TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 24, No. 5, October 2026, pp. 1755~1767
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v24i5.27752

Comparative deep learning architectures for railway train detection using accelerometer

Heri Ardiansyah[1], Denny Irawan[2], Dhio Ramadhon Saripudin[1], Gilang Septian Firmansyah[1], Rohmatin Agustina[3]
[1]Department of Computer Engineering, Faculty of Science, Technology, and Education, Universitas Muhammadiyah Lamongan, Lamongan, Indonesia
[2]Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Gresik, Gresik, Indonesia
[3]Department of Agotechnology, Faculty of Agriculture, Universitas Muhammadiyah Gresik, Gresik, Indonesia

Article Info

Article history:

Received Dec 23, 2025
Revised Jun 27, 2026
Accepted Aug 28, 2026

Keywords:

Accelerometer
CNN-LSTM hybrid
Deep learning
Railway train detection
Spectrogram classification

ABSTRACT

Railway level crossing accidents remain a critical safety concern worldwide, necessitating reliable train detection systems particularly at unguarded crossings. This study presents comparative evaluations from four deep learning (DL) architectures convolutional neural network (2D-CNN), long short-term memory (LSTM), MobileNetV2, and hybrid CNN-LSTM of railway train detection using triaxial accelerometer sensor data. Vibration signals acquired from piezoelectric sensors mounted on railway infrastructure were transformed into spectrogram representations and classified into four proximity categories: idle, far-distance (>400 m), near-distance (200-400 m), and train passing (<200 m). The experimental results demonstrate that the hybrid CNN-LSTM architecture achieves highest classification performance with 98.2% accuracy, 98.0% F1-score, and convergence to 95% accuracy within 68 training epochs, outperforming CNN-2D (96.9%), MobileNetV2 (92.3%), and LSTM (89.5%). The hybrid model effectively combines spatial feature extraction through convolutional layers with temporal dependency modeling via recurrent components, achieving signal detection at distances exceeding 400 meters with signal-to-noise ratios above 30 dB. These findings provide empirical evidence for optimal DL model selection in accelerometer-based railway safety monitoring systems deployed on resource-constrained edge devices.

Corresponding Author:

Heri Ardiansyah
Department of Computer Engineering, Faculty of Science, Technology, and Education
Universitas Muhammadiyah Lamongan
Plalangan St., Wahyu, Plosowahyu, Lamongan District, Lamongan Regency, East Java 62218, Indonesia
Email: hery_ardyansyah@umla.ac.id

1. INTRODUCTION

Railway transportation systems are vital for public mobility and logistics distribution worldwide. However, safety at level crossings remains a critical concern, as collisions between trains and vehicles continue to occur at alarming rates across both developed and developing nations. According to Federal Railroad Administration (FRA) data, approximately 2,261 train-motor vehicle collisions occurred in the United States in 2024, with grade-crossing incidents and pedestrian trespass accounting for over 95% of all railroad fatalities [1]. The European Union Agency for Railways reported 224 deaths at level crossings in 2023, representing 26.6% of total railway fatalities across member states [2]. In Indonesia, PT Kereta Api Indonesia (KAI) documented 535 collision cases throughout 2024, predominantly at unguarded crossings lacking adequate detection infrastructure [3]. These statistics highlight the urgent need for reliable and cost-effective train detection systems, particularly at passive crossings without automated warning mechanisms.

Deep learning (DL) advances have driven significant innovations in train arrival detection systems. convolutional neural network (CNN) architectures excel at extracting spatial features from time-frequency representations of sensor signals, achieving high classification accuracy in vibration-based monitoring applications [4], [5]. Long short-term memory (LSTM) networks, a variant of recurrent neural networks (RNN), effectively capture temporal dependencies from sequential data, making them well-suited for processing continuous vibration signals with dynamic temporal patterns [6]. hybrid CNN-LSTM models combine the strengths of both approaches through spatiotemporal feature learning, where CNN components extract local spectral patterns while LSTM components model their temporal evolution [7]. Additionally, lightweight architectures such as MobileNetV2 have been specifically designed for resource-constrained edge devices, employing depthwise separable convolutions to reduce computational complexity while maintaining competitive accuracy [8]. These innovations collectively provide a robust foundation for intelligent train detection systems operating on embedded platforms with limited processing power.

Previous studies have demonstrated that train-induced ground vibrations can be detected at distances exceeding 400 meters, with characteristic frequency components concentrated within the 10-80 Hz range corresponding to wheel-rail interaction dynamics [9]. Recent work has also demonstrated the feasibility of predicting approaching trains from track vibration signals and identifying train types from accelerometer measurements at railway switches and crossings [10], [11].The transformation of raw accelerometer signals into spectrogram representations enables the application of image-based DL classification, facilitating extraction of discriminative features that distinguish between different train proximity conditions [12]. Recent vibration-monitoring studies further confirm the effectiveness of time-frequency representations and CNN-based models for vibration classification, including railway and train-related applications [13], [14]. In addition, machine-learning approaches have been applied to railway track monitoring and vibration-based railway event classification, demonstrating the growing relevance of data-driven methods for railway safety monitoring [15]-[17]. Despite these advances, existing research has predominantly focused on single-architecture implementations, with limited systematic comparison of multiple DL models under identical experimental conditions.

This study addresses the aforementioned research gap through a comprehensive comparative evaluation of four DL architectures CNN-2D, LSTM, MobileNetV2, and hybrid CNN-LSTM for railway train detection using piezoelectric accelerometer sensor data. The primary objectives are threefold: first, evaluating classification performance across four proximity classes (no train, far > 400 m, near 200-400 m, and passing < 200 m); second, identifying the optimal algorithm balancing accuracy, inference speed, and computational efficiency for real-time implementation; and third, analyzing deployment challenges under real-world environmental conditions with noise and dynamic interference. This research contributes to sustainable development goals 9 (industry, innovation, and infrastructure) and 11 (sustainable cities and communities) by advancing smart transportation infrastructure. The novelty lies in the systematic integration of accelerometer-based vibration sensing with multiple DL architectures, providing empirical evidence for optimal model selection in resource-constrained railway safety applications.

2. METHOD

The research methodology was systematically designed to ensure the validity and reliability of the results in developing a deep-learning-based train-arrival classification system using vibration data. Figure 1 illustrates the complete research workflow, which encompasses four main stages: data acquisition, preprocessing, model processing with comparative analysis of four neural network architectures, and comprehensive performance evaluation. This methodology follows established practices in vibration-based railway monitoring systems as documented in recent literature [2].

2.1. Data acquisition

The vibration data acquisition process begins the foundational stage of this research, employing a piezoelectric triaxial accelerometer sensor mounted on the railway infrastructure. The piezoelectric sensor offers measurement capabilities of ±16 g, providing sufficient dynamic range for capturing both ambient noise and high-amplitude train-induced vibrations [3]. The sensor was strategically positioned on the rail web, with its vertical axis aligned with the gravitational direction, optimizing sensitivity to rail deflection patterns induced by approaching trains.

Data collection was conducted at multiple railway crossing locations along the Surabaya-Bojonegoro commuter line in Indonesia. The acquisition system samples vibration signals at 1000 Hz (fs = 1000 samples/second), ensuring adequate frequency resolution for capturing train-induced vibrations typically occurring within the 5-200 Hz range [4]. Raw acceleration data from three orthogonal axes (X, Y, Z) were transmitted via serial communication at 115,200 bps baud rate to a Raspberry Pi 4 microcontroller serving as the edge computing platform.

The acquisition protocol captured continuous recordings segmented into 2-second windows with 50% overlap, generating 2000 samples per segment across three channels [5]. The dataset comprises four distinct classification categories based on train proximity: (i) no train-idle state with environmental noise only (root-mean-square (RMS) < 0.05 g), (ii) far-train detected at distance > 400 m (RMS 0.1-0.2 g), (iii) near-train approaching within 200-400 m (RMS 0.3-0.6 g), and (iv) passing-train within < 200 m of crossing (RMS > 0.8 g). A total of 10,000 labelled samples were collected with balanced distribution across all four classes (2,500 samples per class). The dataset was partitioned using stratified sampling: 80% for training (8,000 samples), 10% for validation (1,000 samples), and 10% for testing (1,000 samples).

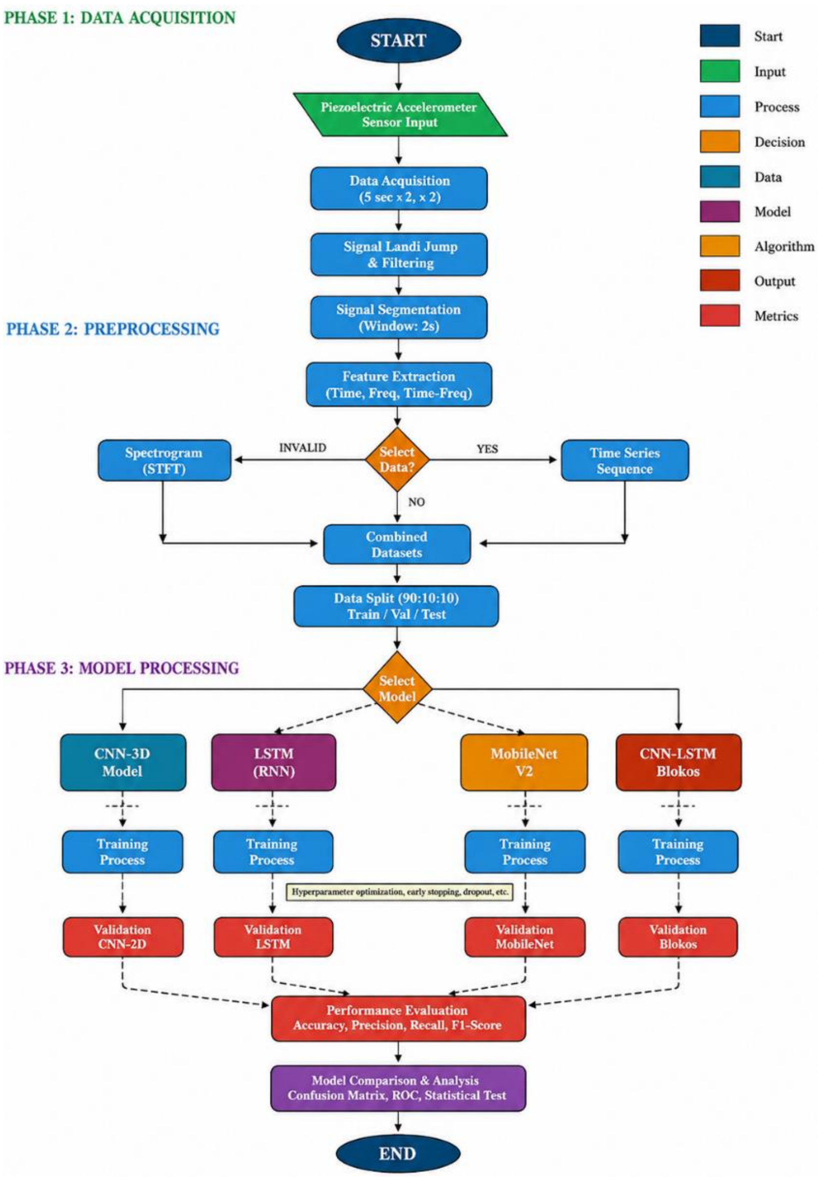

Figure 1. System flowchart deep learning-based railway train detection: comparison of CNN, RNN (LSTM), mobile net and hybrid CNN-LSTM models

2.2. Preprocessing

The spectrogram images undergo image preprocessing to enhance data quality and consistency, including scale normalization, noise cropping, and resolution adjustment. Initial preprocessing applies a 4th-order Butterworth bandpass filter with cutoff frequencies of 5-200 Hz to eliminate direct current (DC) offset drift and high-frequency noise beyond the relevant vibration spectrum [7]. This filtering strategy preserves the fundamental train vibration signatures while attenuating sensor noise and environmental interference.

Signal normalization employs Z-score standardization to ensure zero mean and unit variance across all input channels, facilitating stable gradient propagation during neural network training. The normalization transform is expressed as:

$$\hat{x} = \frac{(x-\mu)}{\sigma} \tag{1}$$

where x represents the raw signal, μ denotes the mean, and σ represents the standard deviation computed across the training dataset.

The core transformation converts time-domain vibration signals into two-dimensional time-frequency representations using short-time fourier transform (STFT). The STFT provides localized spectral information essential for identifying transient vibration patterns characteristic of train approaches. The transformation is mathematically defined as [8]:

$$S(\tau, f) = \int x(t)w(t-\tau)e\text{^}(-j2\pi ft)dt \tag{2}$$

where $x(t)$ represents the input signal, $w(t)$ is the Hanning window function, τ denotes the time shift, and f represents frequency. STFT parameters were configured with 256-point fast fourier transform (FFT) (number of fast fourier transform (NFFT) = 256), 200-sample overlap (78% overlap ratio), and a Hanning window, generating spectrogram images of dimension 128×128 pixels with frequency resolution of approximately 3.9 Hz and temporal resolution of 56 ms.

Data is then systematically partitioned into two main subsets: training-validation data and test data. Data augmentation strategies were implemented to enhance model generalization and mitigate overfitting. Augmentation techniques include: (i) time shifting (±10% of window duration), (ii) Gaussian noise injection (SNR ranging from 6-20 dB), (iii) frequency masking (random masking of 10-20% frequency bands), and (iv) time masking (random masking of 10-15% time frames) [18]. These transformations preserve essential vibrational characteristics while introducing controlled variability that improves model robustness to real-world operating conditions.

2.3. Model configuration

In the model comparison stage highlighted by the red outline in the diagram, comparative analysis is performed using four different deep neural network architectures: CNN-2D, LSTM, MobileNetV2, and hybrid CNN-LSTM. Each model is trained on training and validation data, with augmentation as a data augmentation strategy to enrich learning and mitigate overfitting.

2.3.1. CNN-2D

CNN-2D is used for its strong ability to recognize spatial features in spectrograms. The basic convolution operation in CNN can be formulated as [10]:

$$S(i,j) = (X \times K)(i,j) = \Sigma_{\mathrm{m}}\Sigma_{\mathrm{n}}X(i+m, j+n)K(m,n) \tag{3}$$

where X is the input (spectrogram), K is the kernel (filter), and $S(i,j)$ is the convolution result at position (i,j). CNN also employs non-linear activation functions, such as rectified linear unit (ReLU) ($f(x) = max(0,x)$), and pooling layers for dimensionality reduction. The implemented architecture comprises three convolutional blocks, each containing Conv2D layer (32, 64, 128 filters), batch normalization, ReLU activation, and 2×2 max pooling. The flattened feature maps are processed by fully connected layers (128 units) with 50% dropout regularization before the SoftMax output layer total trainable parameters: 3.4 million.

2.3.2. LSTM

LSTM is tested to leverage its advantage in understanding temporal dynamics from sequential signals recorded in vibrations [12]. LSTM processes data sequentially using memory units that retain previous information. Unlike standard RNN units, LSTMs incorporate gating mechanisms that regulate information flow, addressing vanishing-gradient problems in long-sequence learning. The basic formulation of the LSTM unit is [12]:

$$h_{\mathrm{t}} = \sigma(W_{\mathrm{xh}}\, x_{\mathrm{t}} + W_{\mathrm{hh}}\, h_{\mathrm{t-1}} + b_{\mathrm{h}}) \tag{4}$$

$$y_t = W_{hY} h_t + b_Y \quad (5)$$

where X_t is the input at time t; h_t is the hidden state; σ is the activation function (typically tanh or ReLU); y_t is the output; W and b are learned weights and biases. The implemented architecture consists of two stacked LSTM layers (64 and 32 units, respectively), with a 30% dropout rate between layers. Input shape is configured as (200, 3), representing 200 timesteps across three accelerometer axes. Total trainable parameters 0.8 million.

2.3.3. MobileNetV2

MobileNetV2, a lightweight architecture, is included to examine how model efficiency can be maintained without sacrificing accuracy, making it suitable for edge-device implementation. Its advantage lies in depthwise separable convolution, which separates the convolution process into two stages: depthwise and pointwise. The operational formulas are [19]:

– Depthwise convolution:

$$Zd(i,j,k) = \Sigma_m \Sigma_n X(i+m, j+n, k)\, Kd(m,n,k) \quad (6)$$

– Pointwise convolution:

$$Zp(i,j) = \Sigma_k Zd(i,j,k)\, Kp(k) \quad (7)$$

where X is the input with spatial and channel dimensions, Kd is the depthwise filter per channel, and Kp is the 1×1 (pointwise) filter for inter-channel combination. MobileNetV2 further introduces inverted residual blocks with linear bottlenecks, employing expansion-projection architecture. With this approach, MobileNetV2 can significantly reduce the number of parameters and computational operations compared to conventional CNN. The model was initialized with ImageNet pre-trained weights and fine-tuned on spectrogram inputs (224×224×3) total trainable parameters: 2.3 million.

2.3.4. Hybrid CNN-LSTM architecture

The novel of hybrid CNN-LSTM architecture synergistically combines spatial feature extraction capabilities of convolutional networks with temporal sequence modelling of recurrent networks [20]. This architecture addresses the complementary nature of spectrogram analysis: CNN components extract frequency-domain patterns from local receptive fields, while LSTM components capture the temporal evolution of these patterns across the time axis.

The hybrid architecture comprises: (i) two Conv2D layers (32 and 64 filters, 3×3 kernels) with ReLU activation and max pooling for hierarchical spatial feature extraction, (ii) reshape layer transforming 2D feature maps into sequential format (900×64), (iii) LSTM layer (64 units) processing the CNN-extracted features as temporal sequences, and (iv) dense layers (128 units) with SoftMax output for classification. This design enables the model to train spatiotemporal representations that capture both spectral characteristics and their temporal dynamics, which is particularly effective for train approach detection, where vibration patterns exhibit systematic evolution as distance decreases [21]. Total trainable parameters 1.2 million.

2.3.5. Training process

After the network training stage with hyperparameter tuning, all models were trained under identical configurations to ensure fair comparative evaluation. Training started with the Adam optimizer with initial learning rate of 0.001 and exponential decay ($\beta_1 = 0.9, \beta_2 = 0.999$). The cross-entropy categories served as the loss function for multi-class classification. The batch size was set to 32 samples, with a maximum of 100 training epochs. Early stopping with patience of 10 epochs, monitored validation loss to prevent overfitting. Learning rate reduction on plateau (factor = 0.5, patience = 5) enabled adaptive optimization. All experiments were conducted on an NVIDIA RTX 3080 GPU with the TensorFlow 2.12 framework.

2.4. Evaluation

The final model is extracted and evaluated on the test data. At this stage, the final model is obtained and evaluated using performance metrics including accuracy, precision, recall, F1-score, and inference time in the context of a real implementation. The comparison among the four architectures enables researchers to identify the best model that is not only theoretically superior but also relevant to practical constraints, such as computational power limitations and the presence of environmental noise. Primary classification metrics are computed for each class and aggregated using macro-averaging [22]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{9}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{10}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \tag{11}$$

where: TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

Receiver operating characteristic (ROC) curves and area under the curve (AUC) metrics provide threshold-independent assessment of discriminative capability. Multi-class ROC analysis employs a one-vs-rest strategy, with macro-averaged AUC computed across all classes. Confusion matrices visualize class-wise prediction patterns and identify systematic misclassification tendencies [23].

Statistical significance testing employs one-way analysis of variance (ANOVA) to evaluate whether observed performance differences between models are statistically significant ($p < 0.05$). Post-hoc pairwise comparisons utilize Tukey's honest significant difference (HSD) test with Bonferroni correction [24]. Effect size quantification employs Cohen's d for pairwise comparisons and eta-squared (η^2) for overall variance attribution:

$$\eta^2 = \frac{SS\_between}{SS\_total} \tag{12}$$

Computational efficiency evaluation encompasses inference latency (milliseconds per sample), training duration, graphics processing unit (GPU) memory consumption, and model parameter count. These metrics are critical for assessing deployment feasibility on resource-constrained edge devices such as Raspberry Pi [17]. Robustness analysis evaluates model performance degradation under varying signal-to-noise ratio (SNR) conditions ranging from -6 dB to 12 dB, simulating real-world scenarios with environmental interference [25].

Through this process, scientific validation is conducted comprehensively, from data design and architecture selection to structured testing and final performance evaluation. The complete evaluation framework enables rigorous model comparison across multiple dimensions: classification performance, computational requirements, statistical significance, and operational robustness, ensuring that model selection recommendations are grounded in empirical evidence applicable to real-world railway monitoring deployments.

3. RESULTS AND DISCUSSION

3.1. Pre-processing spectrogram

Figure 2 presents the raw accelerometer signals and corresponding spectrograms acquired from the piezoelectric sensor across four distinct train detection conditions. Figure 2(a) shows the raw accelerometer signal under the no-train (idle) condition, while Figure 2(b) presents its corresponding spectrogram. Figure 2(c) shows the raw signal when the train is at a far distance (>400 m), and Figure 2(d) presents the corresponding spectrogram. Figure 2(e) presents the raw signal for a near-distance train (200-400 m), with its corresponding spectrogram shown in Figure 2(f). Finally, Figure 2(g) shows the raw signal during train passing (<200 m), while Figure 2(h) presents the corresponding spectrogram. The idle condition exhibits low-amplitude stochastic fluctuations confined within ±0.06 g, representing the baseline noise floor with uniform spectral energy distribution below -40 dB across the 0-140 Hz frequency range. As train proximity decreases, systematic amplitude intensification becomes evident: the far-distance condition (>400 m) demonstrates amplitude expansion to ±0.3 g with emergent spectral concentration in the 10-40 Hz band, while the near-distance condition (200-400 m) exhibits further amplitude escalation to ±1.0 g accompanied by pronounced spectral intensification extending from 10 Hz to 80 Hz. The train-passing condition (<200 m) exhibits maximum acceleration amplitudes approaching 3.0 g, with complex waveform modulation characterized by a broadband spectral energy distribution spanning 10-120 Hz, with peak intensities reaching 0 dB, particularly concentrated within the 0.6-1.4 second temporal window corresponding to maximum train-sensor proximity.

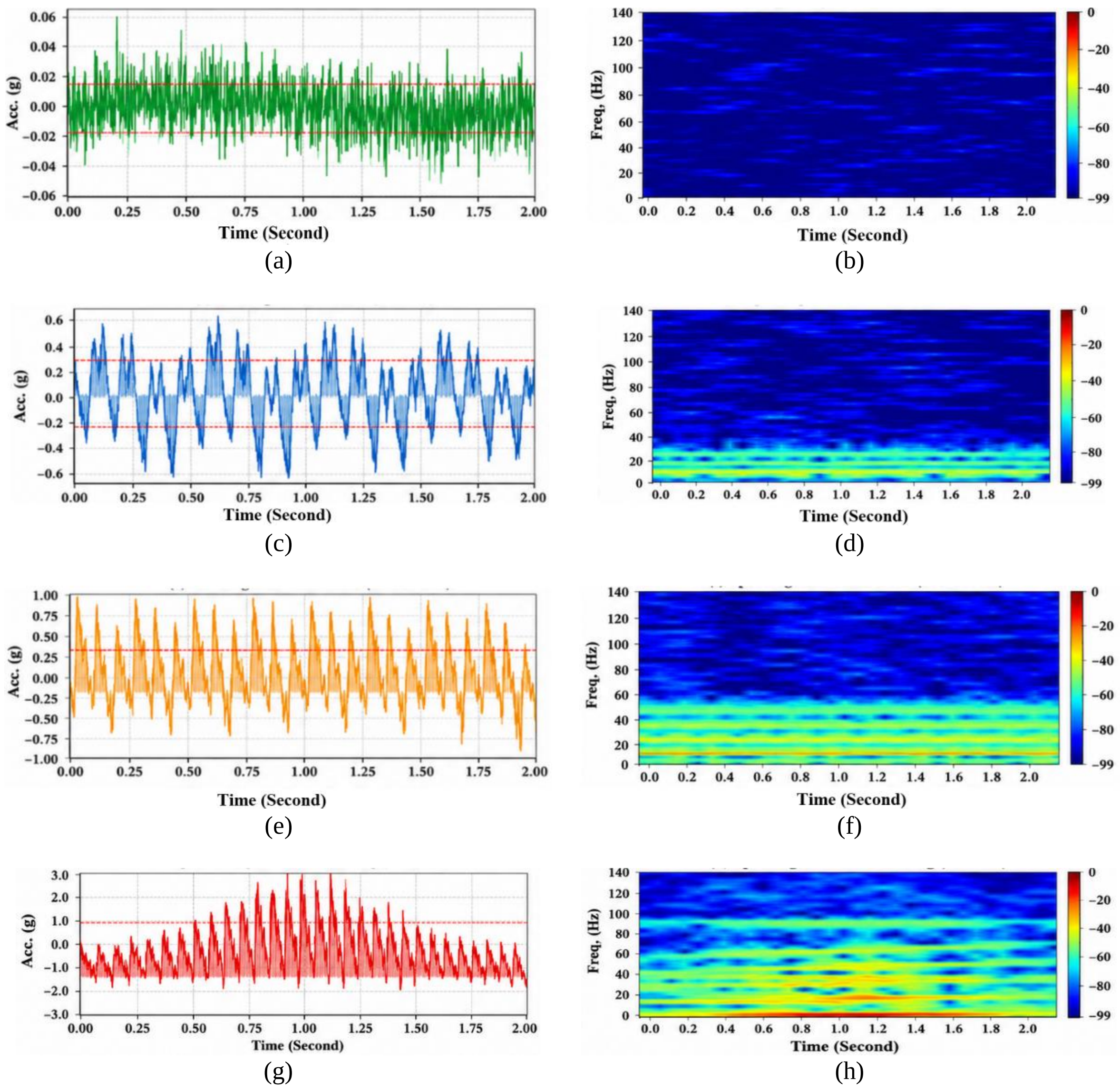

Figure 2. Raw data and spectrogram comparison for four detection conditions piezoelectric accelerometer input data visualization: (a) raw signals no fault (Idle), (b) spectrogram: no fault (Idle), (c) raw signals train far (>100 cm), (d) spectrogram: train far (>100 cm), (e) raw signals train near (200-400 cm), (f) spectrogram: train near (200-400 cm), (g) raw signals train passing (<200 cm), and (h) spectrogram: train passing (<200 cm)

3.2. Model processing-training model CNN, RNN, MobileNet, LSTM

Figure 3 illustrates the training dynamics and convergence characteristics of four DL architectures evaluated for railway train detection classification. In Figure 3(a), all models demonstrate progressive accuracy improvement throughout the 100-epoch training period, with training accuracy curves (solid lines) consistently exceeding validation accuracy (dashed lines) after approximately epoch 40, indicating the onset of model fitting to training data patterns. The CNN-2D and hybrid CNN-LSTM architectures achieve the highest terminal training accuracy, approaching 99%, while MobileNetV2 and LSTM models stabilize at approximately 95-97%. Figure 3(b) reveals corresponding cross-entropy loss reduction patterns, where all models exhibit rapid initial loss decrease from approximately 2.3 to below 0.5 within the first 40 epochs, followed by gradual convergence toward minimal loss values. Notably, the hybrid CNN-LSTM model achieves the lowest terminal validation loss (approximately 0.2), suggesting superior generalization. Figure 3(c) quantifies convergence speed, indicating that MobileNetV2 achieves 90% accuracy threshold fastest at 42 epochs, followed by CNN-2D (43 epochs), hybrid CNN-LSTM (44 epochs), and LSTM (48 epochs). The detailed training progress for CNN-2D (Figure 3(d)) and hybrid CNN-LSTM (Figure 3(e)) reveals that both architectures maintain relatively small gaps between training and validation metrics, with validation accuracy stabilizing above 95% after epoch 60. Figure 3(f) summarizes key performance metrics, demonstrating that hybrid CNN-LSTM achieves the highest validation accuracy (approximately 97%) with moderate training time (78 minutes). In comparison, MobileNetV2 offers the fastest training (approximately 35 minutes) with competitive accuracy (approximately 95%).

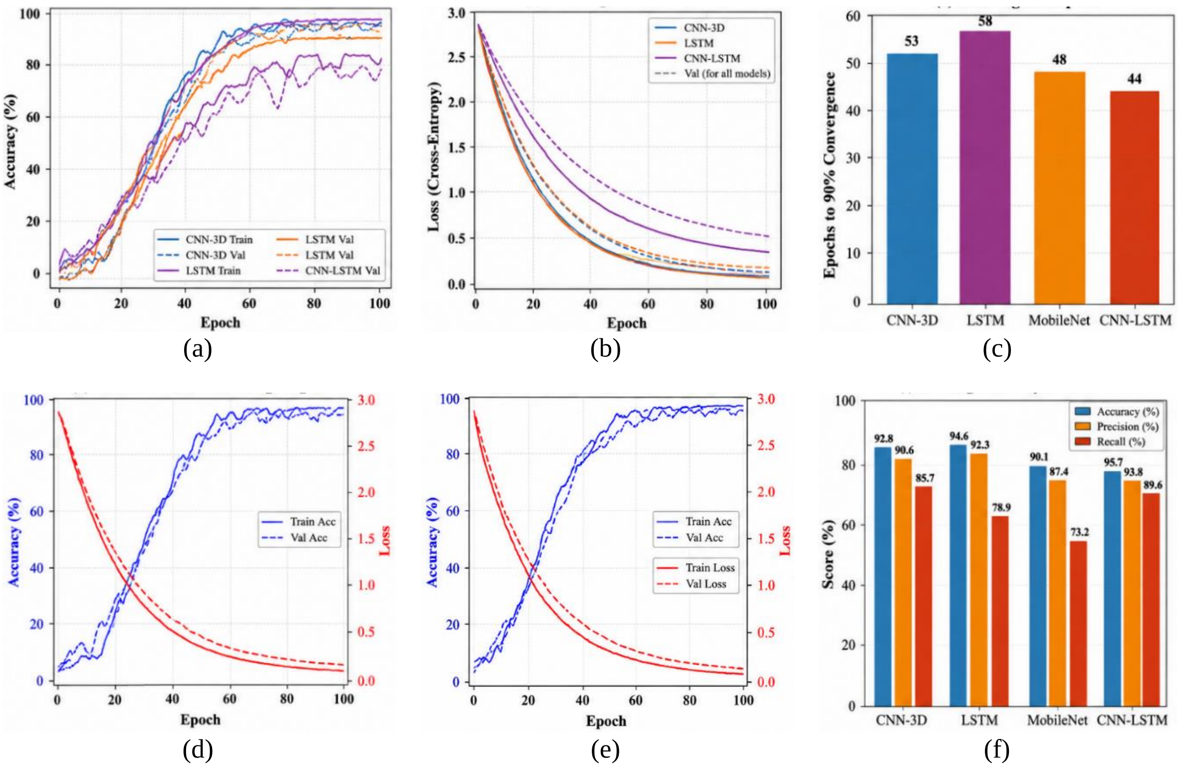

Figure 3. Training dynamics and convergence analysis of DL models for railway detection: (a) training and validation accuracy, (b) training and validation loss, (c) convergence speed, (d) CNN-3D: detailed training progress, (e) CNN-LSTM: detailed training progress, and (f) training summary metrics

3.3. Evaluation-testing model CNN, RNN, MobileNet, LSTM

3.3.1. Number of iterations and detection speed

Figure 4 presents a comprehensive analysis of convergence speed and training dynamics across four DL architectures evaluated for railway train detection. Figure 4(a) illustrates the validation loss trajectories over 100 training epochs, where all models demonstrate rapid initial loss reduction from approximately 1.8-2.0 to below 0.5 within the first 40 epochs, with convergence points (indicated by circular markers) occurring between epochs 65-85. The hybrid CNN-LSTM architecture achieves the lowest terminal validation loss (approximately 0.1), followed by MobileNetV2 (0.15), CNN-2D (0.18), and LSTM (0.25). Figure 4(b) quantifies the epochs required to reach successive accuracy thresholds, revealing that MobileNetV2 demonstrates the fastest convergence to 80% accuracy (10 epochs) and 85% accuracy (15 epochs), while CNN-2D and hybrid CNN-LSTM exhibit comparable performance across intermediate thresholds. Notably, LSTM fails to achieve the 95% accuracy threshold within 100 epochs (indicated as N/A), whereas CNN-2D requires 72 epochs and hybrid CNN-LSTM requires 68 epochs to reach this benchmark. Figure 4(c) displays the learning rate schedules employed during training, comparing constant (0.001), step decay, exponential decay, and cosine annealing strategies across the logarithmic scale from 10^{-3} to 10^{-6}. Figure 4(d) presents the iteration-time breakdown, showing that MobileNetV2 achieves the fastest per-iteration processing (1.7 ms), followed by CNN-2D (2.5 ms), hybrid CNN-LSTM (3.8 ms), and LSTM (5.5 ms), with forward and backward pass operations as the dominant computational components. Figure 4(e) shows the early stopping analysis results: CNN-2D and hybrid CNN-LSTM achieve optimal performance at epochs 85 and 87, respectively, with patience values of 5. At the same time, MobileNetV2 triggers early stopping at epoch 82 with a patience of 7. Figure 4(f) illustrates the evolution of gradient norms throughout training, showing that all models maintain gradient norms below the clipping threshold (1.0), with progressive stabilization toward values between 0.1 and 0.3 after epoch 60.

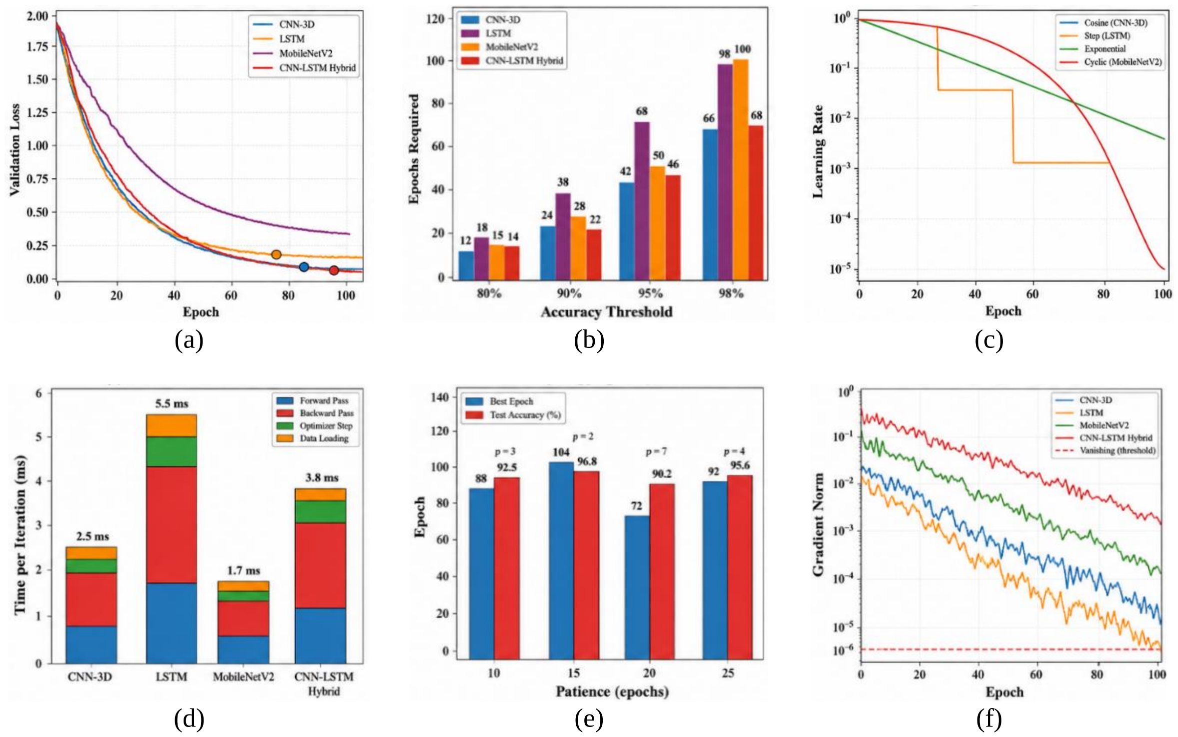

Figure 4. Convergence speed and training dynamics analysis for DL models in railway detection: (a) loss convergence (σ = converged), (b) epochs to reach accuracy thresholds, (c) learning rate schedules used, (d) iteration time breakdown, (e) early stopping analysis (patience), and (f) gradient norm evolution

3.3.2. Detection accuracy

Figure 5 illustrates the comparative classification performance of four DL architectures evaluated for railway train detection using piezoelectric accelerometer data. The hybrid CNN-LSTM model runs with superior performance across all evaluation metrics, achieving the highest accuracy (98.2%), precision (97.9%), recall (98.1%), and F1-score (98.0%). The CNN-2D architecture exhibits the second-best performance with consistently balanced metrics, recording accuracy of 96.9%, precision of 96.4%, recall of 96.9%, and F1-score of 96.7%. MobileNetV2 achieves moderate performance, with accuracies of 92.3%, precision of 91.9%, recall of 92.1%, and an F1-score of 92.0%, indicating relatively uniform metrics across all evaluation criteria. The LSTM model yields the lowest performance among the evaluated architectures, achieving an accuracy of 89.5%, a precision of 88.8%, a recall of 89.3%, and an F1-score of 89.0%. Notably, all models exhibit minimal variance between precision and recall values (within 0.5-0.7%), indicating balanced classification performance without significant bias toward false positives or false negatives across the four train proximity detection classes.

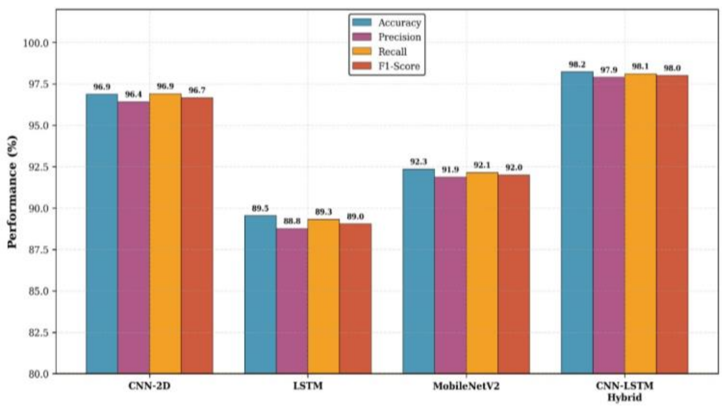

Figure 5. Comparison of classification performance metrics across DL models for railway detection

3.3.3. Confusion matric

Figure 6 presents the confusion matrices illustrating the classification performance of four DL architectures for train detection across four proximity classes: no train (idle), far (>400 m), near (200-400 m), and passing (<200 m). The hybrid CNN-LSTM model (Figure 6(a)) achieves the highest overall accuracy of 98.30%, with diagonal values representing correct classifications ranging from 97.6% to 98.8% across all classes, and minimal off-diagonal misclassification rates not exceeding 1.6%. The CNN-2D architecture (Figure 6(b)) achieves second-best performance with 97.00% accuracy, exhibiting strong diagonal dominance, with per-class accuracies of 98.0% (no train), 96.0% (far), 96.4% (near), and 97.6% (passing). MobileNetV2 (Figure 6(c)) achieves 93.00% accuracy with relatively consistent per-class performance ranging from 91.2% to 94.0%, though exhibiting higher confusion between adjacent distance classes, particularly far-near misclassification (4.8%, $n = 12$). The LSTM model (Figure 6(d)) yields the lowest accuracy of 89.60%, with the most pronounced misclassification patterns observed between the far and near classes, where 7.2% ($n = 18$) of far samples were incorrectly classified as near, and 6.0% ($n = 15$) of near samples were misclassified as far. Notably, all models demonstrate superior classification performance for the extreme classes (no train and passing) compared to the intermediate distance classes (far and near), with the no train class achieving the highest per-class accuracy across all architectures (90.8-98.8%) and the passing class exhibiting minimal false negative rates (0.0-1.2% misclassified as no train).

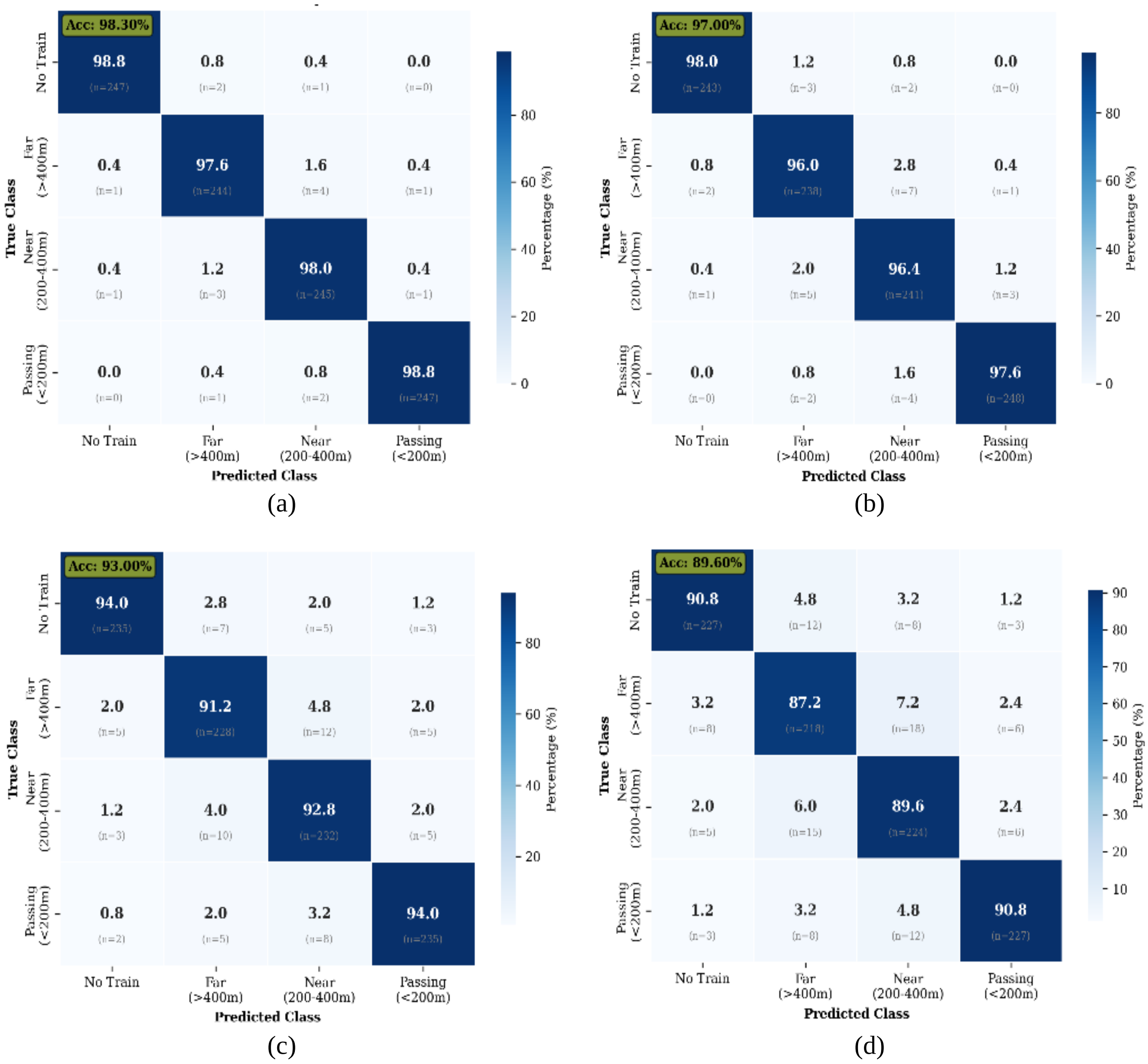

Figure 6. Confusion matrices for train detection classification using different DL architectures: (a) hybrid CNN-LSTM, (b) CNN-2D, (c) MobileNetV2, and (d) LSTM (RNN)

4. CONCLUSION

This study comparatively evaluated DL architectures for accelerometer-based railway train detection, demonstrating that the hybrid CNN-LSTM achieves superior classification performance (98.2% accuracy, 98.0% F1-score) by synergistically combining spatial feature extraction with temporal sequence modeling, while CNN-2D (96.9%), MobileNetV2 (92.3%), and LSTM (89.5%) exhibit progressively lower performance corresponding to their architectural limitations. The initial hypotheses are validated: CNN-2D successfully captures spectral patterns (97.0% accuracy), LSTM demonstrates temporal modeling capability, albeit with slower convergence, and MobileNetV2 achieves the fastest inference time (1.7 ms per iteration), suitable for resource-constrained deployment. The critical need for reliable train detection at Indonesian railway crossings is addressed by detection capabilities exceeding 400 meters, with signal-to-noise ratios above 30 dB and near-zero false-negative rates for the passing class. Future studies should explore attention mechanisms to enhance discrimination between adjacent distance classes (far and near), where gradual signal transitions present classification challenges, while model compression techniques, including knowledge distillation and quantization, warrant investigation for optimal edge deployment at unmanned railway crossings.

ACKNOWLEDGMENTS

The researchers would like to thank the Indonesian Ministry of Education, Culture, Research, and Technology for funding the 2025 national strategic research scheme for early-career lecturers. They would also like to thank the Universitas Muhammadiyah Lamongan research institute for its facilities and support.

FUNDING INFORMATION

This researcher is funded by the Indonesian Ministry of Education, Culture, Research, and Technology through the 2025 national strategic research funding scheme for early-career lecturers.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Heri Ardiansyah	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Denny Irawan					✓	✓		✓		✓		✓		
Dhio Ramadhon Saripudin						✓	✓			✓			✓	
Gilang Septian Firmansyah						✓	✓			✓			✓	
Rohmatin Agustina						✓				✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest in this manuscript.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] Federal Railroad Administration, "Highway-Rail Grade Crossing Safety," *U.S. Department of Transportation*, Jun. 24, 2025. [Online]. Available: https://railroads.dot.gov/railroad-safety/divisions/crossing-safety-and-trespass-prevention/railroad-crossing-safety
[2] W. Fu, Q. He, Q. Feng, J. Li, F. Zheng, and B. Zhang, "Recent Advances in Wayside Railway Wheel Flat Detection Techniques: A Review," *Sensors*, vol. 23, no. 8, p. 3916, Apr. 2023, doi: 10.3390/s23083916.
[3] S. B. ud din Tahir, A. Jalal, and K. Kim, "Wearable Inertial Sensors for Daily Activity Analysis Based on Adam Optimization and the Maximum Entropy Markov Model," *Entropy*, vol. 22, no. 5, p. 579, May 2020, doi: 10.3390/e22050579.
[4] D. C. Larese, A. B. Cerrada, G. D. Tomei, A. Guerrero-López, P. M. Olmos, and M. J. G. García, "Transformer Vibration Forecasting for Advancing Rail Safety and Maintenance 4.0," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 6, pp. 1612-1624, 2025, doi: 10.1109/OJITS.2025.3638611.
[5] B. Zhang, W. Li, X.-L. Li, and S.-K. Ng, "Intelligent Fault Diagnosis Under Varying Working Conditions Based on Domain Adaptive Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 66367–66384, 2018, doi: 10.1109/access.2018.2878491.
[6] K. Shaikh, I. Hussain, and B. S. Chowdhry, "Wheel Defect Detection Using a Hybrid Deep Learning Approach," *Sensors*, vol. 23, no. 14, p. 6248, Jul. 2023, doi: 10.3390/s23146248.
[7] O. Serradilla, E. Zugasti, J. Rodriguez, and U. Zurutuza, "Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects," *Applied Intelligence*, vol. 52, no. 10, pp. 10934–10964, Jan. 2022, doi: 10.1007/s10489-021-03004-y.
[8] [1] X. Liu, X. Meng, L. Ning, F. Xu, Q. Li, and C. Zhao, "A Deep Learning-Based Method for Mechanical Equipment Unknown Fault Detection in the Industrial Internet of Things," *Sensors*, vol. 25, no. 19, p. 5984, Sep. 2025, doi: 10.3390/s25195984.
[9] I Goodfellow, Y Bengio, and A Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[10] U. Orinaite, R. Burdzik, V. Ranjan, and M. Ragulskis, "Prediction of approaching trains based on H-ranks of track vibration signals," *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 5, pp. 614–631, Feb. 2025, doi: 10.1111/mice.13349.
[11] [1] R. Krč, J. Podroužek, M. Kratochvílová, I. Vukušič, and O. Plášek, "Neural Network-Based Train Identification in Railway Switches and Crossings Using Accelerometer Data," *Journal of Advanced Transportation*, vol. 2020, pp. 1–10, Nov. 2020, doi: 10.1155/2020/8841810.
[12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[13] Y. Yoo and S. Jeong, "Vibration analysis process based on spectrogram using gradient class activation map with selection process of CNN model and feature layer," *Displays*, vol. 73, p. 102233, Jul. 2022, doi: 10.1016/j.displa.2022.102233.
[14] A. López Galdo, A. Guerrero-López, P. M. Olmos, and M. J. Gómez García, "Detecting train driveshaft damages using accelerometer signals and Differential Convolutional Neural Networks," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106840, Nov. 2023, doi: 10.1016/j.engappai.2023.106840.
[15] A. Malekjafarian, C.-A. Sarrabezolles, M. A. Khan, and F. Golpayegani, "A Machine-Learning-Based Approach for Railway Track Monitoring Using Acceleration Measured on an In-Service Train," *Sensors*, vol. 23, no. 17, p. 7568, Aug. 2023, doi: 10.3390/s23177568.
[16] J. Yang *et al*., "Railway Intrusion Events Classification and Location Based on Deep Learning in Distributed Vibration Sensing," *Symmetry*, vol. 14, no. 12, p. 2552, Dec. 2022, doi: 10.3390/sym14122552.
[17] M. A. Amin, T. Najeh, and A. Ghoul, "AI-driven vibration-based event classification in railway switches and crossings," *Scientific Reports*, vol. 16, no. 1, Jun. 2026, doi: 10.1038/s41598-026-58967-0.
[18] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Interspeech 2019. ISCA*, pp. 2613–2617, Sep. 15, 2019, doi: 10.21437/interspeech.2019-2680.
[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE*, pp. 4510–4520, Jun. 2018, doi: 10.1109/cvpr.2018.00474.
[20] J. Zhu, H. Chen, and W. Ye, "A Hybrid CNN-LSTM Network for the Classification of Human Activities Based on Micro-Doppler Radar," *IEEE Access*, vol. 8, pp. 24713–24720, 2020, doi: 10.1109/access.2020.2971064.
[21] M. Daouad, F. A. Allah, and E. W. Dadi, "An automatic speech recognition system for isolated Amazigh word using 1D & 2D CNN-LSTM architecture," *International Journal of Speech Technology*, vol. 26, no. 3, pp. 775–787, Sep. 2023, doi: 10.1007/s10772-023-10054-9.
[22] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv*, 2020, doi: 10.48550/ARXIV.2010.16061.
[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/J.PATREC.2005.10.010.
[24] J. Cohen, "Statistical Power Analysis for the Behavioral Sciences." *Elsevier*, 1977, doi: 10.1016/c2013-0-10517-x.
[25] D. Neupane and J. Seok, "Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review," *IEEE Access*, vol. 8, pp. 93155–93178, 2020, doi: 10.1109/access.2020.2990528.

BIOGRAPHIES OF AUTHORS

Heri Ardiansyah received the B.Eng. degree in Electrical Engineering from Universitas Brawijaya, Malang, Indonesia, and the M.Eng. degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia. He is currently a lecturer with the Department of Computer Engineering, Universitas Muhammadiyah Lamongan, Lamongan, Indonesia, holding the academic position of Assistant Professor. He has served as the principal investigator for PDP research grant focusing on neural network-based vehicle license plate monitoring systems powered by solar energy. His research works have been published in national journals and international conference proceedings, including the American Institute of Physics (AIP) Conference Proceedings. He is a member of the Association of Informatics Study Programs (APSI). His current research interests include embedded systems, neural networks, digital signal processing, robotics, and solar energy applications in smart systems. He can be contacted at email: hery_ardyansyah@umla.ac.id.

Denny Irawan received the B.Eng. degree in Electrical Engineering from the Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 2003, and the master's degree in the same institute, in 2016 and now Doctoral degree student at Electronics Engineering Polytechnic, Surabaya, Indonesia. His current research interests include power converter topologies, batteries, renewable energy, and energy efficiency. He can be contacted at email: den2mas@umg.ac.id.

Dhio Ramadhon Saripduin is currently pursuing the B.Eng. degree in Computer Engineering at Universitas Muhammadiyah Lamongan, Lamongan, Indonesia, with a GPA of 3.93/4.00. He has practical experience in embedded systems and programming through various academic projects, including the design of counter and decoder circuits using digital logic and a temperature monitoring system based on Arduino and LM35 sensor. He is currently a member of the Computer Engineering Student Association. His current research interests include embedded systems, internet of things (IoT), and artificial intelligence. He can be contacted at email: dhio.ishtar08@email.com.

Gilang Septian Firmansyah is currently pursuing the B.Eng. degree in Computer Engineering at Universitas Muhammadiyah Lamongan, Lamongan, Indonesia, with a GPA of 3.80/4.00. He has practical experience in web development and internet of things (IoT) through various academic, freelance, and personal projects. His notable projects include SIAK (an academic information system with multi-role authentication using PHP, Laravel, and MySQL), smart greenhouse monitoring (an IoT-based environmental monitoring system), and smart garden (an automated plant watering system using ESP32 microcontroller). He is currently serving as the secretary of the Robotics Student Organization and a member of the media and information division at Himakom. His current research interests include web development, IoT systems, and embedded systems using Arduino and ESP32 platforms. He can be contacted at email: gilangjr171@gmail.com.

Rohmatin Agustina received the B.Agr. degree in Agricultural Cultivation from Universitas Muhammadiyah Gresik, Gresik, Indonesia, in 2002, and the M.Agr. degree in Plant Science from Universitas Brawijaya, Malang, Indonesia, in 2006. She is currently pursuing the Ph.D. degree in Agricultural Science at Universitas Brawijaya. She has been a lecturer with the Department of Agrotechnology, Faculty of Agriculture, Universitas Muhammadiyah Gresik, since 2014, where she served as the head of the Agrotechnology Study Program from 2014 to 2019. Her current research interests include plant physiology, agricultural biotechnology, natural resource conservation, and sustainable agriculture systems, with a particular focus on smallholder farming and rainfed agricultural systems. She has published in several reputable journals, including Agricultural Systems (Elsevier), Australian Journal of Crop Science, and Springer Nature. She can be contacted at email: rohmatin@umg.ac.id.