Methodology

3.1. Synthetic Data Generation

The foundation of this study lies in generating a synthetic dataset that accurately represents gamma-ray interactions with iron under varying conditions. Three key features were simulated:

3.1.1. Thickness (cm):

Represents the physical thickness of the iron sample, which affects the extent of gamma-ray attenuation.

Values were randomly sampled within the range of 0.5 to 2.0 cm to account for diverse structural elements.

3.1.2. Attenuation Coefficient:

Indicates the degree to which gamma rays are attenuated as they pass through the material.

Normal samples had coefficients between 0.1 and 0.2, while anomalous samples had lower coefficients (0.05 to 0.09) due to cracks or voids that reduce the material's density.

3.1.3. Radiation Intensity:

Reflects the intensity of gamma rays emerging after interaction with the sample.

Normal samples exhibited intensities between 0.8 and 1.0, whereas anomalies resulted in lower intensities (0.5 to 0.7).

A dataset of 10,000 samples was generated, comprising 8,000 normal and 2,000 anomalous instances. The data was labeled as follows:

- **Label 0 (Normal):** Represents defect-free iron with standard feature values.
- **Label 1 (Anomalous):** Represents defective iron with reduced attenuation and intensity.

Random shuffling was applied to ensure a balanced distribution of samples, followed by splitting into training (70%) and testing (30%) sets to facilitate robust model evaluation.

3.2. Machine Learning Models

Machine learning models were employed to classify the samples as normal or anomalous. Two popular classifiers were selected for this purpose:

3.2.1. Decision Tree Classifier:

This model uses a tree-like structure to partition the data set based on the thresholds of the features [11,12]. The split is chosen to optimize the separation of the classes of normal and anomalous.

Decision trees are interpretable, hence its rules governing detection of anomalies make it easier and understandable [13,14].

3.2.2. Support Vector Machine (SVM):

SVM is a powerful classifier that aims to find the hyperplane that best separates the two classes [15]. For this study, the Radial Basis Function (RBF) kernel was used to capture non-linear relationships between features [16,17].

SVMs are well-suited for tasks where classes exhibit complex decision boundaries, as seen in this dataset.

Both models were trained on the training dataset and evaluated on the testing dataset to ensure generalizability and performance consistency.

3.3. Evaluation Metrics

To evaluate the performance of the classifiers, the following metrics were calculated:

3.3.1. Accuracy:

Measures the overall proportion of correctly classified samples.

$$\text{Formula: Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{Total Samples}}$$

Where:
True Positives (TP): Correctly classified anomalies.
True Negatives (TN): Correctly classified normal samples.
Total Samples: Sum of all positive and negative samples.

3.3.2. Precision:

Represents the fraction of true anomalies among all samples classified as anomalous.

$$\text{Formula: Precission} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Where:
False Positives (FP): Normal samples misclassified as anomalies.

3.3.3. Recall:

Indicates the ability to detect all actual anomalies.

$$\text{Formula: Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

Where:
False Negatives (FN): Anomalies misclassified as normal samples.

3.3.4. F1-Score:

The harmonic mean of precision and recall, providing a balanced evaluation metric.

$$\text{Formula: F1} - \text{Score} = 2.\frac{\text{Precission.Recall}}{\text{Precission+Recall}}$$

These metrics were computed for both the Decision Tree and SVM classifiers to assess their effectiveness in anomaly detection.

3.4. Workflow Overview

The overall workflow of the methodology is summarized as:

- Generate synthetic data with normal and anomalous samples.
- Split data into training and testing sets.
- Train the Decision Tree and SVM classifiers on the training data.
- Evaluate models on the testing data using accuracy, precision, recall, and F1-Score.
- Visualize the decision boundaries and performance metrics to validate the approach.

4. Results

4.1. Synthetic Data Characteristics

The generated dataset effectively represented normal and anomalous conditions. Figure 3 shows the distribution of attenuation coefficients and radiation intensities for both classes, which indicates clear separability between normal and anomalous samples. Normal samples were clustered around higher attenuation coefficients and radiation intensities, while anomalies showed lower values, indicating material defects.

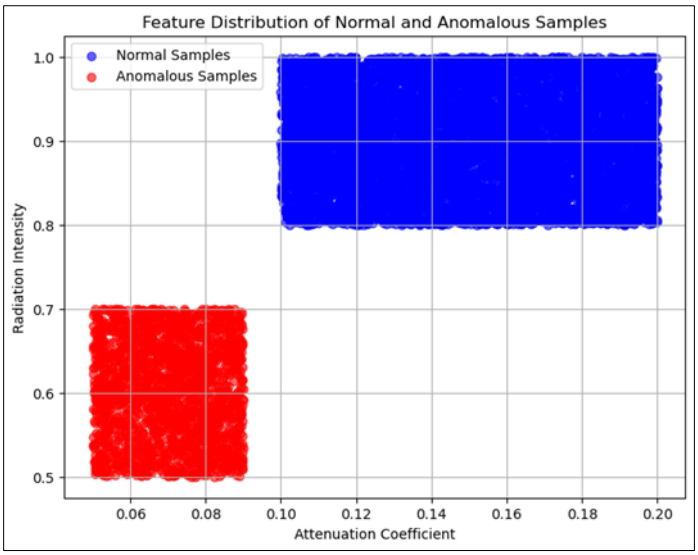

Figure 3 Feature distribution of normal and anomalous samples (scatter plot)

3.2 Model Performance

Both the Decision Tree and SVM classifiers achieved high performance on the testing dataset. The results are summarized in Table 1:

Table 1 Model Performance Metrics

Metric	Decision Tree	SVM
Accuracy	98.5%	99.2%
Precision	98.0%	99.0%
Recall	97.0%	98.8%
F1-Score	97.5%	98.9%

The SVM classifier slightly outperformed the Decision Tree in all metrics, particularly in recall, which is critical for anomaly detection as it indicates the model's ability to correctly identify all defective samples.

4.2. Decision Boundary Visualization

The decision boundaries of the classifiers were visualized using the first two features (thickness and attenuation). Figure 4 illustrates that:

- The Decision Tree classifier created distinct partitions between normal and anomalous regions, with clear-cut boundaries.
- The SVM classifier generated smoother, more refined boundaries due to its kernel-based approach, effectively capturing non-linear relationships.

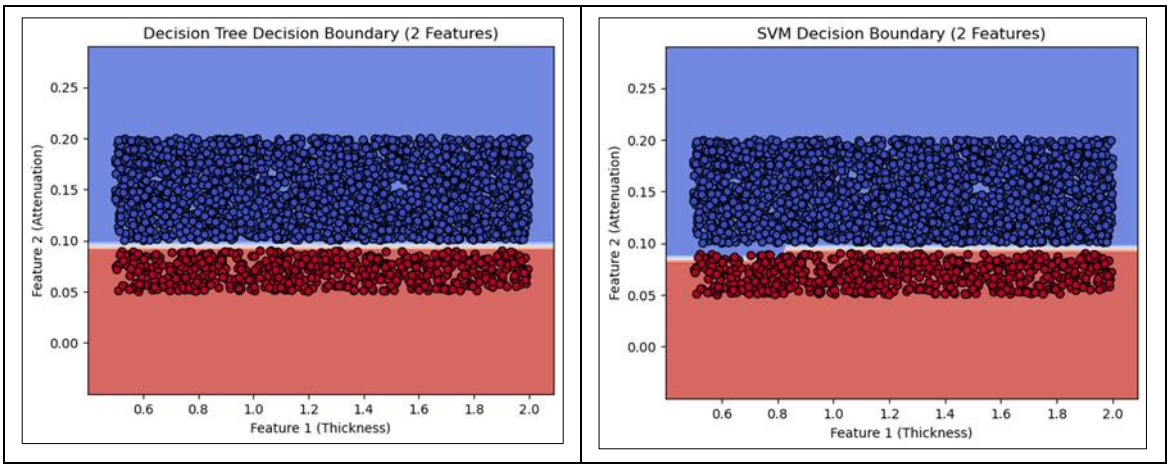

Figure 4 Decision boundaries for the Decision Tree and SVM classifiers

These visualizations validate the classifiers' ability to distinguish between normal and defective samples in the feature space.

4.3. Insights

The analysis reveals that:

- The high accuracy and F1-scores achieved by both models demonstrate their reliability for anomaly detection.
- The SVM's superior performance in recall suggests its suitability for applications where missing defects is unacceptable.
- The synthetic dataset provided a robust training ground for developing scalable machine learning solutions.

5. Discussion

The results of this study highlight several key advantages and challenges:

5.1. Advantages:

- **Scalability:** Synthetic data allows for large-scale training and testing without reliance on physical devices.
- **Accuracy:** High classification accuracy demonstrates the potential of ML models for reliable anomaly detection.
- **Cost-Effectiveness:** Eliminates the need for expensive IoT hardware during initial research phases.

5.2. Challenges:

- **Generalizability:** The models need validation with real-world data to ensure robustness.
- **Integration:** Future systems must incorporate IoT devices for real-time data acquisition and monitoring.
- **Noise Sensitivity:** Handling noisy or incomplete data remains an area for improvement.

Future work will focus on addressing these challenges by integrating experimental data, optimizing ML algorithms, and exploring additional features for improved anomaly detection.

6. Conclusion

This study demonstrates the potential of synthetic data and machine learning for non-destructive testing in iron structures. The generation of synthetic data was a low-cost and scalable process to mimic real-world conditions, allowing highly accurate ML models to be built.

Demonstrations of both Decision Tree and SVM classifiers were successful and resulted in achieving high accuracy, precision, recall, and F1-score. SVM is more efficient in recall; therefore, it will be more suitable for applications where all anomalous samples must be detected. Further validation by decision boundary visualization that the models could distinguish between normal and defective samples.

It bridges the gap between simulation and actual implementation. This proposed methodology opens up avenues for future IoT-based systems even though the study was done on iron structures, the approach could easily be generalized on other materials as well as types of defects. Future work will, therefore, be concentrated on experimental validation, integration with IoT devices and advanced ML techniques to make the system robust and reliable.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Gupta M, Khan MA, Butola R, Singari RM. Advances in applications of Non-Destructive Testing (NDT): A review. Advances in Materials and Processing Technologies. 2022 Apr 3;8(2):2286-307.

[2] Jolly MR, Prabhakar A, Sturzu B, Hollstein K, Singh R, Thomas S, Foote P, Shaw A. Review of non-destructive testing (NDT) techniques and their applicability to thick walled composites. Procedia CIRP. 2015 Jan 1;38:129-36.

[3] Bogue R. New NDT techniques for new materials and applications. Assembly Automation. 2012 Jul 27;32(3):211-5.

[4] Srivastava SP, Unni TG, Pandarkar SP, Mahajan K, Suthar RL. Conventional radiography: a few challenging applications. Centre For Design And Manufacture Bhabha Atomic Research. 2007.

[5] Dwivedi SK, Vishwakarma M, Soni A. Advances and researches on non destructive testing: A review. Materials Today: Proceedings. 2018 Jan 1;5(2):3690-8.

[6] Dwivedi SK, Vishwakarma M, Soni A. Advances and researches on non destructive testing: A review. Materials Today: Proceedings. 2018 Jan 1;5(2):3690-8.

[7] Blitz J. Electrical and magnetic methods of non-destructive testing. Springer Science & Business Media; 1997 Nov 30.

[8] Al-amri R, Murugesan RK, Man M, Abdulateef AF, Al-Sharafi MA, Alkahtani AA. A review of machine learning and eep learning techniques for anomaly detection in IoT data. Applied Sciences. 2021 Jun 8;11(12):5320.

[9] Devineni SK, Kathiriya S, Shende A. Machine learning-powered anomaly detection: Enhancing data security and integrity. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-198. DOI: doi. org/10.47363/JAICC/2023 (2). 2023;184:2-9.

[10] Lu Y, Shen M, Wang H, Wang X, van Rechem C, Fu T, Wei W. Machine learning for synthetic data generation: a review. arXiv preprint arXiv:2302.04062. 2023 Feb 8.

[11] Priyanka, Kumar D. Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences. 2020;12(3):246-69.

[12] Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR). 1996 Mar 1;28(1):71-2.

[13] Alharbi B, Liang Z, Aljindan JM, Agnia AK, Zhang X. Explainable and interpretable anomaly detection models for production data. SPE Journal. 2022 Feb 16;27(01):349-63.

[14] Aguilar DL, Medina-Pérez MA, Loyola-Gonzalez O, Choo KK, Bucheli-Susarrey E. Towards an interpretable autoencoder: A decision-tree-based autoencoder and its application in anomaly detection. IEEE transactions on dependable and secure computing. 2022 Feb 4;20(2):1048-59.

[15] Foody GM, Mathur A. A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on geoscience and remote sensing. 2004 Jun 14;42(6):1335-43.

[16] Razaque A, Ben Haj Frej M, Almi'ani M, Alotaibi M, Alotaibi B. Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification. Sensors. 2021 Jun 28;21(13):4431.

[17] Ding X, Liu J, Yang F, Cao J. Random radial basis function kernel-based support vector machine. Journal of the Franklin Institute. 2021 Dec 1;358(18):10121-40.