electronics

MDPI

Article

Pantograph Wear Classification via Dual-Backbone Feature-Fusion Ensemble Network

Naeem Ullah [1], Yasir Iqbal [2], Shamim Ibne Shahid [2], Muhammad Yaqoob [2], Javed Ali Khan [2,*] and Alexios Mylonas [2]

[1] Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80125 Naples, Italy; naeem.ullah@unina.it
[2] School of Physics, Engineering, and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK; y.iqbal2@herts.ac.uk (Y.I.); titu2297@yahoo.com (S.I.S.); m.yaqoob3@herts.ac.uk (M.Y.); a.mylonas@herts.ac.uk (A.M.)
* Correspondence: j.a.khan@herts.ac.uk

Abstract

Vision-based pantograph wear recognition plays a critical role in the safety and reliability of railway power supply systems. Although recent studies report promising deep learning-based results, these models solely depend on the integrity of the dataset. Data integrity is a critical yet often overlooked factor in research and production, and neglecting it may lead to inconsistencies and compromised operational safety. In the proposed approach, we demonstrate that a widely used pantograph wear dataset contains severe redundancy and label inconsistencies, including duplicate images appearing within classes and across different wear categories. These issues undermine supervised learning, reduce model generalisation, compromise predictive reliability, and may weaken the safety of rail infrastructure systems. This work (i) preprocesses the dataset by employing MD5-based cryptographic hashing and manual verification, where 626 redundant samples were identified from a dataset of 909 images; subsequently, a manual relabelling procedure is used to correct inherited annotation errors and consistent class definitions. (ii) It devises a Dual-Backbone Feature Fusion Ensemble Network (DBFF-Net) for small and challenging datasets by integrating frozen ShuffleNetV2 and DeiT-tiny as the best individual performing classifiers using various fusion strategies, including concat, weighted sum, Bilinear, Cross-Attention, and Gated. Amongst the different fusion approaches, we obtain the best results with the Gated approach. We reproduced the comparatively improved pantograph wear classification results and conducted extensive experiments to demonstrate that dataset sanitization improves the stability and reproducibility of the model. Moreover, it has been shown that DBFF-Net outperforms individually employed pretrained CNNs and transformer models and achieves an accuracy of 96.46% even with limited but sanitised data.

Keywords: vision-based inspection; pantograph wear detection; dataset sanitisation; dual-backbone feature fusion; ensemble deep learning

Academic Editor: Kefeng Ji

Received: 1 April 2026
Revised: 25 April 2026
Accepted: 30 April 2026
Published: 6 May 2026

1. Introduction

Pantograph catenary systems are among the most critical components of railway electrical infrastructure, which provide the electrical power supply in railway high-speed train operation [1–3]. With the increasing need for high-speed railway systems, pantographs may directly affect railway reliability and safety during operation in different environmental conditions [2]. The pantograph structures are exposed to constant stress, friction,

aerodynamic forces, and environmental interference during operation, thus increasing the tendency towards deformation, cracks, wear, and misalignment of the structure [1,3,4]. Pantograph wear may affect system performance in supplying electrical power directly to railway trains, which increases the likelihood of electrical short circuits or even fires during operation [1,3]. Conventional inspection methods, which rely on manual visual assessment, are inefficient, laborious, and error-prone, although monitoring methods such as sensors, including fibre-optic sensors or ultrasonic sensors, face their own problems in the form of complex installation, environmental sensitivity, and interference in the structure of the pantograph [1,5]. All of this emphasises the need for vision-based intelligent systems of diagnostics. Consequently, vision-based solutions have emerged as a requirement in automatically recognising different pantograph wear states [4].

Recently, computer vision and deep learning (DL) models have proven to be strong tools for pantograph wear inspection. Convolutional neural networks (CNNs), dual-backbone models, and other computer vision-based approaches have outperformed traditional knowledge-based approaches in recognising pantograph wear characteristics in images [6,7]. In the literature, there has been an increasing effort to improve the accuracy of pantograph wear detection. In contrast, limited attention has been given to validating the data integrity and quality to generate more reliable and generalised results. Also, existing approaches to pantograph wear detection implicitly assume that publicly available datasets are reliable and suitable for supervised learning. However, these machine learning-based approaches primarily depend on the quality and integrity of the training data. In contrast, in industrial settings, the dataset size is comparatively limited and may be prone to redundancy, duplication, human annotation errors, and distributional bias, which can adversely affect machine learning model training and evaluation.

Furthermore, pantograph wear images exhibit complex, multi-scale visual characteristics, making it challenging for existing single-model supervised learning architectures to capture them efficiently. Therefore, in real industrial settings, relying on a single supervised learning algorithm might result in suboptimal generalisation and limited robustness required for a critical system. Although the existing CNN-based pantograph wear approaches are good performance at extracting local texture and edge information, they struggle to capture deeper dependencies and global contextual relationships in the input. Unlike DL-based approaches, dual-backbone approaches provide improved global representation capabilities; however, they require large amounts of training data and may be prone to overfitting with limited data. These potential limitations in the existing approaches motivated us to develop a unified holistic approach that overcomes existing limitations and model representational constraints for the pantograph wear detection and classification problem.

Moreover, CNN-transformer hybrid architectures have been investigated in general computer vision tasks; their direct adoption does not address the specific challenges of pantograph wear recognition, where datasets are small, challenging, noisy, and safety-critical. Therefore, the novelty of this work lies not in a generic combination of CNN and transformer classifiers, but in the formulation of a unified framework that jointly addresses dataset issues and robust feature learning. In particular, we design a lightweight dual-backbone feature fusion mechanism that adaptively models interactions between complementary CNN and transformer representations under limited-data conditions, making it suitable for practical industrial deployment.

To overcome the limitations of existing pantograph wear detection approaches, we proposed a systematic dataset sanitisation method that eliminates redundancies among instances and improves annotation before the dataset is used to model training. For this purpose, we employed MD5 cryptographic hashing to identify exact duplicates, followed by a manual re-annotation to ensure consistent class definitions. The annotation process

was manually validated by the first and fourth authors of the paper. This process yields a compact yet comparatively reliable dataset that enables meaningful evaluation of DL classifiers. Furthermore, we proposed a Dual-Backbone Feature Fusion Ensemble Network (DBFF-Net) to improve pantograph wear classification using the preprocessed sanitised dataset. The proposed DBFF-Net approach combines frozen ShuffleNetV2 and DeiT-tiny, the best-performing individual classifiers, using various fusion strategies. Unlike conventional ensemble methods that combine predictions at the decision level, DBFF-Net enables adaptive interaction between heterogeneous feature representations and achieves improved robustness and generalization on pantograph wear recognition compared to existing single-model or data-agnostic approaches. Moreover, a stratified five-fold cross-validation approach was used to train the proposed DBFF-Net. The proposed DBFF-Net approach consistently outperforms individual CNN and transformer models, which shows the utility of dual-backbone feature fusion for industrial visual inspection tasks.

The methodological contributions of the proposed DBFF-Net approach can be summarised as follows:

- We propose DBFF-Net, a dual-backbone feature fusion network by combining a frozen DeiT-Tiny (192-d) and ShuffleNetV2 (1024-d) with five fusion strategies (Concat, Weighted Sum, Bilinear, Cross-Attention, and Gated) to integrate complementary transformer and CNN representations for improved pantograph wear classification.
- We conduct an ablation across all five fusion strategies under stratified five-fold cross-validation, and show that lightweight gated fusion (478 K params) achieves competitive accuracy with substantially lower computational cost than attention-based alternatives. We show that projection-based alignment of asymmetric feature dimensions, combined with selective per-feature gating, outperforms other fusion strategies.
- We conducted a detailed analysis and preprocessing on a small, challenging pantograph wear dataset, and demonstrated improved results by fine-tuning various CNN and transformer architectures and by applying feature fusion approaches for critical and resource-intensive industries.

2. Related Work

Recently, researchers have focused on pantograph condition monitoring for reliable power transmission and operational safety in high-speed railway systems. Existing studies can be grouped into several categories, including vision-based defect detection, signal-based and physics-driven modelling, data-driven predictive approaches, and intelligent control frameworks.

2.1. Vision-Based Pantograph Defect Detection

The vision-based approaches primarily rely on image processing and CNNs to detect wear and surface defects on current collector strips. To detect wear on collector strips, Karaduman and Akin [2] proposed a CNN-based framework that incorporates the Hough Transform and power-law preprocessing. Their approach demonstrated improved performance compared to classical architectures, i.e., ResNet50 and VGG16. Similarly, Tastimur et al. [8] employed CNNs with image preprocessing to improve wear detection performance and highlighted the importance of feature enhancement in visual inspection tasks. Unlike simple pantograph defect classification, several studies have developed more structured visual detection pipelines. For example, Chen et al. [9] proposed a two-stage deep visual neural network comprising a pantograph detection network and a contact point detection module, aimed at improving detection accuracy in complex backgrounds. Similarly, Yang et al. [10] suggested a multi-strategy contact point detection framework using correlation filters, regression networks, and Kalman filtering. Furthermore, Liu et al. [11]

integrated YOLOv5-based detection with UNet-based semantic segmentation to identify pantograph structural anomalies without relying on abnormal data, thereby improving generalisation capability.

More recently, researchers have proposed studies to address specific challenges, such as occlusion and complex field conditions, associated with pantograph wear detection. For example, Yao et al. [12] proposed a robust pantograph slide plate (PSP) wear monitoring method that integrates YOLO-based localisation with morphological and edge-based reconstruction techniques. Meanwhile, Na et al. [13] extended visual monitoring to multiple pantograph components, including contact strips and horns, combining image processing and deep learning to assess panhead conditions in real time. Despite these advances, vision-based methods often exhibit limited robustness to visually ambiguous defect patterns, class imbalance, and insufficient global contextual understanding, thereby restricting their applicability in real-world environments. Apart from research into specific pantograph problems, there have been developments within the field of computer vision in using causal reasoning to enhance image restoration in cases of complicated degradation. Lu et al. [14] proposed CausalSR, a new method of super-resolution that utilises a structural causal model along with counterfactual reasoning. This study suggested that incorporating knowledge of causal structures and degradation factors could improve image restoration performance, even under challenging imaging conditions. Although developed for super-resolution rather than for defect classification, these ideas are relevant to railway inspection scenarios where image-quality degradation may affect downstream recognition performance.

2.2. Physics-Based Modelling and Simulation Approaches

In addition to vision-based approaches, many studies have adopted numerical simulations and physical modelling to study pantograph behaviour. Song et al. [15] added real-world wear measurement data to the pantograph–catenary interaction models, enabling a detailed analysis of the strip imperfections' influence on dynamic contact force, with big dispersion in the interaction behaviour being observed. Similarly, Zhou et al. [16] proposed a heuristic wear prediction model developed by combining mechanistic analysis and data-driven methods; current and arcing effects were identified as important influencing factors. Signal-based approaches have also been explored to estimate pantograph performance indirectly. Furthermore, Gregori et al. [17] used acceleration measurements and ANNs to predict current collection quality and achieved lower prediction errors. Moreover, it was found that there was limited capability to generalise conditions other than those in which the model had been trained, a major drawback in the model itself. Although important insights into the behaviour of the pantograph have been gained through physics-based and signal-based methods, these techniques sometimes involve sophisticated measurements, modelling processes, and scalability issues.

2.3. Intelligent Control and Reinforcement Learning Approaches

Recent studies have increasingly focused on intelligent control strategies to deal with the issues of interactions between the pantograph and the catenary. Han et al. [18] recently presented a new method based on hybrid deep reinforcement learning combined with feedback control to reduce the fluctuations of the contact force between the pantograph and the catenary. Also, Han et al. [19] introduced a new compensation control framework based on the optimised LQR control combined with behaviour cloning and SAC learning policies. Fuzzy and neuro-intelligent control methods were also studied. Sharma et al. [20] presented a fuzzy-based fractional-order PID controller optimised by an Aquila algorithm and supervised by an LSTM neural network to reduce the oscillations of contact forces in different working conditions. Although intelligent control methods have produced

satisfactory outcomes in terms of the optimisation of system dynamic performance, they mainly emphasise control and stability, but not visual defects.

2.4. Research Gaps and Motivation

Based on the above literature, some critical limitations of these approaches, to our knowledge, are as follows: Most of these vision-based approaches rely on single-stage CNN architectures, which are limited in handling global contextual relations between objects. The approaches developed are limited in handling visually overlapped classes of defects. Also, these approaches are limited in handling class ambiguity and imbalance issues in a generic form. The physics-based and control-based approaches are limited in their integration with vision-based systems for scalability. The experiments conducted on using both local and global features in vision-based systems are limited. A detailed methodological comparative study of the proposed DBFF-Net approach against existing approaches is presented in Table 1.

Driven by these limitations, the proposed approach presents a hybrid method that combines CNNs and transformers, employing various feature fusion methods to enrich global feature representations and address class confusion in pantograph defect classification. The proposed wear classification approach contributes to the connection between local image features and global context reasoning and provides an improved robustness mechanism based on the decision hierarchy.

Table 1. Comparison of existing pantograph monitoring and classification approaches.

Paper	Main Objective	Methodology	Key Contributions	Limitations
[2]	Wear detection on current collector strips	CNN with Hough Transform and power-law preprocessing	Custom CNN architecture outperforming ResNet50 and VGG16 and enhanced dataset quality	Limited global feature modelling
[15]	Impact of strip wear on pantograph–catenary interaction	Numerical simulation with real measurement data	Quantified dispersion in contact force due to strip imperfections	Not applicable to real-time visual defect detection
[9]	Contact point detection in complex environments	Two-stage deep visual neural network (DPDN + IVFE)	Improved detection accuracy and real-time performance	Limited robustness under visual ambiguity and class imbalance
[17]	Prediction of current collection quality	ANN-based signal modelling using acceleration data	Achieved prediction error below 10% using ANN models	Poor generalization across varying operating conditions
[16]	Wear prediction of pantograph strips	Hybrid mechanistic and data-driven wear model	Identified key wear factors and proposed heuristic wear model	Not designed for image-based defect classification
[12]	PSP wear monitoring under occlusion	YOLOv5-based localization + edge reconstruction	Robust detection under occlusion with 0.6 mm accuracy	High dependency on handcrafted reconstruction rules
[13]	Real-time monitoring of panhead and horn condition	Image processing + deep learning	Integrated monitoring of multiple pantograph components	Limited scalability and ambiguity handling
[18]	Suppression of extreme contact force fluctuations	Hybrid DRL with feedback control	Reduced PCCF standard deviation by up to 41.85%	Focused on control rather than defect classification

Table 1. *Cont.*

Paper	Main Objective	Methodology	Key Contributions	Limitations
[19]	Active compensation control for PCCF oscillations	CPO-LQR + BC-SAC reinforcement learning	Achieved over 77% reduction in PCCF fluctuations	Computational complexity and limited visual integration
[10]	High-precision contact point detection	Multi-module framework with CPRR-Net and Kalman filter	Achieved 97.07% accuracy within 3 pixels at 65 FPS	Limited analysis of defect-level classification
[20]	Intelligent control of contact force	Fuzzy fractional-order PID with LSTM and Aquila optimizer	Reduced oscillations under varying operating conditions	Not applicable to visual defect detection tasks
[11]	Pantograph structural anomaly detection	YOLOv5 + UNet semantic segmentation	High-speed anomaly detection without abnormal data dependency	Limited capability in fine-grained defect classification
[8]	Wear detection using CNN with preprocessing	CNN with image enhancement techniques	Improved classification accuracy using preprocessed images	Lack of global contextual modelling and ambiguity resolution
Proposed Method	Ambiguity-aware pantograph defect classification	CNN–Transformer hybrid + two-stage hierarchical classification	Joint local–global feature modelling; ambiguity resolution via refinement; improved classification accuracy	Requires further validation on large-scale datasets

3. Methodology

This work proposes an end-to-end DL framework for pantograph contact strip wear classification under limited and noisy data conditions. The overall pipeline consists of three stages: (i) dataset preprocessing and sanitization, (ii) candidate backbone evaluation using stratified cross-validation, and (iii) the proposed DBFF-Net for adaptive feature fusion and final classification. Figure 1 illustrates the proposed workflow.

3.1. Dataset Preprocessing and Sanitization

Before developing wear classification models, data preprocessing was performed to improve training quality and prevent bias during model evaluation. The analysis of the initial dataset revealed duplicates, the same image appearing in other classes, and inconsistent labelling. Thus, the data was sanitised using standard practices before the training process.

In this step of the proposed DBFF-Net approach, we first conducted a detailed manual inspection and analysis of the dataset. This identifies a comparatively large number of visually identical samples, including instances in which identical images were assigned to different wear classes. These inconsistencies add challenges to the supervised learning approaches and can affect model convergence and generalisation. To overcome this, we used the Message Digest (MD5) Algorithm 1 cryptographic hash function to systematically identify exact duplicates. MD5 maps an arbitrary-length input to a fixed 128-bit hash value. For each image x_i, an MD5 hash signature is computed as

$$h_i = \mathcal{H}_{\text{MD5}}(x_i), \tag{1}$$

where $\mathcal{H}_{\text{MD5}}(\cdot)$ denotes the MD5 hashing function. Two images x_i and x_j are considered exact duplicates if

$$h_i = h_j. \tag{2}$$

Algorithm 1 MD5-Based Exact Deduplication of Pantograph Images

Require: Dataset root directory $\mathcal{R}$ with splits and class folders
Ensure: Clean file set $\mathcal{D}_{\text{unique}}$ and duplicate log $\mathcal{L}$
Initialize hash map $\mathcal{H} \leftarrow \emptyset$ (hash $\rightarrow$ list of file paths)
Initialize log $\mathcal{L} \leftarrow \emptyset$
for all image files f under $\mathcal{R}$ **do**
    Read bytes $b \leftarrow \text{read}(f)$
    Compute MD5 hash $h \leftarrow \text{MD5}(b)$
    Append f to $\mathcal{H}[h]$
end for
$\mathcal{D}_{\text{unique}} \leftarrow \emptyset$
for all hashes h in $\mathcal{H}$ **do**
    Let $\mathcal{F} \leftarrow \mathcal{H}[h]$
    if $|\mathcal{F}| = 1$ **then**
        Add $\mathcal{F}[1]$ to $\mathcal{D}_{\text{unique}}$
    else
        Choose canonical file $f^\star \in \mathcal{F}$ (e.g., prefer train split)
        Add $f^\star$ to $\mathcal{D}_{\text{unique}}$
        Add $(f^\star, \mathcal{F} \setminus \{f^\star\})$ to log $\mathcal{L}$
    end if
end for
return $\mathcal{D}_{\text{unique}}, \mathcal{L}$

We grouped images based on their hash values to identify duplicates across the training and testing sets. For each group of identical images, only one representative image was kept, and all other duplicates were removed. This step was taken to avoid information leakage between the training and testing data and to prevent inconsistent or misleading supervision during model learning. Firstly, it must be pointed out that MD5 hashes were employed as a fast initial step in searching for exact duplicates. The image files that had the same hashes were then manually checked in order to make sure that the images were indeed copies of each other and not separate ones. There have been no hash collision cases throughout this analysis process. Though cryptographic hash collisions can occur, the chance of their happening on such a data scale is so low as to be unimportant for the search for duplicates. Moreover, all the duplicated files were valid and readable images, and some duplicates occurred in different filenames or classes. Therefore, it can be concluded that the occurrence of duplicates was due to the dataset compilation procedures. To further assess the severity of duplication in the original dataset, we analysed the distribution of duplicate files across labels and found several visually identical samples were associated with different class labels.

Next, we performed a manual relabelling procedure to correct inherited label noise. Each remaining image is independently reviewed by the first and fourth authors of the paper and assigned to one of four wear categories based on visual inspection of the pantograph contact strip. The conflicts between the annotators were resolved through discussion, if there were any. Also, the fifth author of the paper, with extensive experience in data annotation, helped resolve the conflicts between the two annotators. Moreover, the manual relabelling still needs to be validated by a domain expert with experience in railway component inspection and visual quality assessment of pantograph systems. Since the dataset originates from an industrial inspection context, it has been widely used for wear classification. Therefore, the annotators manually reviewed it and re-annotated it into the correct wear classes. Also, we did not add new wear images to the dataset; we only identified the correct wear category based on visual patterns. Furthermore, to improve consistency and reduce subjectivity, a structured annotation guideline was followed. Specifically, each class was defined using the following operational criteria.

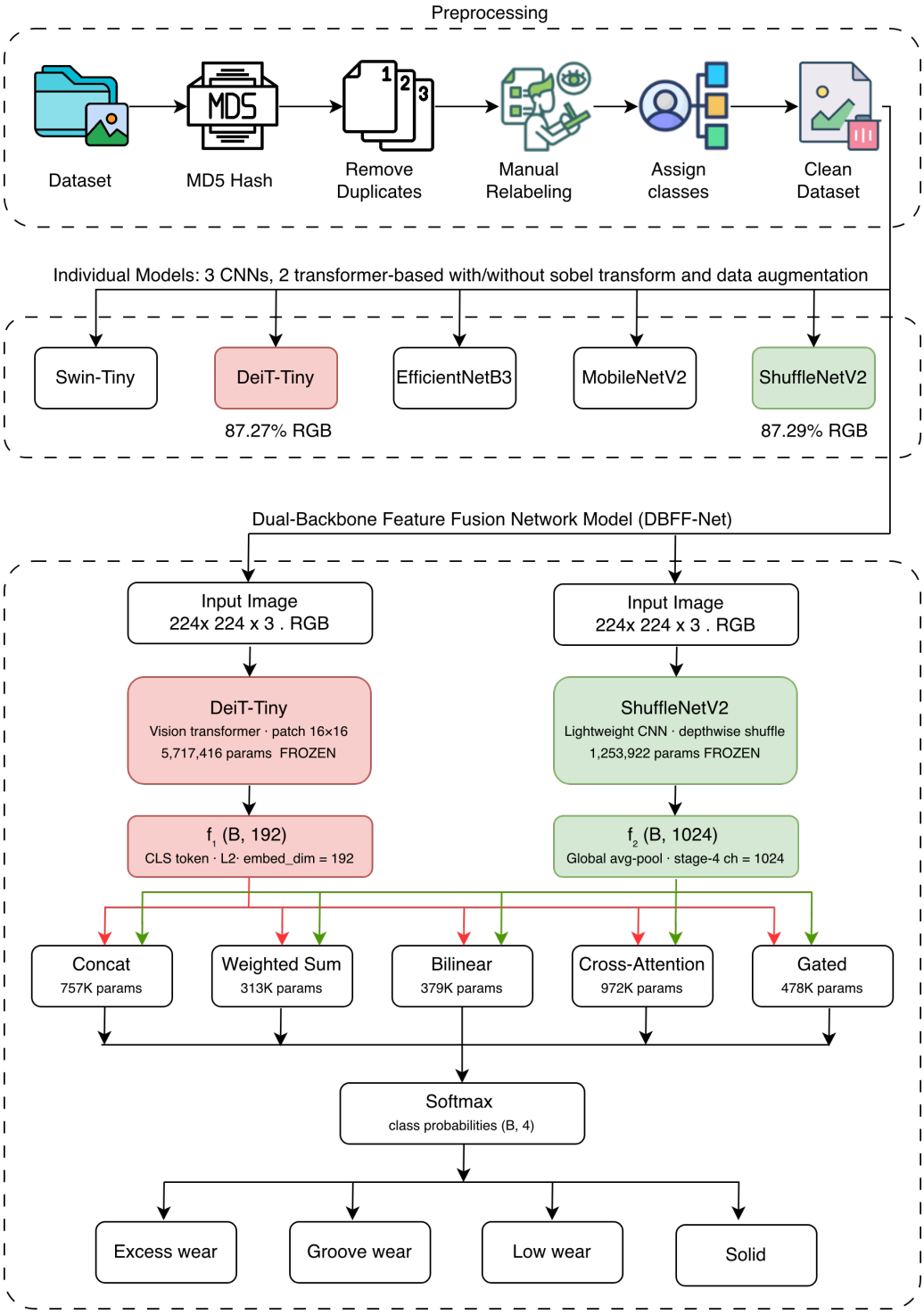

Figure 1. Proposed DBFF-Net framework for pantograph wear classification. The coloured arrows indicate the feature inputs from the two backbone networks into the fusion module, where red represents Diet-Tiny and green represents ShuffleNetV2.

Each image was assigned a label based on the dominant visual wear characteristics observed on the pantograph contact strip. The solid class corresponds to samples with no visible material degradation, smooth surface texture, and well-defined edges without noticeable abrasion. The low wear class represents early-stage deterioration characterised by slight surface roughness or minor edge degradation without structural deformation. The excess wear class includes images showing pronounced material loss, visible surface damage, and irregular or rough textures indicative of severe degradation. Finally, the

groove wear class refers to cases where clear longitudinal grooves or channel-like patterns are present along the contact surface. To ensure internal consistency of labelling, all images were reviewed in a single annotation pass, and ambiguous cases were re-evaluated to reach a consistent final decision. This procedure reduces intra-annotator inconsistency and ensures a uniform labelling standard across the sanitized dataset.

Let $\tilde{y}_i$ denote the corrected label. The cleaned dataset is therefore

$$\mathcal{D}_{\text{clean}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_u}, \tag{3}$$

where N_u is the number of unique and verified samples.

This process corrected inherited label noise and produced a reliable dataset for training and evaluation. Let $\tilde{y}_i$ denote the corrected label, yielding the cleaned dataset $\mathcal{D}_{\text{clean}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_u}$.

3.2. Edge-Aware Preprocessing Using Sobel Transform

Wear-related patterns on the pantograph contact strip, i.e., grooves, boundaries, and surface irregularities, are often subtle and difficult for models to detect directly from raw images. To address this, we conducted an ablation study across three pretrained CNNs (MobileNetV2, ShuffleNetV2, EfficientNet) and two transformer-based architectures (Deit-tiny and Swin-tiny), comparing their performance with and without edge-aware preprocessing using the Sobel transform.

Given an input image I, we compute the horizontal and vertical intensity gradients using the Sobel operator:

$$G_x = S_x * I, \quad G_y = S_y * I, \tag{4}$$

where S_x and S_y denote the Sobel kernels and $*$ represents the convolution operation. The gradient magnitude is then calculated as

$$G = \sqrt{G_x^2 + G_y^2}. \tag{5}$$

To evaluate the impact of the Sobel operator, the resulting edge-enhanced and original RGB images are both normalised and provided as inputs to the three pretrained CNNs (MobileNetV2, ShuffleNetV2, EfficientNet) and two transformer-based architectures (Deit-tiny and Swin-tiny) models.

Data Augmentation

Given the limited sample size after cleaning, we used augmentation during training to reduce overfitting. The augmentation set includes random cropping, horizontal/vertical flipping, small rotations, colour jitter, affine translation, and random erasing. Let $t(\cdot)$ denote the stochastic augmentation operator; the training sample becomes $x_i' = t(x_i)$.

3.3. Deep Learning Models and Transfer Learning Strategy

In this step, we evaluated 5 baseline pretrained architectures, i.e., MobileNetV2, ShuffleNetV2, EfficientNet-B3, DeiT-Tiny, and Swin-Tiny, to identify the most robust classification model (Table 2). These models are selected based on their complementary design philosophies, ranging from lightweight CNNs to transformer-based architectures. For each model, we replaced the original classification head with a four-class output layer. Transfer learning was applied by initialising backbone parameters with pretrained weights, and the task-specific layers were fine-tuned on the cleaned dataset. Dropout regularisation was used to improve generalisation.

Given logits $z \in \mathbb{R}^4$, class probabilities are computed using the softmax function:

$$p_k = \frac{\exp(z_k)}{\sum_{j=1}^{4} \exp(z_j)}. \tag{6}$$

The training objective is the categorical cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k=1}^{4} \mathbf{1}[y = k] \log(p_k). \tag{7}$$

For an input image x, the network outputs logits $z = f_\theta(x)$, and the model parameters are optimized by minimizing the loss defined in Equation (7).

Model parameters are optimised using the AdamW optimiser with learning rate η and decoupled weight decay λ. Learning-rate scheduling is applied using cosine annealing for transformer-based models and ReduceLROnPlateau for CNN-based models. Early stopping with a predefined patience is used to prevent overfitting (Table 2).

Table 2. Pretrained CNN/transformer-based model size, training time on RTX3060, training strategy, and hyperparameters.

Model	Parameters	Time	Strategy	Hyperparameters
ShuffleNetV2	∼1.26 M	∼2.5 min	Full fine-tuning from epoch 1; new FC head	LR: $3 \times 10^{-4}/1 \times 10^{-3}$, Weight decay: 0.01, Batch size: 8, Dropout: 0.3, Label smoothing: 0.1, Warmup: 6, Cosine, Clip: 1.0
MobileNetV2	∼2.27 M	∼3.8 min	Full fine-tuning from epoch 1; new FC head	LR: $3 \times 10^{-4}/1 \times 10^{-3}$, Weight decay: 0.01, Batch size: 8, Dropout: 0.3, Label smoothing: 0.1, Warmup: 6, Cosine, Clip: 1.0
DeiT-Tiny	∼5.72 M	∼7.1 min	Freeze for 3 epochs, then unfreeze; head reinit; Drop Path 0.1	LR: $3 \times 10^{-5}/2 \times 10^{-4}$, Weight decay: 0.05/0.02, Batch size: 8, Dropout: 0.2, Label smoothing: 0.1, Warmup: 6, Cosine, Clip: 1.0
EfficientNet-B3	∼12.23 M	∼9.6 min	Full fine-tuning from epoch 1; timm head reset; Dropout 0.4	LR: $3 \times 10^{-4}/1 \times 10^{-3}$, Weight decay: 0.01, Batch size: 8, Dropout: 0.4, Label smoothing: 0.1, Warmup: 6, Cosine, Clip: 1.0
Swin-Tiny	∼28.29 M	∼9.1 min	Freeze for 3 epochs, then unfreeze; head reinit; Drop Path 0.15	LR: $3 \times 10^{-5}/2 \times 10^{-4}$, Weight decay: 0.05/0.02, Batch size: 8, Dropout: 0.2, Label smoothing: 0.1, Warmup: 6, Cosine, Clip: 1.0

3.4. Proposed Dual-Backbone Feature Fusion Network (DBFF-Net)

Considering the limitations of individual CNN and transformer algorithms, we propose a dual-backbone feature-fusion network for improved rail pantograph wear classification. We integrated the best performing two frozen pretrained models, DeiT-Tiny (feature dimension $d_1 = 192$) and ShuffleNetV2 (feature dimension $d_2 = 1024$), to extract complementary representations from input images. We then applied a gated fusion mechanism to combine these heterogeneous features for final classification adaptively (Algorithm 2). The overall pipeline consists of dual-stream preprocessing, frozen feature extraction, feature projection into a shared embedding space ($d = 256$), adaptive fusion, and classification. Different fusion strategies were evaluated in terms of trainable parameters, architectural complexity, model size, and estimated training time, as summarized in Table 3.

Algorithm 2 Gated Multi-Modal Fusion for Pantograph Wear Classification

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, $y_i \in \{1, \ldots, C\}$, $C = 4$, frozen backbones: DeiT-Tiny Φ_D (RGB-trained), ShuffleNetV2 Φ_S (RGB-trained), embedding dimension $d = 256$, folds $K = 5$, epochs $E = 80$, patience $P = 20$.

Ensure: Mean accuracy μ and standard deviation σ over K folds, trained gated fusion model $\mathcal{M}^*$.

Preprocessing

Define RGB transform $\mathcal{T}_D(x) = \text{Augment}(x)$ for DeiT branch

Define RGB transform $\mathcal{T}_S(x) = \text{Augment}(x)$ for ShuffleNet branch

Forward Pass (Gated Fusion)

for each input x **do**

$\mathbf{f}_D = \Phi_D(\mathcal{T}_D(x)), \quad \mathbf{f}_S = \Phi_S(\mathcal{T}_S(x))$

$\mathbf{p}_D = \text{LayerNorm}(W_D \mathbf{f}_D + b_D), \quad \mathbf{p}_S = \text{LayerNorm}(W_S \mathbf{f}_S + b_S)$

$\mathbf{g} = \sigma(W_g[\mathbf{p}_D; \mathbf{p}_S] + b_g)$

$\mathbf{z} = \mathbf{g} \odot \mathbf{p}_D + (\mathbf{1} - \mathbf{g}) \odot \mathbf{p}_S$

$\hat{y} = \text{softmax}(\Psi(\mathbf{z}))$ ▷ Ψ: MLP with GELU, LayerNorm, dropout

end for

Training

Construct K stratified folds $\{\mathcal{D}^{(k)}\}_{k=1}^{K}$

for $k = 1$ to K **do**

Initialize fusion head parameters Θ_k

for $e = 1$ to E **do**

for mini-batch $\{(x_j, y_j)\} \subset \mathcal{D} \setminus \mathcal{D}^{(k)}$ **do**

Compute predictions $\hat{y}_j$ via forward pass

$\mathcal{L} = \text{CrossEntropy}_{\text{ls}}(\hat{y}_j, y_j)$ ▷ label smoothing $\epsilon = 0.05$

$\Theta_k \leftarrow \text{AdamW}(\nabla_{\Theta_k} \mathcal{L})$ ▷ $\eta = 3 \times 10^{-4}$, $\lambda = 10^{-2}$

end for

Evaluate accuracy a_e on $\mathcal{D}^{(k)}$

if $a_e > a^*$ **then** $a^* \leftarrow a_e$; save Θ_k

end if

if no improvement for P epochs **then break**

end if

end for

$\mathcal{A}_k \leftarrow a^*$

end for

$$\mu = \frac{1}{K} \sum_{k=1}^{K} \mathcal{A}_k, \quad \sigma = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (\mathcal{A}_k - \mu)^2}$$

return (μ, σ), $\mathcal{M}^*$

Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, two independent transformation pipelines are applied to construct modality-specific inputs for each backbone.

For the dual-backbone branch (DeiT-Tiny), standard data augmentation is applied directly to the RGB image without edge enhancement:

$$x^{(1)} = \mathcal{T}_{\text{DeiT}}(x) = \text{Augment}(x) \tag{8}$$

where $\text{Augment}(\cdot)$ includes random resized cropping (to 224×224), horizontal/vertical flips (probability 0.5 and 0.3), rotation ($\pm 15^\circ$), colour jitter, random affine translation, and random erasing ($p = 0.3$).

For the CNN-based branch (ShuffleNetV2), standard data augmentation is applied directly to the RGB image without edge enhancement:

$$x^{(2)} = \mathcal{T}_{\text{Shuffle}}(x) = \text{Augment}(x) \tag{9}$$

where the same augmentation operations are applied, excluding Sobel transformation.

Table 3. Comparison of fusion strategies in terms of trainable parameters, architecture, model size, and estimated training time on an RTX 3060 GPU.

Fusion Strategy	Trainable Params	Architecture Summary	Size	Est. Time
Concat	756,996	Linear (1216 $\rightarrow$ 512) + LN + GELU · Linear (512 $\rightarrow$ 256) + LN + GELU · Linear (256 $\rightarrow$ 4)	Medium	$\sim$3.7 min
Weighted Sum	313,350	proj_1 (192 $\rightarrow$ 256) · proj_2 (1024 $\rightarrow$ 256) · softmax weights · LN + GELU · Linear (256 $\rightarrow$ 4)	Light	$\sim$1.5 min
Bilinear	379,140	proj_1 + ReLU · proj_2 + ReLU · Hadamard $\rightarrow$ sign $\cdot\sqrt{\cdot}\cdot$ L2 · Linear (256 $\rightarrow$ 256) + LN + GELU · Linear (256 $\rightarrow$ 4)	Light	$\sim$1.9 min
Cross-Attention	972,036	$\text{proj}_{1,2}$ ($\rightarrow$256) · bidirectional MHA (4 heads) · LN $\times$2 · Linear (512 $\rightarrow$ 256) + LN + GELU · Linear (256 $\rightarrow$ 4)	Heavy	$\sim$10.5 min
Gated	478,340	proj_1 + LN · proj_2 + LN · sigmoid gate · $g \odot p_1 + (1-g) \odot p_2$ · LN · Linear (256 $\rightarrow$ 128) + LN · Linear (128 $\rightarrow$ 4)	Light	$\sim$2.4 min
Total	2,899,862	Backbones frozen (no trainable backbone parameters)	–	$\sim$20 min

Note: "$\rightarrow$" denotes mapping between input and output dimensions, and "·" denotes sequential operation composition.

This design ensures that the transformer branch focuses on structural edge cues while the CNN branch captures texture and appearance information. We used two pretrained models as fixed feature extractors: DeiT-Tiny transformer $f_{\theta_1}(\cdot)$ producing $d_1 = 192$-dimensional features and ShuffleNetV2 CNN $f_{\theta_2}(\cdot)$ producing $d_2 = 1024$-dimensional features. Both backbones are completely frozen during training:

$$\theta_1, \theta_2 = \text{constant} \quad \text{(no gradient update)} \tag{10}$$

The extracted features are computed as

$$\mathbf{f}_1 = f_{\theta_1}(x^{(1)}) \in \mathbb{R}^{192}, \quad \mathbf{f}_2 = f_{\theta_2}(x^{(2)}) \in \mathbb{R}^{1024} \tag{11}$$

No gradient is propagated through either backbone, ensuring that only the fusion module is trained. Since the two backbones produce features of different dimensionalities, learned linear projections with layer normalization are used to map them into a common embedding space of dimension $d = 256$:

$$\mathbf{p}_1 = \text{LayerNorm}(W_1\mathbf{f}_1 + b_1), \quad \mathbf{p}_2 = \text{LayerNorm}(W_2\mathbf{f}_2 + b_2) \tag{12}$$

where $W_1 \in \mathbb{R}^{256\times192}$ and $W_2 \in \mathbb{R}^{256\times1024}$ are learnable projection matrices, and $b_1 \in \mathbb{R}^{256}$ and $b_2 \in \mathbb{R}^{256}$ are bias terms.

To dynamically control the contribution of each modality at the feature dimension level, a feature-wise gating mechanism is introduced. The gating vector is computed by conditioning on both projected features:

$$\mathbf{g} = \sigma\left(W_g[\mathbf{p}_1; \mathbf{p}_2] + b_g\right) \tag{13}$$

where $[\cdot;\cdot] \in \mathbb{R}^{512}$ denotes concatenation, $W_g \in \mathbb{R}^{256\times512}$ and $b_g \in \mathbb{R}^{256}$ are learnable parameters, and $\sigma(\cdot)$ is the element-wise sigmoid activation function. The gating vector satisfies $\mathbf{g} \in (0,1)^{256}$. The fused representation is obtained via element-wise adaptive weighting:

$$\mathbf{f}_{\text{fusion}} = \mathbf{g} \odot \mathbf{p}_1 + (\mathbf{1} - \mathbf{g}) \odot \mathbf{p}_2 \tag{14}$$

where $\odot$ denotes the Hadamard (element-wise) product. This formulation allows the model to selectively emphasize either dual-backbone structural features or CNN-based texture features on a per-dimension basis, with the mixing weights determined dynamically from the input. The fused feature vector is passed through a multi-layer perceptron (MLP) classifier with dropout regularization ($p = 0.4$):

$$\hat{\mathbf{y}} = \text{softmax}(W_{\text{out}} \cdot \text{GELU}(\text{LayerNorm}(W_{\text{mid}}\mathbf{f}_{\text{fusion}} + b_{\text{mid}})) + b_{\text{out}}) \tag{15}$$

where $W_{\text{mid}} \in \mathbb{R}^{128\times256}$, $b_{\text{mid}} \in \mathbb{R}^{128}$, $W_{\text{out}} \in \mathbb{R}^{4\times128}$, and $b_{\text{out}} \in \mathbb{R}^4$, and GELU is the Gaussian Error Linear Unit activation function. Dropout ($p = 0.4$) is applied after each GELU activation. The output $\hat{\mathbf{y}} \in [0,1]^4$ represents class probabilities for $C = 4$ wear categories.

The model is optimized using cross-entropy loss with label smoothing ($\epsilon = 0.05$):

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C}\left[(1-\epsilon)\cdot\mathbf{1}_{y_i=c} + \frac{\epsilon}{C}\right]\log\hat{y}_{i,c} \tag{16}$$

where B is the batch size, and $\mathbf{1}_{(\cdot)}$ is the indicator function. Only the parameters of the projection layers, gating network, and classification head are updated:

$$\Theta = \{W_1, b_1, W_2, b_2, W_g, b_g, W_{\text{mid}}, b_{\text{mid}}, W_{\text{out}}, b_{\text{out}}\} \tag{17}$$

The backbones remain frozen throughout training. Optimization is performed using AdamW ($\eta = 3\times10^{-4}$, weight decay $= 10^{-2}$) with cosine annealing learning-rate scheduling ($\eta_{\min} = 10^{-6}$) and gradient clipping ($\|\nabla\Theta\|_2 \le 1.0$).

For robust evaluation, stratified 5-fold cross-validation is employed. The dataset is partitioned into $K = 5$ disjoint subsets preserving class distribution, and the model is trained and evaluated K times. Early stopping with patience $P = 20$ epochs is applied to prevent overfitting. The final performance is reported as

$$\mu = \frac{1}{K}\sum_{k=1}^{K}\text{Acc}^{(k)}, \qquad \sigma = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\text{Acc}^{(k)} - \mu\right)^2} \tag{18}$$

where $\text{Acc}^{(k)}$ denotes validation accuracy for the k-th fold. The mean accuracy μ and standard deviation σ are reported along with the confusion matrix and per-class F1-scores.

The proposed gated fusion strategy enables dynamic, input-dependent weighting of transformer and CNN representations at the individual feature dimension level, which is strictly more expressive than scalar weighting schemes. In the case of wear classification, gated fusion enables structural defects (captured by DeiT-Tiny's edge-based features) and texture variations (captured by ShuffleNetV2's RGB features) to jointly contribute to discriminative learning. The per-dimension gating allows the model to rely on structural features for detecting groove wear while depending on texture features for identifying excess wear.

3.5. Stratified K-Fold Cross-Validation

Due to the limited dataset size, we performed the model evaluation using K-fold cross-validation for statistically reliable performance estimation (Algorithm 3). The dataset is partitioned into stratified $K = 5$ folds to preserve class distribution. For each fold k, we trained the model on $\mathcal{D} \setminus \mathcal{D}^{(k)}$ and validated it on $\mathcal{D}^{(k)}$. The mean and standard deviation of accuracy are computed as

$$\mu = \frac{1}{K}\sum_{k=1}^{K} \mathrm{acc}^{(k)}, \quad \sigma = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathrm{acc}^{(k)} - \mu\right)^2}. \tag{19}$$

Algorithm 3 Stratified K-Fold Training for Candidate Backbones

Require: Clean dataset $\mathcal{D}_{\text{clean}}$, candidate models $\{\mathcal{M}_j\}_{j=1}^{J}$, folds K
Ensure: Cross-validated performance summaries $\{(\mu_j, \sigma_j)\}_{j=1}^{J}$

```
Create stratified folds {D^(k)}_{k=1}^K
for j = 1 to J do
    Initialize accuracy list A ← [ ]
    for k = 1 to K do
        D_train ← D_clean \ D^(k),  D_val ← D^(k)
        Instantiate model f_θ ← M_j and replace head with 4-class classifier
        Train f_θ on D_train using augmentation, AdamW, scheduler, early stopping
        Evaluate accuracy acc^(k) on D_val; append to A
    end for
    Compute μ_j ← (1/K) Σ_{k=1}^K acc^(k) and σ_j as in Equation (19)
end for
return {(μ_j, σ_j)}_{j=1}^J
```

3.6. Evaluation Metrics

The performance of the proposed DBFF-Net approach is reported using confusion matrices and per-class precision, recall, and F1-score. The confusion matrices were evaluated to analyse class-wise performance. For a class c, precision and recall are

$$\mathrm{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \mathrm{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{20}$$

and the F1-score is

$$\mathrm{F1}_c = \frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}. \tag{21}$$

The overall accuracy of the proposed DBFF-Net approach is computed as

$$\mathrm{Accuracy} = \frac{\sum_{c=1}^{4} TP_c}{\sum_{c=1}^{4}(TP_c + FP_c + FN_c + TN_c)}. \tag{22}$$

4. Results and Discussion

4.1. Dataset Description

The pantograph wear data used in this study is originally derived from real-world railway inspection data published by Karaduman and Akin [2]. The dataset comprises pantograph images of current collector strips acquired from operational railway systems using a camera-based non-contact monitoring setup. The pantograph images were manually annotated by domain experts according to the degree and type of wear on the collector strips, resulting in four distinct wear categories. For experimental evaluation, the dataset is structured into separate training and testing subsets, ensuring a clear separation between model learning and performance assessment. Each class contains a dedicated train and test set, enabling a balanced evaluation of classification performance across wear categories [2].

The four classes are groove-shaped wear excess wear, low wear, and solid, corresponding to different wear patterns of pantograph components. Moreover, excess and groove wear images are visually similar. Table 4 summarises the distribution of images across training and testing sets for each class.

Table 4. Distribution of pantograph dataset across classes.

Class	Training Images	Testing Images	Total Images
Excess wear	156	32	188
Groove wear	131	19	150
Low wear	329	60	389
Solid	155	27	182
Total	771	138	909

As shown in Table 4, the dataset contains 909 pantograph images, with 771 used for training and 138 reserved for testing. Although the dataset shows a moderate class imbalance, it suffers from significant inconsistencies due to duplicate and cross-class mislabelled images, as summarised in Tables 5 and 6, based on our analyses and understanding. The duplication problem is overwhelmingly concentrated in the training set with 547 of the 626 removed images, with the low wear class being the most affected (188 removed images across 132 duplicate groups). In contrast, the test set shows comparatively milder duplication, with 79 removed images that are more evenly distributed across classes, averaging about 10 duplicate groups per class. As shown in Table 6, substantial label inconsistencies existed between class pairs, most notably between excess wear and groove wear (54 overlapping images in both directions) and between excess wear and low wear (37 shared images). These overlaps indicate that a non-trivial portion of the dataset contains images simultaneously assigned to multiple conflicting wear categories, which can negatively impact model learning and class separability.

Table 5. Duplicate image counts (removed images).

Class	Train (Within)	Test (Within)
Excess wear	111	15
Groove wear	108	35
Low wear	188	12
Solid	140	17
Total	547	79

Table 6. Label inconsistencies and mislabelled images between class pairs.

Class	Duplicate Copies Found			
	Excess Wear	Groove Wear	Low Wear	Solid
Excess wear	–	54	37	0
Groove wear	54	–	21	0
Low wear	37	21	–	18
Solid	0	0	18	–

After removing duplicate images and resolving label inconsistencies, the dataset was cleaned and consolidated to improve overall reliability. The sanitised dataset distribution across classes is excess wear (93), groove wear (109), low wear (56), and solid (25) remaining after preprocessing. Following this cleaning stage, the training and test sets were combined

into a single unified dataset. The entire dataset was then manually relabelled and validated by the first and fourth authors for consistent and correct class assignments by a uniform coding guideline.

4.2. Reproducibility and Implementation Details

To identify pantograph wear damage with the proposed DBFF-Net approach, we implemented all experiments in PyTorch (2.2.2). For reproducibility, random seeds are fixed for Python (3.12.2), NumPy (1.26.4), and PyTorch (2.2.2), and stratified splits are used with a fixed random state. Models are trained for up to 100 epochs with early stopping (patience 12–20 epochs). Batch sizes are set to 8 for training and 16 for validation to accommodate limited GPU memory. Input images are resized to $224 \times 224 \times 3$ after augmentation and normalized using ImageNet mean and standard deviation. The entire pipeline, including dataset cleaning and model training, is executed under identical experimental conditions to guarantee a fair comparison.

Furthermore, the proposed dual-backbone feature fusion fine-tuning in this study refers exclusively to training the fusion head parameters from scratch. At the start of each fold, all fusion head weights are randomly initialised, and only these parameters receive gradient updates throughout training. The AdamW optimiser with a learning rate of 3×10^{-4}, cosine annealing schedule, and weight decay of 10^{-2} is applied solely to the fusion head, while backbone parameters are explicitly excluded from the optimiser's parameter groups. Early stopping with a patience of 20 epochs monitors validation loss to halt training when the fusion head ceases to improve. Thus, each strategy is evaluated under identical and fair conditions across all five folds.

4.3. Performance Evaluation of the Proposed Dual-Backbone Feature Fusion Ensemble Model

The proposed DBFF-Net is quantitatively assessed for the classification of pantograph contact strip wear using a comparatively challenging, limited, and noise-mitigated dataset. We evaluated the ensemble model (DBFF-Net) using stratified five-fold cross-validation and assessed its ability to achieve balanced and reliable classification across four wear categories: excess wear, groove wear, low wear, and solid surface conditions. Figure 2 illustrates the confusion matrix of the Gated DBFF-Net ensemble model. Table 7 reports the mean and standard deviation of various feature fusion approaches. It also reports the per-fold accuracy of each feature fusion approach for classifying wear images into various defect types, demonstrating more stable and generalised results. In contrast, Table 8 reports the per-class precision, recall, and F1-score, along with macro-averaged metrics to account for class imbalance.

Table 7. Accuracy (%) of different feature fusion strategies using frozen backbones. Mean and standard deviation are reported along with per-fold results.

Fusion	Mean ± Std	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Concat	96.13 ± 3.02	94.74	91.23	98.25	100.00	96.43
Weighted Sum	95.41 ± 3.80	92.98	91.23	100.00	100.00	92.86
Bilinear	95.39 ± 2.41	98.25	92.98	98.25	94.64	92.86
Cross-Attention	95.77 ± 3.62	92.98	91.23	100.00	100.00	94.64
Gated	96.45 ± 2.78	98.25	98.25	98.25	96.43	91.07

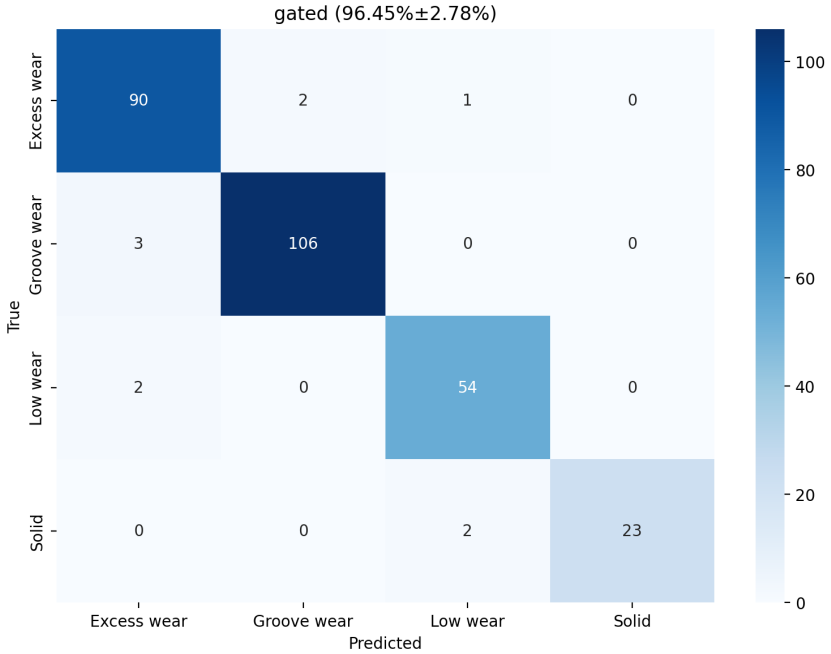

Figure 2. Confusion matrix of the proposed Gated DBFF-Net ensemble model.

Table 8. Class-wise precision, recall, and F1-score for different fusion strategies, along with overall accuracy.

Fusion	Class	Precision	Recall	F1-Score	Accuracy
Concat	Excess wear	0.9670	0.9462	0.9565	0.9611 DeiT = +7.45% Shuffle = +7.79%
	Groove wear	0.9636	0.9725	0.9680	
	Low wear	0.9636	0.9464	0.9550	
	Solid	0.9259	1.0000	0.9615	
Weighted Sum	Excess wear	0.9362	0.9462	0.9412	0.9541 DeiT = +6.73% Shuffle = +7.07%
	Groove wear	0.9630	0.9541	0.9585	
	Low wear	0.9815	0.9464	0.9636	
	Solid	0.9259	1.0000	0.9615	
Bilinear	Excess wear	0.9271	0.9570	0.9418	0.9541 DeiT = +6.71% Shuffle = +7.05%
	Groove wear	0.9717	0.9450	0.9581	
	Low wear	0.9474	0.9643	0.9558	
	Solid	1.0000	0.9600	0.9796	
Cross-Attention	Excess wear	0.9565	0.9462	0.9514	0.9576 DeiT = +7.09% Shuffle = +7.43%
	Groove wear	0.9633	0.9633	0.9633	
	Low wear	0.9636	0.9464	0.9550	
	Solid	0.9259	1.0000	0.9615	
Gated	Excess wear	0.9474	0.9677	0.9574	0.9647 DeiT = +7.77% Shuffle = +8.11%
	Groove wear	0.9815	0.9725	0.9770	
	Low wear	0.9474	0.9643	0.9558	
	Solid	1.0000	0.9200	0.9583	

As shown in Table 8, all the feature fusion strategies perform comparatively well in identifying fine-grained pantograph wear types. However, the Gated DBFF-Net achieved an overall classification accuracy of 96.47%, which is slightly higher than the remaining feature fusion approaches, demonstrating its generalisation capability on the clean and noise-mitigated dataset. The proposed DBFF-Net ensemble model achieves robust performance on all wear classes, thus validating that the proposed feature-level gated per-feature fusion extracts meaningful features from both backbones. The best performance is reported

for the groove wear class, achieving a recall of 0.9725 and an F1-score of 0.9770. Hence, this proves that the proposed model is correctly detecting longitudinal groove patterns. The excess and low wear and solid classes achieve well-balanced precision and recall values of approximately 0.94 and 0.96, which demonstrate the ability of the proposed gated per-feature fusion DBFF-Net to distinguish between severely damaged and pristine contact strips. Furthermore, Figure 3 shows the confusion matrices for the Concat, Weighted Sum, Bilinear, and Cross-Attention feature fusion approaches. It demonstrates that all feature fusion approaches outperform single-CNN and transformer-based approaches. Figure 4 further illustrates the fold-accuracy distributions of the different fusion strategies compared with the top-performing individual backbone models.

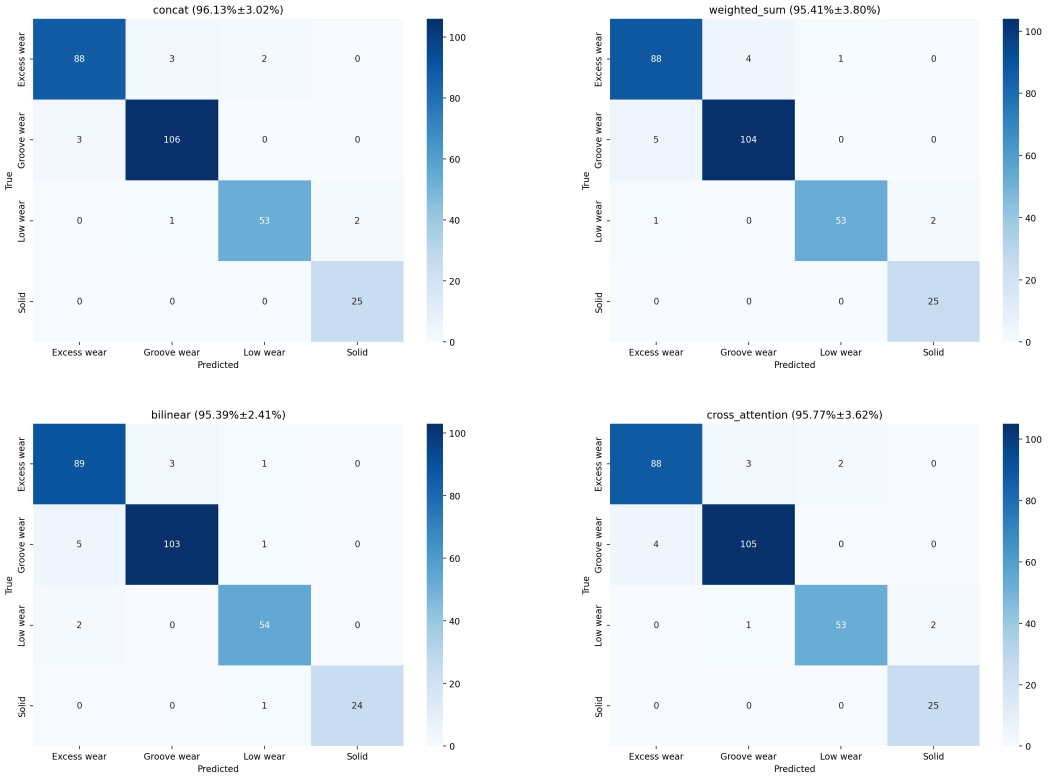

Figure 3. Confusion matrices of the remaining 4 fusion strategies: Concat, Weighted Sum, Bilinear, Cross-Attention.

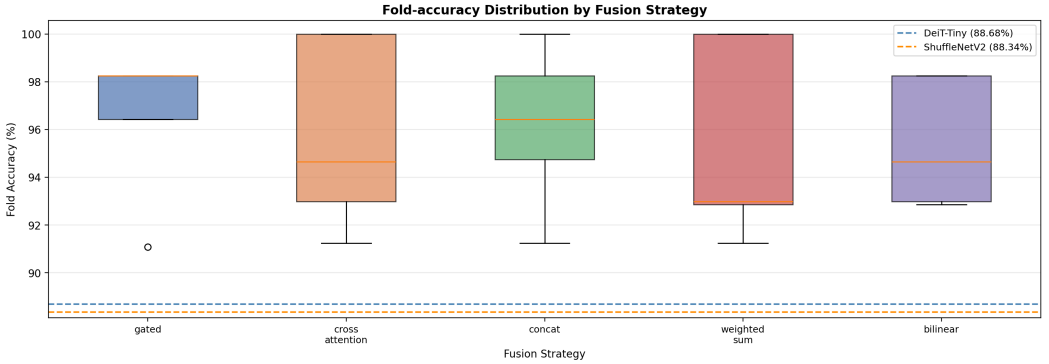

Figure 4. Average accuracy of the top-performing individual models versus fold-accuracy distributions for different fusion strategies. The orange horizontal lines inside the boxplots represent the median fold accuracy across the five cross-validation folds.

4.4. Comparative Analysis of Individual Models

Table 9 summarises the overall performance of pretrained CNN and transformer-based models on the sanitised pantograph wear dataset with and without edge-aware preprocessing using the Sobel transform. Among all evaluated models, ShuffleNetV2 achieved the best performance, with an accuracy of 87.29% followed by DeiT-Tiny (87.29% accuracy) without applying the Sobel transform. The model's performance improvement from using edge-aware preprocessing using the Sobel transform is minimal. Therefore, we chose ShuffleNetV2 and DeiT-Tiny without the Sobel transform as the frozen backbones in the proposed DBFF-Net model.

Although ShuffleNetV2 and DeiT-Tiny achieved a standalone accuracy of 87.29% and 87.27%, the proposed Gated DBFF-Net further improved accuracy to 96.47%, corresponding to an absolute gain of approx 8 percentage points. On a four-class fine-grained industrial dataset containing only 283 unique samples, such an improvement is practically meaningful. More importantly, the benefit of DBFF-Net is not limited to overall accuracy. The proposed model also achieved a balanced F1-score of 0.96 and improved recognition of visually similar wear categories through gated per-feature fusion between complementary CNN and transformer backbones.

Table 9. Model performance (%) across 5-fold cross-validation with and without Sobel preprocessing. Mean and standard deviation are reported along with per-fold accuracy.

Setting	Model	Mean ± Std	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Without Sobel operator	ShuffleNetV2	87.29 ± 1.67	87.72	84.21	87.72	87.50	89.29
	DeiT-Tiny	87.27 ± 2.13	85.96	89.47	89.47	83.93	87.50
	MobileNetV2	86.58 ± 2.60	89.47	85.96	82.46	85.71	89.29
	Swin-Tiny	71.03 ± 4.98	66.67	78.95	64.91	73.21	71.43
	EfficientNet-B3	62.60 ± 9.84	63.16	43.86	64.91	69.64	71.43
With Sobel operator	MobileNetV2	87.28 ± 2.32	85.96	91.23	84.21	87.50	87.50
	ShuffleNetV2	86.57 ± 1.40	87.72	87.72	84.21	85.71	87.50
	DeiT-Tiny	86.55 ± 2.99	87.72	89.47	89.47	83.93	82.14
	Swin-Tiny	55.14 ± 7.12	49.12	45.61	64.91	55.36	60.71
	EfficientNet-B3	46.27 ± 3.94	52.63	49.12	43.86	42.86	42.86

Furthermore, the statistical tests indicate that all evaluated fusion strategies show significant performance differences compared to both DeiT-Tiny and ShuffleNetV2 across the five folds. For Concat, Weighted Sum, Cross-Attention, Bilinear, and Gated fusion methods, the p-values against both baselines are consistently below the conventional significance threshold of 0.05, as shown in Table 10. This suggests that the observed improvements are unlikely to have occurred due to random variation across folds.

Table 10. Paired t-test p-values comparing fusion strategies against single-backbone models.

Fusion	DeiT-Tiny	ShuffleNetV2
Concat	0.0173	0.0015
WeightedSum	0.0286	0.0112
CrossAttention	0.0225	0.0064
Bilinear	0.0076	0.0033
Gated	0.0048	0.0106

Among the methods, Gated fusion (p = 0.0048 vs. DeiT-Tiny, p = 0.0106 vs. ShuffleNetV2) and Bilinear fusion (p = 0.0076 vs. DeiT-Tiny, p = 0.0033 vs. ShuffleNetV2) show clear statistical evidence of improvement. Concat and Cross-Attention also show significant improvement across both baselines. These results indicate that all fusion approaches consistently outperform the individual backbones, with more sophisticated methods such as Gated and Bilinear fusion achieving the most reliable gains.

These findings demonstrate that single-model architectures are insufficient for robust pantograph wear classification under limited and noisy industrial data conditions. The substantial performance gap between CNN-based and dual-backbone models leads to the proposed Dual-Backbone Feature Fusion Ensemble Network (DBFF-Net), which aims to integrate complementary local and global feature representations. Furthermore, the results confirm that dataset sanitization is a critical prerequisite for achieving stable, reproducible, and generalizable deep learning performance in industrial visual inspection tasks.

4.5. Performance Analysis on the Original Dataset

To assess the robustness of the proposed DBFF-Net ensemble approach under realistic, noisy conditions, we also evaluated its performance on the original dataset. Unlike the sanitized dataset, the original data contains duplicated samples, label inconsistencies, and higher inter-class visual similarity. We performed this evaluation on a held-out test set from the original dataset, comprising 138 images across four wear categories (Table 4).

Figure 5 presents the confusion matrix obtained on the original dataset. Compared to the cleaned dataset, a noticeable degradation in classification performance is observed, particularly for visually overlapping wear categories. This outcome is expected, as the original dataset includes redundant samples and inconsistent annotations that impact class separability. Table 11 provides the precision, recall, and F1-score values for individual classes. A significant drop in performance is observed in the original data, highlighting the comparatively low performance impact of noisy and redundant data on the model's overall performance in fine-grained classification of different types of wear.

Table 11. Performance metrics of the proposed DBFF-Net ensemble model on the original (non-cleaned) dataset.

Class	Precision	Recall	F1-Score	Support
Excess wear	0.37	0.54	0.44	28
Groove wear	0.61	0.38	0.47	50
Low wear	0.59	0.70	0.64	33
Solid	1.00	1.00	1.00	27
Macro Avg.	0.64	0.65	0.64	–
Accuracy		0.6087		

It is observed that the solid class achieves better precision and recall, implying that the visual characteristics of intact contact strips are strong and distinctive, even in noisy data. For other types of wear, a high degree of inter-class confusion is noted. For the groove wear type, a comparatively low recall of 0.38 is observed, with many instances misclassified as excess or low wear types, possibly due to difficulty capturing the visual characteristics of longitudinal groove patterns amid redundant images and noisy data. The excess and low wear classes exhibit moderate recall (0.54 and 0.70, respectively) but low precision, indicating overlapping decision boundaries across the stages of progressive wear. These results show the ineffectiveness of severity-based visual cues for discrimination in noisy, challenging datasets.

Furthermore, the ensemble model demonstrated an accuracy of 60.87% and a macro F1-score of 0.64 on the original dataset, which is significantly lower than its performance on the cleaned dataset. This shows the importance of preprocessing and sanitising the dataset for improved model performance on smaller, more challenging datasets.

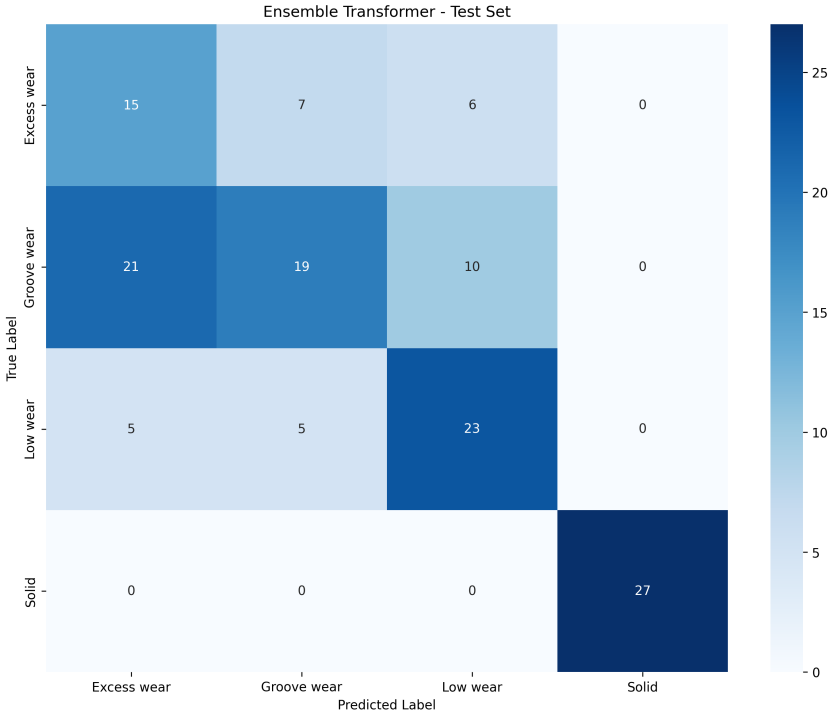

Figure 5. Confusion matrix of the proposed DBFF-Net evaluated on the original (non-cleaned) dataset.

4.6. Contextual Reference to Prior Work

To provide a contextual reference point, we report results from the Hough-enhanced single-stage CNN proposed by Karaduman et al. [2] alongside the proposed DBFF-Net (Table 12). Both studies originate from the same underlying dataset; however, the experimental settings differ significantly due to dataset refinement and evaluation protocol differences introduced in this work. The baseline method in [2] is evaluated on the original dataset containing 771 images using a single train-test split. In contrast, the proposed DBFF-Net is evaluated on a sanitised subset of 283 unique images using a stratified five-fold cross-validation protocol to ensure robustness on the small, sanitised pantograph dataset.

Because of these fundamental differences in dataset structure, sample sizes, and testing procedures, a direct quantitative comparison is not scientifically valid. Consequently, the findings cannot be considered an evaluative benchmark test or an attempt to measure superiority; rather, they must be considered merely relative performance indicators based on varying testing conditions. Under their respective evaluation settings, the Hough-enhanced CNN reports an accuracy of 70.14% with a macro F1-score of 0.71, whereas the proposed Gated DBFF-Net achieves 96.47% accuracy with a macro F1-score of 0.96 on the sanitised dataset.

This performance improvement must be taken into consideration, considering the data processing methods used in the proposed approach, i.e., cleaning and deduplication of the data and the application of a dual-backbone approach for the feature fusion stage, and not compared to other existing methods. This performance gap shows the importance of understanding the dataset and the evaluation method.

Table 12. Contextual reference to prior pantograph wear classification results under different experimental settings.

Method	Accuracy (%)	Macro F1	# Images	Evaluation Protocol
Hough-Enhanced CNN [2]	70.14	0.71	771	Single split
Proposed DBFF-Net	96.47	0.96	283	Stratified 5-fold CV

5. Conclusions and Future Work

In this paper, we explored two challenges related to vision-based pantograph wear recognition, i.e., (i) critical analysis of the dataset to understand input data, and (ii) the comparatively low performance of existing single-model architectures in classifying pantograph detections into fine-grained types. For the dataset analysis, we conduct an in-depth investigation of the commonly used pantograph contact strip dataset and identify various anomalies, including redundancies and label consistencies, based on our knowledge. For the latter challenge, we propose a novel DBFF-Net, a dual-backbone feature fusion network that combines a frozen DeiT-Tiny (192-d) and ShuffleNetV2 (1024-d) with five fusion strategies (Concat, Weighted Sum, Bilinear, Cross-Attention, and Gated) to integrate complementary transformer and CNN representations for improved pantograph wear classification.

To validate the performance of DBFF-Net, we conducted experiments on both re-annotated and original datasets, using various pretrained CNN and transformer models (ShuffleNetV2, DeiT-Tiny, MobileNetV2, Swin-Tiny, and EfficientNet-B3) with and without the Sobel operator. We obtained comparatively better accuracies of 87.29% and 87.27% with ShuffleNetV2 and DeiT-Tiny algorithms, outperforming other individual algorithms. Furthermore, we used a dual-backbone feature fusion approach by combining frozen DeiT-Tiny and ShuffleNetV2 with various strategies (Concat, Weighted Sum, Bilinear, Cross-Attention, and Gated) to evaluate their performance on a small and challenging wear dataset comprising 283 images. The results show that Gated DBFF-Net achieves an overall 8% improvement in pantograph wear classification accuracy compared to individual models.

Despite the positive result, there are some limitations. First of all, though the sanitisation process provides more reliable data, it also leads to a reduction in the overall number of data samples. In addition, the proposed approach relies on manual data relabelling, which is time-consuming, challenging, and error-prone. Also, it causes potential bias, as the author involved in the experimentation is also involved in the data re-annotation. Although we evaluated the re-annotation with multiple annotators who had prior experience. However, it still needs validation from the concerned experts. In addition to this, the evaluation is restricted to a single pantograph wear dataset, and the generalisation of the proposed approach to other railway components or inspection scenarios has not yet been validated.

Future work will focus on extending the proposed framework in several directions. Methods for automated or semi-supervised label verification could be investigated to reduce the need for manual data relabelling while maintaining annotation quality. For this purpose, vision transformers could be used to annotate the dataset as an alternative to human annotation. The inclusion of self-supervised or contrastive learning methods could further mitigate data scarcity by leveraging unlabelled data. Moreover, domain adaptation and cross-dataset analysis should be explored to evaluate the transferability of the proposed DBFF-Net framework to other data acquisition settings, camera configurations, and railway environments. Lastly, the utilisation of temporal information from sequential inspections and the deployment of the proposed framework in real-time settings are promising future directions.

Author Contributions: Conceptualization, N.U., M.Y. and J.A.K.; methodology, N.U. and M.Y.; software, M.Y. and S.I.S.; validation, J.A.K., M.Y., Y.I. and N.U.; formal analysis, S.I.S.; investigation, J.A.K. and Y.I.; resources, J.A.K. and A.M.; data curation, M.Y.; writing—original draft preparation, N.U.; writing—review and editing, M.Y., J.A.K., Y.I. and A.M.; visualization, N.U. and S.I.S.; supervision, M.Y., J.A.K. and A.M.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and code used for this research are available in the Git repository: https://github.com/yaqoobcs/rail-pantograph-classification (updated on 30 April 2026).

Acknowledgments: The authors acknowledge the use of the writing assistance tool (Grammarly v.1.2.208) to improve the writing quality of this paper. Following its use, the authors thoroughly reviewed and revised the content, and they take full responsibility for the final version of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chang, L.; Liu, Z.; Shen, Y. On-line detection of pantograph offset based on deep learning. In *Proceedings of the 2018 IEEE 3rd Optoelectronics Global Conference (OGC), Shenzhen, China, 4–7 September 2018* ; IEEE: Piscataway, NJ, USA, 2018; pp. 159–164.
2. Karaduman, G.; Akin, E. A deep learning based method for detecting of wear on the current collector Strips' surfaces of the pantograph in railways. *IEEE Access* **2020**, *8*, 183799–183812. [CrossRef]
3. Wei, X.; Jiang, S.; Li, Y.; Li, C.; Jia, L.; Li, Y. Defect detection of pantograph slide based on deep learning and image processing technology. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 947–958. [CrossRef]
4. Li, D.; Liu, Z.; Lu, S.; Chang, L. A robust 3-D abrasion diagnosis method of pantograph slipper based on stereo vision. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 9072–9086. [CrossRef]
5. Du, P.J.; Zhang, M.Z. Computer Vision Aided Pantograph Fault Identification Method for Multiple Units. *J. Comput.* **2023**, *34*, 145–152.
6. Kumar, A.; Harsha, S. A systematic literature review of defect detection in railways using machine vision-based inspection methods. *Int. J. Transp. Sci. Technol.* **2025**, *18*, 207–226. [CrossRef]
7. Olivier, B.; Guo, F.; Qian, Y.; Connolly, D.P. A Review of Computer Vision for Railways. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 11034–11065. [CrossRef]
8. Tastimur, C.; Karaduman, G.; Akin, E. A novel method based on deep learning and image processing techniques for wearing inspection on the pantograph surface. In *Proceedings of the 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 6–8 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–7.
9. Chen, R.; Lin, Y.; Jin, T. High-speed railway pantograph-catenary anomaly detection method based on depth vision neural network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1502710. [CrossRef]
10. Yang, X.; Zhou, N.; Liu, Y.; Quan, W.; Lu, X.; Zhang, W. Online pantograph-catenary contact point detection in complicated background based on multiple strategies. *IEEE Access* **2020**, *8*, 220394–220407. [CrossRef]
11. Liu, L.; Liu, Q.; Wang, W.; Yu, Z.; Zhao, X. Pantograph structure anomaly detection based on computer vision. In *Proceedings of the 2023 6th International Conference on Electronics Technology (ICET), Chengdu, China, 12–15 May 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1145–1150.
12. Yao, X.; Liu, W.; Zhang, Z.; Xing, Z.; Sheng, A. PSPMoni: A Robust Wear Monitoring Method of Pantograph Slide Plate. *Measurement* **2025**, *239*, 115460. [CrossRef]
13. Na, K.M.; Lee, K.; Kim, H. Condition monitoring of railway pantograph using r-CNN and image processing. *J. Electr. Eng. Technol.* **2023**, *18*, 2407–2416. [CrossRef]
14. Lu, Z.; Lu, B.; Wang, F. CausalSR: Structural causal model-driven super-resolution with counterfactual inference. *Neurocomputing* **2025**, *646*, 130375. [CrossRef]
15. Song, Y.; Mei, G.; Liu, Z.; Gao, S. Assessment of railway pantograph-catenary interaction performance with realistic pantograph strip imperfection. *Veh. Syst. Dyn.* **2024**, *62*, 2355–2374. [CrossRef]
16. Zhou, N.; Zhi, X.; Cheng, Y.; Sun, Y.; Wang, J.; Gu, Z.; Li, Z.; Zhang, W. Contact strip of pantograph heuristic wear model and its application. *Tribol. Int.* **2024**, *194*, 109546. [CrossRef]

17. Gregori, S.; Tur, M.; Gil, J.; Fuenmayor, F. Assessment of catenary condition monitoring by means of pantograph head acceleration and Artificial Neural Networks. *Mech. Syst. Signal Process.* **2023**, *202*, 110697. [CrossRef]
18. Han, Z.; Feng, Q.; Liu, W.; Yang, H.; Cui, Y.; Li, H.; Shao, Y. FC-DRL-based framework considering wheel-rail excitation: A speed-interval-oriented active control scheme for high-speed railway pantographs. *Appl. Intell.* **2025**, *55*, 1110. [CrossRef]
19. Han, Z.; Feng, Q.; Liu, W.; Liu, Y.; Yang, H.; Li, H.; Xu, M.; Xiao, S. An Optimized Active Compensation Control Framework for High-Speed Railway Pantograph via Imitation-Guided Deep Reinforcement Learning. *Machines* **2025**, *13*, 769. [CrossRef]
20. Sharma, R.; Mahajan, P.; Garg, R. Deep-reinforcement-learning-based controller design for pantograph and catenary system. *Sādhanā* **2025**, *50*, 46. [CrossRef]