IJIE

INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

Vol. 18 No. 4 (2026) 199-216
https://publisher.uthm.edu.my/ojs/index.php/ijie

Multiclass Railway Wheel Defect Classification Using Machine Learning Algorithms: A Comparative Study

Nurfitri Muhamad Nasir[1], Siti Armiza Mohd Aris[1]*, Tay Yung Taw[1]

[1] *Faculty of Artificial Intelligence,*
Universiti Teknologi Malaysia, 54100 Kuala Lumpur, MALAYSIA

*Corresponding Author: armiza.kl@utm.my
DOI: https://doi.org/10.30880/ijie.2026.18.04.014

Article Info

Received: 26 November 2025
Accepted: 17 Mac 2026
Available online: 17 June 2026

Keywords

Railway defects, machine learning, synthetic data, multiclass classification

Abstract

Railway wheel condition monitoring plays a critical role in ensuring safe and reliable rail operations, as defects in wheel geometry can lead to excessive wear, vibration, and potential derailment. However, real-world inspection data commonly exhibit severe class imbalance, where several defect types are underrepresented, making accurate multiclass classification challenging. This study aims to develop a machine learning–based framework for identifying seven defective wheel conditions using three geometric profile parameters which is flange thickness, flange height, and flange inclination (Qr). Kernel Density Estimation (KDE) with a Gaussian kernel (bandwidth = 0.2) was employed to generate synthetic data samples for minority classes. This augmentation increased the dataset from a skewed distribution to a balanced set of 1827 samples, preserving original statistical properties while providing a more robust foundation for model training. Three supervised classifiers namely, Random Forest, Linear-Support Vector Machine, and K-Nearest Neighbors were implemented and evaluated on the augmented dataset. Model performance was validated using a 5-fold cross-validation technique to ensure accuracy and generalizability. While all models achieved high predictive performance with accuracies exceeding 90%, the Random Forest model yielded the highest validated performance with a mean accuracy of 93.1% and a macro F1-score of 93.5%. Confusion matrix analysis confirmed that most defect types were correctly classified with minimal misclassification. These findings demonstrate that KDE-based augmentation effectively mitigates class imbalance, significantly enhancing the reliability of multiclass wheel defect classification in predictive maintenance systems.

1. Introduction

Railway transportation serves as one of the vital modes for mass transit, ensuring safe and efficient mobility for millions of passengers daily. The reliability and safe functioning of train operations are highly dependent on the condition of the wheelsets, which directly influence ride comfort, maintenance cost, and operational stability. Wheel defects such as flange wear, excessive flange height, and variations in flange inclination (Qr) can lead to uneven load distribution and wheel-rail contact degradation, resulting in service interruptions, premature wear, or in severe cases, derailments [1, 2]. Early and accurate identification of such defects is therefore essential for maintaining optimal wheel health and ensuring passenger safety.

Traditional wheel inspection methods, which rely on manual measurement and scheduled maintenance intervals, are increasingly limited in effectiveness. These approaches are labor-intensive, subject to human error, and incapable of continuous monitoring under real-time operational conditions [3, 4]. The rapid expansion of urban rail networks and high-frequency train services further amplify the demand for data driven solutions that can support automated and predictive defect detection. As railway systems evolve under the framework of Industry 4.0, the integration of artificial intelligence (AI) and data analytics into maintenance strategies offers new opportunities for intelligent, condition-based monitoring [5, 6].

Recent advancements have positioned machine learning (ML) as a powerful and widely adopted approach for defect detection and predictive maintenance in the railway domain. ML algorithms can automatically extract patterns from multi-dimensional measurement data, enabling the classification of wheel conditions with high accuracy and minimal manual intervention [6-8]. Studies have reported the successful application of supervised models namely Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) in identifying wheel defects, axle fatigue, and track irregularities [8-10]. These models have demonstrated promising results in separating defective and non-defective components.

Beyond traditional ML classifiers, a wide range of data-driven and signal-processing approaches has been explored across different engineering fields for fault detection [6, 7, 11]. Time–frequency analysis techniques, including Short-Time Fourier Transform (STFT), Fast Fourier Transform (FFT), and Wavelet Transform, have been widely adopted to extract vibration signatures linked to crack propagation and structural degradation [12-14]. In the domain of image-based inspection, edge detection, morphological filtering, and texture descriptors have proven effective for identifying surface wear and crack defects [15-17].

In recent years, deep learning (DL) architecture has gained strong traction across safety-critical applications [8, 18, 19]. Convolutional Neural Network (CNN) have been applied to wheel profile imaging and ultrasonic testing, while Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) have been successfully used for time-series sensor data in traction motor monitoring and gearbox diagnostics [15, 20, 21]. Hybrid and ensemble frameworks which combine handcrafted features with DL models have been further improved diagnostic robustness. In addition, advanced approaches such as Autoencoders, Generative Adversarial Network (GAN), and Variational Autoencoders (VAE) have also been investigated for anomaly detection, feature extraction, and synthetic fault data generation in domains such as manufacturing, power systems, and medical imaging [3, 22].

In actual railway maintenance practice, wheel degradation occurs in multiple distinct forms. A wheelset may exhibit failure in a wide range of defect conditions, which are commonly evaluated through key geometric parameters including flange height, flange thickness, and flange inclination (Qr). Therefore, a multiclass classification framework may offer a practical way to capture the range of possible defect types. This approach allows for more granular fault identification, enabling maintenance teams to prioritize interventions based on defect severity and type [5, 23].

Another critical challenge is data imbalance. In real-world railway datasets, non-defective samples typically dominate, while certain defect classes occur far less frequently. This imbalance biases model training and reduces performance in minority classes. Hence, generating synthetic data has proven to be feasible approach for mitigating this problem. Among various techniques, Kernel Density Estimation (KDE) offers a statistically grounded method for producing realistic synthetic samples that preserve the distributional properties of the original dataset [24-26]. Recent research has emphasized the transition from manual inspections to automated data-driven maintenance. For instance, Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) have been successfully applied to identify track irregularities and axle fatigue. However, many existing studies are limited to binary classification (defective vs. non-defective) [27-29] , which offers low practical value for specific maintenance planning. Furthermore, real-world railway datasets are frequently plagued by severe class imbalance.

While previous studies have demonstrated the effectiveness of AI techniques in railway defect detection, most implementations remain limited in capturing the full range of wheel degradation conditions encountered in practice [8, 9, 28-30]. In operational settings, defect categories are diverse and unevenly distributed, posing challenges for reliable classification. Although synthetic data generation methods such as GAN and KDE have shown potential in other engineering domains [3, 22, 24-26, 31, 32], their structured integration within multiclass railway wheel geometry classification remains relatively underexplored. Strengthening this integration is essential for developing more robust and practically deployable intelligent maintenance systems.

2. Methodology

The proposed framework aims to construct reliable ML models for the multiclass classification of railway wheel defects based on three key parameters which are flange thickness, flange height, and flange inclination (Qr). The selected parameters were chosen because they directly reflect the wear condition of the wheel and are commonly used in railway maintenance inspections [17, 33]. The methodology was structured to ensure that data

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

preprocessing, model design, and evaluation followed a systematic process. An overview of the process is shown in Fig. 1, which summarizes the research design, starting from data collection and preprocessing up to the final classification results and comparative analysis.

This multi-stage design workflow was designed to emulate a realistic predictive maintenance environment, where railway wheelsets are first screened for defects and subsequently classified according to defect type. In practice, railway wheelsets are not only detected for the presence of defects but also require detailed classification to determine the exact defect type. The methodological structure also ensures that the effect of data imbalance and synthetic data generation is systematically analysed. This ensures that the reliability of the models is evaluated under conditions that resemble real-world operational challenges, particularly when certain defect types are rare in actual datasets.

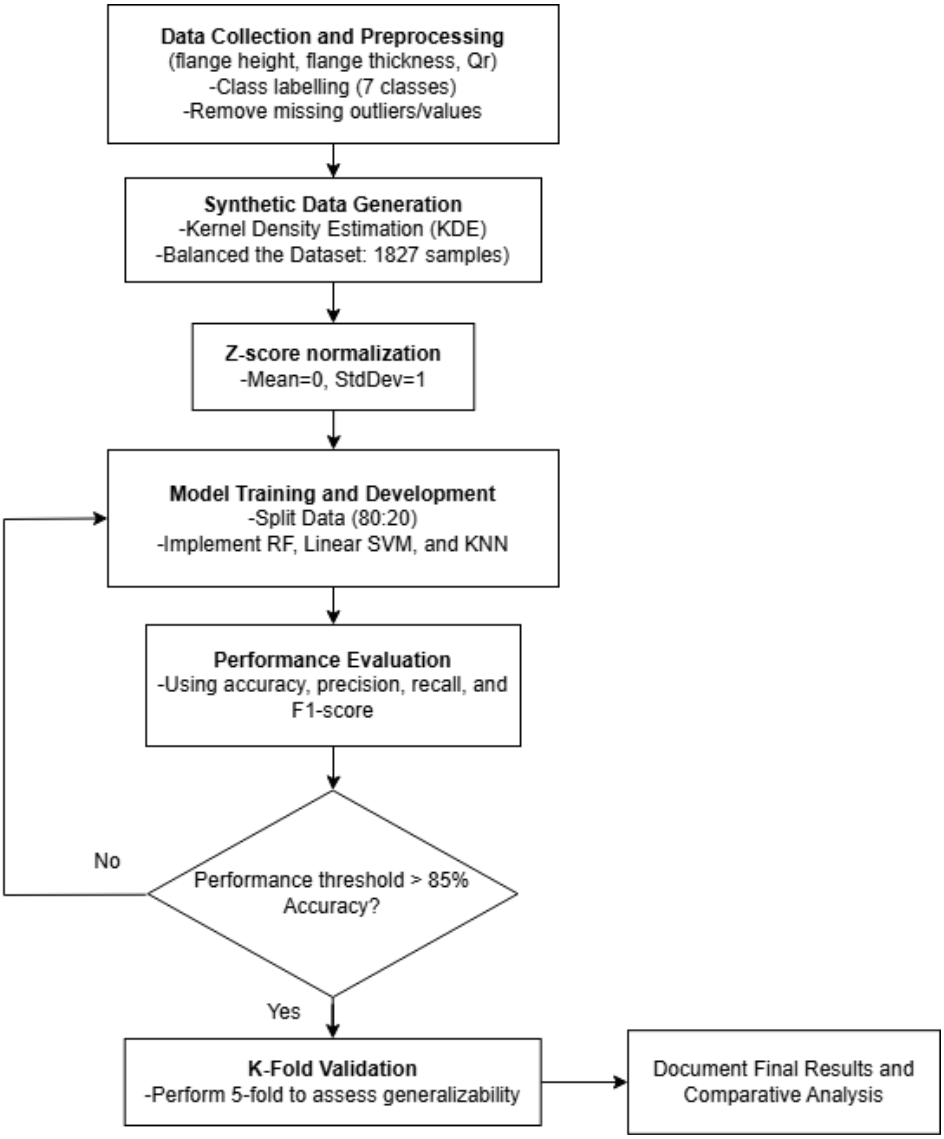

Fig. 1 *Research framework*

2.1 Data Collection and Preprocessing

The dataset utilized for this research was obtained from wheel profile measurements of Mass Rapid Transit (MRT) wheelsets collected during scheduled maintenance activities. Each measurement record includes the three primary wheel profile parameters which are thickness, height, and Qr, which represent the flange geometry and inclination of the wheel tread as shown in Fig. 2. These attributes are widely used in railway maintenance as key indicators of wheel wear, surface deformation, and contact stability with the rail [27, 34]. To capture different operational conditions, the data was categorized into seven defect classes. Each category represents a unique combination of dimensional deviations beyond allowable tolerance limits as illustrated in Table I. The data were originally stored in comma-separated value (CSV) files, with each file corresponding to a specific defect class.

During preprocessing, duplicate entries and missing values were removed. This process guaranteed that each feature had an equal influence on the distance driven models such as KNN and kernel-based method like SVM.

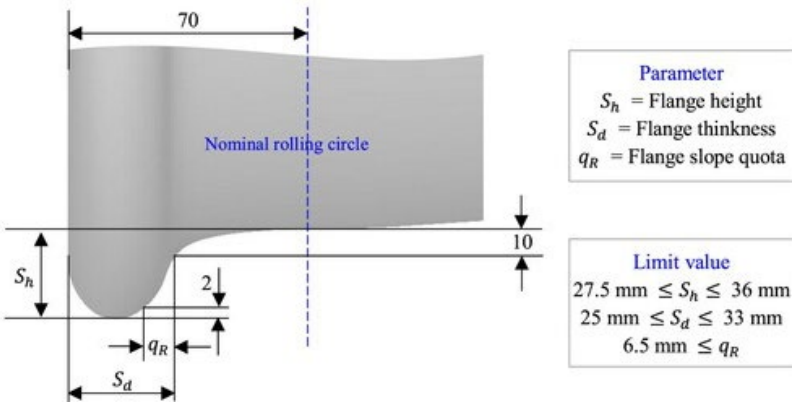

Fig. 2 *Three critical dimensional parameters of wheel profile [33]*

Table 1 *Distribution of defective's wheel data collection*

Defective Types	Group Label	Total Data
All fail	1	53
Thickness fails	2	261
Height fail	3	98
Qr fail	4	210
Thickness and Qr fail	5	176
Thickness and Height fail	6	32
Height and Qr fail	7	20

Defect type definition:

1. All fail: All three parameters fall outside acceptable range, representing wheels in severe wear conditions.
2. Thickness fail: Only flange thickness is out of range, indicating excessive lateral wear typically caused by braking.
3. Height fail: Only flange height exceeds tolerance limits, which can affect rail gauge contact.
4. Qr fail: Only Qr is defective, reflecting incorrect contact geometry between the wheel and rail.
5. Thickness and Qr fail: Both thickness and Qr out from the range values.
6. Thickness and Height fail: Both thickness and height exceed thresholds, typically associated with high flange wear.
7. Height and Qr fail: Both height and Qr parameters are out of range, indicating a wheel nearing its operational safety limit.

The original dataset exhibited noticeable class imbalance, with the largest defect category which is Thickness fail containing 261 samples. Meanwhile the smallest sample is only 20 for Height and Qr fail. Synthetic data were generated for each underrepresented class until defect categories reach and equal amount of 261 samples, producing a balanced dataset of 1827 samples across seven defect types.

2.2 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric technique employed to approximate the probability density function of a data set based solely on observed samples. The technique is widely used in data-driven engineering applications because it does not assume any predefined distribution shape, making it suitable for modelling real-world measurement data [32].

The estimated density function can be applied to analyze various attributes of the random variables. Let $(x_1, x_2, .., x_n)$ be an independent identically distributed sample selected from an unknown probability density distribution $\hat{f}$. The KDE of $\hat{f}$ is defined as shown in Equation 1:

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

$$\hat{f} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \quad (1)$$

For Equation 2, where K is the kernel function, h is the bandwidth parameter, and $K_h(t)$ is the similarity measure between a pair of points.

$$K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right) \quad (2)$$

Bandwidth value is essential in density estimation. The choice of bandwidth heavily influences the smoothness of the estimated distribution and is commonly selected through empirical testing or cross-validation rather than using a single prescribed value [31] .In this study, a Gaussian kernel was used with a bandwidth value of 0.2. The bandwidth value was determined empirically by comparing multiple bandwidths and selecting the one that produced the most realistic density estimates without excessive smoothing or noise to balance between over smoothing and undersmoothing. KDE was applied to each minority defect class to generate synthetic data that mimic the original statistical distributions measurements.

KDE was selected as the augmentation strategy because it is a non-parametric technique that approximates the probability density function without assuming a predefined distribution shape. By generating continuous-valued synthetic samples that preserve the original statistical properties (mean, variance, and spread), KDE mitigates the training bias toward majority classes. This ensures that the classifiers can learn the subtle geometric signatures of rare defects, such as 'Height and Qr fail,' thereby improving overall multiclass accuracy.

The fidelity of the synthetic data was verified through probability density plots, which showed strong overlap between original and synthetic distribution for all three parameters. Thus, it is confirmed that KDE preserved the underlying statistical properties of the real measurements without introducing bias or unrealistic values.

2.3 Model Development

Three machine learning algorithms were implemented to perform the multiclass classification task: RF, Linear SVM, and KNN. RF was selected for it assemble stability and resistance to overfitting. Linear SVM was chosen for its effectiveness in small-to-medium datasets and its ability to find the maximum-margin linear boundary. KNN was included to capture the local geometric structure of the defect types in the feature space. All models were trained using the KDE augmented dataset to ensure equal representation of all defect types and improve generalization performance. Prior to training, z-score normalization was applied to standardize the features to maintain consistency across feature scales. As calculated in Equation 3, this process transforms the data into a mean (μ) of 0 and a standard deviation (σ) of 1:

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

This step ensures that distance-based models like KNN and SVM are not biased by the differing scales of flange height and Qr measurements. The data then were divided into 80% training and 20% testing sets to maintain class proportions using stratified sampling. This ratio was selected as a standard empirical practice to ensure the models have sufficient training data to capture complex defect patterns while maintaining a representative test set for unbiased evaluation. All models were implemented using scikit-learn and trained using the KDE augmented dataset of 1827 samples.

2.3.1 Random Forest

The RF classifier was chosen due to its capability to capture non-linear relationships and handle interactions among high-dimensional features effectively. As an ensemble-based approach, RF generates numerous decision trees throughout the training process and aggregates the prediction through majority voting. Each tree is trained using bootstrap samples from the dataset, with random subsets of features selected at each split thus minimizing overfitting as well as correlation among the trees as illustrated in Fig. 3. This ensemble approach reduces overfitting, stabilizes predictions, and improves the classifier's ability to model complex relationships between wheel defect features.

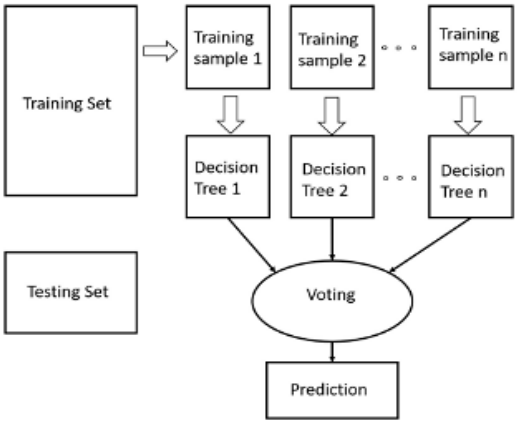

Fig. 3 *The architecture of Random Forest classifier [35]*

RF is widely used in fault detection and industrial monitoring due to its high stability and resistance to overfitting. In this study, 100 decision trees were implemented using the Gini criterion. RF was expected to perform well because wheel defect classification involves interacting variables e.g Height-Thickness correlation that ensemble averaging captures effectively. Moreover, RF's feature importance scores help identify which geometric attributes contribute the most to defect differentiation, making it robust for predictive maintenance.

2.3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm designed to find the best separating boundary that partitions data points into their respective classes. The algorithm is particularly effective for small to medium sized datasets and robust against outliers. The core idea of SVM is to determine a hyperplane that best divides the feature space into regions corresponding to different classes.

Among the many possible separating hyperplanes, SVM chooses the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points for each class. These nearest examples are the support vectors, which determine the position and orientation of the optimal hyperplane.

A hyperplane is a linear decision surface of one dimension less than the feature space, and it functions as a binary classifier by partitioning the space into two regions. By maximizing the margin, SVM improves classification stability and generalization, producing a robust linear boundary suitable for multiclass classification when extended through strategies such as one-vs-rest (OvR).

Fig. 4 illustrates a simple example in a two-dimensional space where the samples are associated with the two classes: stars (Class 1) and circles (Class 2). The objective of SVM is to identify a hyperplane, shown as the red line, that separates these two groups. The samples that lie closest to this boundary are known as support vectors and are responsible for shaping the hyperplane's orientation and position.

The fundamental idea of SVM is to construct a decision boundary $f(x) = 0$ that best separates the classes in a multi-dimensional feature space. The algorithm labels any sample x as belonging to the positive class when $f(x) > 0$ and negative when $f(x) < 0$. The objective is to choose a boundary that maximizes the margin (Gap) while keeping the training error as small as possible.

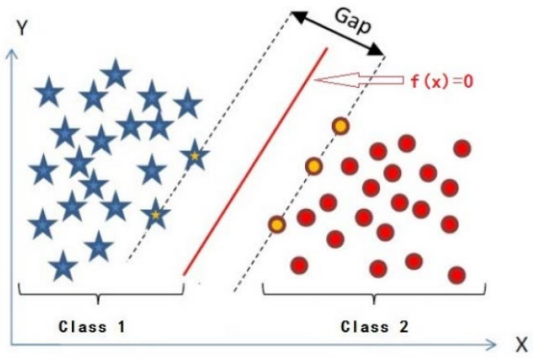

Fig. 4 *Schematic diagram of Support Vector Machine in 2D space [36]*

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

The decision function used by SVM is formulated as shown in Equation 4:

$$f(x) = sign\left(\sum_{x_{i\epsilon s_v}}^{n} a_i y_i k(x_i x) + b\right) \quad (4)$$

where,

- n: The number of samples
- x_i: The positive examples
- y_i: The negative examples
- s_v: Support vector
- a_i: Lagrange multiplier
- $k(x_i x)$: Kernel function
- b: Threshold

2.3.3 K-Nearest Neighbor

KNN is used as a non-parametric, instance-based classifier that directly measures geometric proximity between data points. Unlike model-based approaches, KNN makes no assumptions about data distribution, making it useful for exploratory defect classification where class boundaries may be irregular. The algorithm classifies a new sample by examining the classes of its K nearest neighbors in the training dataset. The value of k is a positive integer selected prior to classification and directly influences model performance.

The similarity between a query point and training samples is typically measured using Euclidean distance, expressed in Equation 5:

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \quad (5)$$

The KNN classification procedure can be summarized as follows:

- Determine the value k.
- Compute the distance between the query points and all samples in the training dataset.
- Identify the k closest samples based on ascending distance.
- Determine the classes of these k neighbors.
- Assigned the query point to the class that occurs most frequently among the neighbors.

In this study, $k = 5$ was selected and distance was computed using the Euclidean metric, implemented as Minkowski distance with $p = 2$. KNN performs effectively when data are well-scaled and balanced, conditions achieved through standardization and KDE augmentation. Although the algorithm is simpler than RF and SVM, KNN provides intuitive insight into the local structure of the dataset. The inclusion of this study for including KNN lies in its local decision-making ability where it reflects the similar structure of defect types in the geometric feature space, complementing the global decision models produced by RF and KNN.

2.4 Model Evaluation

The classification performance of the three algorithms was assessed using four standard metrics: Accuracy (A), Precision (P), Recall (R), and F1-score (F1). In this study, Sensitivity (Recall) and Selectivity (Specificity) are implemented to provide a comprehensive assessment of the classifiers' diagnostic performance. Sensitivity is utilized to measure the model's capability to detect defects by showing the percentage of true positives relative to all actual positives. Selectivity is employed to determine the proportion of actual negatives correctly identified, representing the model's ability to avoid false alarms across the seven defect categories. While these metrics provide insights into class-specific accuracy, the F1-score is also analyzed to offer a robust measure of performance that captures the trade-off between precision and sensitivity, ensuring reliable classification across the balanced dataset. The metrics are defined based on the following components:

- True Positive (TP): The number of defective samples correctly identified as the specific defect class.
- True Negative (TN): The number of samples correctly identified as not belonging to the specific defect class.
- False Positive (FP): The number of samples incorrectly identified as belonging to a defect class when they actually do not (Type I error).
- False Negative (FN): The number of defective samples that the model failed to identify as belonging to the specific defect class (Type II error).

The metrics are expressed in Equation 6 to 10:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = (2 \times \frac{TP + TN}{TP + TN + FP + FN}) \tag{9}$$

$$Selectivity = \frac{TN}{TN + FP} \tag{10}$$

where,

- Accuracy: The proportion of correctly classified instances among all predictions, offering an overall measure of model performance.
- Precision: The percentage of true positives among all positive predictions made by the model, indicating the reliability of defect detection.
- Recall (Sensitivity): The percentage of true positives relative to all actual positives, measuring the model's capability to detect defects.
- F1-score: The harmonic means of precision and recall, offering a single metric that reflects the trade-off between false positive and false negative errors.
- Selectivity (Specificity): The proportion of actual non-defective samples (for a specific class) correctly identified as negative, representing the model's ability to avoid false alarms.

2.5 Model Validation using K-Fold Cross-Validation

A 5-fold cross-validation was implemented to ensure the robustness and generalizability of the classifiers. The balanced dataset, consisting of 1827 samples, was divided into five equal-sized partitions to implement the cross-validation procedure. In this configuration, four subsets served as the training foundation, while the remaining partition was set aside for model validation. This sequence was executed in five iterations, ensuring that every subsample functioned as the validation set exactly once. Finally, the performance metrics from all five cycles were averaged to yield a statistically stable assessment of the model's overall accuracy.

3. Result and Discussion

This section outlines the findings produced by evaluating the three ML classifiers which are RF, SVM, and KNN to both the original and KDE-synthetic datasets. The analysis focuses on model performance in terms of accuracy, precision, recall, and F1-score, with additional observations from confusion matrices results.

The primary goal of this analysis is to assess the performance of different ML algorithms in accurately classifying multiple railway wheel defect types based on wheel profile parameters. The inclusion of KDE-generated synthetic data was introduced as a supporting technique to address class imbalance and assess its contribution to improve overall classification reliability. All experiments were performed under consistent preprocessing, parameter tuning, and validation protocols as described in the next section.

3.1 Z-score Normalization

The first stage of the results analysis involves evaluating the effectiveness of the data standardization process. Prior to normalization, the three geometric profile parameters namely, flange thickness, flange height, and flange inclination (Qr) exhibited significantly different numerical ranges and scales. As shown in the Fig. 5 for "Original Wheel Profile Measurements" box plot, the flange height values were predominantly concentrated between 30 mm and 40 mm, while Qr values were much lower, mostly falling below 10 mm.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

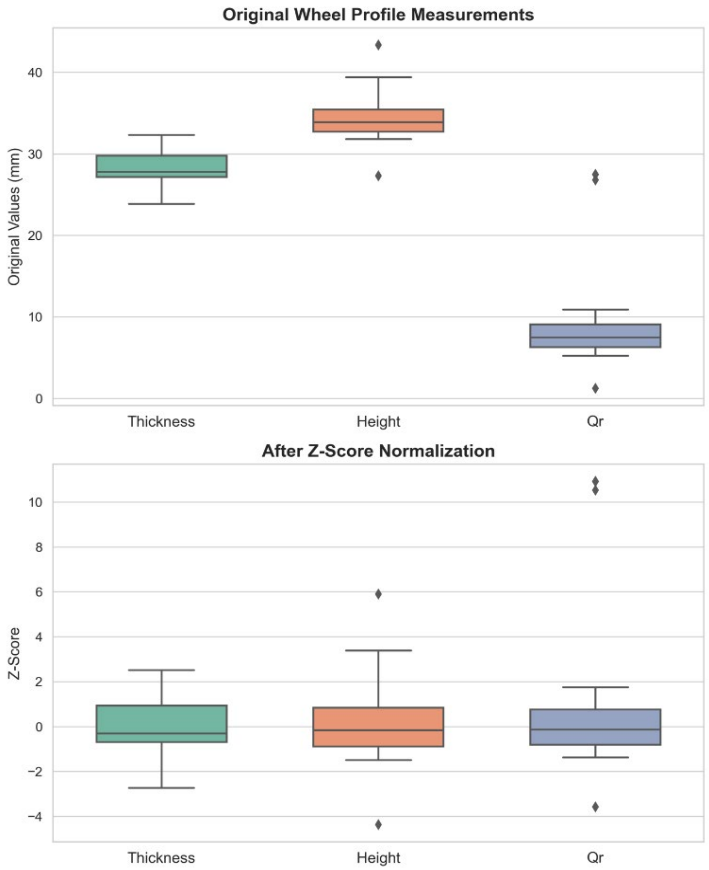

Fig. 5 *Before and after Z-score normalization*

To address these discrepancies, Z-score normalization was applied to ensure that each feature contributed equally to the classification models, particularly for distance-driven algorithms like KNN and Linear SVM. The results of this transformation are illustrated in the Fig.5 "After Z-Score Normalization" box plot:

- Mean Centering: All three parameters were successfully shifted to a common center, with the mean (μ) of each distribution effectively aligned at 0.
- Standardization of Scale: The features were scaled to a standard deviation (σ) of 1, bringing most of the data points for thickness, height, and Qr into a comparable range between -2 and +2.
- Outlier Preservation: The normalization process maintained the relative distribution of the data, including the identification of outliers, which are critical for characterizing severe defect conditions.

3.2 Synthetic Data Generation

As illustrated in Fig. 6 to Fig. 9, the visual comparison between the original and KDE-synthetic datasets is provided for selected defect types. Each figure displays the probability density distribution of the three key wheel-profile parameters which are flange thickness, flange height, and flange inclination (Qr) for both datasets. In addition to the visual comparison, Table 2 summarizes the number of data samples before and after augmentation for each defect type, showing how KDE was used to increase the minority classes to achieve a balanced dataset.

Table 2 *Data distribution before and after KDE augmentation*

Defective Types	Group Label	Original Data	Augmented Data
All fail	1	53	261
Thickness fails	2	261	261
Height fail	3	98	261
Qr fail	4	210	261
Thickness and Qr fail	5	176	261
Thickness and Height fail	6	32	261
Height and Qr fail	7	20	261

Fig. 6(a) presents the probability density distributions for the original and KDE-generated synthetic data of Defect Type 1, showing that the synthetic samples closely match the shape and spread of the original distributions for flange thickness, flange height, and flange inclination (Qr). This indicates that KDE successfully captured the statistical characteristics of this defect type. In contrast, Fig. 6(b) displays only the original distributions for Defect Type 2, as this class already contained sufficient samples and did not require augmentation. The distributions in Fig. 6(b) provide the baseline reference for this defect category and show clear, consistent patterns across all three wheel-profile parameters.

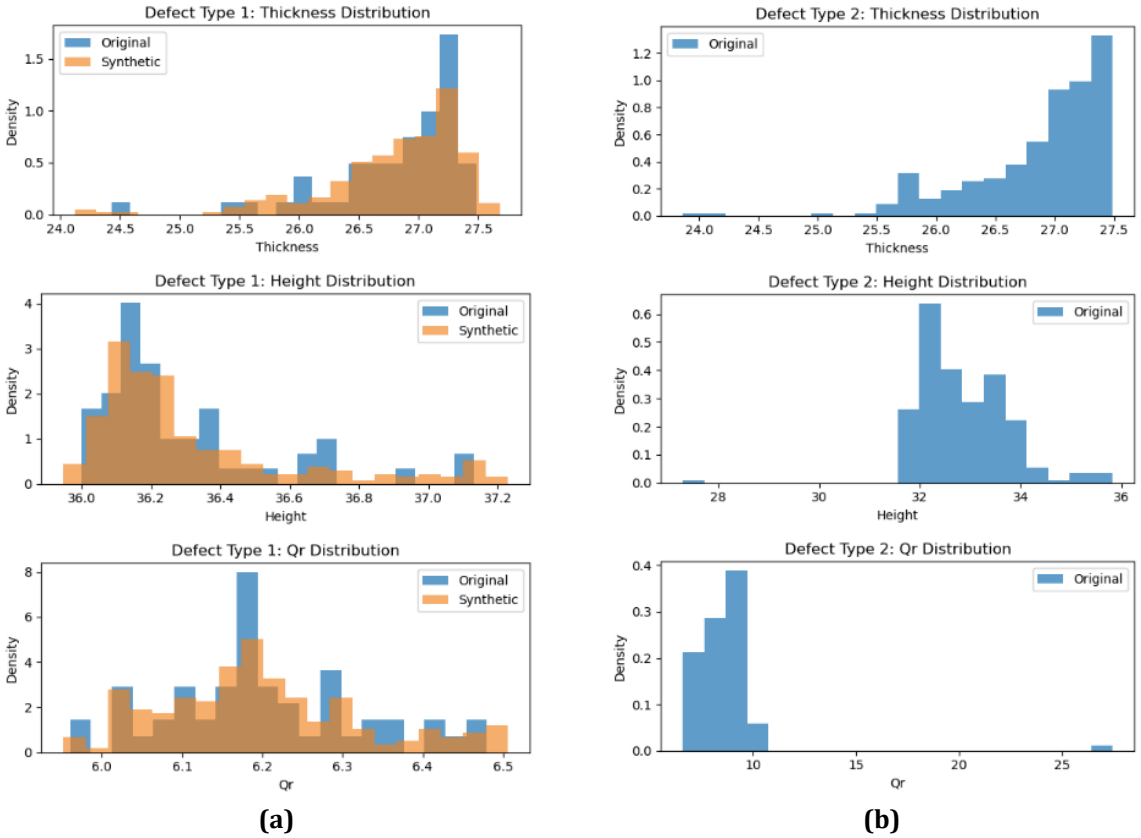

Fig. 6 *Data distribution between original and synthetic dataset (a) Defect type 1; (b) Defect type 2*

Fig. 7(a) shows the probability density distributions of the original data and KDE-generated synthetic data for Defect Type 3. The distributions for flange thickness, flange height, and flange inclination (Qr) demonstrate strong overlap between the original and synthetic samples, indicating that KDE was able to closely reproduce the overall shape and variability of this defect category. Meanwhile, Fig. 7(b) presents the original and synthetic distributions for Defect Type 4. Like Defect Type 3, the synthetic samples follow the same general distribution patterns for all three wheel-profile parameters, although slight deviations appear at the distribution tails due to the limited number of original samples. Overall, both figures confirm that KDE successfully captured the statistical behavior of Defect Type 3 and Defect Type 4, allowing these classes to be effectively balanced for model training.

Fig. 8(a) presents the probability density distributions of the original and KDE-generated synthetic data for Defect Type 6. The distributions for flange thickness, flange height, and flange inclination (Qr) show strong overlap, indicating that KDE successfully captured the main trends and variability of this defect class. The synthetic samples follow the general shape of the original distributions, although small differences appear at the tails due to the limited number of original samples. Similarly, Fig. 8(b) displays the original and synthetic distributions for Defect Type 5, where the KDE-generated data also replicates the overall structure of the original measurements across all three parameters. This consistency confirms that the KDE augmentation process produced realistic and representative samples for both Defect Type 5 and Defect Type 6, supporting effective dataset balancing for subsequent model training.

Fig. 9 illustrates the probability density distributions of the original and KDE-generated synthetic data for Defect Type 7. Across all three wheel-profile parameters, the synthetic samples closely follow the overall pattern of the original data. The main peaks and distribution shapes are well preserved, although slight differences can be seen in some regions because of the small sample size associated with this defect category. Despite these minor variations, the KDE-generated data remain consistent with the statistical behavior of Defect Type 7, confirming

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

that the augmentation process successfully produced realistic samples to support the dataset balancing required for model training.

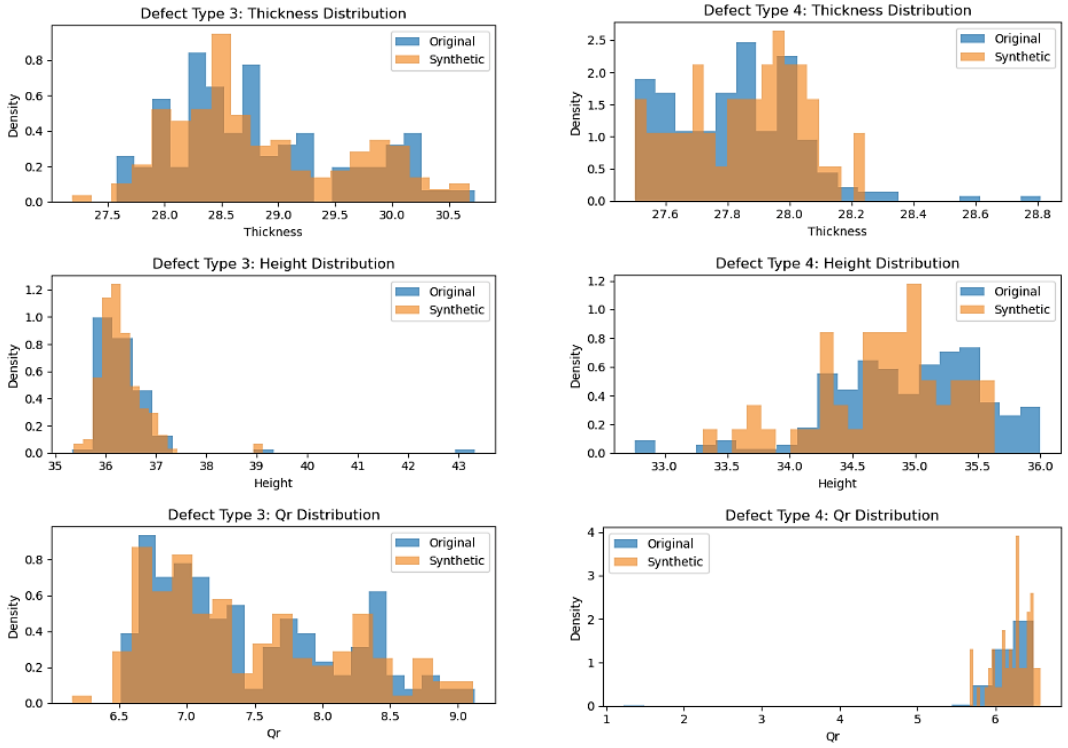

Fig. 7 *Data distribution between original and synthetic dataset (a) Defect type 3; (b) Defect type 4*

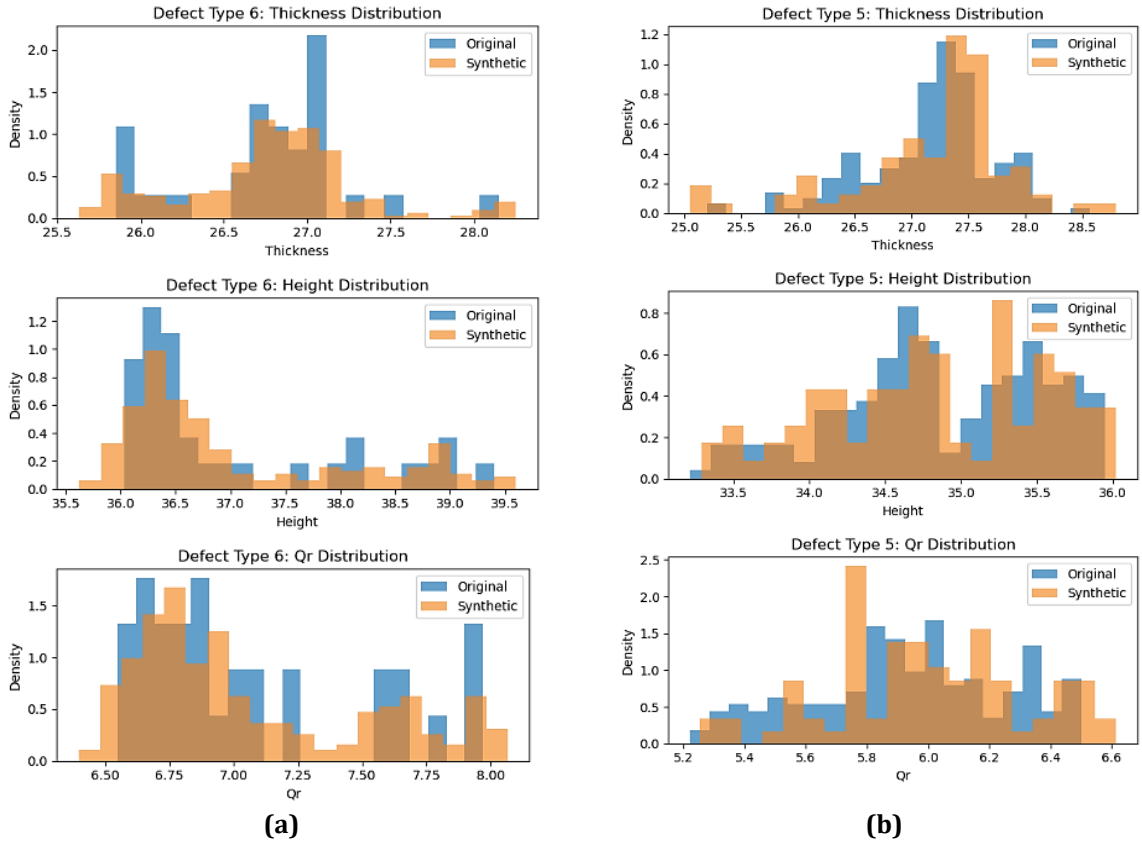

Fig. 8 *Data distribution between original and synthetic dataset (a) Defect type 6; (b) Defect type 7*

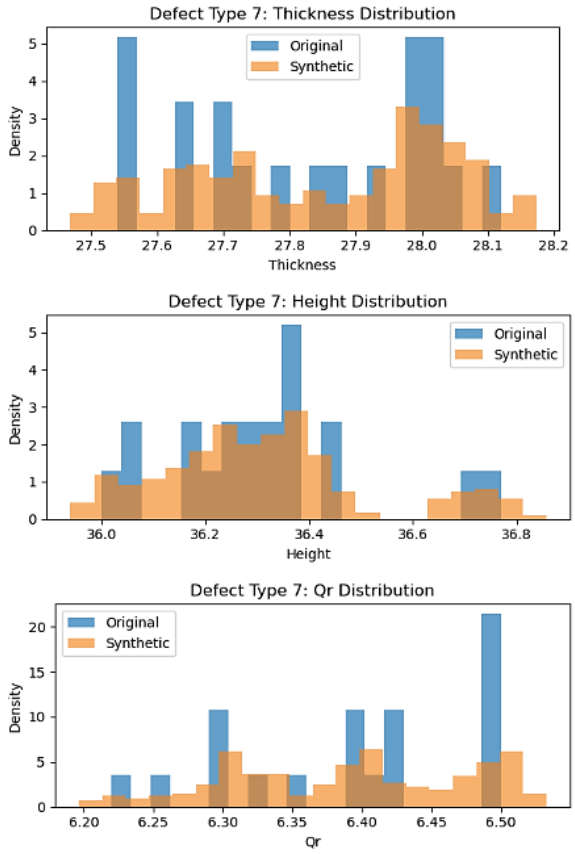

Fig. 9 *Data distribution between original and synthetic dataset for Defect type 8*

Overall, the KDE-based augmentation approach demonstrated strong consistency across all seven defect types. For each class, the synthetic samples closely followed the key patterns of the original measurements, including the general shape, central tendency, and spread of the distributions for flange thickness, flange height, and flange inclination (Qr). While minor deviations appeared in classes with very limited original samples, these differences remained within an acceptable range and did not distort the underlying distribution characteristics. The strong overlap between original and synthetic density curves confirms that KDE successfully generated realistic, continuous-valued samples suitable for balancing the dataset. In summary, KDE-based data synthesis substantially improved the dataset balance while maintaining fidelity to the original feature distributions. This enhancement supports the construction of the models that are more generalized and unbiased for the defect detection of the railway system.

3.3 Model Performance

Three supervised classifiers were trained on the balanced dataset and evaluated for their ability to differentiate among the seven defect types. Each model produced high overall accuracy values exceeding 90%, confirming the effectiveness of KDE augmentation in enabling fair and robust classification. The following subsections present the results and discussion for RF, SVM, and KNN classifiers respectively. This breakdown provides insight into how each algorithm handles overlapping defect distributions and class-specific variations in the geometric parameters.

3.3.1 Random Forest

The confusion matrix in Fig. 10 shows that the RF classifier performed well across most defect categories. Since the dataset was balanced before training, the results reflect the actual separability of the defect patterns rather than differences in sample size. Group 2 (Thickness Fail) and Group 7 (Height and Qr Fail) were classified perfectly, achieving 100% accuracy, indicating that these defect types have distinctive geometric signatures that the model can easily separate. Group 1 (All Fail) and Group 3 (Height Fail) also showed strong performance with accuracies of 94.2% and 92.4%, respectively.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

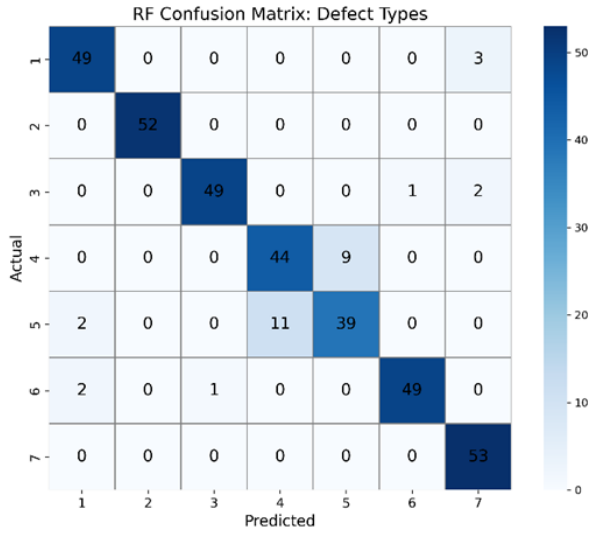

Fig. 10 *Confusion matrix for Random Forest model*

The most challenging categories were Group 4 (Qr Fail) and Group 5 (Thickness and Qr Fail), which obtained accuracies of 83.0% and 75.0%. Most of the misclassifications occurred between these two groups, where several Group 4 samples were predicted as Group 5 and vice versa. This pattern is expected because both defect types share similar Qr characteristics, and their KDE-augmented distributions have considerable overlap, making them harder to distinguish. A few minor confusions also appeared between Group 3 and Group 7 due to similarities in height values.

Overall, the RF model yielded an accuracy of 91.5%, and the low misclassification rates across most groups demonstrate that RF can effectively learn to distinguish patterns of wheel-profile parameters. These results confirm that Random Forest is robust for multiclass railway wheel defect classification, especially when classes are balanced and feature variations are well represented.

3.3.2 Support Vector Machine

The confusion matrix in Fig. 11 shows that the Linear SVM classifier achieved strong performance across most defect categories. Since the dataset was balanced prior to training, the results reflect the intrinsic separability of the defect patterns rather than class-size effects. Group 2 (Thickness Fail) and Group 7 (Height and Qr Fail) achieved 100% accuracy, indicating that these defect types align well with the linear decision boundary. Group 1 (All Fail) and Group 6 (Thickness and Height Fail) also performed well, each achieving 94.2% accuracy, while Group 3 (Height Fail) reached 90.4%.

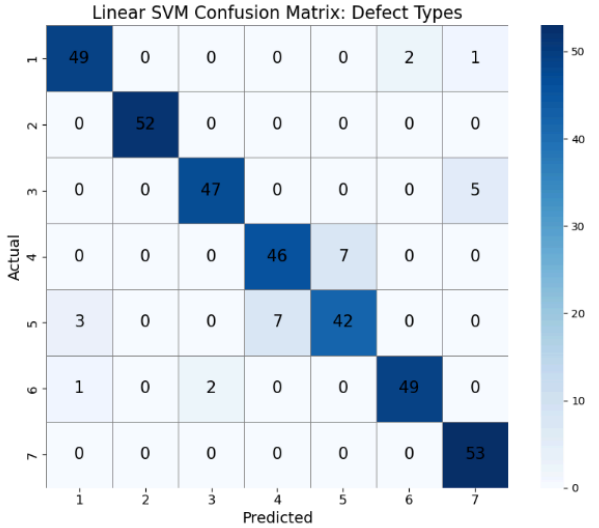

Fig. 11 *Confusion matrix for Linear-Support Vector Machine model*

The most challenging defect types for SVM were the Qr-related categories. Group 4 (Qr Fail) obtained 86.8% accuracy, while Group 5 (Thickness and Qr Fail) was lower at 80.8%. Misclassifications for these two groups primarily occurred between each other, consistent with their overlapping Qr characteristics and similar geometric distributions. Additional minor confusion occurred in Group 5, with several samples misclassified into Groups 1 and 4. Overall, Linear SVM yielded an accuracy of 92.4%, slightly outperforming the RF model. These results demonstrate that SVM is suitable for multiclass wheel defect classification, although it is more sensitive to classes with overlapping feature distributions, particularly those involving Qr-related defects.

3.3.3 K-Nearest Neighbors

The confusion matrix in Fig. 12 shows that the KNN classifier also achieved strong performance on the balanced dataset. Group 2 (Thickness Fail) and Group 7 (Height and Qr Fail) were classified with 100% accuracy, indicating that these defect types form well-separated clusters in the feature space. Group 1 (All Fail) and Group 6 (Thickness and Height Fail) also performed very well, with accuracies of 98.1% and 94.2%, respectively, and only a few samples misclassified into neighboring groups.

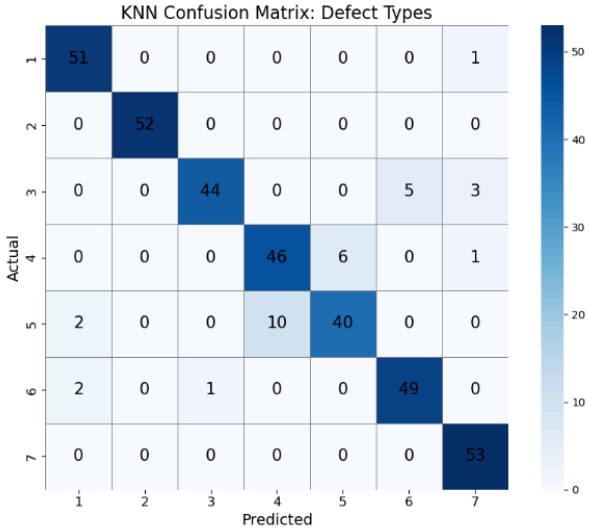

Fig. 12 *Confusion matrix for K-Nearest Neighbours model*

The more challenging defect types for KNN were Group 3 (Height Fail), Group 4 (Qr Fail), and especially Group 5 (Thickness and Qr Fail). Group 3 reached an accuracy of 84.6%, with misclassifications mainly into Group 6 and Group 7, which share similar characteristics. Group 4 (Qr Fail) achieved 86.8% accuracy, with errors primarily assigned to Group 5 and Group 7, reflecting overlap in Qr patterns. Group 5 (Thickness and Qr Fail) had the lowest accuracy at 76.9%, with several samples predicted as Group 1 (All Fail) and Group 4 (Qr Fail). This indicates that KNN is more sensitive to local overlaps between classes that differ only by one parameter, particularly when their neighborhoods in the feature space are close. Nonetheless, the KNN model achieved an accuracy of 91.5%, compared to Random Forest. The strong performance for several classes confirms that the KDE-balanced dataset preserves meaningful local structure, while the remaining misclassifications highlight the limitations of distance-based methods in regions where defect types have highly overlapped geometric features.

3.4 Comparative Analysis

All three classifiers achieved strong and consistent performance, with accuracy exceeding 91%. The model performance accuracy was primarily influenced by four key factors: (i) class distribution balance, (ii) feature scaling, (iii) intrinsic geometric separability among defect categories, and (iv) algorithm-specific learning behavior. The balanced dataset generated through KDE augmentation, combined with Z-score normalization, contributed significantly to this stability by ensuring equitable class representation and proportional feature influence in the decision space.

The confusion matrices (Figs. 10, 11 and 12) reveal that certain defect categories, such as Thickness Fail and Height and Qr Fail, are geometrically well separated, resulting in perfect classification. In contrast, slightly lower accuracy in Qr-related categories reflects geometric overlap between Qr Fail and Thickness and Qr Fail, indicating intrinsic feature similarity rather than model weakness.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 199-216

As summarized in Table 3, Linear SVM achieved the highest overall accuracy (92.4%), with macro-averaged precision, recall, and F1-score all exceeding 92%. This suggests that the normalized feature space exhibits near-linear separability, allowing the maximum-margin boundary to effectively distinguish among the seven defect classes. Random Forest and KNN both achieved 91.5% accuracy, demonstrating competitive and consistent performance. The small performance differences among models further indicate that class imbalance was effectively mitigated, enabling fair evaluation of each algorithm's intrinsic learning behavior. High Selectivity values (above 98%) across all models confirm strong discrimination capability, with minimal misclassification between defect categories. In particular, the Linear SVM achieved 98.7% specificity, reducing the likelihood of incorrect defect labeling, which is an important consideration for minimizing unnecessary maintenance actions.

Table 3 *Performance comparison of the models*

Model	Recall (Sensitivity)	Selectivity (Specificity)	False Positive Rate	Precision	F1-score
RF	0.915	0.986	0.014	0.916	0.915
Linear SVM	0.924	0.987	0.013	0.924	0.923
KNN	0.915	0.985	0.015	0.917	0.914

A closer examination of the model behavior provides insight into why the Linear SVM achieved the highest overall performance. SVM optimizes a maximum-margin hyperplane, which makes it highly effective when class boundaries are approximately linear, as observed in the KDE feature space. Through maximizing the separation between classes, SVM reduces overlap and minimizes misclassification in regions where defect groups exhibit subtle geometric differences. In contrast, Random Forest relies on multiple decision trees that partition the feature space into discrete regions. While this ensemble strategy offers robustness, it can struggle with fine linear boundaries and may over-segment smooth distributions. KNN, being a distance-based classifier, is more sensitive to local overlaps between classes, particularly when defect types differ by only one parameter. Although KNN and RF still performed well, the Linear SVM benefited most from the augmented and well-structured dataset, enabling it to form clearer class boundaries and achieve the highest accuracy among the three models.

3.5 Cross-Validation (5-Folds)

A 5-fold cross-validation was conducted to ensure the robustness and generalizability of the classifiers beyond a single train-test split. The results were summarized in Table 4, providing a statistically stable estimation of model performance across the entire KDE-augmented dataset of 1827 samples.

Table 4 *Final 5-fold cross-validation results*

Model	Accuracy (Mean)	Accuracy (Std)	Macro F1 (Mean)	Macro F1 (Std)
RF	**0.931**	**0.014**	**0.935**	**0.021**
Linear SVM	0.905	0.018	0.883	0.015
KNN	0.905	0.016	0.885	0.014

As shown in Table 4, the RF model outperformed the other classifiers during cross-validation, achieving the highest mean accuracy of 93.1% and a Macro F1-score of 93.5%. This shift in performance indicates that while Linear SVM was highly effective on the initial split, the ensemble structure of RF provided superior stability and generalization across different data partitions. The low standard deviation across all models (ranging from 0.0136 to 0.0176) confirms that the KDE-based augmentation successfully produced a statistically consistent dataset, as the performance remained stable regardless of which fold was used for validation. This suggests that RF's ensemble structure is more robust to the varied geometric signatures produced during the KDE augmentation process.

4. Conclusion

This study presents a robust machine learning–based framework for the multiclass classification of seven railway wheel defect conditions. A key contribution lies in the application of Kernel Density Estimation (KDE) to effectively address severe class imbalance, enabling statistically realistic augmentation of minority defect categories and facilitating unbiased model training.

Feature normalization further enhanced classification stability by ensuring proportional contribution of geometric parameters within the decision space. The evaluation results demonstrate strong predictive performance across all classifiers, with Random Forest achieving the highest generalization capability during 5-

fold cross-validation, reaching a mean accuracy of 93.1% and a Macro F1-score of 93.5%. The low performance variance across folds confirms the robustness and consistency of the proposed framework.

Overall, the findings underscore the effectiveness of combining statistically grounded data balancing with structured multiclass learning for practical railway wheel defect identification. The framework provides a scalable foundation for intelligent, condition-based maintenance systems. Future research will expand the feature set to incorporate additional geometric parameters and multi-source real-world datasets, while exploring advanced deep learning architectures to further enhance representation capability and real-time deployment potential.

Acknowledgement

The authors would like to thank Universiti Teknologi Malaysia for the sponsorship of this study under the UTMER research grant no.: Q.K130000.3856.31J95, as well as to all collaborating partners.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Nurfitri Muhamad Nasir, Siti Armiza Mohd Aris;* ***data collection:*** *Tay Yung Taw;* ***analysis and interpretation of results:*** *Nurfitri Muhamad Nasir, Siti Armiza Mohd Aris;* ***draft manuscript preparation:*** *Nurfitri Muhamad Nasir, Siti Armiza Mohd Aris. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Y. Zhao *et al.*, "A Review on Rail Defect Detection Systems Based on Wireless Sensors," *Sensors*, vol. 22, no. 17, 2022 doi: 10.3390/s22176409.

[2] J. Lai, J. Xu, T. Liao, Z. Zheng, R. Chen, and P. Wang, "Investigation on train dynamic derailment in railway turnouts caused by track failure," *Engineering Failure Analysis,* vol. 134, p. 106050, 2022/04/01/ 2022, doi: https://doi.org/10.1016/j.engfailanal.2022.106050.

[3] A. Shangguan, G. Xie, R. Fei, L. Mu, and X. Hei, "Train wheel degradation generation and prediction based on the time series generation adversarial network," *Reliability Engineering & System Safety,* vol. 229, p. 108816, 2023/01/01/ 2023, doi: https://doi.org/10.1016/j.ress.2022.108816.

[4] Y. U. P. Yusuf, "A Review of Wheel Wear Damage In Railway Vehicle," *Jurnal Perkeretaapian Indonesia (Indonesian Railway Journal),* vol. 8, no. 1, pp. 42-52, 2024.

[5] M. Mohammadi, A. Mosleh, C. Vale, D. Ribeiro, P. Montenegro, and A. Meixedo, "An Unsupervised Learning Approach for Wayside Train Wheel Flat Detection," (in English), *SENSORS,* vol. 23, no. 4, FEB 2023, Art no. 1910, doi: 10.3390/s23041910.

[6] N. Davari, B. Veloso, G. D. Costa, P. M. Pereira, R. P. Ribeiro, and J. Gama, "A Survey on Data-Driven Predictive Maintenance for the Railway Industry," *Sensors*, vol. 21, no. 17, 2021 doi: 10.3390/s21175739.

[7] J. A. P. Braga and A. R. Andrade, "Data-driven decision support system for degrading assets and its application under the perspective of a railway component," *Transportation Engineering,* vol. 12, p. 100180, 2023/06/01/ 2023, doi: https://doi.org/10.1016/j.treng.2023.100180.

[8] K. Shaikh, I. Hussain, and B. S. Chowdhry, "Wheel Defect Detection Using a Hybrid Deep Learning Approach," *Sensors*, vol. 23, no. 14, 2023 doi: 10.3390/s23146248.

[9] J.-H. Lee *et al.*, "A Study on Wheel Member Condition Recognition Using Machine Learning (Support Vector Machine)," *Sensors*, vol. 23, no. 20, 2023 doi: 10.3390/s23208455.

[10] Y. Santur, M. Karaköse, and E. Akin, "Random forest based diagnosis approach for rail fault inspection in railways," in *2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO)*, 1-3 Dec. 2016 2016, pp. 745-750.

[11] Y. Zeng, D. Song, W. Zhang, B. Zhou, M. Xie, and X. Tang, "A new physics-based data-driven guideline for wear modelling and prediction of train wheels," *Wear,* vol. 456-457, p. 203355, 2020/09/15/ 2020, doi: https://doi.org/10.1016/j.wear.2020.203355.

[12] P. D. A. Aziz, M. F. Azizol, N. H. Jabarullah, and S. Z. M. Noor, "Defective wheel detection via fast fourier transform for rolling stock," in *AIP Conference Proceedings*, 2025, vol. 3349, no. 1: AIP Publishing LLC, p. 040006.

[13] C. Ghosh, A. Verma, and P. Verma, "Real time fault detection in railway tracks using Fast Fourier Transformation and Discrete Wavelet Transformation," *International Journal of Information Technology,* vol. 14, no. 1, pp. 31-40, 2022.

[14] A. Mosleh, A. Meixedo, D. Ribeiro, P. Montenegro, and R. Calçada, "Early wheel flat detection: an automatic data-driven wavelet-based approach for railways," *Vehicle System Dynamics,* vol. 61, no. 6, pp. 1644-1673, 2023.

[15] S. A. Singh and K. A. Desai, "Automated surface defect detection framework using machine vision and convolutional neural networks," *Journal of Intelligent Manufacturing,* vol. 34, no. 4, pp. 1995-2011, 2023/04/01 2023, doi: 10.1007/s10845-021-01878-w.

[16] W. Li, H. Hacid, E. Almazrouei, and M. Debbah, "A comprehensive review and a taxonomy of edge machine learning: Requirements, paradigms, and techniques," *AI,* vol. 4, no. 3, pp. 729-786, 2023.

[17] H. Soleimani, M. Moavenian, R. Masoudi Nejad, and Z. Liu, "An applied method for railway wheel profile measurements due to wear using image processing techniques," *SN Applied Sciences,* vol. 3, no. 2, p. 147, 2021/01/18 2021, doi: 10.1007/s42452-021-04180-9.

[18] J. Sresakoolchai and S. Kaewunruen, "Wheel flat detection and severity classification using deep learning techniques," *Insight-Non-Destructive Testing and Condition Monitoring,* vol. 63, no. 7, pp. 393-402, 2021.

[19] K. Oh *et al.*, "A Review of Deep Learning Applications for Railway Safety," *Applied Sciences,* vol. 12, no. 20, p. 10572, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/20/10572.

[20] Q. Wang, S. Bu, and Z. He, "Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN," *IEEE Transactions on Industrial Informatics,* vol. 16, no. 10, pp. 6509-6517, 2020.

[21] D. Ye *et al.*, "Prediction of Key Parameters of Wheelset Based on LSTM Neural Network," *Applied Sciences*, vol. 13, no. 21*,* 2023 doi: 10.3390/app132111935.

[22] V. C. Pezoulas *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and Structural Biotechnology Journal,* vol. 23, pp. 2892-2910, 2024/12/01/ 2024, doi: https://doi.org/10.1016/j.csbj.2024.07.005.

[23] J. Sresakoolchai and S. Kaewunruen, "Detection and Severity Evaluation of Combined Rail Defects Using Deep Learning," *Vibration,* vol. 4, no. 2, pp. 341-356, 2021, doi: 10.3390/vibration4020022.

[24] J. Heine, E. E. Fowler, A. Berglund, M. J. Schell, and S. Eschrich, "Techniques to produce and evaluate realistic multivariate synthetic data," *Scientific Reports,* vol. 13, no. 1, p. 12266, 2023.

[25] M. Abufadda and K. Mansour, "A Survey of Synthetic Data Generation for Machine Learning," in *2021 22nd International Arab Conference on Information Technology (ACIT)*, 21-23 Dec. 2021 2021, pp. 1-7, doi: 10.1109/ACIT53391.2021.9677302.

[26] E. Plesovskaya and S. Ivanov, "An Empirical Analysis of KDE-based Generative Models on Small Datasets," *Procedia Computer Science,* vol. 193, pp. 442-452, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.procs.2021.10.046.

[27] J. Sresakoolchai and S. Kaewunruen, "Railway defect detection based on track geometry using supervised and unsupervised machine learning," *Structural Health Monitoring,* vol. 21, no. 4, pp. 1757-1767, 2022, doi: 10.1177/14759217211044492.

[28] I. Soleimanmeigouni, A. Ahmadi, A. Nissen, and X. Xiao, "Prediction of railway track geometry defects: a case study," *Structure and Infrastructure Engineering,* vol. 16, no. 7, pp. 987-1001, 2020.

[29] L. Zhuang, H. Qi, and Z. Zhang, "The automatic rail surface multi-flaw identification based on a deep learning powered framework," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 8, pp. 12133-12143, 2021.

[30] A. Mosleh, A. Meixedo, D. Ribeiro, P. Montenegro, and R. Calçada, "Machine learning approach for wheel flat detection of railway train wheels," *Transportation Research Procedia,* vol. 72, pp. 4199-4206, 2023/01/01/ 2023, doi: https://doi.org/10.1016/j.trpro.2023.11.354.

[31] F. Kamalov, S. Moussa, and J. Avante Reyes, "KDE-Based Ensemble Learning for Imbalanced Data," *Electronics,* vol. 11, no. 17, p. 2703, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/17/2703.

[32] F. Kamalov, "Kernel density estimation based sampling for imbalanced class distribution," *Information Sciences,* vol. 512, pp. 1192-1201, 2020.

[33] Y. Qi, H. Dai, P. Wu, F. Gan, and Y. Ye, "RSFT-RBF-PSO: a railway wheel profile optimisation procedure and its application to a metro vehicle," *Vehicle System Dynamics,* vol. 60, no. 10, pp. 3398-3418, 2022.

[34] R. Ghiasi, M. A. Khan, D. Sorrentino, C. Diaine, and A. Malekjafarian, "An unsupervised anomaly detection framework for onboard monitoring of railway track geometrical defects using one-class support vector machine," *Engineering Applications of Artificial Intelligence,* vol. 133, p. 108167, 2024.

[35] L. Tuan, M.-T. Vo, T. Pham, and S. Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic," *IEEE Access,* vol. 9, pp. 7869-7884, 12/22 2020, doi: 10.1109/ACCESS.2020.3047942.

[36] X. Yao, "Application of optimized SVM in sample classification," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 6, 2022.